



GENERATIVE AI MODELS

Edited by: Jovan Pehcevski

GENERATIVE AI MODELS

GENERATIVE AI MODELS

Edited by:

Jovan Pehcevski



www.arclerpress.com

Generative AI Models

Jovan Pehcevski

Arcler Press

224 Shoreacres Road

Burlington, ON L7L 2H2

Canada

www.arcлерpress.com

Email: orders@arcлерeducation.com

© 2024

ISBN: 978-1-77469-996-6 (e-book)

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated and copyright remains with the original owners. Copyright for images and other graphics remains with the original owners as indicated. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data. Authors or Editors or Publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The authors or editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

© 2024 Arcler Press

ISBN: 978-1-77469-920-1 (Hardcover)

Arcler Press publishes wide variety of books and eBooks. For more information about Arcler Press and its products, visit our website at www.arcлерpress.com

DECLARATION

Some content or chapters in this book are open access copyright free published research work, which is published under Creative Commons License and are indicated with the citation. We are thankful to the publishers and authors of the content and chapters as without them this book wouldn't have been possible.

ABOUT THE EDITOR



Jovan Pehcevski obtained his PhD in Computer Science from RMIT University in Melbourne, Australia in 2007. His research interests include modern data center technologies (XaaS), big data, machine learning and artificial intelligence, and information retrieval. He has published over 30 journal and conference papers and he also serves as a journal and conference reviewer. Jovan has extensive academic and research experience, coupled with practical expertise in the IT industry. He is currently working as a Senior Technology Consultant at Dell Technologies, covering South Eastern Europe.

TABLE OF CONTENTS

| | |
|------------------------------------|-----|
| <i>List of Contributors</i> | xv |
| <i>List of Abbreviations</i> | xix |
| <i>Preface</i> | xxi |

Section 1 Image Generation Techniques

| | |
|--------------------------------------------------------------------------------------------------------|-----------|
| Chapter 1 Research on Image Generation and Style Transfer Algorithm Based on Deep Learning..... | 3 |
| Abstract | 3 |
| Introduction | 4 |
| Related Model Analysis..... | 6 |
| The Method of this Paper | 6 |
| Main Results | 11 |
| Conclusion | 15 |
| References | 16 |
| Chapter 2 An Overview of Image Caption Generation Methods | 19 |
| Abstract | 19 |
| Introduction | 20 |
| Feature Extraction Methods..... | 21 |
| Attention Mechanism..... | 24 |
| Dataset and Evaluation | 34 |
| Conclusion | 39 |
| Acknowledgments | 40 |
| References | 41 |
| Chapter 3 Application of an Improved DCGAN for Image Generation | 51 |
| Abstract | 51 |
| Introduction | 52 |

| | |
|-----------------------------------------------------------------------------------------------------------|------------|
| Improved Design of the Structure of the Dcgan..... | 54 |
| Construction of the Image Generation Models Based on the DCGAN and Gans | 57 |
| Comparative Analysis and Assessment of Image Generation Quality | 68 |
| Discussion and Conclusions | 71 |
| Acknowledgments | 72 |
| References | 73 |
| Chapter 4 Private Face Image Generation Method Based on Deidentification in Low Light | 77 |
| Abstract | 77 |
| Introduction..... | 78 |
| Related Work..... | 80 |
| The Proposed Method..... | 82 |
| Experiments and Analysis..... | 87 |
| Conclusion | 93 |
| Authors' Contributions..... | 94 |
| Acknowledgments | 94 |
| References | 95 |
| Chapter 5 Application of Remote Sensing Image Data Scene Generation Method in Smart City | 99 |
| Abstract | 99 |
| Introduction..... | 100 |
| The Proposed Method..... | 103 |
| Experiments | 111 |
| Discussion | 115 |
| Conclusions | 122 |
| Acknowledgments | 122 |
| References | 123 |
| Section 2 Video Generation Techniques | |
| Chapter 6 Realistic Speech-Driven Talking Video Generation with Personalized Pose..... | 127 |
| Abstract | 127 |
| Introduction..... | 128 |
| Related Work..... | 130 |

| | |
|------------------------------------------------------------------------------------------------------------|------------|
| Methods | 132 |
| Experiments | 136 |
| Conclusion and Future Work | 140 |
| Acknowledgments | 141 |
| References | 142 |
| Chapter 7 Video Transformation in Big Video Era and its Impact on Content Editing | 147 |
| Abstract | 147 |
| Introduction | 148 |
| Literature Review | 148 |
| Analysis of the Positive Impact of the Big Video ERA on Video Orientation..... | 150 |
| Analysis of the Negative Impact of the Big Video ERA on Video Orientation..... | 152 |
| New Requirements of Video Content Editing in the Big Video ERA..... | 154 |
| Conclusion | 156 |
| References | 157 |
| Chapter 8 A Fast Depth-Map Generation Algorithm Based on Motion Search from 2D Video Contents | 159 |
| Abstract | 159 |
| Introduction | 160 |
| The Proposed Depth-Map Generation Algorithm | 161 |
| Experimental Results..... | 166 |
| Conclusions | 167 |
| References | 168 |
| Chapter 9 Adaptive Content Management for UGC Video Delivery in Mobile Internet Era | 171 |
| Abstract | 171 |
| Introduction..... | 172 |
| Framework Design..... | 175 |
| Key Algorithms | 178 |
| Simulation and Evaluation | 183 |
| Related Work | 187 |

| | |
|-----------------------|-----|
| Conclusions..... | 187 |
| Acknowledgments | 188 |
| References | 189 |

Section 3 Voice and Speech Generation

| | |
|------------------------------------------------------------------------------------------------------------|------------|
| Chapter 10 Generating the Voice of the Interactive Virtual Assistant | 193 |
| Abstract | 193 |
| Introduction..... | 194 |
| Speech Processing Fundamentals..... | 195 |
| Text Processing | 197 |
| Acoustic Modelling..... | 198 |
| Open Resources And Tools | 206 |
| Quality Measurements..... | 211 |
| Conclusions And Open Problems | 211 |
| References | 213 |
| Chapter 11 Voice Quality Modelling for Expressive Speech Synthesis..... | 225 |
| Abstract | 225 |
| Introduction..... | 226 |
| Speech Material | 227 |
| Expressive Speech Style Transformation..... | 230 |
| Perceptual Assessment | 238 |
| Conclusions and Future Work..... | 247 |
| References | 249 |
| Chapter 12 Prosodically Rich Speech Synthesis Interface Using Limited Data of Celebrity Voice | 253 |
| Abstract | 253 |
| Introduction..... | 254 |
| Parametric Speech Synthesis Based on HMMS | 256 |
| Speech Materials with a Rich Prosodic Personality..... | 256 |
| Prosodic Personality Improvement with Limited Data | 259 |
| Experiments | 263 |
| Conclusion | 269 |
| Acknowledgements | 269 |

| | |
|-----------------------------------------------------------------------------------------------------------------------|------------|
| Notes | 269 |
| References | 270 |
| Chapter 13 Resources for Development of Hindi Speech Synthesis System: An Overview | 273 |
| Abstract | 273 |
| Introduction | 274 |
| Hindi Text Encoding | 275 |
| Corpora Development in Hindi Language | 275 |
| Challenges in Database Preparation | 280 |
| Conclusions | 280 |
| References | 282 |
| Section 4 Societal and Ethical Issues | |
| Chapter 14 How AI-Human Symbionts May Reinvent Innovation and What the New Centaurs Will Mean for Cities | 287 |
| Abstract | 287 |
| Introduction: Chess Doomsday | 288 |
| Understanding Centaurs | 289 |
| Centaurs and Innovation | 294 |
| Centaurs and the City | 299 |
| Conclusion: Who Would Ever Want to Be a Centaur? | 305 |
| Acknowledgements | 306 |
| Notes | 306 |
| References | 308 |
| Chapter 15 AI, Automation and New Jobs..... | 311 |
| Abstract | 311 |
| Introduction | 312 |
| Relevant Literature | 314 |
| Automation and New Jobs | 315 |
| Skill, Technologies and New Jobs | 317 |
| Theoretical Model..... | 318 |
| Conclusion | 321 |
| Notes 323 | |
| References | 324 |

| | |
|-------------------------------------------------------------------------------------------|------------|
| Chapter 16 Discussion on the Development of Artificial Intelligence in Taxation... | 327 |
| Abstract | 327 |
| Introduction..... | 328 |
| How is AI Applied in Taxation? | 329 |
| The Development of AI in China..... | 330 |
| Global Developments..... | 331 |
| Conclusions..... | 332 |
| References | 335 |
| Chapter 17 AI and Zen: AI Films as Reflections on Reality and Illusion..... | 337 |
| Abstract | 337 |
| Introduction..... | 338 |
| The Question of Reality and Zen..... | 338 |
| Reflection on the Reality of the World | 340 |
| Reflection on the Reality of Human Distinctiveness | 342 |
| Reflection on the Reality of “Self” | 344 |
| Findings..... | 345 |
| Summary | 346 |
| References | 347 |
| Chapter 18 Ecologically Sound Procedural Generation of Natural Environments.... | 349 |
| Abstract | 349 |
| Introduction..... | 350 |
| Related Work..... | 352 |
| Basic Approach..... | 356 |
| Vegetation Model | 359 |
| Visualization Model..... | 368 |
| Implementation | 373 |
| Results And Discussion..... | 376 |
| Conclusion | 383 |
| Acknowledgments | 384 |
| References | 385 |
| Index | 389 |

LIST OF CONTRIBUTORS

Ruikun Wang

School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China.

Haoran Wang

College of Information Science and Engineering, Northeastern University, China

Yue Zhang

College of Information Science and Engineering, Northeastern University, China

Xiaosheng Yu

Faculty of Robot Science and Engineering, Northeastern University, China

Bingqi Liu

School of Mechanical Engineering, Chengdu University, Chengdu 610106, China
Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

Jiwei Lv

Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

Xinyue Fan

Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

Jie Luo

Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

Tianyi Zou

Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

Beibei Dong

School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China

Zhenyu Wang

Sifang College, Shijiazhuang Tiedao University, Shijiazhuang 051132, China

Zhihao Gu

School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China

Jingjing Yang

School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China

Yuanjin Xu

Institute of Mathematical Geology and Remote Sensing Geology, School of Earth Resources, China University of Geosciences, 388 Lumo Road, Wuhan 430074, China

Xu Zhang

Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China

Liguo Weng

Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China

Mingzhi Yin

School of Foreign Languages, North China Electric Power University, Beijing, China.

Weiwei Wang

Communication and information Security Lab

Yuesheng Zhu

Shenzhen Graduate School, Peking University, China

Qilin Fan

National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, China

Hao Yin

National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, China

Zexun Jiang

National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, China

Haojun Huang

Department of Communication Engineering, Wuhan University, China

Yan Luo

Department of Electrical and Computer Engineering, University of Massachusetts Lowell, USA

Xu Zhang

National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, China

Adriana Stan

Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Beáta Lórincz

Technical University of Cluj-Napoca, Cluj-Napoca, Romania

“Babeş-Bolyai” University, Cluj-Napoca, Romania

Carlos Monzo

Computer Science, Multimedia and Telecommunication Studies, Universitat Oberta de Catalunya (UOC), Rambla del Poblenou 156, 08018 Barcelona, Spain

Ignasi Iriondo

Grup de Recerca en Tecnologies Mèdia (GTM), Universitat Ramon Llull, La Salle, Quatre Camins 2, 08022 Barcelona, Spain

Joan Claudi Socoró

Grup de Recerca en Tecnologies Mèdia (GTM), Universitat Ramon Llull, La Salle, Quatre Camins 2, 08022 Barcelona, Spain

Takashi Nose

Department of Communication Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan.

Taiki Kamei

Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Japan.

Archana Balyan

Department of Electronics and Communication, Maharaja Surajmal Institute of Technology, Affiliated to GGSIPU, New Delhi, India.

Emmanuel Muller

University of Applied Sciences, Kehl, Germany.
University of Strasbourg, Strasbourg, France.
Fraunhofer ISI, Karlsruhe, Germany.

Jaures Badet

Department of Economics, Necmettin Erbakan University, Konya, Turkey

Zhuowen Huang

Nanfang College of Sun Yat-sen University, Guangzhou, China

Jun Yu

Department of Film and Television Arts, Shanghai Publishing and Printing College, Shanghai, China.

Bo Zhang

Department of Film and Television Arts, Shanghai Publishing and Printing College, Shanghai, China.
Academy for Engineering & Technology, Fudan University, Shanghai, China.

Benny Onrust

¹Computer Graphics and Visualization Group, Delft University of Technology, Delft, Netherlands

Rafael Bidarra

Computer Graphics and Visualization Group, Delft University of Technology, Delft, Netherlands

Robert Rooseboom

Department of Spatial Ecology, Royal Netherlands Institute for Sea Research, Yerseke, Netherlands

Johan van de Koppel

Department of Spatial Ecology, Royal Netherlands Institute for Sea Research, Yerseke, Netherlands

LIST OF ABBREVIATIONS

| | |
|------|--------------------------------------|
| AI | Artificial intelligence |
| ASR | Automatic speech recognition |
| CBR | Case-based reasoning |
| CRF | Conditional random field |
| CDNs | Content delivery networks |
| CNN | Convolutional neural network |
| DNN | Deep neural network |
| DDSR | Dilated Depthwise Separable Residual |
| ESS | Expressive speech synthesis |
| GANs | Generative adversarial networks |
| HNM | Harmonic plus noise model |
| HMM | Hidden Markov model |
| IS | Inception score |
| IA | Intelligence Augmentation |
| IVA | Interactive Virtual Assistant |
| LRU | Least recently used |
| LCC | Leipzig Corpora Collection |
| LPC | Linear predictive coding |
| MOS | Mean opinion score |
| MDL | Minimum description length |
| MIL | Multi-instance learning |
| NFs | Normalising flows |
| PGC | Provider generated content |
| RMS | Root mean square |
| TTS | Text-to-speech |
| UGC | User generated content |
| VAEs | Variational AutoEncoders |
| VOD | Video on demand |

PREFACE

The world is changing rapidly, and scientific and technological progress plays a key role in this. Artificial intelligence (AI), which has permeated all spheres of social, economic, scientific research and everyday life, has a special influence. Today, the spotlight is on generative artificial intelligence, which has the potential to change the world in the coming years, in terms of development, commercial and social perspectives.

Within a few months, ChatGPT has become the most prominent representative of the new generation of generative artificial intelligence systems. Others are called LaMDA, DALL-E or Stable Diffusion. These programs produce fundamentally new texts, codes, images or even videos. The results are so convincing that it is often impossible to tell whether they were created by human or machine.

The generative AI is especially good and applicable in 3 major areas:

- Text generation - applications in law (drafting contracts), medicine (diagnostics), journalism (news production), education (production of educational materials), science (search and generation of scientific papers), etc.
- Image and video generation - application in marketing (advertisement creation), media (virtual host), art, architecture, design, social networks, etc.
- Voice and sound generation - application in the film industry (special effects), music industry, customer support (virtual references), surveying (automation of telephone surveys), etc.

Generative artificial intelligence is expected to revolutionize the way people work or find information online. But it also raises numerous ethical questions, as rarely has any technology done so far. There are fears that millions of people could lose their jobs, or that the system could be misused for disinformation, or even that the world, as we know it, will end. This sparked a debate about the necessary rules and regulation of AI.

This book edition covers different topics of generative AI models, including: image generation techniques, video generation techniques, speech / voice generation techniques, and societal and ethical issues of these models.

Section 1 focuses on image generation techniques, describing image generation and style transfer algorithm based on deep learning; an overview of image caption generation methods; an application of an improved DCGAN for image generation; a private face image generation method based on deidentification in low light; and a remote sensing image data scene generation method in smart city.

Section 2 focuses on video generation techniques, describing realistic speech-driven talking video generation with personalized pose; video transformation in big video

era and its impact on content editing; a fast depth-map generation algorithm based on motion search from 2D video contents; and adaptive content management for UGC video delivery in mobile internet era.

Section 3 focuses on voice and speech generation, describing generating the voice of the interactive virtual assistant; voice quality modelling for expressive speech synthesis; prosodically rich speech synthesis interface using limited data of celebrity voice; and an overview of resources for development of Hindi speech synthesis system.

Section 4 focuses on societal and ethical issues of generative AI models, describing AI-human symbiotes reinventing innovation and what the new centaurs will mean for cities; the impact of AI and automation on new jobs; a development of artificial intelligence in taxation; AI films as reflections on reality and illusion; and an ecologically sound procedural generation of natural environments.

SECTION 1

IMAGE GENERATION TECHNIQUES

CHAPTER 1

Research on Image Generation and Style Transfer Algorithm Based on Deep Learning

Ruikun Wang

School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China.

ABSTRACT

Aiming at the current process of artistic creation and animation creation, there are a lot of repeated manual operations in the process of conversion from sketch to the stylized image. This paper presented a solution based on a deep learning framework to realize image generation and style transfer. The method first used the conditional generation to resist the network, optimizes the loss function of the training mapping relationship, and generated the actual image from the input sketch. Then, by defining and optimizing the perceptual loss function of the style transfer model, the

Citation: Wang, R. (2019), "Research on Image Generation and Style Transfer Algorithm Based on Deep Learning". Open Journal of Applied Sciences, 9, 661-672. doi: 10.4236/ojapps.2019.98053.

Copyright: © 2019 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

style features are extracted from the image, thereby forming the actual The conversion between images and stylized art images. Experiments show that this method can greatly reduce the work of coloring and converting with different artistic effects, and achieve the purpose of transforming simple stick figures into actual object images.

Keywords:- Deep Learning, Image Generation, Style Transfer

INTRODUCTION

At present, the art creation and animation creation process mainly uses sketching first, and then through a series of processes such as coloring to form an actual picture. When the style needs to be converted, most of them need to be re-colored, which leads to a large number of repeated manual operations in the process. This paper uses the advantages of deep neural networks, combined with conditional confrontation networks and convolutional neural networks, to automatically implement the process of sketching to physical and style conversion. CNNs are the main methods to solve various image recognition and detection. CNNs minimize the loss function by learning features [1]. Although the feature learning process is automated, it still requires a lot of manpower to design its tags. In contrast, generating anti-network GANs, using the generation model and the discriminant model, while minimizing loss, can then use the loss function to generate a new picture.

Style transfer is the process of migrating from one reference style to another to generate another image. The feedforward image conversion task has been widely used. Many conversion tasks use the pixel-by-pixel differential method to train the deep convolutional neural network, which spans the pixel-by-pixel difference [2] , by putting the CRF as an RNN, train with other parts of the network. The structure of our conversion network was inspired by [3] and [4] , using down-sampling in the network to reduce the spatial extent of the feature map, followed by up-sampling in a network to produce the final output image. Some methods change the pixel-by-pixel difference to a penalty image gradient or use the CRF loss layer to force the output image to be consistent. A feedforward model in [5] is trained with a loss function of pixel-by-pixel difference for coloring grayscale images. There are a number of papers that use optimized methods to produce images, their objects are perceptual, and perceptuality depends on the high-level features extracted from CNN. Mahendran and Vedaldi reversed features from convolutional networks, reconstructing loss

functions by minimizing features, in order to understand image information stored in different network layers; similar methods were also used to invert local binary descriptors [6] and HOG features [7]. The work of Dosovitskiy and Brox is most relevant to us. They train a feedforward neural network to invert the convolution feature and quickly approximate the outcome of the proposed optimization problem. However, their feedforward network uses pixel by pixel. Reconstruct the loss function to train, and our network directly uses the feature reconstruction loss function used in [8]. Gatys et al. show artistic style conversion [9] [10], combining a content map and another style map. By minimizing the cost function reconstructed according to features, the cost function for style reconstruction is also based on the advanced from the pre-training model. Features; a similar method was previously used for texture synthesis. Their approach yields a high-quality record, but the computational cost is very expensive because each iteration of the optimization requires a feedforward, feedback-pre-trained network. In order to overcome the burden of such a computational load, this paper trains a feedforward neural network to quickly obtain a feasible solution.

Our network consists of two parts: a picture conversion network f_w and a loss network φ , where the picture conversion network is a deep residual network [11], the parameter is the weight W, which converts the input picture x by mapping $y = f_w(x)$. To output the picture y , each loss function calculates a scalar value $l_i(y, y_i)$, which measures the difference between the output y and the target image y_i . The picture conversion network is trained with SGD so that the weighted sum of a series of loss functions remains degraded. This paper implements the task of generating stylized art images from sketches. First, use conditional generation to combat the network [12], optimize the loss function of the training mapping relationship to generate the actual image from the input sketch. This paper trains a feedforward network for image conversion tasks, and does not use pixel-by-pixel difference to construct the loss function, and instead uses the perceptual loss function to extract advanced features from the pre-trained network. In the process of training, the perceptual loss function is more suitable than the pixel-by-pixel loss function to measure the degree of similarity between images. After training, the effect of sub-network image translation achieves the expected effect, and because of the characteristics of the anti-network, we no longer need to manually design the mapping function like the ordinary CNN network. Experiments have shown that reasonable results can be achieved even without manually setting the loss function.

RELATED MODEL ANALYSIS

Structure-Generated Image Modeling Structure Loss

The structure loss image conversion problem of image generation image modeling is usually expressed as the classification or regression problem of each pixel [13] , and the output space is regarded as “unstructured”, and each pixel of the output is regarded as independent of all other pixels of the input image as appropriate. Instead, conditional GANs learn the structured loss. Structured loss penalizes the node construction of the output. Most types of literature consider this type of loss, such as conditional random fields [14] , SSIM metrics [15] , feature matching [16] , nonparametric loss [17] , convolutional pseudo-prior [18] , and loss based on matching covariance statistics [19] . Our conditional GAN differs from these learned losses and can theoretically penalize any possible structure different from the output and target.

Condition GANs

This paper is not the first to apply GANs to conditional settings. There have been previous works to constrain GANs with discrete tags [20], text, and the like. Image-based GANs have solved image restoration [21] , predicting images from normal maps [22] , editing images based on user constraints, video predictions, state predictions, and generating merchandise and style transitions from photos [23] [24] . These methods have all changed based on specific applications, and our methods are simpler than most of them. Our approach to the choice of several structures in the generator and discriminator is also different from the previous work. Unlike the previous one, our generator used the “U-Net” structure [25] , and the discriminator used the convolution “PatchGAN” classifier. Previously, a similar PatchGAN structure was proposed to capture local style statistics.

THE METHOD OF THIS PAPER

Image Generation

GANs is a generation model for learning the mapping of random noise vector zz to output image yy : $G: z \rightarrow y$ $G: z \rightarrow y$. Conversely, the conditional GANs learn the mapping of the observed image xx and the random noise vector zz to yy . The formula is:

$$G : \{x, z\} \rightarrow y \quad G : \{x, z\} \rightarrow y \quad (1)$$

The training generator GG generates an image in which the discriminator D cannot discriminate, and the training discriminator DD detects the “falsified” image of the generator as much as possible.

Image Generated Objective Function

The objective function of the condition GAN is calculated as:

$$L_{cGAN}(G, D) = E_{x, y \sim p_{data}(x, y)} [\log D(x, y)] + E_{x, y \sim p_{data}(x), z \sim p_z(z)} [\log 1 - D(x, G(x, z))] \quad (2)$$

GG wants to minimize the value of this function, DD wants to maximize the value of this function, that is, in order to test the importance of the condition to the discriminator, we compare the variant form without the discriminator without xx input, condition GAN previous method found Using the traditional loss is beneficial to the hybrid GAN target equation: the work of the discriminator remains the same, but the generator not only deceives the discriminator, but also generates real images as much as possible. Based on this consideration, the L_1 distance is used instead of the L_2 distance. Because L_1 encourages less blur, the formula is:

$$L_{l1}(G) = E_{x, y \sim p_{data}(x, y), Z \sim P_z(Z)} [\|y - G(x, z)\|_1] \quad (3)$$

The final target formula is:

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \gamma L_{l1}(G) \quad (4)$$

Network Structure

This paper uses the structure of the generator and discriminator in [9], both of which use the convolution unit form of “conv-BatchNorm-ReLu”. The appendix provides details of the network structure. Below we only discuss the main features.

Construct a generator with jumpers

One feature of the image conversion problem is the mapping of high resolution input meshes to a high resolution output mesh. In addition, for the problem we are considering, the input and output are different in appearance, but they are consistent with the underlying structure. Therefore, the structure of the input can be roughly aligned with the structure of the output. We design the generator structure based on these considerations. We mimicked “U-Net” to

add jumper connections. In particular, we add jumpers between each of the i^{th} and $n-in-i$ layers, where n is the total number of layers in the network. Each jumper simply connects the feature channels of the i^{th} layer and the $n-in-i$ layer.

The discriminator for constructing the Markov process (PatchGAN)

It is well known that L_1 and L_2 loss have ambiguities in image generation problems. The discriminator structure we designed only penalizes the structure of the patch size. The discriminator classifies each $N \times NN \times N$ as true or false. We run this discriminator (sliding window) on the entire image and finally take the average as the final output of DD. Such a discriminator models the image as a Markov random field, assuming that the pixels segmented by the patch diameter are directly independent of each other. This finding has been studied and is a commonly used hypothesis in texture and style models. Our PatchGAN can therefore be understood as a form of texture/style loss.

Optimization and reasoning

To optimize the network, we use the standard method: alternate training DD and GG. We use minibatch SGD and apply the Adam optimizer. In the reasoning, we run the generator in the same way as the training phase.

Style Transfer

The system consists of two parts: a picture conversion network f_w and a loss network φ (used to define a series of loss functions $[l_1, l_2, l_3]$). The picture conversion network is a deep residual network, and the parameters are weights W . It converts the input image x into the output image y by mapping $y = f_w(x)$, and each loss function calculates a scalar value $l_i(y, y_i)$, which measures the difference between the output y and the target image y_i . The picture conversion network is trained by SGD, and the effect diagram is shown in Figure 1. The purpose is to calculate the weighted sum of a series of loss functions by operation, and the formula is:

$$W^* = \arg \min E_{x, \{y_i\}} \left[\sum_{i=1} \gamma_i l_i(f_w(x), y_i) \right] \quad (5)$$

We used a pre-trained network φ for image classification to define our loss function. We then train our deep convolutional transformation network using a loss function that is also a deep convolutional network, as shown in Figure 2.

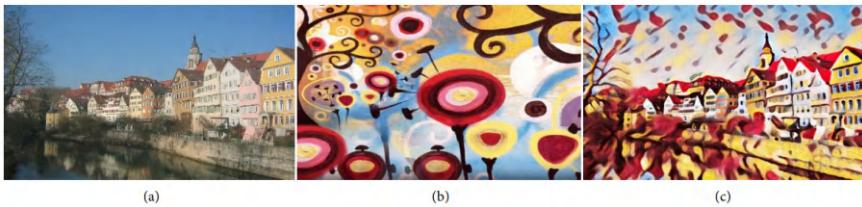


Figure 1. Style transfer effect chart. (a) Content (b) Style (c) Result.

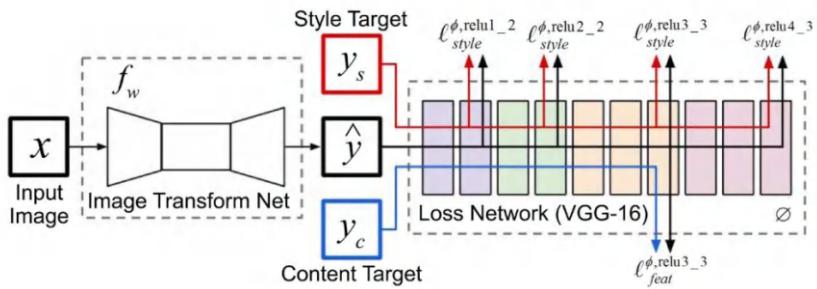


Figure 2. Training network diagram.

The loss network φ is able to define a feature (content) loss l_{feat} and a style loss l_{style} , respectively measuring the difference in content and style. For each input image x we have a content target y_c a style target y_s , for style conversion, the content target y_c is the input image x , the output image y , the style y_s should be combined to the content $x = y_c$. We train a network for each target style.

Construction of Image Conversion Network

Instead of any pooling layer, we use a convolution or micro-step convolution instead. Our neural network consists of five residual blocks. All non-residual convolutional layers follow a spatial batch-normalization, and the nonlinear layer of the RELU, with the exception of the last output layer. The last layer uses a scaled Tanh to ensure that the pixels of the output image are between [0, 255]. Except for the first and last layers with a 9×9 kernel, all other convolutional layers use 3×3 kernels.

Input and Output: For style conversion, both input and output are color images, size $3 \times 256 \times 256$. For super-resolution reconstruction, there is an upsampling factor f , the output is a high resolution image $3 \times 288 \times 288$, the input is a low resolution image $3 \times 288/f \times 288/f$, because the image

conversion network is completely convolved, so during the test, it can be applied to images of any resolution.

Downsampling and Upsampling: For super-resolution reconstruction, there is an upsampling factor f , and we use several residual blocks followed by the $\text{Log}2f$ volume and the network (stride = 1/2). This process is different from [1]. Double-cubic interpolation is used to upsample this low-resolution input before putting the input into the network. Without relying on any fixed upsampling interpolation function, the microstep convolution allows the upsampling function to be trained along with the rest of the network. For image conversion, our network uses two contention = 2 convolutions to downsample the input, followed by several residual blocks, followed by two convolution layers (stride = 1/2) upsampling.

Perceptual Loss Function

We define two perceptual loss functions to measure the high level of perceptual and semantic differences between two images. Use a pre-trained network model for image classification. In our experiments this model was VGG-16 [25], using Imagenet's dataset for pre-training.

Feature (content) loss: We do not recommend pixel-by-pixel comparison, but use VGG to calculate the advanced feature (content) representation. This method is the same as the original style using VGG-19 [26] to extract style features. The formula is:

$$l_{feat}^{\phi_j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \left\| \phi_j(\hat{y}) - \phi_j(y) \right\|_2^2 \quad (6)$$

Style Loss: Feature (content) loss penalizes the output image (when it deviates from the target y), so we also want to punish style deviations: color, texture, common patterns, and so on. In order to achieve such an effect, Gatys et al. proposed a loss function for the following style reconstruction. Let $\phi_j(x)$ represent the j th layer of the network ϕ , and the input is x . The shape of the feature map is $C_j \times H_j \times W_j$, and the definition matrix $G_j(x)$ is $C_j \times C_j$ matrix (characteristic matrix). The elements are derived from the following formula:

$$G_j^\phi(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \quad (7)$$

If we understand $\phi_j(x)$ as a feature of the C_j dimension, and the size of each feature is $H_j \times W_j$, then the left $G_j(x)$ is proportional to the non-

central covariance of the C_j dimension. Each grid location can be used as a separate sample. This can therefore capture which feature can drive other information. The gradient matrix can be calculated in a very funny time by adjusting the shape of $\varphi_j(x)$ to a matrix ψ , the shape is $C_j \times H_j W_j$, and then $G_j(x)$ is $\psi\psi^T/C_j H_j W_j$. The loss of style reconstruction is well defined, even when the output and target have different sizes, because with the gradient matrix, the two will be adjusted to the same shape.

MAIN RESULTS

Conditional Confrontation Network Model

To optimize the versatility of GANs, we tested the method on a variety of tasks and data sets, including graphics tasks (such as photo generation) and visual tasks (such as semantic segmentation). We have found that very good results are often obtained on small data sets. The training data set we used contains only 400 images, and training can be made very fast with this size of training set. Some of the super parameters are shown in Table 1.

Qualitative results: the completed model is displayed, and the actual generated effect is displayed. Below we list three sets of pictures, as shown in Figure 3, the input of the figure, the second column is the output (model generation result), and the third column is the actual result. Equation (8) is the calculation formula used. A lot of experiments show that our average is around 0.4.

$$\text{rate} = \frac{\text{time}}{\text{max_steps}} \quad (8)$$

Table 1. Training hyperparameter selection and result numerical mapping ratio

| Hyperparameter | Value |
|----------------|--------|
| aspect_ratio | 1.0 |
| gan_weight | 1.0 |
| l1_weight | 100.0 |
| lr | 0.0002 |
| scale_size | 286 |

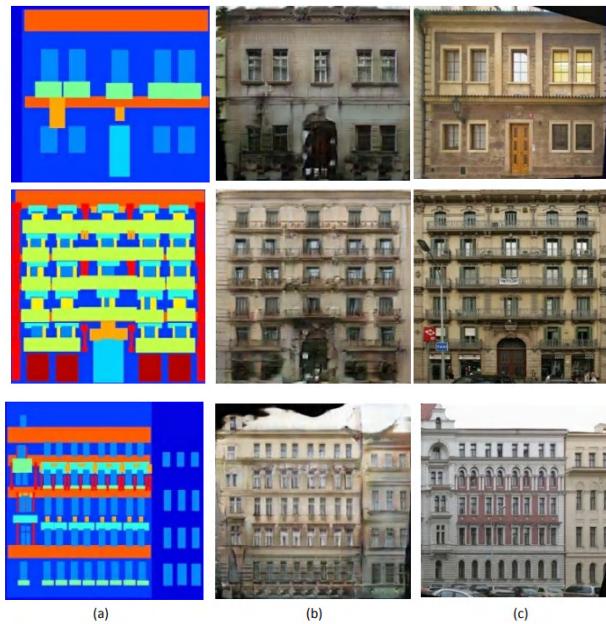


Figure 3. Conditional confrontation network model implementation rendering.
 (a) input (b) Model generation result (c) result.

Style Migration

The goal of style conversion is to produce a picture with both the content information of the content map and the style information of the style map. As a baseline, we reproduce the method of Gatys et al., giving the style and content goals y_s and y_c , layer i and J represent feature and style reconstruction. The implementation formula is:

$$\hat{y} = \arg \min \mu_c l_{feat}^{\phi,j}(y, y_c) + \mu_s l_{style}^{\phi,j}(y, y_s) + \mu_{TV} l_{TV}(y) \quad (9)$$

In the formula, u starts with parameters, y is initialized to white noise, and is optimized with LBFGS. We found that unconstrained optimization equations usually cause the pixel values of the output image to go beyond $[0, 255]$ to make a more fair comparison. For the baseline, we use L-BFGS projection, and adjust the image y to each iteration. $[0, 255]$, in most cases, the computational optimization converges to satisfactory results within 500 iterations, which is slower because each LBFGS iteration requires feedforward feedback and feedback through the VGG16 network.

Training details: Our style conversion network is trained with COCO datasets. We adjust each image to 256×256 , a total of 80,000 training charts, batch-size = 4, iterations 40,000 times, and about two rounds. Optimized with Adam, the initial learning rate is 0.001. The output graph is normalized by the whole variable (strength between 1e-6 and 1e-4), selected by cross-validation set. There is no weight attenuation or dropout because the model has no overfitting in these two rounds. For all style conversion experiments we take the relu2_2 layer for content, relu1_2, relu2_2, relu3_3 and relu4_3 as styles. For the VGG-16 network, our experiments used Torch and cuDNN, and the training took about 4 hours on a GTX Titan X GPU.

Qualitative results: For the model after training, we performed the actual effect test. We screened out four sets of images, as shown in Figure 4. In the figure, column a provides content features for content images, and column b provides style textures for style images. We train different models to migrate effects for different styles. In column 4 of column 4, compared with the optimized method, our network produces comparable quality results, but can achieve three orders of magnitude speed increase. This optimization is of great significance for practical applications. After a lot of experiments, the average time we took the picture was around 10 seconds.

Model Combination

We combine the conditional confrontation network model and the style transfer model to achieve a good combination effect. The specific results are shown in Figure 5. The a column is the sketch, the b column is the generated result, and the c column is the effect after the style transfer.

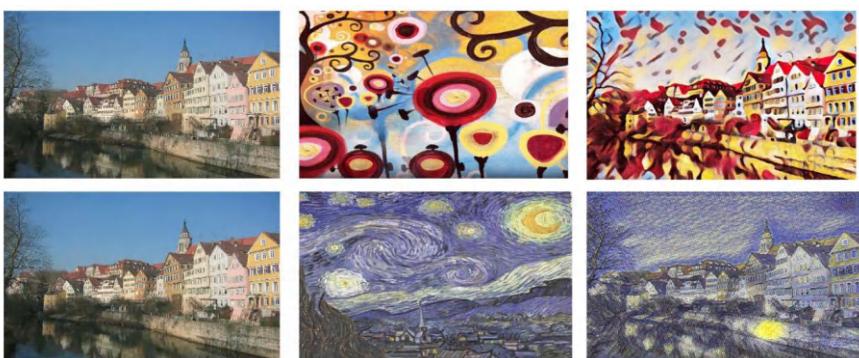




Figure 4. Schematic diagram of style transfer results. (a) Content (b) Style (c) Result.



Figure 5. Schematic diagram of the results of the model. (a) input (b) output1 (c) output2.

CONCLUSION

In this paper, we take advantage of the feedforward network and the optimization-based approach to achieve a good performance and speed by training the feedforward network with a perceptual loss function. We use the conditional confrontation network to implement the function of image translation. Finally, we combine the two models to achieve the application effect in a specific scenario. But, the migration of details is not in place. The lack of detail in the depiction of different image styles will follow the following two aspects to improve the network's capabilities: First, for the already trained model, the generated image has reached a very fast speed, but the training model still takes several hours. I hope that the training process of the model can be optimized and the training time of the model can be improved. Second, for more research on the details of the image, you can add more detail extraction to the network to transfer the style of the image, achieve more realistic comic style migration effects, and imitate different painter strokes and for buildings and characters adapt to different parameters.

REFERENCES

1. Krizhevsky, A., Sutskever, I., Hinton, G.E., et al. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 141, 1097-1105.
2. Zheng, S., Jayasumana, S., Romeraparedes, B., et al. (2015) Conditional Random Fields as Recurrent Neural Networks. *International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1529-1537. <https://doi.org/10.1109/ICCV.2015.179>
3. Long, J., Shelhamer, E., Darrell, T., et al. (2015) Fully Convolutional Networks for Semantic Segmentation. *Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. <https://doi.org/10.1109/CVPR.2015.7298965>
4. Noh, H., Hong, S., Han, B., et al. (2015) Learning Deconvolution Network for Semantic Segmentation. *International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1520-1528. <https://doi.org/10.1109/ICCV.2015.178>
5. Cheng, Z., Yang, Q., Sheng, B., et al. (2015) Deep Colorization. *International Conference on Computer Vision*, Santiago, 7-13 December 2015, 415-423. <https://doi.org/10.1109/ICCV.2015.55>
6. Dangelo, E., Alahi, A., Vandergheynst, P., et al. (2012) Beyond Bits: Reconstructing Images from Local Binary Descriptors. *International Conference on Pattern Recognition*, Tsukuba, Japan, 11-15 November 2012, 935-938.
7. Vondrick, C., Khosla, A., Malisiewicz, T., et al. (2013) HOGgles: Visualizing Object Detection Features[C]. *International Conference on Computer Vision*, Sydney, 1-8 December 2013, 1-8. <https://doi.org/10.1109/ICCV.2013.8>
8. Mahendran, A. and Vedaldi, A. (2015) Understanding Deep Image Representations by Inverting Them. *Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 5188-5196. <https://doi.org/10.1109/CVPR.2015.7299155>
9. Gatys, L.A., Ecker, A.S. and Bethge, M. (2015) Texture Synthesis Using Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*, Neural Information Processing Systems Foundation, Quebec, 262-270. <https://doi.org/10.1109/CVPR.2016.265>

10. Gatys, L.A., Ecker, A.S. and Bethge, M. (2015) A Neural Algorithm of Artistic Style. *Computer Science*, 11, 510-519.
11. He, K., Zhang, X., Ren, S., et al. (2016) Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. <https://doi.org/10.1109/CVPR.2016.90>
12. Isola, P., Zhu, J., Zhou, T., et al. (2017) Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5967-5976. <https://doi.org/10.1109/CVPR.2017.632>
13. Xie, S. and Tu, Z. (2015) Holistically-Nested Edge Detection. *International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1395-1403. <https://doi.org/10.1109/ICCV.2015.164>
14. Chen, L., Papandreou, G., Kokkinos, I., et al. (2015) Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations*, San Diego, CA, 9 April 2015.
15. Wang, Z., Bovik, A.C., Sheikh, H.R., et al. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13, 600-612. <https://doi.org/10.1109/TIP.2003.819861>
16. Dosovitskiy, A. and Brox, T. (2016) Generating Images with Perceptual Similarity Metrics Based on Deep Networks. *Neural Information Processing Systems*, Barcelona, Spain, 5-10 December 2016, 658-666.
17. Li, C. and Wand, M. (2016) Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. *Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2479-2486. <https://doi.org/10.1109/CVPR.2016.272>
18. Xie, S., Huang, X., Tu, Z., et al. (2016) Top-Down Learning for Structured Labeling with Convolutional Pseudoprior. *European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 302-317. https://doi.org/10.1007/978-3-319-46493-0_19
19. Johnson, J., Alahi, A., Feifei, L., et al. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

20. Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets.
21. Pathak, D., Krahenbuhl, P., Donahue, J., et al. (2016) Context Encoders: Feature Learning by Inpainting. Computer Vision and Pattern Recognition, Las Vegas, 27-30 June 2016, 2536-2544. <https://doi.org/10.1109/CVPR.2016.278>
22. Wang, X. and Gupta, A. (2016) Generative Image Modeling Using Style and Structure Adversarial Networks. European Conference on Computer Vision, Amsterdam, 8-16 October 2016, 318-335. https://doi.org/10.1007/978-3-319-46493-0_20
23. Yoo, D., Kim, N., Park, S., et al. (2016) Pixel-Level Domain Transfer. European Conference on Computer Vision, Amsterdam, 8-16 October 2016, 517-532. https://doi.org/10.1007/978-3-319-46484-8_31
24. Li, C. and Wand, M. (2016) Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. European Conference on Computer Vision, Amsterdam, 8-16 October 2016, 702-716. https://doi.org/10.1007/978-3-319-46487-9_43
25. Ronneberger, O., Fischer, P., Brox, T., et al. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer Assisted Intervention, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
26. Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, San Diego, 7-9 May 2015.

CHAPTER 2

An Overview of Image Caption Generation Methods

Haoran Wang¹, Yue Zhang¹, and Xiaosheng Yu²

¹College of Information Science and Engineering, Northeastern University, China

²Faculty of Robot Science and Engineering, Northeastern University, China

ABSTRACT

In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the field of artificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive

Citation: Haoran Wang, Yue Zhang, Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, Computational Intelligence and Neuroscience, vol. 2020, Article ID 3062706, 13 pages, 2020. <https://doi.org/10.1155/2020/3062706>.

Copyright: © 2020 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

and significant, for example, the realization of human-computer interaction. This paper summarizes the related methods and focuses on the attention mechanism, which plays an important role in computer vision and is recently widely used in image caption generation tasks. Furthermore, the advantages and the shortcomings of these methods are discussed, providing the commonly used datasets and evaluation criteria in this field. Finally, this paper highlights some open challenges in the image caption task.

INTRODUCTION

The development of the image description system may help the visually impaired people “see” the world in the future. Recently, it has drawn increasing attention and become one of the most important topics in computer vision [1–11]. Early image description generation methods aggregate image information using static object class libraries in the image and modeled using statistical language models.

Aker and Gaizauskas [12] use a dependency model to summarize multiple web documents containing information related to image locations and propose a method for automatically tagging geotagged images. Li et al. [13] propose a n-gram method based on network scale, collecting candidate phrases and merging them to form sentences describing images from zero. Yang et al. [14] propose a language model trained from the English Gigaword corpus to obtain the estimation of motion in the image and the probability of colocated nouns, scenes, and prepositions and use these estimates as parameters of the hidden Markov model.

The image description is obtained by predicting the most likely nouns, verbs, scenes, and prepositions that make up the sentence. Kulkarni et al. [15] propose using a detector to detect objects in an image, classifying each candidate region and processing it by a prepositional relationship function and finally applying a conditional random field (CRF) prediction image tag to generate a natural language description. Object detection is also performed on images. Lin et al. [16] used a 3D visual analysis system to infer objects, attributes, and relationships in an image and convert them into a series of semantic trees and then learn the grammar to generate text descriptions for these trees.

Some indirect methods have also been proposed for dealing with image description problems, such as the query expansion method proposed by Yagcioglu et al. [17], by retrieving similar images from a large dataset and using the distribution described in association with the retrieved images.

The expression is used to create an extended query, and then the candidate descriptions are reordered by estimating the cosine between the distributed representation and the extended query vector, and finally, the closest description is taken as a description of the input image. In summary, the methods described are brainstorming and have their own characteristics, but all have the common disadvantage that they do not make intuitive feature observations on objects or actions in the image, nor do they give an end-to-end mature general model to solve this problem. The efficiency and popularization of neural networks have made breakthroughs in the field of image description and saw new hopes until the advent of the era of big data and the outbreak of deep learning methods.

In this paper, we review the development process of image description methods in recent years and summarize the basic framework and some improved methods. Then, we analyze the advantages and shortcomings of existing models and compare their results on public large-scale datasets. Finally, we summarize some open challenges in this task.

This paper is organized as follows. The second part details the basic models and methods. The third part focuses on the introduction of attention mechanism to optimize the model and make up for the shortcomings. The fourth part introduces the common datasets come up by the image caption and compares the results on different models. Different evaluation methods are discussed. The fifth part summarizes the existing work and proposes the direction and expectations of future work.

FEATURE EXTRACTION METHODS

Image caption models can be divided into two main categories: a method based on a statistical probability language model to generate handcraft features and a neural network model based on an encoder-decoder language model to extract deep features. The specific details of the two models will be discussed separately.

Handcraft Features with Statistical Language Model

This method is a Midge system based on maximum likelihood estimation, which directly learns the visual detector and language model from the image description dataset, as shown in Figure 1. Fang et al. [18] first analyze the image, detect the object, and then generate a caption. Words are detected by applying a convolutional neural network (CNN) to the image area [19] and

integrating the information with MIL [20]. The structure of the sentence is then trained directly from the caption to minimize the priori assumptions about the sentence structure. Finally, it turns an image caption generation problem into an optimization problem and searches for the most likely sentence.

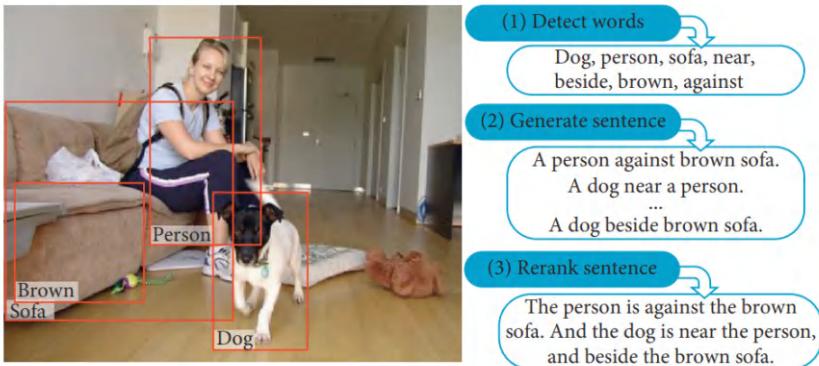


Figure 1. Method based on the visual detector and language model.

The implementation steps are as follows:(1) Detect a set of words that may be part of the image caption. We detect the words from the given vocabulary according to the content of the corresponding image based on the weak monitoring method in multi-instance learning (MIL) in order to train the detectors iteratively.(2) Running a fully convolutional network on an image, we get a rough spatial response graph. Each position in the response map corresponds to a response obtained by applying the original CNN to the region of the input image where the shift is shifted (thus effectively scanning different locations in the image to find possible objects). By upsampling the image, we get a response map on the final fully connected layer and then implement the noisy-OR version of MIL on the response map for each image. Each word produces a single probability.(3) The process of caption generation is searching for the most likely sentence under the condition of the visually detected word set. The language model is at the heart of this process because it defines the probability distribution of a sequence of words. Although the maximum entropy language model (ME) is a statistical model, it can encode very meaningful information. For example, “running” is more likely to follow the word “horse” than “speaking.” This information can help identify the wrong words and encode commonsense knowledge. (4) There are similar ways to use the combination of attribute detectors and

language models to process image caption generation. Devlin et al. [21] used a combination of CNN and k-NN methods and a combination of a maximum entropy model and RNN to process image description generation tasks. Kenneth Tran proposed an image description system, [22] using CNN as a visual model to detect a wide range of visual concepts, landmarks, celebrities, and other entities into the language model, and the output results are the same as those extracted by CNN. The vectors together are used as input to the multichannel depth-similar model to generate a description.

Deep Learning Features with Neural Network

The recurrent neural network (RNN) [23] has attracted a lot of attention in the field of deep learning. It was originally widely used in the field of natural language processing and achieved good results in language modeling [24]. In the field of speech, RNN converts text and speech to each other [25–31], machine translation [32–37], question and answer session [38–43], and so on. Of course, they are also used as powerful language models at the level of characters and words. Currently, word-level models seem to be better than character-level models, but this is certainly temporary. RNN is also rapidly gaining popularity in computer vision. For example, frame-level video classification [44–46], sequence modeling [47, 48], and recent visual question-answer tasks.

As shown in Figure 2, the image description generation method based on the encoder-decoder model is proposed with the rise and widespread application of the recurrent neural network [49]. In the model, the encoder is a convolutional neural network, and the features of the last fully connected layer or convolutional layer are extracted as features of the image. The decoder is a recurrent neural network, which is mainly used for image description generation. Because RNN training is difficult [50], and there is a general problem of gradient descent, although it can be slightly compensated by regularization [51], RNN still has a fatal flaw that it can only remember the contents of the previous limited time unit, and LSTM [52] is a special RNN architecture that can solve problems such as gradient disappearance, and it has long-term memory. In recent years, the LSTM network has performed well in dealing with video-related context [53–55]. Similar with video context, the LSTM model structure in Figure 3 is generally used in the text context decoding stage.

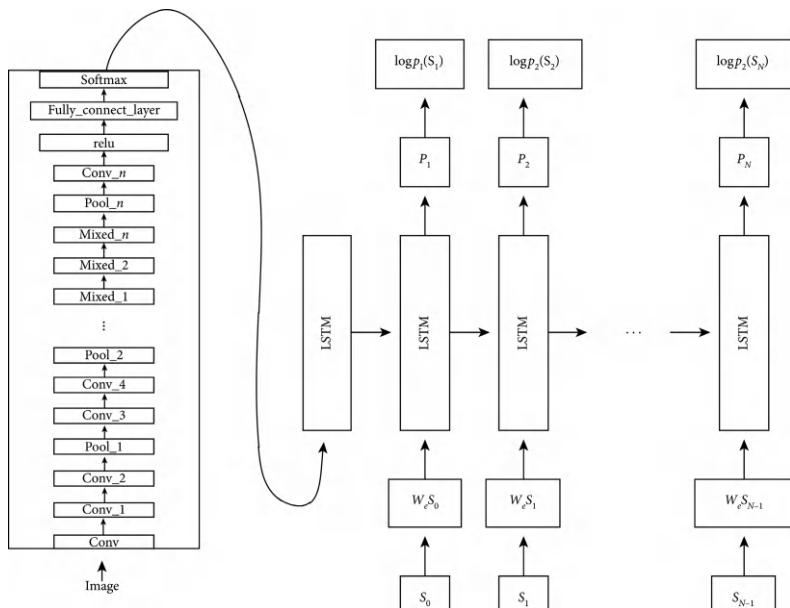


Figure 2. Model based on encoder-decoder.

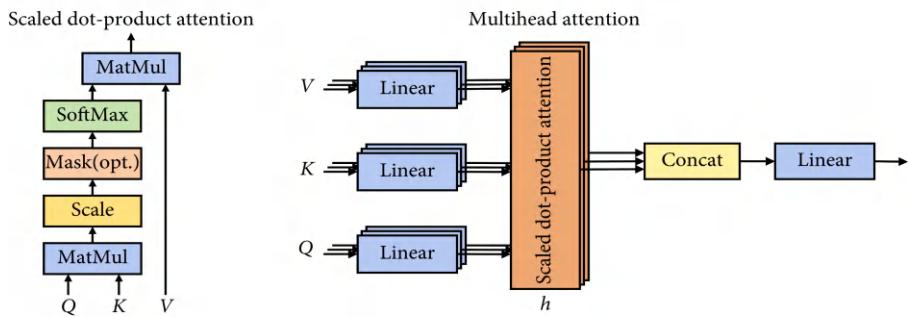


Figure 3. (a) Scaled dot-product attention. (b) Multihead attention.

ATTENTION MECHANISM

Attention mechanism, stemming from the study of human vision, is a complex cognitive ability that human beings have in cognitive neurology. When people receive information, they can consciously ignore some of the main information while ignoring other secondary information. This ability of self-selection is called attention. This mechanism was first proposed to

be applied to the image classification in the field of visual images using the attention mechanism on the RNN model [56]. In natural language processing, when people read long texts, human attention is focused on keywords, events, or entities. A large number of experiments have proved that the attention mechanism is applied in text processing, for example, machine translation [35, 57], abstract generation [58, 59], text understanding [60–63], text classification [64–66], visual captioning [67, 68], and other issues, the results achieved remarkable, and the following describes the application of different attention mechanism methods in the image description basic framework introduced in the second part, so that its effect is improved.

In neural network models, the realization of the attention mechanism is that it allows the neural network to have the ability to focus on its subset of inputs (or features)—to select specific inputs or features. The main part of the attention mechanism is the following two aspects: the decision needs to pay attention to which part of the input; the allocation of limited information processing resources to the important part. At present, the mainstream attention mechanism calculation formulas are shown in equations (1) and (2); the design idea is to link the target module m_t with the source module m_s through a function and finally normalize to obtain the probability distribution:

$$a_t = \text{align}(m_t, m_s) = \frac{\exp(f(m_t, m_s))}{\sum_s \exp(f(m_t, m_{s'}))}, \quad (1)$$

$$f(m_t, m_s) = \begin{cases} m_t^T m_s, & \text{dot,} \\ m_t^T W_a m_s, & \text{general,} \\ W_a [m_t; m_s], & \text{concat,} \\ v_a^T \tanh(W_a m_t + U_a m_s), & \text{perception.} \end{cases} \quad (2)$$

Based on the advantages of the attention mechanism mentioned above, this chapter details the various achievements of the attention mechanism algorithm and its application in image description generation.

Soft Attention

Dzmitry et al. [57] first proposed the soft attention model and applied it to machine translation. In fact, “soft” refers to the probability distribution of attention distribution. For any word in the input sentence S , the probability is given according to the context vector Z_t [69]. Finally, the weighted sum of all regions is calculated to get the probability distribution:

$$E_{P(s_t|a)}[\hat{z}_t] = \sum_{i=1}^L \alpha_{t,i} a_i. \quad (3)$$

A deterministic attention model is formulated by computing a soft attention weighted attention vector [57]:

$$\Phi(\{a_i\}, \{\alpha_i\}) = \sum_i^L \alpha_i a_i. \quad (4)$$

The objective function can be written as follows:

$$L = -\log(P(y|x)) + \lambda \sum_i^L \left(1 - \sum_t^C \alpha_{t,i} \right)^2. \quad (5)$$

Soft attention is parameterized and therefore can be embedded and modeled for direct training. Gradient can be passed back through the attention mechanism module to other parts of the model.

Hard Attention

Unlike the soft attention mechanism, which focuses on calculating the weighted sum of all regions, hard attention only focuses on one location and is a process of randomly selecting a unique location. It samples the hidden state of the input by probability, rather than the hidden state of the entire encoder. The context vector Z_t [69] is calculated as follows:

$$p(s_{t,i} = 1 | a) = \alpha_{t,i}, \\ \hat{z}_t = \sum_{i=1}^L s_{t,i} a_i, \quad (6)$$

where $s_{t,i}$ refers to whether to select the i -th position in the L feature maps, if selected, set to 1, set to 0, otherwise the opposite.

In order to achieve gradient backpropagation, Monte Carlo sampling is needed to estimate the gradient of the module. One disadvantage of hard attention is that information is selected based on the method of maximum sampling or random sampling. Therefore, the functional relationship between the final loss function and the attention distribution is not achievable, and training in the backpropagation algorithm cannot be used.

Multihead Attention

In general, we can represent input information in a key-value pair format, where “key” is used to calculate the attention distribution and “value” is used to generate the selected information. The multiheaded attention mechanism uses a plurality of keys, values, and queries to calculate a plurality of information selected from the input information in parallel for linear projection. As shown in Figure 3, each attention focuses on different parts of the input information to generate output values, and finally, these output values are concatenated and projected again to produce the final value [70]:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \\ \text{where } \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \quad (7)$$

Scaled Dot-Product Attention

Scaled dot-product attention [70] performs a single attention function using keys, values, and query matrices:

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (8)$$

Additional attention is paid to the compatibility function using a feedforward network with a single hidden layer. In practice, the scaled-down dot product is faster and more space-efficient than the multiheaded attention mechanism because it can be implemented using a highly optimized matrix multiplication code.

Global Attention

The main idea of global attention [71] is to consider the hidden layer state of all encoders. It obtains the attention weight distribution by comparing the current decoder hidden layer state with the state of each encoder hidden layer. It is similar to soft; that is, in the process of decoding, each time step needs to calculate the attention weight of each word in the encoding and then weights the context vector. The overall flow is shown in Figure 4. Since it chooses to focus on all the encoder inputs when calculating each decoder state, the amount of calculation is relatively large.

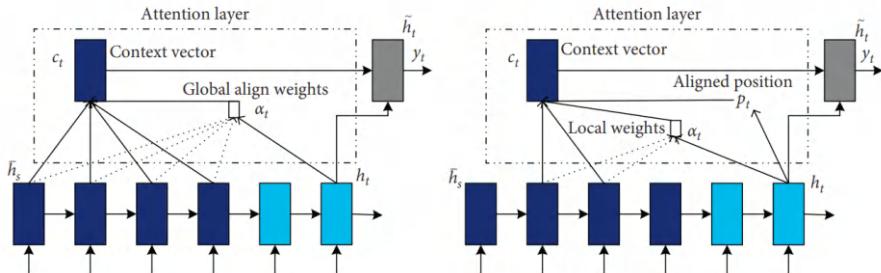


Figure 4. (a) Global attention model and (b) local attention model.

Local Attention

Local attention [71] first finds an alignment position and then calculates the attention weight in the left and right windows where its position is located and finally weights the context vector.

This is actually a mixed compromise between soft and hard. The main advantage of local attention is to reduce the cost of the attention mechanism calculation.

In the calculation, the local attention is not to consider all the words on the source language side, but to predict the position of the source language end to be aligned at the current decoding according to a prediction function and then navigate through the context window, considering only the words within the window.

Adaptive Attention with Visual Sentinel

For most of the attention models used for image caption and visual question and answer, regardless of which word is generated next, the image is focused on in each time step [72–74].

However, not all words have corresponding visual signals. The adaptive attention mechanism and the visual sentinel [75] solve the problem of when to add attention mechanisms and where to add them in order to extract meaningful information for sequence words. As shown in Figure 5, the context vector is considered to be the residual visual information of the LSTM hidden state. It reduces the uncertainty and supplements the informational of the next word prediction in the current hidden state. The calculation is as follows:

$$\begin{aligned}
 Ct = g(V, ht) &= \sum_{i=1}^k \alpha_{ti} v_{ti} = \text{soft max}(z_t) \cdot v_{ti} \\
 &= \text{soft max}\left(w_h^T \tanh\left(W_V V + (W_g h_t) I^T\right)\right) \cdot v_{ti}, \\
 \hat{c}_t &= \beta_t s_t + (1 - \beta_t) c_t,
 \end{aligned} \tag{9}$$

where the adaptive context vector is defined as \hat{c}_t , which is modeled as a mixture of spatial image features (i.e., the context vector of the spatial attention model) and the visual sentinel vector β_t . It determines how much new information the network takes into account from the image and what it already knows in decoding the memory.

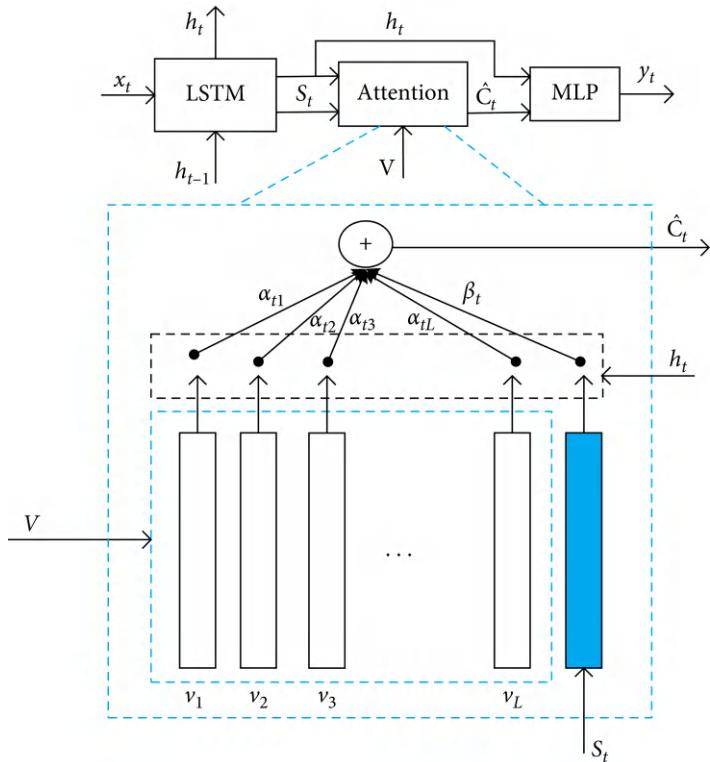


Figure 5. Adaptive attention model with visual sentinel.

Semantic Attention

Semantic attention [76] selectively handles semantic concepts and fuses them into the hidden state and output of LSTM. Selection and fusion form

feedback that connects top-down and bottom-up calculations. First, multiple top attribute and bottom-up features are extracted from the input image using multiple attribute detectors (AttrDet), and then all visual features are input as attention weight to a recurrent neural network (RNN) input and state calculation. The implementation is as follows:

$$\begin{aligned} x_0 &= \Phi_0(v) = W^{x,v}v, \\ h_t &= \text{RNN}(h_{t-1}, x_t), \\ Y_t \sim p_t &= \varphi(h_t, \{A_i\}), \\ x_t &= \phi(Y_{t-1}, \{A_i\}), \quad t > 0, \end{aligned} \tag{10}$$

The entire model architecture is shown in Figure 6.

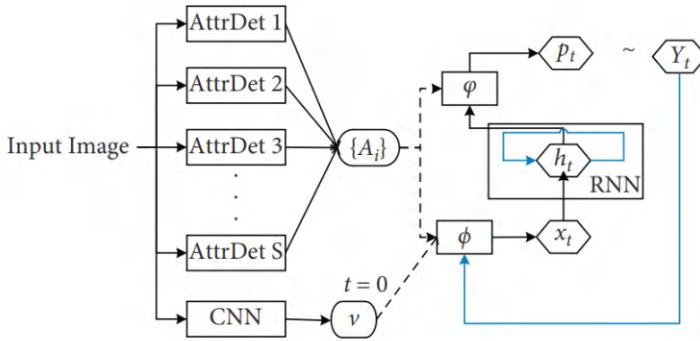


Figure 6. Semantic attention.

Spatial and Channel-Wise Attention

Spatial and channel attention [77] is the process of selecting semantic attributes according to the needs of the sentence context as shown in Figure 7. It uses the attention mechanism according to the extracted semantics in the encoding process, in order to overcome the general attention mechanism in decoding. Pay attention to the problem of overrange when using the last layer of the process. For example, when we want to predict “cake,” channel-wise attention (e.g., in the “convolution 5_3/convolution 5_4 feature map”) will be based on “cake,” “fire,” “light,” and “candle” and equivalent shape semantics, and more weight is assigned on the channel. Secondly, since the feature map depends on its underlying feature extraction, it is natural to apply attention in multiple layers; this allows obtaining visual attention on multiple semantic abstractions.

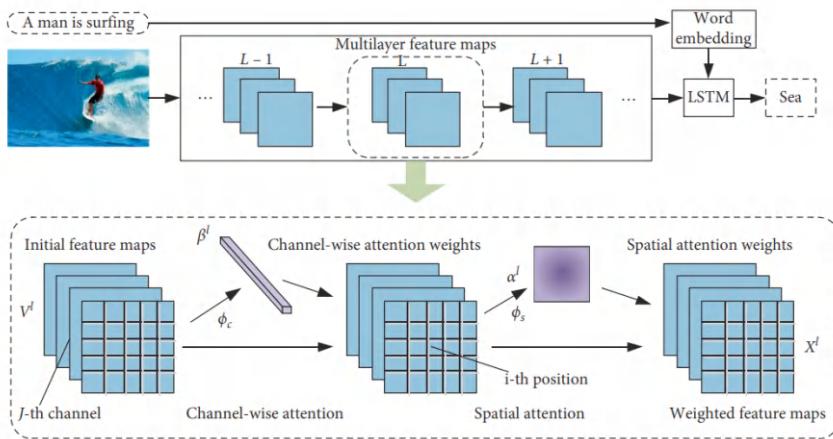


Figure 7. SCA-CNN model.

Areas of Attention

Pedersoli et al. [4] proposed a note-taking model (Figure 8). The method uses three pairs of interactions to implement an attention mechanism to model the dependencies between the image region, the title words, and the state of the RNN language model.

Compared with the previous method of associating only the image region with the RNN state, this method allows a direct association between the title word and the image region, not only considering the relationship between the state and the predicted word, but also considering the image [78]. The relationship between the region and the word and state is more comprehensive.

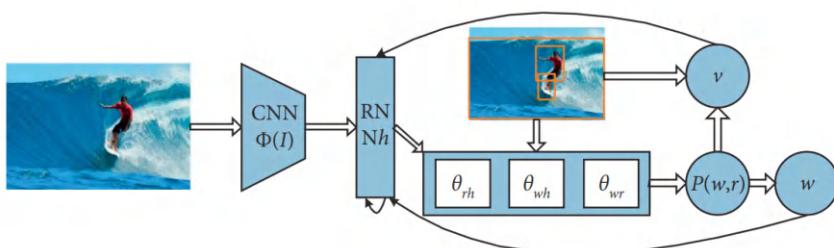


Figure 8. Areas of attention.

Deliberate Attention

Gao et al. [79] proposed a deliberate attention model (Figure 9). The method is proposed by observing people's daily habits of dealing with things, such as a common behavior of improving or perfecting work in people's daily writing, painting, and reading.

In the paper, the authors present a novel Deliberate Residual Attention Network, namely DA, for image captioning. The first-pass residual-based attention layer prepares the hidden states and visual attention for generating a preliminary version of the captions, while the second-pass deliberate residual-based attention layer refines them. Since the second-pass is based on the rough global features captured by the hidden layer and visual attention in the first-pass, the DA has the potential to generate better sentences. They also further equip the DA with discriminative loss and reinforcement learning to disambiguate image/caption pairs and reduce exposure bias.

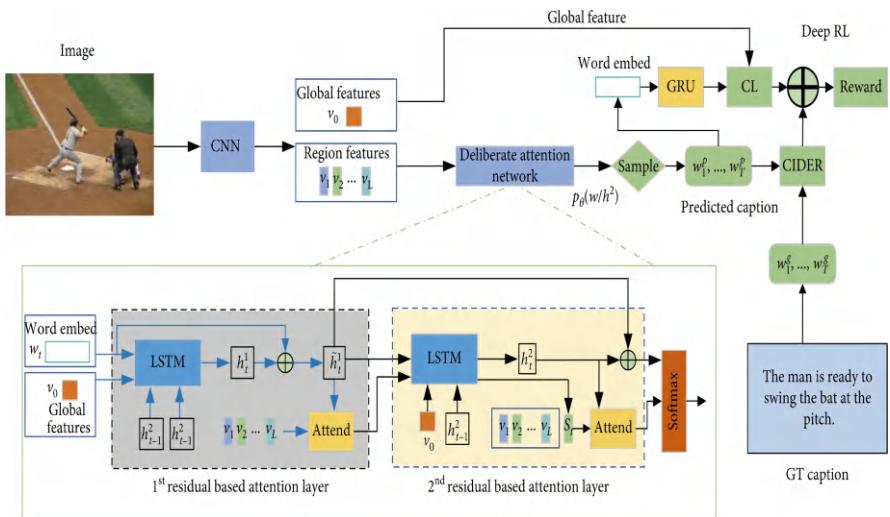


Figure 9. Deliberate attention framework.

This chapter analyzes the algorithm models of different attention mechanisms. Table 1 summarizes the application of attention mechanism in image description and points out the comments of different attention mechanisms and the way they add models, which is convenient for readers to choose appropriate in future research. The attention mechanism improves the model's effect.

Table 1. Comparison of attention mechanism modeling methods

| Ref. | Attention name | Method | Comment |
|------|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| [69] | Soft attention | Give a probability according to the context vector for any word in the input sentence when seeking attention probability distribution | Parameterization Derivative enable Definitely |
| [69] | Hard attention | Focus only on a randomly chosen location using Monte Carlo sampling to estimate the gradient | Randomly On the base of probability Simple |
| [70] | Multihead attention | Linearly projecting multiple pieces of information selected from the input in parallel using multiple keys, values, and queries | Linear projection Parallel Focus on information from different representation subspaces in different locations Multiple attention head |
| [70] | Scaled dot-product attention | Execute a single attention function using keys, values, and query matrices | High speed Save space |
| [71] | Global attention | Considering the hidden layer state of all encoders, the weight distribution of attention is obtained by comparing the current decoder hidden layer state with the state of each encoder hidden layer | Comprehensive Time-consuming Large amount of calculation |
| [71] | Local attention | First find a location for it, then calculate the attention weight in the left and right windows of its location, and finally weight the context vector | Reduce the cost of calculations |
| [75] | Adaptive attention | Define a new adaptive context vector which is modeled as a mixture of the spatially attended image features and the visual sentinel vector. This trades off how much new information the network is considering from the image with what it already knows in the decoder memory | Solve when and where to add attention in order to extract meaningful information for sequence words |

| | | | |
|------|------------------------------------|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| [76] | Semantic attention | Select semantic concepts and incorporate them into the hidden state and output of the LSTM | Optional Merge From top to bottom From bottom to top |
| [77] | Spatial and channel-wise attention | Select semantic attributes based on the needs of the sentence context | Multiple semantics In order to overcome the problem of overrange when using the general attention |
| [4] | Areas of attention | Modeling the dependencies between image regions, title words, and the state of the RNN language model | Interaction Comprehensive |

DATASET AND EVALUATION

This chapter mainly introduces the evaluation methods of open-source datasets and generated sentences in this field. Data, computational power, and algorithms are the three major elements of the current development of artificial intelligence. The three complement each other and enhance each other. It can be said that a good dataset can make the algorithm or model more effective. The image description task is similar to machine translation, and its evaluation method extends from machine translation to form its own unique evaluation criteria.

Dataset

Data are the basis of artificial intelligence. People are increasingly discovering that many laws that are difficult to find can be found from a large amount of data. In the image description generation task, there are currently rich and colorful datasets, such as MSCOCO, Flickr8k, Flickr30k, PASCAL 1K, AI Challenger Dataset, and STAIR Captions, and gradually become a trend of contention. In the dataset, each image has five reference descriptions, and Table 2 summarizes the number of images in each dataset. In order to have multiple independent descriptions of each image, the dataset uses different syntax to describe the same image. As illustrated in the example in Figure 10, different descriptions of the same image focus on different aspects of the scene or are constructed using different grammars. *MSCOCO*. Microsoft COCO Captions dataset [80], developed by the Microsoft Team that targets scene understanding, captures images from complex daily scenes and can be used to perform multiple tasks such as image recognition,

segmentation, and description. The dataset uses Amazon’s “Mechanical Turk” service to artificially generate at least five sentences for each image, with a total of more than 1.5 million sentences. The training set contains 82,783 images, the validation set has 40,504 images, and the test set has 40,775 images. Its 2014 version of the data has a total of about 20G pictures and about 500M of annotation files which mark the correspondence between one image and its descriptions. *Flickr8k/Flickr30k* [81, 82]. Flickr8k image comes from Yahoo’s photo album site Flickr, which contains 8,000 photos, 6000 image training, 1000 image verification, and 1000 image testing. Flickr30k contains 31,783 images collected from the Flickr website, mostly depicting humans participating in an event. The corresponding manual label for each image is still 5 sentences. *PASCAL 1K* [83]. A subset of the famous PASCAL VOC challenge image dataset, which provides a standard image annotation dataset and a standard evaluation system. The PASCAL VOC photo collection consists of 20 categories, and for its 20 categories, 50 images were randomly selected for a total of 1,000 images. Then, Amazon’s Turkish robot service is used to manually mark up five descriptions for each image. The dataset image quality is good and the label is complete, which is very suitable for testing algorithm performance. *AIC*. The Chinese image description dataset, derived from the AI Challenger, is the first large Chinese description dataset in the field of image caption generation. The dataset contains 210,000 pictures of training sets and 30,000 pictures of verification sets. Similar to MSCOCO, each picture is accompanied by 5 Chinese descriptions, which highlight important information in the image, covering the main characters, scenes, actions, and other contents. Compared with the English datasets common to similar scientific research tasks, Chinese sentences usually have greater flexibility in syntax and lexicalization, and the challenges of algorithm implementation are also greater. *STAIR*. The Japanese image description dataset [84], which is constructed based on the images of the MSCOCO dataset. STAIR consists of 164,062 pictures and a total of 820,310 Japanese descriptions corresponding to each of the five pictures. It is the largest Japanese image description dataset.

Table 2. Summary of the number of images in each dataset

| Dataset name | Size | | |
|--------------|-------|-------|-------|
| | Train | Valid | Test |
| MSCOCO | 82783 | 40504 | 40775 |
| Filckr8k | 6000 | 1000 | 1000 |

| | | | |
|-----------|--------|-------|-------|
| Flickr30k | 28000 | 1000 | 1000 |
| PASCAL 1K | — | — | 1000 |
| AIC | 210000 | 30000 | 30000 |
| STAIR | 82783 | 40504 | 40775 |



A man is skate boarding down a path and a dog is running by his side.
A person riding a skate board with a dog following beside.
This man is riding a skateboard behind a dog.

Figure 10. An example in MSCOCO dataset image.

Evaluation Criteria

In the evaluation of sentence generation results, BLEU [85], METEOR [86], ROUGE [87], CIDEr [88], and SPICE [89] are generally used as evaluation indexes. For five indicators, BLEU and METEOR are for machine translations, ROUGE is for automatic summary, and CIDEr and SPICE are present for image caption. They measured the consistency of the n-gram between the generated sentences, which was affected by the significance and rarity of the n-gram. At the same time, all four indicators can be directly calculated by the MSCOCO title assessment tool. The source code is publicly available. *BLEU*. It is the most widely used evaluation indicator; the original intention of the design is not for the image caption problem, but for the machine translation problem based on the accuracy rate evaluation. It is used to analyze the correlation of n-gram between the translation statement to be evaluated and the reference translation statement. Its core idea is that the closer the machine translation statement is to a human professional translation statement, the better the performance. In this task, the processing is the same as machine translation: multiple images are equivalent to multiple source language sentences in the translation. The

advantage of BLEU is that the granularity it considers is an n-gram rather than a word, considering longer matching information. The disadvantage of BLEU is that no matter what kind of n-gram is matched, it will be treated the same. For example, the importance of verb matching should be intuitively greater than the article. The higher the BLEU score, the better the performance. *METEOR*. METEOR is also used to evaluate machine translation, which aligns the translation generated from the model with the reference translation and matches the accuracy, recall, and *F*-value of various cases. What makes METEOR special is that it does not want to generate very “broken” translations and the method is based on the precision of one gram and the harmonic mean of the recall. The weight of the recall is a bit higher than the precision. This criterion also has features that are not available in others. It is designed to solve some of the problems with BLEU. It is highly relevant to human judgment and, unlike BLEU, it has a high correlation with human judgment not only at the entire collection but also at the sentence and segment level. The higher the METEOR score, the better the performance. *ROUGE*. ROUGE is a set of automated evaluation criteria designed to evaluate text summarization algorithms. The higher the RUGE score, the better the performance. *CIDEr*. CIDEr is specifically designed for image annotation problems. It measures the consistency of image annotation by performing a Term Frequency-Inverse Document Frequency (TF-IDF) weight calculation for each n-gram. This indicator treats each sentence as a “document,” represents it in the form of a TF-IDF vector, and then calculates the cosine similarity of the reference description to the description generated by the model as a score. In other words, it is the vector space model. This indicator compensates for one of the disadvantages of BLEU, that is, all words on the match are treated the same, but in fact, some words should be more important. Again, the higher the CIDEr score, the better the performance. *SPICE*. It is a semantic evaluation indicator for image caption that measures how image titles effectively recover objects, attributes, and relationships between them. On the natural image caption dataset, SPICE is better able to capture human judgments about the model’s subtitles, rather than the existing n-gram metrics.

Table 3 shows the scores of the attention mechanisms introduced in part 3. From Table 3, we found that the scores on different evaluation criteria for different models’ performance are not the same. Although there are differences in some evaluation criteria, if the improvement effect of an attention model is very obvious, in general, all evaluation indicators are relatively high for its rating.

Table 3. Scores of attention mechanisms based on the evaluations above

| Ref. | Attention model | BLEU-4 | METE-OR | ROUGE-L | CI-DEr |
|------|------------------------------|--------|---------|---------|--------|
| [69] | Soft attention | 24.3 | 23.9 | — | — |
| [69] | Hard attention | 25.0 | 23.0 | 51.6 | 86.5 |
| [70] | Multihead/scaled dot-product | 28.4 | — | — | — |
| [71] | Global/local attention | 25.9 | — | — | — |
| [75] | Adaptive attention | 33.2 | 26.6 | 55.0 | 108.5 |
| [76] | Semantic attention | 30.4 | 24.3 | 53.5 | 94.3 |
| [77] | Spatial and channel-wise | 31.1 | 25.4 | 53.0 | 94.3 |
| [4] | Areas of attention | 31.9 | 25.2 | — | 98.1 |
| [79] | Deliberate attention | 37.5 | 28.5 | 58.2 | 125.6 |

Based on the NIC model [49] as state-of-the-art performance, Xu et al. [69] describe approaches to caption generation that attempt to incorporate a form of attention with two variants: a “hard” attention mechanism and a “soft” attention mechanism. Encouraged by recent advances in caption generation and inspired by recent success in employing attention in machine translation [57] and object recognition [90, 91], they investigate models that can attend to a salient part of an image while generating its caption.

Existing approaches are either top-down, which start from a gist of an image and convert it into words, or bottom-up, which come up with words describing various aspects of an image and then combine them. You et al. [89] propose a new algorithm that combines both approaches through a model of semantic attention. The algorithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. The method is slightly more effective than the “soft” and “hard” attention.

Visual attention models are generally spatial only. Chen et al. [77] introduce a novel convolutional neural network dubbed SCA-CNN that incorporates spatial and channel-wise attentions in a CNN. In the task of image captioning, SCA-CNN dynamically modulates the sentence generation context in multilayer feature maps, encoding where and what the visual attention is. Pedersoli and Lucas [89] propose “Areas of Attention,”

the approach models the dependencies between image regions, caption words, and the state of an RNN language model, using three pairwise interactions, this method allows a direct association between caption words and image regions. Both two methods mentioned above together yield results mentioned earlier on the MSCOCO dataset.

Lu et al. [75] propose a adaptive attention model with a visual sentinel. The model not only decides whether to attend to the image or to the visual sentinel but also decides where, in order to extract meaningful information for sequential word generation. This sets the new state-of-the-art by a significant margin so far.

CONCLUSION

In this overview, we have compiled all aspects of the image caption generation task, discussed the model framework proposed in recent years to solve the description task, focused on the algorithmic essence of different attention mechanisms, and summarized how the attention mechanism is applied. We summarize the large datasets and evaluation criteria commonly used in practice.

Although image caption can be applied to image retrieval [92], video caption [93, 94], and video movement [95] and the variety of image caption systems are available today, experimental results show that this task still has better performance systems and improvement.

It mainly faces the following three challenges: first, how to generate complete natural language sentences like a human being; second, how to make the generated sentence grammatically correct; and third, how to make the caption semantics as clear as possible and consistent with the given image content. For future work, we propose the following four possible improvements:(1)An image is often rich in content.

The model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.(2)For corpus description languages of different languages, a general image description system capable of handling multiple languages should be developed.(3)Evaluating the result of natural language generation systems is a difficult problem. The best way to evaluate the quality of automatically generated texts is subjective assessment by linguists, which is hard to achieve. In order to improve system performance, the evaluation indicators should be optimized to make them more in line

with human experts' assessments.(4)A very real problem is the speed of training, testing, and generating sentences for the model should be optimized to improve performance.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (61603080 and 61701101), the Fundamental Research Funds for the Central Universities of China (N182608004), and Doctor Startup Fund of Liaoning Province (201601019).

REFERENCES

1. P. Anderson, X. He, C. Buehler et al., “Bottom-up and top-down attention for image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
2. J. Aneja, A. Deshpande, and S. Alexander, “Convolutional image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
3. T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 4904–4912, Las Vegas, NV, USA, June 2016.
4. M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 1251–1259, Venice, Italy, October 2017.
5. H. R. Tavakoli, R. Shetty, B. Ali, and J. Laaksonen, “Paying attention to descriptions generated by image captioning models,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 2506–2515, Venice, Italy, October 2017.
6. A. Mathews, L. Xie, and X. He, “SemStyle: learning to generate stylised image captions using unaligned text,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
7. T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, “Show, adapt and tell: adversarial training of cross-domain image captioner,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision and Pattern Recognition*, pp. 521–530, Honolulu, HI, USA, July 2017.
8. C. C. Park, B. Kim, and G. Kim, “Towards personalized image captioning via multimodal memory networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, p. 1, 2018.
9. X. Chen, Ma Lin, W. Jiang, J. Yao, and W. Liu, “Regularizing RNNs for caption generation by reconstructing the past with the present,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

10. R. Zhou, X. Wang, N. Zhang, X. Lv, and L.-J. Li, “Deep reinforcement learning-based image captioning with embedding reward,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1151–1159, Honolulu, HI, USA, July 2017.
11. Q. You, Z. Zhang, and J. Luo, “End-to-end convolutional semantic embeddings,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5735–5744, Salt Lake City, UT, USA, June 2018.
12. A. Aker and R. Gaizauskas, “Generating image descriptions using dependency relational patterns,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, vol. 49, no. 9, pp. 1250–1258, Uppsala, Sweden, July 2010.
13. S. Li, G. Kulkarni, T. L. Berg, and Y. Choi, “Composing simple image descriptions using web-scale N-grams,” in *Proceeding of Fifteenth Conference on Computational Natural Language Learning*, pp. 220–228, Association for Computational Linguistics, Portland, OR, USA, June 2011.
14. Y. Yang, C. L. Teo, H. Daume, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in *Proceeding of the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454, Edinburgh, UK, July 2011.
15. G. Kulkarni, V. Premraj, V. Ordonez et al., “Babytalk: understanding and generating simple image descriptions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.
16. D. Lin, C. Kong, S. Fidler, and R. Urtasun, “Generating multi-sentence lingual descriptions of indoor scenes,” pp. 2333–9721, 2015, <http://arxiv.org/abs/1503.00064> Computer Science.
17. S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici, “A distributed representation based query expansion approach for image captioning,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 10, no. 3115, Beijing, China, July 2015.
18. H. Fang, S. Gupta, F. Iandola et al., “From captions to visual concepts and back,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015.

19. R. Girshick, J. Donahue, D. Trevor, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.
20. C. Zhang, J. C. Platt, and V. Paul, “Multiple instance boosting for object detection,” in *Advances in Neural Information Processing Systems 18*, pp. 1417–1424, MIT Press, London, UK, 2005.
21. J. Devlin, H. Cheng, H. Fang, S. Gupta, Li Deng, and X. He, “Language models for image captioning: the quirks and what works,” 2015, <http://arxiv.org/abs/1505.01809> Computer Science.
22. K. Tran, X. He, L. Zhang, and J. Sun, “Rich image captioning in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 434–441, Las Vegas, NV, USA, June 2016.
23. P. Razvan, G. Caglar, K. Cho, and B. Yoshua, “How to construct deep recurrent neural networks,” 2014, <http://arxiv.org/abs/1312.6026> Computer Science.
24. T. Mikolov, M. Karafiat, L. Burget, J. “Honza” Cernocky, and S. Khudanpur, “Recurrent neural network based language model,” in *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, DBLP*, pp. 1045–1048, Chiba, Japan, September 2010.
25. C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pp. 146–152, Sunnyvale, CA, USA, September, 2016.
26. S. O. Arik, M. Chrzanowski, A. Coates, and G. Diamos, “Deep voice: real-time neural text-to-speech,” 2017, <http://arxiv.org/abs/1702.07825>.
27. S. O. Arik, M. Chrzanowski, A. Coates, and G. Diamos, “Deep voice 2: multi-speaker neural text-to-speech,” 2017, <http://arxiv.org/abs/1705.08947>.
28. T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580–587, 2015.

29. T. Hughes and K. Mierle, “Recurrent neural networks for voice activity detection, Acoustics,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7378–7382, Vancouver, Canada, May 2013.
30. P. Wei, K. Peng, G. Andrew, and J. Miller, “Deep voice 3: 2000-speaker neural text-to-speech,” 2017, <http://arxiv.org/abs/1710.07654>.
31. X. Wang, S. Takaki, and J. Yamagishi, “An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis,” in *Proceedings of the Interspeech 2017*, pp. 1059–1063, Stockholm, Sweden, August 2017.
32. K. Cho, B. van Merriënboer, C. Gulcehre, and F. Bougares, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” 2014, <http://arxiv.org/abs/1406.1078> Computer Science.
33. K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: encoder-decoder approaches,” 2014, <http://arxiv.org/abs/1409.1259> Computer Science.
34. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, <http://arxiv.org/abs/1409.0473> Computer Science.
35. L. Minh-Thang, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” 2015, <http://arxiv.org/abs/1508.04025> Computer Science.
36. G. Klein, K. Yoon, Y. Deng, and A. M. Rush, “OpenNMT: open-source toolkit for neural machine translation,” 2017, <http://arxiv.org/abs/1701.02810>.
37. Y. Wu, M. Schuster, Z. Chen, and J. Dean, “Google’s neural machine translation system: bridging the gap between human and machine translation,” 2016, <http://arxiv.org/abs/1609.08144>.
38. H. Zhang, H. Yu, and W. Xu, “Listen, interact and talk: learning to speak via interaction,” 2017, <http://arxiv.org/abs/1705.09906>.
39. B. Sherman and Z. Hammoudeh, “Make deep learning great again: character-level RNN speech generation in the style of Donald Trump,” 2017.
40. S. Mehri, K. Kumar, L. Gulrajani, and Y. Bengio, “SampleRNN: an unconditional end-to-end neural audio generation model,” 2016, <http://arxiv.org/abs/1612.07837>.

41. F. Tian, B. Gao, Di He, and T.-Y. Liu, “Sentence level recurrent topic model: letting topics speak for themselves,” 2016, <http://arxiv.org/abs/1604.02038>.
42. S.-H. Chen and C.-C. Ho, “A hybrid statistical/RNN approach to prosody synthesis for Taiwanese TTS,” in *Proceedings of the Sixth International Conference on Spoken Language Processing*, pp. 613–616, Takamatsu, Japan, October-November 2000.
43. W. Hinoshita, T. Ogata, H. Kozima, H. Kanda, T. Takahashi, and H. G. Okuno, “Emergence of evolutionary interaction with voice and motion between two robots using RNN Intelligent robots and systems,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4186–4192, St. Louis, MO, USA, October 2009.
44. Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, “Modeling spatial-temporal clues in a hybrid deep learning framework for video classification,” in *Proceedings of the 23rd ACM International Conference on Multimedia*, vol. 99, pp. 461–470, Brisbane, Australia, October 2015.
45. X. Yang, P. Molchanov, and J. Kautz, “Multilayer and multimodal fusion of deep neural networks for video classification,” in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 978–987, Amsterdam, Netherlands, October 2016.
46. Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, “Multi-stream multi-class fusion of deep networks for video classification,” in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 791–800, Amsterdam, Netherlands, October 2016.
47. S. Ilya, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014, <http://arxiv.org/abs/1409.3215>.
48. A. Graves, “Generating sequences with recurrent neural networks,” 2013, <http://arxiv.org/abs/1308.0850> Computer Science.
49. O. Vinyals, T. Alexander, S. Bengio, and D. Erhan, “Show and tell: a neural image caption generator,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Columbus, OH, USA, June 2014.
50. R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” *International Conference on Machine Learning*, vol. 52, no. 3, pp. 1310–1318, 2012.

51. W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization,” 2014, <http://arxiv.org/abs/1409.2329>.
52. K. Andrej, J. Johnson, and F.-F. Li, “Visualizing and understanding recurrent networks,” 2015, <http://arxiv.org/abs/1506.02078>.
53. X. Wang, L. Gao, and P. Wang, “Two-stream 3D convNet fusion for action recognition in videos with arbitrary size and length,” *Proceedings of the IEEE Transactions on Multimedia*, vol. 20, no. 3, 2017.
54. J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, “Self-supervised video hashing with hierarchical binary auto-encoder,” *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3210–3221, 2018.
55. X. Wang, L. Gao, J. Song, and H. Shen, “Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2016.
56. V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” *Advances in Neural Information Processing Systems*, vol. 3, pp. 2204–2212, 2014.
57. B. Dzmitry, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, <http://arxiv.org/abs/1409.0473> Computer Science.
58. M. Rush Alexander, S. Chopra, and J. Weston, “A neural attention model for abstractive sentence summarization,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.
59. M. Allamanis, H. Peng, and C. Sutton, “A convolutional attention network for extreme summarization of source code,” in *Proceedings of the Thirty-Third International Conference on Machine Learning*, New York, NY, USA, June 2016.
60. K. M. Hermann, T. Kočiský, E. Grefenstette et al., “Teaching machines to read and comprehend,” in *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, Canada, December 2015.
61. W. Yin, H. Schütze, B. Xiang, and B. Zhou, “Attention-based convolutional neural network for machine comprehension,” in *Proceedings of the Workshop on Human-Computer Question Answering*, San Diego, CA, USA, June 2016.
62. R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, “Text understanding with the attention sum reader network,” in *Proceedings*

- of the International Conference On Learning Representations, San Juan, Puerto Rico, May 2016.
- 63. B. Dhingra, H. Liu, Z. Yang, and W. William, “Cohen, and ruslan salakhutdinov, gated-attention readers for text comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1832–1846, Berlin, Germany, August 2016.
 - 64. L. Wang, C. Zhu, G. de Melo, and Z. Liu, “Relation classification via multi-level attention CNNs,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1298–1307, Berlin, Germany, August 2016.
 - 65. P. Zhou, W. Shi, J. Tian et al., “Attention-based bidirectional long short-term memory networks for relation classification,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 16, no. 2, pp. 207–212, Berlin, Germany, August 2016.
 - 66. Z. Yang, D. Yang, C. Dyer, X. He, Alex Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, San Diego, CA, USA, June 2016.
 - 67. J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. Shen, “From deterministic to generative: multi-modal stochastic RNNS for video captioning,” *IEEE Transaction on Neural Networks and Learning System*, vol. 30, no. 10, pp. 3047–3058, 2018.
 - 68. J. Song, X. Li, L. Gao, and H. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” 2018, <http://arxiv.org/abs/1812.11004>.
 - 69. K. Xu, J. Ba, K. Ryan et al., “Show, attend and tell: neural image caption generation with visual attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2048–2057, Boston, MA, USA, June 2015.
 - 70. A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Proceedings of the Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.
 - 71. L. Minh-Thang, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.

72. Z. Yang, X. He, J. Gao, Li Deng, and Alex Smola, “Stacked attention networks for image question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.
73. C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in *Proceedings of the International Conference on Machine Learning*, pp. 21–29, IEEE Computer Society, New York, NY, USA, June 2016.
74. J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 289–297, Barcelona, Spain, December 2016.
75. J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: adaptive attention via a visual sentinel for image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3242–3250, Las Vegas, NV, USA, June-July 2016.
76. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, Las Vegas, NV, USA, June-July 2016.
77. L. Chen, H. Zhang, J. Xiao et al., “SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–6306, Las Vegas, NV, USA, June-July 2016.
78. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiari, “Visual saliency for image captioning in new multimedia services,” in *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 309–314, Hong Kong, China, July 2017.
79. L. Gao, K. Fan, J. Song, X. Liu, X. Xu, and H. Shen, “Deliberate attention networks for image captioning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8320–8327, Honolulu, HI, USA, January-February 2019.
80. X. Chen, H. Fang, T.-Yi Lin et al., “Microsoft COCO captions: data collection and evaluation server,” 2015, <http://arxiv.org/abs/1504.00325> Computer Science.

81. M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.
82. B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 123, pp. 74–93, Boston, MA, USA, June 2015.
83. C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmainer, “Collecting image annotations using Amazon’s Mechanical Turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pp. 139–147, Los Angeles, CA, USA, June 2010.
84. Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, “Stair captions: constructing a large-scale Japanese image caption dataset,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 417–421, Vancouver, Canada, July 2017.
85. P. Kishore, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, PA, USA, July 2002.
86. S. Banerjee and L. Alon, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, pp. 65–72, Ann Arbor, MI, USA, June 2005.
87. C.-Y. Lin, “ROUGE: a package for automatic evaluation of summaries,” in *Proceedings of the Text Summarization Branches Out, Workshop on Text Summarization Branches Out*, Barcelona, Spain, July 2004.
88. R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: consensus-based image description evaluation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, Boston, MA, USA, June 2015.
89. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: semantic propositional image caption evaluation,” in *Computer Vision—ECCV 2016*, vol. 11, no. 4, pp. 382–398, Springer, Cham, Switzerland, 2016.

90. J. L. Ba, M. Volodymyr, and K. Koray, “Multiple object recognition with visual attention,” 2014, <http://arxiv.org/abs/1412.7755> Computer Science.
91. M. Volodymyr, H. Nicolas, A. Graves, and K. Koray, “Recurrent models of visual attention,” *Neural Information Processing Systems*, vol. 3, pp. 2204–2212, 2014.
92. F. Qiao, C. Wang, X. Zhang, and H. Wang, “Large scale near-duplicate celebrity web images retrieval using visual and textual features,” *The Scientific World Journal*, vol. 2013, Article ID 795408, 11 pages, 2013.
93. S. Lei, G. Xie, and G. Yan, “A novel key-frame extraction approach for both video summary and video index,” *Recent Advances on Internet of Things*, vol. 2014, Article ID 695168, 9 pages, 2014.
94. S. Lee and I. Kim, “Multimodal feature learning for video captioning,” *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 3125879, 8 pages, 2018.
95. A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, “Vision-based fall detection with convolutional neural networks,” *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 9474806, 16 pages, 2017.

CHAPTER 3

Application of an Improved DCGAN for Image Generation

Bingqi Liu^{1,2}, Jiwei Lv², Xinyue Fan², Jie Luo², and Tianyi Zou²

¹School of Mechanical Engineering, Chengdu University, Chengdu 610106, China

²Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

ABSTRACT

With the rapid development of deep learning, image generation technology has become one of the current hot research areas. A deep convolutional generative adversarial network (DCGAN) can better adapt to complex image distributions than other methods. In this paper, based on a traditional generative adversarial networks (GANs) image generation model, first, the fully connected layer of the DCGAN is further improved. To solve the

Citation: B. Liu, J. Lv, X. Fan, J. Luo, T. Zou, “Application of an Improved DCGAN for Image Generation”, Mobile Information Systems, vol. 2022, Article ID 9005552, 14 pages, 2022. <https://doi.org/10.1155/2022/9005552>.

Copyright: © 2022 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

problem of gradient disappearance in GANs, the activation functions of all layers of the discriminator are LeakyReLU functions, the output layer of the generator uses the Tanh activation function, and the other layers use ReLU. Second, the improved DCGAN model is verified on the MNIST dataset, and simple initial fraction (ISs) and complex initial fraction (ISc) indexes are established from the two aspects of image quality and image generation diversity, respectively. Finally, through a comparison of the two groups of experiments, it is found that the quality of images generated by the DCGAN model constructed in this paper is 2.02 times higher than that of the GANs model, and the diversity of the images generated by the DCGAN is 1.55 times higher than that of GANs. The results show that the improved DCGAN model can solve the problem of low-quality images being generated by the GANs and achieve good results.

INTRODUCTION

With the introduction of the concepts of cloud computing and big data and the rapid development of computer hardware facilities, deep learning has undergone rapid development and has been used in many applications in recent years [1]. However, the development of image generation technology is slow in several branches of deep learning. Before GANs were proposed, the main image generators were automatic regression models [2] and variational autoencoders [3]. At the same time, based on the improvements in GANs theory, GANs have been applied in image conversion, image feature extraction, and other fields.

There are many models used in image generation and modeling research, including BPT-CNN [4], the BERT-based deep spatial-temporal network [5], GeNet of deep convolutional neural network [6], and RNN-LSTM [4]. However, as a new type of image generation model, GANs have attracted the attention of many researchers, who have gradually improved and provided a large number of mature image generation frameworks (such as DCGAN, CGAN, Pix2Pix, etc.). In terms of theoretical research on GANs, in 2014, Goodfellow et al. [7] first described a new image generation model, the GANs, which is composed of a generator and a discriminator [7]. In the same year, Mirza and Osindero [8] were inspired by the introduction of convolutional neural networks on the basis of GANs and proposed the CGAN, which solved the unstable training behavior problem of GANs by adding category labels [8]. To solve the instability of GANs, in 2016, Goodfellow et al. [7] and Salimans et al. [9] proposed stabilizing the training process of DCGAN

with feature matching, small batch recognition, and historical averaging, and this work provided a basis for follow-up research [9]. With regard to the applications of GANs, Isola et al. [10] implemented image conversion using Pix2Pix and paired training data [10]; Zhang et al. [11] proposed StackGAN, which first generates basic images and text descriptions based on the original image information and then improved the process to generate high-resolution images [11]. In 2017, Zhu et al. [12] proposed CycleGAN, which solved the problem of Pix2Pix needing paired data and proposed the cycle-consistency loss function to realize image conversion from a horse to a zebra [12]. Karras et al. [13] proposed StyleGAN in 2018 to accelerate and stabilize the training speed of the network by gradually increasing the numbers of generators and discriminators [13]. This method uses natural style-conversion technology, such as adaptive instance normalization (AdaIN), for reference purposes and realizes the real-time transmission of any style [14]. BigGAN was proposed by Andrew Brock in 2019. BigGAN adopts a self-attention mechanism and spectral normalization, and it is a good model for image generation on ImageNet at present [15]. In summary, there have been some comprehensive methods proposed to solve the problems of GANs generation and resolution in recent years.

Alec Radford et al. [16] used the CNN structure [17, 18] to implement the GANs model for the first time and proposed the DCGAN [16]. Liu et al. compared the unconstrained DCGAN and the constrained DCGAN, and the results showed that after adding constraints during the training phase, the DCGAN model significantly improved upon the results of the virtual face generation model, thus demonstrating the enhanced ability of the generator and discriminator [19].

Mahmoud and Guo [20] used the DCGAN to extract depth features for strongly representing TSR images [20]. Fang et al. [21] proposed a new gesture recognition algorithm based on a convolutional neural network and the DCGAN and applied this method to expression recognition, calculation, and text output, achieving good results in all cases [21]. The actual image and noise vector of the DCGAN were trained, and smoke image training was used to generate a discriminator, thereby showing that the DCGAN can effectively monitor smoke images [22]. The biggest differences between the DCGAN and original GANs are that the DCGAN uses a convolutional neural network (CNN) to replace the multilayer perceptron in the original GANs, removes the pooling layer, and uses a convolution with a defined step size to replace the upper sampling layer for improving the stability of the training process.

To solve the problem of the easily disappearing gradient in GANs, this paper further improves the DCGAN fully connected layer. For the activation functions of all layers of the discriminator, generator output, and other layers, LeakyReLU, Tanh, and ReLU functions are used, respectively, and the open source MNIST dataset is used. For verification, we use the improved DCGAN and GANs for comparison and establish ISs and ISc from the two aspects of image quality and image generation, respectively. According to the numerical value, it can be concluded that the improved DCGAN has better image processing ability. The remainder of the paper is organized as follows: Section 1 summarizes the progress of research with regard to GANs and the DCGAN; Section 2 mainly introduces the principles of the improved DCGAN algorithm and designs the network structure; Section 3 constructs the image generation models, with one based on GANs and the other based on the DCGAN; Section 4 introduces two image generation quality evaluation methods and analyzes the effects of the two models on image generation quality and image diversity. The study's discussion and conclusions are presented in Section 5.

IMPROVED DESIGN OF THE STRUCTURE OF THE DCGAN

Principles of the DCGAN Algorithm

The adversarial training process of the DCGAN model established in this paper was calculated by using

$$\begin{aligned} \min_G \max_D V(D, G) = & E_{x \sim p_{\text{data}}(x)} \log[D(x)] \\ & + E_{Z \sim P_Z(Z)} [\log(1 - D(G(z)))] \end{aligned} \quad (1)$$

where x is the distribution of the real data, x is the sample image data, P_Z represents arbitrarily distributed noise, Z expresses the number of random vectors in P_Z , and E expresses expectations.

The first step is to find the minimum cross-entropy of the discriminator D under the condition where a generator G is given. The objective function was calculated by using

$$\text{Obj}(D) = -E_{x \sim p_{\text{data}}} \log[D(x)] - E_{Z \sim P_Z} \log[1 - D(G(Z))], \quad (2)$$

where $\log[D(x)]$ is used to judge the sample data, $\log[1 - D(G(Z))]$ represents the judgment of the generated sample data, that is, the closeness

of the distribution of the sample data output by the discriminator P_{data} and the data distribution generated by $GP_{G(x)}$, and x is a sample from the real data, according to

$$\begin{aligned} \text{Obj}(\theta_D, \theta_G) &= - \int_x P_{\text{data}}(x) \log(D(x)) dx \\ &\quad - \int_z P_Z(z) \log(1 - D(G(z))) dz \\ &= \int_x [P_{\text{data}} \log(D(x)) + P_G(x) \log(1 - D(x))] dx. \end{aligned} \quad (3)$$

At this time, because the data and generator have been given, they can be regarded as constants. Assuming that the data and generator are replaced, then

$$f(D) = c_1 \log D + c_2 \log(1 - D). \quad (4)$$

Let $f(D) = 0$ in (1); then it is the maximum point:

$$D * (x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)}. \quad (5)$$

The second step is to fix the discriminator D . At this time, the optimization function for the generator G was calculated by using

$$V(G, D) = E_{x \sim P_{\text{data}}} \log D(x) + E_{x \sim P_G} \log[1 - D(x)]. \quad (6)$$

Furthermore, using (2), $D *$ brings $V(G, D)$ as the optimal solution of the generator, which can be calculated by using

$$\begin{aligned} \min_G V(G, D) &= V(G, D *) \\ &= E_{x \sim P_{\text{data}}}(x) \left[\log \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)} \right] \\ &\quad + E_{x \sim P_G}(x) \left[\log \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)} \right]. \end{aligned} \quad (7)$$

In practical training, the discriminator D is usually trained first. Then, the discriminator D is fixed and the generator G is trained. Next, we continue to fix G and train the discriminator D , performing iterative optimization training until $P_{\text{data}} = P_G$, at which point global optimization is achieved.

Design of the Structure of the DCGAN

Compared with traditional GANs, the salient feature of the DCGAN is that a CNN is used to replace the multilayer perceptron. The pooling layer and

sampling layer are removed in the CNN model. The convolution layer is used to discriminate the image in the discriminator, and the deconvolution layer is used to generate the image in the generator. The specific structure of the DCGAN generator is as follows: the input layer is followed by a batch normalization layer (which can hasten the convergence of the model), and the reshaping layer is used to normalize the preliminary data; then, an upsampling layer, a Conv2DTranspose layer, and a batch normalization layer are used to sample, deconvolute, and normalize the data, respectively. In this paper, the DCGAN adds three groups of the above structures to increase the depth of the network. The main framework of the network architecture of the generator is shown in Figure 1.

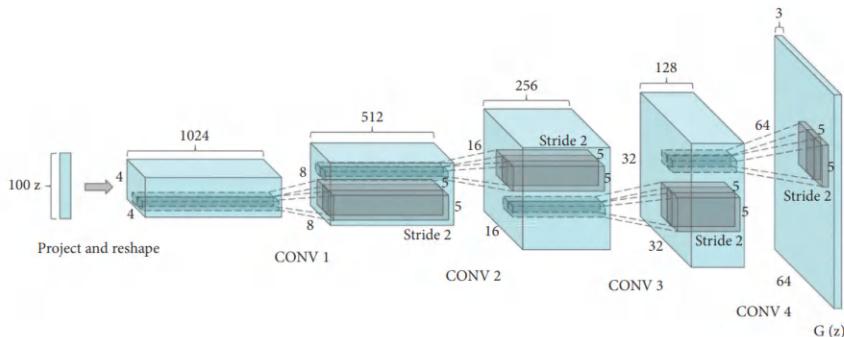


Figure 1. Network structure of the DCGAN generator [23].

In this paper, to solve the problem in which gradients disappear easily, the LeakyReLU activation function is used in all layers of the discriminator, and the *Tanh* activation function is used in the output layer of the generator, where definition of this function is

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}. \quad (8)$$

For the generator, except for the activation function of the last layer, the ReLU activation function is used, and its definition is

$$f(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (9)$$

In addition, the concrete structure of the DCGAN discriminator is a Conv2D layer (2D convolution layer), a batch normalization layer, and

a dropout layer (after the image is convoluted, normalization process is continued, and the dropout layer is added to increase the generalization ability of the model). These three layers form a group, and four groups are added. Finally, a flattening layer and a fully connected layer are used to flatten the data and output the probability of whether it is sample data or generated data. Except for that of the last layer, the activation function of the other layers is the LeakyReLU function. The definition of this function is

$$f(x) = \begin{cases} x, & x \geq 0, \\ ax, & x < 0, \end{cases} \quad (10)$$

where x is a sample from the real data and a is the relevant parameter.

The activation function of the last layer adopts the sigmoid function. The definition of this function is

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \quad (11)$$

CONSTRUCTION OF THE IMAGE GENERATION MODELS BASED ON THE DCGAN AND GANS

Data Sources

This paper uses the MNIST dataset, which is free of charge, is open source, contains small pixels, and includes a large number of points [24]. It is composed of 250 handwritten digits (0–9, a total of 10 digits); it is relatively mature in image processing and image quality processing as shown in Figure 2.



Figure 2. Example from the MNIST dataset.

To universalize the dataset, 50% of the data in the dataset are from high school students, and 50% are from Census Bureau staff. At the same time, this paper uses the Keras framework with Tensorflow as the back end. The Keras framework is an open source artificial neural network library written in

Python. The code structure is written with an object-oriented method, which is completely modular and extensible. It is suitable for the implementation framework of the code in this experiment.

Determination of the DCGAN Model Parameters

The collected basic data is used in the DCGAN model for experimental research, and the network parameters are determined through continuous testing. The situation is as follows: the model generator accepts 1×100 random, normally distributed noise data. Considering that the layout of the pixels of the image generated by deconvolution is 28×28 (i.e., MNIST dataset image pixels), the input layer is set as a fully connected layer, and the number of neurons is $7 \times 7 \times 256$. After that, a batch normalization layer and reshaping layer are added, and then the size of the data pixel is expanded through the upsampling layer. Using a 5×5 convolution kernel, the conv2dspread layer uses the “same” border model to preserve the convolution results at the boundary, so that the input and output dimensions are the same. The upsampling layer is added after the first module to make the pixel size of the output image 28×28 . The activation function in this layer uses the ReLU function, and the last layer uses the Tanh function.

In the discriminator, a Conv2D layer (2D convolution layer), a batch normalization layer, and a dropout layer are added as a module, and four modules are added. Finally, a flattening layer and a fully connected layer are used as the back end. The Conv2D layer uses a 5×5 convolution kernel, the boundary mode of the convolution layer is the same as that in the generator, and the activation function is the LeakyReLU function. The dropout layer in the module makes the activation value of a certain neuron have a certain probability p when it propagates forward. This can make the model independent of some local features and enhance the generalization ability of the model. The last activation function uses the sigmoid function, which can output the probability of discrimination, that is, the probability that the discriminator thinks the image belongs to the real image and not a generated image.

The discriminator uses the Adam optimizer with a learning rate of 0.0002, and the GANs use the RMSprop optimizer with a learning rate of 0.0001. The RMSprop optimizer combines the exponential moving average of the square of the gradient to adjust the learning rate. It can converge effectively under an unstable objective function and yields good results with the DCGAN model. The batch size is 32, and the training time of each DCGAN iteration is 10000. The binary cross-entropy function is selected as the loss function.

The training process of the DCGAN is slightly different from that of the GANs because it takes more time to train the discriminator of the DCGAN model. First, when training the discriminator, the input data size is $2 \times$ batch, where the input data contains the real data and generated data from one batch; second, the combined data of size $2 \times$ batch are used as the input data to train the discriminator 5 times; finally, the whole GAN model is trained once with the random noise data from one batch as the input data, and only the generator is updated. A complete training cycle is a batch (epoch) in which the ratio of discriminator training iterations to generator training iterations is 1 : 5 to realize the alternating iterative training process. The main parameter configuration of the DCGAN is shown in Table 1.

Table 1. Main parameters of the DCGAN

| Parameter | Value |
|--------------------------------|------------|
| BATCHSIZE | 32 |
| EPOCH | 100,000 |
| LEARNINGRATE (GANs) | 0.0001 |
| DECAY (GANs) | 0.00000003 |
| LEARNINGRATE (Discriminator) | 0.0002 |
| DECAY (Discriminator) | 0.00000006 |
| ALPHA (LeakyReLU) | 0.2 |
| MOMENTUM (Batch normalization) | 0.9 |
| DROPOUT | 0.3 |
| STRIDES (Conv2D) | 2 |
| OUTPUTWIDTH | 28 |
| OUTPUTHEIGHT | 28 |
| OUTPUTCHANNEL | 1 |
| LATENTSIZE | 100 |

Determination of the Parameters of the GAN Model

The noise data received by the generator are the same as above, and a basic multilayer perceptron is used to add these three modules as a fully connected layer, a LeakyReLU layer, and a batch normalization layer; the model ends with a fully connected layer and a reshaping layer. The activation function of the last layer uses the Tanh function. The discriminator uses a flattening layer to flatten the data and then adds two fully connected layers. The activation function also uses the LeakyReLU function, and the last activation function

uses the sigmoid function to output the discrimination probability. Both the GANs and the discriminator use Adam as their optimizer, and the learning rate is 0.002. To calculate the update step size, the Adam optimizer comprehensively considers the first-order moment estimation (average value of the gradient) and second-order moment estimation (noncentral variance of the gradient). The batch size is 32. Due to the slow convergence speeds of GANs, the model can avoid falling into local optimal solutions and undergo training 100K times iteratively. The binary cross-entropy function is selected as the loss function. The main parameter configuration of the GANs is shown in Table 2.

Table 2. Main parameters of the GANs

| Parameter | Value |
|--------------------------------|-------------|
| BATCHSIZE | 32 |
| EPOCH | 100,000 |
| LEARNINGRATE (GANs) | 0.0002 |
| LEARNINGRATE (Discriminator) | 0.0002 |
| DECAY (Discriminator) | 0.000000009 |
| ALPHA (LeakyReLU) | 0.2 |
| MOMENTUM (batch normalization) | 0.8 |
| OUTPUTWIDTH | 28 |
| OUTPUTHEIGHT | 28 |
| OUTPUTCHANNEL | 1 |
| LATENTSIZE | 100 |

During training, the discriminator is trained once, and the input data are half true and half false; i.e., the input dataset is composed of half real data and half batch-generated data. Such a complete combined dataset is used as input data to train the discriminator once. Then, the whole GAN model is trained once with a batch of random noise data as input, and only the generator is updated. Such a training cycle is a batch (epoch), which is performed to realize the alternating iterative training process.

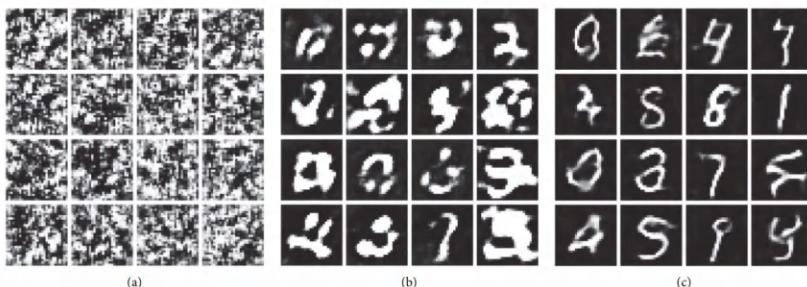
Experimental Results of the Image Generation Models Based on the GANs and DCGAN

There are experimental groups in the training experiments of the image generation models based on the GANs and DCGAN. Each experimental

group is divided into Experiment 1 (composed of 250 handwritten digits (0–9, a total of 10 digits)) and Experiment 2 (only using the number “6” in the dataset). The experiment under the unified experimental group is conducted to observe the learning effects of different networks with different image distribution complexity and image generation quality; the experiment with different experimental groups using the same dataset is done to compare the advantages and disadvantages of the GANs and DCGAN with regard to image generation. The experiments covered in this article require the use of the following: CPU frequency: 2.5 GHz, memory capacity: 16 GB, graphics chip: NVIDIA GeForce RTX 3070, and hard disk capacity: 512 GB.

Training Results of the Image Generation Model Based on the DCGAN

(1) *Experiment 1.* Set checkpoints in the training process, run the training process for 10 h in the local environment, and observe the output results once every 50 training iterations. The results are shown in Figure 3, in which Figures 3(a)–3(d) are the results of the original noise image, the results after 1K training iterations, the results after 5K training iterations, and the results after 10K training iterations, respectively. By observing the output at each checkpoint, we can find that the original noise data are disordered. After 1K iterations, the image is gradually regionalized, and the only content is in the central area. However, most of the digital contours cannot be clearly recognized, and obvious characteristics of the deconvolution layer can be found. The image learning process includes regionalization and characterization rather than pixel learning, similar to the process of the GANs. Training 5K times can yield a gradually clearer line for which the numbers can be identified but not recognized clearly. After training 10K times, each number can be clearly identified, clear images are generated, and the complex image distribution is successfully fitted.



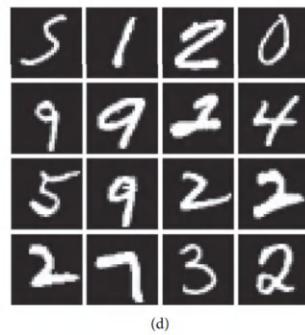
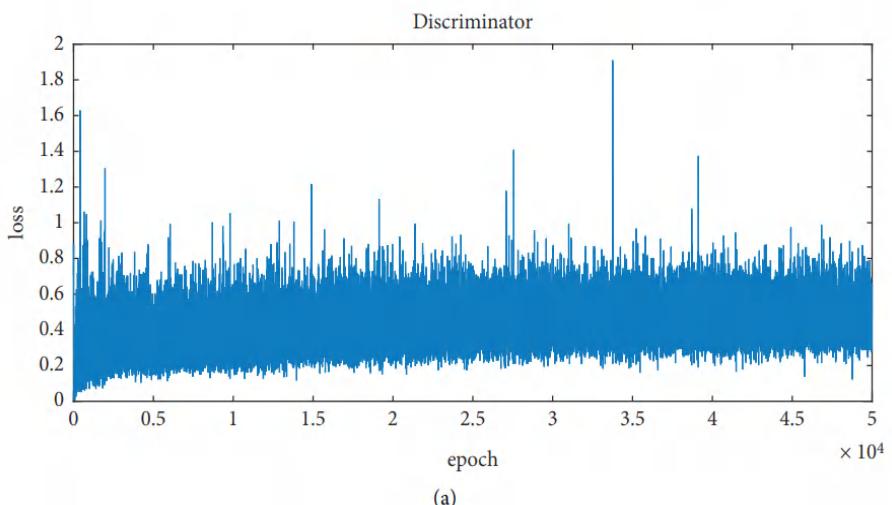


Figure 3. DCGAN partial results (Experiment 1). (a) Training 0 times. (b) Training 1000 times. (c) Training 5000 times. (d) Training 10000 times.

From the loss images of the discriminator (Figure 4(a)) and the generator (Figure 4(b)), it can be seen that the discriminator stabilizes at approximately 0.5 for a very short batch. For the 1K batch, the loss of the generator decreases to 2, then approaches 1 slowly, and finally stabilizes at approximately 1.5, but there is a downward trend. Because of this configuration, the experiment cannot continue; even if the Nash equilibrium point is not reached, the effect of the output image is still very good, and this shows that the DCGAN can obtain better experimental results than those of other methods under the premise of satisfying the computational power requirements.



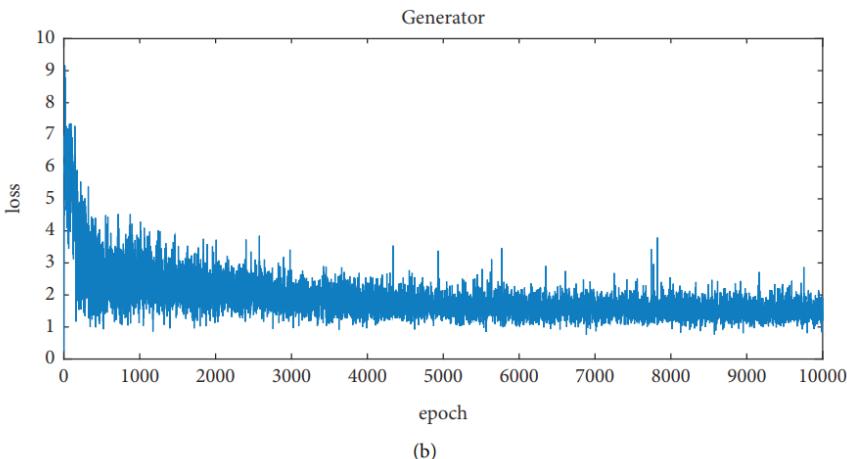
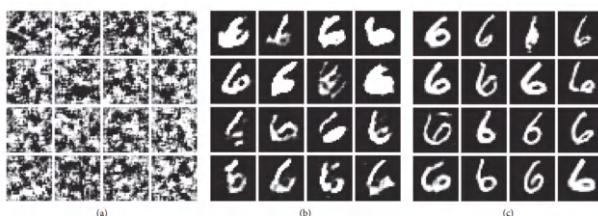
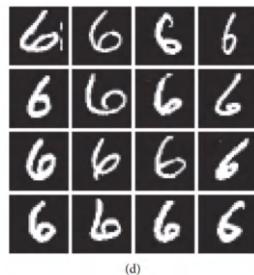


Figure 4. Loss curves of the DCGAN discriminator (a). Generator (b) (Experiment 1).

(2) *Experiment 2.* Run the same training process as in Experiment 1 in the local environment for 10 h; the results are shown in Figure 5, in which Figures 5(a)–5(d) are the results of original noise image, the results after training 1K times, the results after training 5K times, and the results after training 10K times, respectively. By observing the output of each checkpoint, it can be found that the original noise is the same as in Experiment 1. When iterating 1K times, the number “6” can be identified, such as the first number from the left in the second row and the fourth number from the left in the third row; most of the number “6” can be recognized after iterating 5K times. All the numbers can be recognized after 10K iterations, but the numbers are not standardized, mainly because the MNIST dataset contains handwritten numbers from adults and children.





(d)

Figure 5. Partial results of the DCGAN (Experiment 2). (a) Training 0 times. (b) Training 1000 times. (c) Training 5000 times. (d) Training 10000 times.

According to the loss images of the discriminator (Figure 6(a)) and generator (Figure 6(b)), the discriminator D starts to fluctuate at approximately 0.5 at 1000 epochs (the ratio of the training times of the discriminator and generator is 5 : 1); the generator also starts to stabilize at approximately 2 at 1K epochs, gradually converges to 1, and finally fluctuates at approximately 1 with a small fluctuation range.

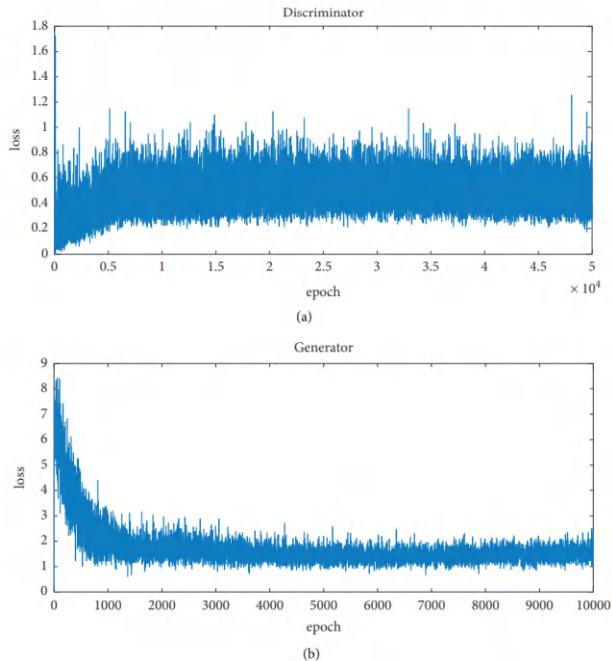


Figure 6. Loss curves of the DCGAN discriminator (a). Generator (b) (Experiment 2).

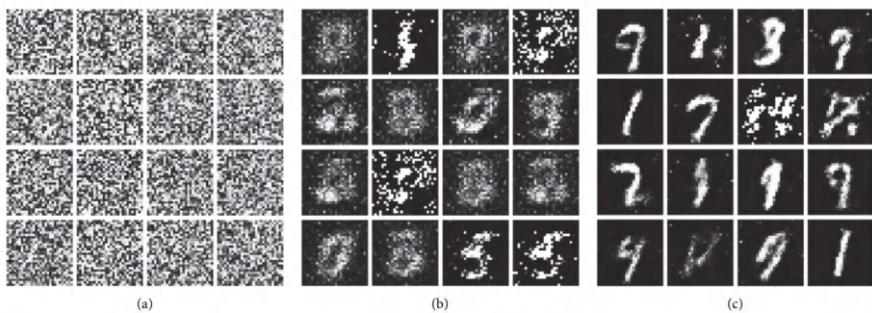
Through the comparative training processes of the two experiments, it is found that the DCGAN model can adapt to both complex image distributions and simple image distributions, and it can converge earlier than other methods. Its learning law is characterized and regionalized. At the beginning, it concentrates on the central area, then learns the line features, and finally learns the location characteristics of the lines.

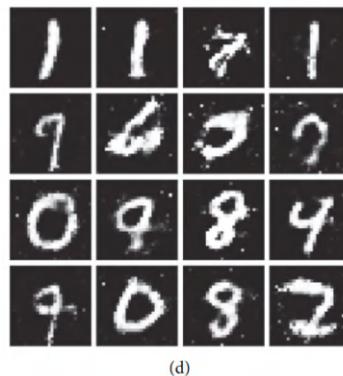
Training Results of the Image Generation Model Based on GANs

(1) *Experiment 1.* Set checkpoints in the training process, run the training process for 5 h in the experimental environment, and observe the output results once every 500 training iterations.

The results are shown in Figure 7, in which Figures 7(a)–7(d) are the results of the original noise image, the results after training 10K times, the results after training 50K times, and the results after training 100K times, respectively. By observing the output at each checkpoint, we can find that the original noise data are disordered.

After 10K iterations, the image is gradually focused in the central area rather than at scattered points in each position. After iterating 50K times, some figures have a preliminary outline, indicating that the generator is gradually learning the image distribution of the original dataset. Continuing to train for a total of 100K iterations, one can find that most of the figures are clear and distinguishable. This situation lasts nearly 20K batches during the training process, indicating that the GANs maintain stability for a long time but only reach the pseudo-Nash equilibrium point. It is found that adding a hidden layer does not affect the experimental results but rather increases the experimental time.

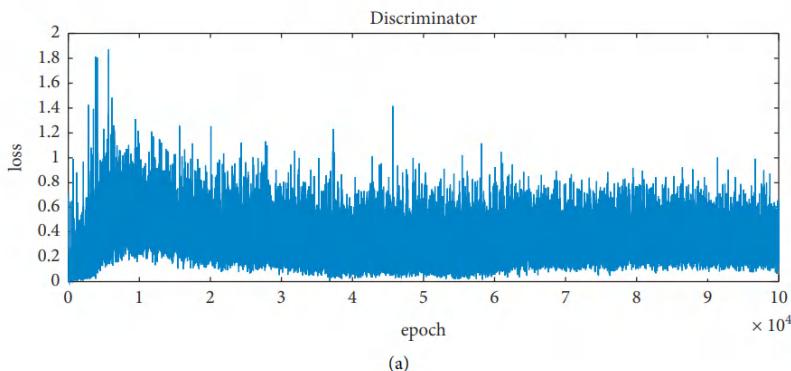




(d)

Figure 7. Partial results of the GANs (Experiment 1). (a) Training 0 times. (b) Training 10000 times. (c) Training 50000 times. (d) Training 100000 times.

By observing the loss function images of the discriminator (Figure 8(a)) and generator (Figure 8(b)), it is not difficult to see that the loss of the generator decreases rapidly in each of the first 10K batches, then gradually approaches 1, fluctuates greatly between 30K and 70K iterations, and finally stabilizes near 1. This shows that the image generated by the generator can make the discriminator think that it is true. However, the discriminator gradually becomes stable at approximately 0.5 starting with the 15000th batch, and then it fluctuates up and down. As the number of batches increases, the fluctuation range does not decrease, and the overall loss is slightly less than 0.5, indicating that the discriminator has a high probability of correct discrimination (recognition of the real image). After 100K iterations, the Nash equilibrium point cannot be reached, but the model cannot be further converged.



(a)

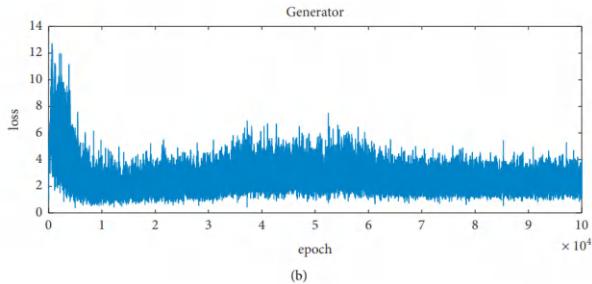


Figure 8. Loss curves of the GANs discriminator (a). Generator (b) (Experiment 1).

(2) *Experiment 2.* The running time of the training process is 5 h in the local environment, and the results are shown in Figure 9, where Figures 9(a)–9(d) are the results of the original noise image, the results after training 10K times, the results after training 50K times, and the results after training 100K times, respectively. The original noise is the same as in Experiment 1, and the learning process of the GANs can still be seen in (b) and (c), but the speed of the pixel setting is much higher than that in Experiment 1. Compared with Figures 9(d) and 7(d) of Experiment 1, it can be found that when iterating 100K times, GANs can effectively fit the simple image distribution. The second digit on the left in the second row of Figure 7(d) generates “6”, but this is not as effective as simply learning “6”. In Figure 9(d) of Experiment 2, all the numbers “6” can be clearly identified.

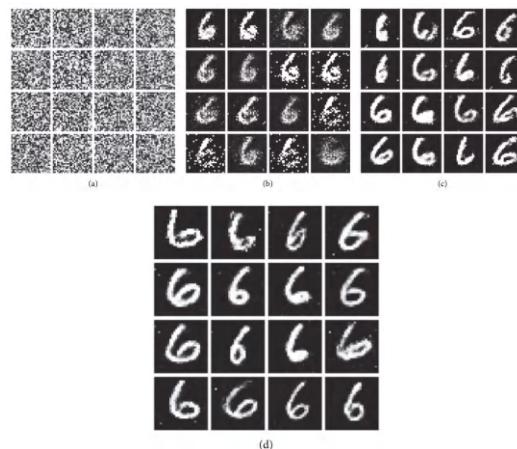


Figure 9. Partial results of the GANs (Experiment 2). (a) Training 0 times. (b) Training 10000 times. (c) Training 50000 times. (d) Training 100000 times.

Comparing the loss function images of the discriminator (Figure 10(a)) and generator (Figure 10(b)), it is not difficult to find that the generator loss rapidly decreases in each of the first 20K batches, gradually approaches 1, and stabilizes near 1. This shows that the image gradually generated by the generator can make the discriminator think it is true. Since the 20000th batch, the discriminator gradually stabilizes at approximately 0.5 and gradually reduces its fluctuation range, but the overall loss is slightly higher than 0.5, which indicates that the discriminator cannot correctly judge that the generated image is real, and the loss curve still cannot converge at 0.5; that is, it cannot reach the real Nash equilibrium point.

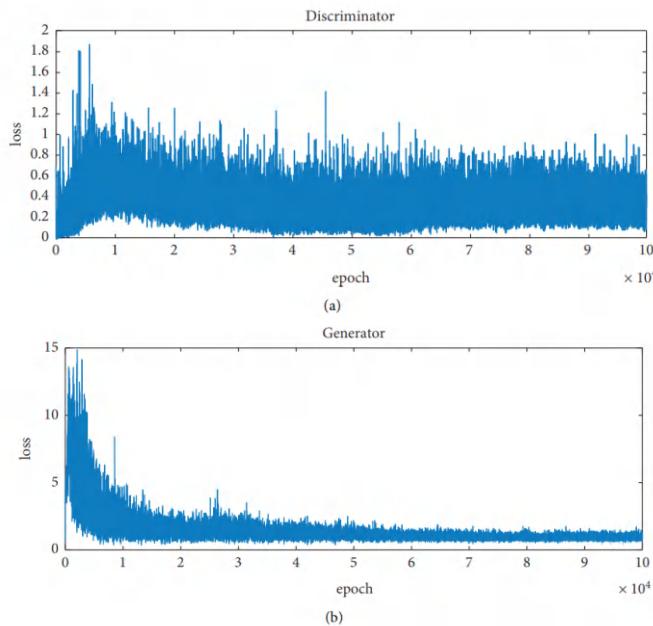


Figure 10. Loss curves of the GANs discriminator (a). Generator (b) (Experiment 2).

COMPARATIVE ANALYSIS AND ASSESSMENT OF IMAGE GENERATION QUALITY

Construction of the Evaluation Index for Image Generation

When comparing and analyzing the quality of images generated by models, two factors are generally considered: one is the textures of the images,

and the other is the diversity of image generation. At present, the popular quantitative evaluation method for image texture is the inception score (IS). In this paper, the simple initial fraction (ISs) is used to evaluate image quality, and the complex initial fraction (ISc) is used to evaluate image diversity.

Image Quality Assessment

The calculation formula of the ISs evaluation index using the simple initial score method is

$$IS_s(G) = \exp\left(E_{x \sim p_G} D_{KL}(p(y|x) \| p(y))\right), \quad (12)$$

where $x \sim p_G$ represents the image generated by the generator to be evaluated, $p(y|x)$ represents the probability that the picture belongs to each category, and $p(y)$ is the probability distribution of the image to be evaluated.

The following formula can be obtained by further derivation:

$$\ln(IS_s(G)) = H(y) - H(y|x). \quad (13)$$

It can be seen from (13) that the larger the ISs evaluation index, the greater the differences between the $p(y|x)$ and $p(y)$ distributions.

If only the ISs comprehensive analysis method is used, there may be a large error. In this paper, a concept classification network is proposed to eliminate the error as much as possible. In the MNISTdataset, there are 10 numbers from 0 to 9, and each image is a black and white image (the value is composed of 0s and 1s). For any two random variables X, Y, the correlation coefficient is calculated as follows

$$\tau(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}, \quad (14)$$

where $\text{Cov}(X, Y)$ in (14) is the covariance of X and Y, $\text{Var}[X]$ is the variance of X, and $\text{Var}[Y]$ is the variance of Y.

According to the principle of the concept classification network, it is not difficult to know that if the image correlation is weak, the classifier easily divides it into two categories; otherwise, it is easier to divide it into one category. When considering the diversity of images, we should consider the distribution of the labels. If the distribution of complex images is complex, we naturally want the labels to be evenly distributed instead of generating a certain kind of image. For example, when all the numbers in the MNIST dataset are used, a good situation is that the generated 0 s–9 s are evenly

distributed, rather than the distribution containing more of certain numbers than others. In this case, we need to consider the edge probability $p(y)$. In the ideal state, the expansion can yield $p(y_1) = p(y_2) = \dots = p(y_n) = 1/n$, where n is the number of classes in the original training data; the greater the entropy of $p(y)$, the better the situation. If the same kind of data is used, that is, a single number is used, although these data would be highly correlated and belong to the same class, because of the characteristics of the concept classification network, the set would still be divided into several categories, but $\lim p(y_i) = 1$ and $\lim p(y_i) = 0$ where $i = 2, 3, \dots, n$. At this time, n is the number of classification categories for the concept classification network.

Image Diversity Assessment

The complex initial fraction (ISc) is based on the simple initial fraction method. According to the image classification value of the simple distribution, the entropy of the simple image distribution under discrete conditions is calculated, where H in (15) represents entropy, and the calculation formula is as follows:

$$H(p(y_i)) = \sum_{i=1}^n p(y_i) \times \log(p(y_i)). \quad (15)$$

The edge probability value of the simple image distribution is included when $i=1$; using (16), we calculated the potential:

$$H(p(y_i)) = H(p(y_1)) = 0. \quad (16)$$

When $i = 2, 3, \dots, n$, $\lim p(y_i) = 0^+$:

$$H(p(y_i)) = \sum_{i=1}^n p(y_i) \times \log(p(y_i)) = 0. \quad (17)$$

Thus, the entropy limit of the concept classification network is obtained when the image tends to be simple distribution, i.e., when $\lim H(p(y)) = 0$ in (17). +en, the formula of the ISc evaluation

$$\ln(IS_C(G)) = H(y|x), \quad (18)$$

where $H(y|x)$ in (18) represents the probability that an image belongs to a certain category. The higher the value, the higher the image quality. Therefore, when the image distribution is simple, the ISc under a complex distribution can be used to comprehensively evaluate the diversity of the images. The diversity of image generation can be combined with two experiments and a

comprehensive ISc analysis. The two groups of experiments discuss image generation under different complexities of the image distribution. If the ISc value can still reach a high value under the condition of a complex image distribution, this shows that the diversity of image generation is very high.

Comparative Analysis of Experimental Results

According to the image quality evaluation method proposed in Section 4.1, two groups of experiments are compared and analyzed. For the MNIST dataset, first, 10K groups of images are generated by the generator, all data are divided into 10 pieces, and each piece is calculated and averaged. Second, the concept classification network is built to calculate the ISc and ISs of the experimental GANs and the DCGAN, respectively. The experimental results are shown in Table 3. The first experiment is conducted to evaluate the image diversity, and the second is performed to evaluate the image quality.

Table 3. Experimental results of the image quality comparison

| Experimental group | Model | IS (28×28) |
|--------------------|-------|-----------------------|
| Group 1 | GANs | ISc: 4.55 |
| | DCGAN | ISc: 6.10 |
| Group 2 | GANs | ISs: 4.80 |
| | DCGAN | ISs: 6.82 |

According to the experimental results in Table 3, in terms of image quality, the ISs value of the DCGAN model is 2.02 higher than that based on the GANs model; in terms of image generation diversity, the ISc value based on the DCGAN model reaches 6.10, which is 1.55 higher than that based on the GANs model. The results show that the improved DCGAN model has more advantages than the GANs model, can effectively solve the problem of low-quality images being output by the GANs model, and achieves good results.

DISCUSSION AND CONCLUSIONS

Based on the GANs model and the improved DCGAN model, this paper uses the MNIST dataset as experimental data for experiments on the algorithms and evaluates the quality and diversity of the generated images based on the ISs and ISc metrics:

(1)This paper compares and analyzes the different performances of traditional GANs and the DCGAN in two groups of experiments. The DCGAN is a model based on a combination of GANs and a convolutional neural network. The fully connected layer is replaced by a convolution layer and deconvolution layer. The structure of the DCGAN layers is redesigned: An upsampling layer is used in the output layer of the generator to expand the data, and a dropout layer is added in each layer of the discriminator. To solve the problem that the gradient easily disappears in GANs, the generator output process uses a beneficial *Tanh* function; the *ReLU* function is used in the other layers of the generator, and the *LeakyReLU* function is used in all layers of the discriminator. In the two groups of comparative experiments, it can be concluded that the images of the GANs model are generated in a column because of the flattening layers and reshaping layers. The image generation method using the DCGAN constructed in this paper involves generating regional and characteristic images by using a conv2dspread layer (two-dimensional anticonvolution layer), which fundamentally solves the problem of low-quality images being generated by the GANs. (2)In this paper, through the optimized DCGAN model, it is proven that the variables of a simple image distribution tend to be independent of each other, thereby overcoming the error caused by the independence of traditional indexes. The ISs value under the simple image distribution is taken as the image quality evaluation result, and the ISc value of the complex image distribution is combined with it to comprehensively evaluate the diversity and quality of the generated image. Through two groups of experiments, it can be concluded that the image quality evaluation index ISs of the DCGAN model is 6.82, which is 2.02 higher than that of the GANs model with the same image distribution. The image diversity index ISc of the DCGAN is 6.10, while that of the GANs is only 4.55. The reason for this is that the image quality evaluation index ISs of the DCGAN is 6.82, which is 2.02 higher than that of the GANs model with the same image distribution. In addition, the DCGAN has more advantages than the GANs in terms of its model framework and model detail parameters.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No:4210040255), the Sichuan Science and Technology Program (2021JDRC0108), Opening Fund of Geomathematics Key Laboratory of Sichuan Province (scsxdz2020yb05) and Chengdu University of Technology Development Funding Program for Young and Middleaged Key Teachers (10912-JXGG2020-06251)”.

REFERENCES

1. J. Wang, Y. Yang, T. Wang, R. Sherratt, and J. Zhang, “Big data service architecture: a survey,” *Journal of Internet Technology*, vol. 21, no. 2, pp. 393–405, 2020.
2. A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” in *Proceedings of the ICML’16 Proceedings of the 33rd International Conference on International Conference on Machine Learning*, vol. 48, pp. 1747–1756, 2016.
3. D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the ICLR 2014 International Conference on Learning Representations (ICLR)*, 2014.
4. J. Chen, K. Li, K. Bilal, Xu Zhou, K. Li, and P. S. Yu, “A Bi-layered parallel training architecture for large-scale convolutional neural networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 5, pp. 965–976, 2019.
5. D. Cao, K. Zeng, J. Wang et al., “BERT-Based deep spatial-temporal network for taxi demand prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 2021, 13 pages, 2021.
6. N. Shobha Rani and B. J Nair, “A deep convolutional architectural framework for radiograph image processing at bit plane level for gender & age assessment,” *Computers, Materials & Continua*, vol. 62, no. 2, pp. 679–694, 2020.
7. I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, and S. Ozair, “Generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.
8. M. Mirza and S. Osindero, *Conditional Generative Adversarial Nets*, vol. 1411, 2014, <https://arxiv.org/abs/1411.1784?context=cs>.
9. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in *Proceedings of the NIPS’16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242, 2016.
10. P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, IEEE, Honolulu, HI, USA, July 2017.
11. H. Zhang, T. Xu, and H. Li, “StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks,” in *Proceedings*

- of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5908–5916, IEEE, Venice, Italy, October 2017.
- 12. J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, IEEE, Venice, Italy, October 2017.
 - 13. T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in *Proceedings of the ICLR 2018 International Conference on Learning Representations*, 2018, <https://arxiv.org/abs/1812.04948>.
 - 14. X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, IEEE, Venice, Italy, 22-29 October 2017.
 - 15. A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *Proceedings of the ICLR 2019 7th International Conference on Learning Representations*, 2019, <https://arxiv.org/abs/1809.11096>.
 - 16. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proceedings of the ICLR 2016 International Conference on Learning Representations*, 2016.
 - 17. B. Pu, K. Li, S. Li, and N. Zhu, “Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7771–7780, 2021.
 - 18. C. Chen, K. Li, S. G. Teo, X. Zou, K. Li, and Z. Zeng, “Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks,” *ACM Transactions on Knowledge Discovery from Data*, vol. 14, no. 4, pp. 42–51, 2020.
 - 19. S. Liu, M. Yu, M. Li, and Q. Xu, “The research of virtual face based on deep convolutional generative adversarial networks using TensorFlow,” *Physica A: Statistical Mechanics and Its Applications*, vol. 521, pp. 667–680, 2019.
 - 20. M. A. B. Mahmoud and P. Guo, “A novel method for traffic sign recognition based on DCGAN and MLP with PILAE algorithm,” *IEEE Access*, vol. 7, Article ID 74602, 2019.

21. W. Fang, Y. Ding, F. Zhang, and J. Sheng, “Gesture recognition based on CNN and DCGAN for calculation and text output,” *IEEE Access*, vol. 7, Article ID 28230, 2019.
22. S. Aslan, U. Gudukbay, B. U. Toreyin, and A. E. Cetin, “Early wildfire smoke detection based on motion-based geometric image transformation and deep convolutional generative adversarial networks,” in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8315–8319, IEEE, Brighton, UK, May 2019.
23. J. Viola, Y. Q. Chen, and J. Wang, “Faultface: deep convolutional generative adversarial network (DCGAN) based ball-bearing failure detection method,” *Information Sciences*, vol. 542, no. 4, pp. 195–211, 2021.
24. K. Y. Cheng, R. Tahir, L. K. Eric, and M. Z. Li, “An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset,” *Multimedia Tools and Applications*, vol. 79, no. 19–20, Article ID 13725, 2020.

CHAPTER 4

Private Face Image Generation Method Based on Deidentification in Low Light

Beibei Dong¹, Zhenyu Wang², Zhihao Gu¹, and Jingjing Yang¹

¹School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China

²Sifang College, Shijiazhuang Tiedao University, Shijiazhuang 051132, China

ABSTRACT

The existing face image recognition algorithm can accurately identify underexposed facial images, but the abuse of face image recognition technology can associate face features with personally identifiable information, resulting in privacy disclosure of the users. The paper puts forward a method for private face image generation based on deidentification

Citation: B. Dong, Z. Wang, Z. Gu, J. Yang, “Private Face Image Generation Method Based on Deidentification in Low Light”, Computational Intelligence and Neuroscience, vol. 2022, Article ID 5818180, 11 pages, 2022. <https://doi.org/10.1155/2022/5818180>.

Copyright: © 2022 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

under low light. First of all, the light enhancement and attenuation networks are pretrained using the training set, and low-light face images in the test set are input into the light enhancement network for photo enhancement. Then the facial area is captured by the face interception network, and corresponding latent code will be created through the latent code generation network and feature disentanglement will be done. Tiny noise will be added to the latent code by the face generation network to create deidentified face images which will be input in a light attenuation network to generate private facial images in a low-lighting style. At last, experiments show that, compared with other state-of-the-art algorithms, this method is more successful in generating low-light private face images with the most similar structure to original photos. It protects users' privacy effectively by reducing the accuracy of the face recognition network, while also ensuring the practicability of the images.

INTRODUCTION

At present, face image recognition technology, based on deep learning technology, has become one of the first choices for identifying and verifying individual identity due to its convenience, efficiency, and maturity [1], and it has been widely applied in the Internet of Things (IoT) and cloud computing [2, 3]. In addition, in the fields of target detection [4], social media data mining [5], and autonomous driving [6, 7], face images are constantly being collected. Face images, however, representing individual characteristics, are of uniqueness and invariability. If they are posted by users or collected passively without any protection of face characters, those characters will inevitably be illegally collected and analyzed [8], thereby resulting in serious identity theft and information fraud, for example, the privacy disclosure incident covering more than 50 million users of Facebook [9] and the illegal profit-making issue of Alipay (a mobile payment software) by forging face images [10]. As shown in Figure 1, face image acquisition devices and applications collect a large number of face images under various lights, and the use of face recognition algorithms and data mining algorithms by criminals will lead to user privacy leakage and identity theft. User privacy leak is detrimental to social stability. The leakage of private data has become a major global social problem in the Internet era, which is universal, frequent, and explosive. Enterprises and users are harassed and violated. Leakage of private data often triggers explosive incidents. Once an incident occurs, it will have serious consequences, with high levels of damage, often producing resonance effects, triggering social dissatisfaction

and turbulence, and having a wide range of impacts. In terms of time, it may continue for several years, and it is difficult to eliminate the impact in a short time. Therefore, the European Union formally implemented General Data Protection Regulation in May 2018, clarifying the data rights of citizens and the basis of privacy protection. At the same time, frequent privacy leakage events made users averse to face image recognition technology and they refused to enter places installed with face image acquisition equipment. The above incidents have seriously hindered the application and development of artificial intelligence technology and the Internet of Things.

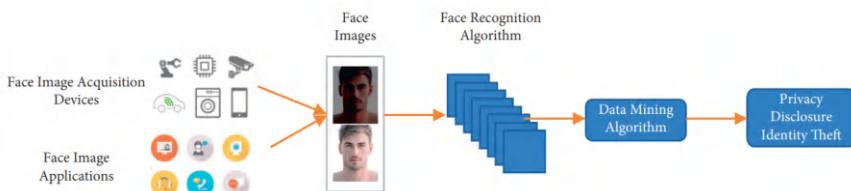


Figure 1. Schematic diagram of privacy leakage of users.

It has been a hotspot to study how to prevent the abuse of face image recognition technology, remove the association between facial features and personal identity information, and avoid the disclosure of user privacy on the premise of ensuring the practicability of face images. Deidentification of the face has become a potential solution to this problem [11]. Although related studies on face deidentification can already mislead face image recognition algorithms in identity recognition, its effectiveness usually relies on sufficient light [12, 13]. In the low-light environment, changes in ambient light and differences in the object's surface material often result in uneven brightness, unclear image texture, and low contrast of local features. All of these problems will bring great challenges to existing face image deidentification methods. However, existing face image recognition algorithms have long been able to accurately identify underexposed face images [14, 15]. After our experiments on existing face deidentification methods in low-light environments, the success rate of generating private face images cannot be guaranteed due to the failed generation or the generated images being too dark. If low-light private face images could not be generated, then users' privacy will not be well protected due to the existence of low-light face recognition algorithms. Therefore, it is crucial to overcome the impact of low light on face image deidentification. The existing deidentification methods have a low success rate in generating low-light-style private face images. We should look for and achieve a new method for low-light private image

generation based on face image deidentification. It can eliminate or reduce unfavorable factors caused by low light, allowing generated face images to show more details and features. It can also generate low-light, deidentified face images, which are extremely similar to the brightness and contrast of the original images. Otherwise, the faces will be not natural enough in low-light scenes, and the user experience will be deteriorated.

The paper proposes a face privacy image generation method based on deidentification and under low light. It has a higher privacy protection capacity with fewer processing traces and good visual quality. The main contributions of the paper are as follows:(1)The method in this paper designs the light enhancement network based on the Retinex theory. The low-light face image is enhanced by the light enhancement network and then face deidentification is performed, which overcomes the adverse effects of low light and improves the success rate of generating deidentified private face images under low light.(2)The method in this paper trains the light attenuation network with the opposite training strategy of the light enhancement network to generate low-light style face privacy images. The face privacy images are real and natural under low light, which improve the user's experience.(3)Although the features of the face region of the face privacy images generated by this method are obviously different from the original face images, they still maintain the basic appearance of the original face images. The method in this paper ensures the practicability of generating face privacy images and, at the same time, misleads face image recognition network recognition and protects user privacy.

The structure of the thesis is as follows: The second part introduces relevant studies, the third part explains the method proposed in this paper, the fourth part is about experiments and analysis, and the fifth part summarizes the whole paper.

RELATED WORK

Generative Adversarial Networks

The classic deidentification method is based on cryptography. However, a large number of computing resources are required, which is not conducive to real-time transmission. In the current popular research, private face image generation methods based on deidentification are divided into the deidentification method based on face disturbance [16], the deidentification method based on face mixing [17], and the deidentification method based on

deep learning. Thereinto, face privacy images generated by the deep learning-based deidentification method are of higher image quality with stronger privacy protection capacity, so it has become a hot research topic. The basis of deidentification methods is the generation of virtual faces. The generation of virtual faces is mainly realized by using the GAN (Generative Adversarial Networks) proposed by Goodfellow et al. [18]. GAN is structurally inspired by the two-person zero-sum game in game theory. It sets the two parties participating in the game as a generator and a discriminator. The purpose of the generator is to learn and capture the truth as much as possible, learn the potential distribution of data samples, and generate new data samples. The discriminator is a binary classifier whose purpose is to correctly judge whether the input data comes from real data or the generator. To win the game, these two game participants need to be continuously optimized. Each improves its own generation ability and discriminative ability. This learning optimization process is a Minimax game problem. The purpose is to find a Nash equilibrium between the two so that the generator can estimate the distributed data samples. Various GAN-based derivative models are proposed to improve the structure of the model and further expand the theory and apply it. Arjovsky et al. [19] proposed Wasserstein GAN (WGAN), which solves the problem of gradient disappearance caused by discontinuity of the optimization target. Radford et al. [20] proposed DCGAN (Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks), which uses convolutional neural networks for supervised learning and GAN for unsupervised learning to generate images and obtains relatively good results to verify the generated image feature representation expressive ability. GAN can generate images, videos, etc., and has a very wide range of applications. In this paper, GAN is used to generate private images of human faces.

Deidentification Method Based on Deep Learning

Among the method of deidentification methods based on deep learning, Karras et al. [21] proposed PGGAN which makes the generation of high-quality and high-resolution images possible through a progressive approach. PGGAN proposes the concept of layer-by-layer training, but it also increases the complexity of training. Then they proposed the epoch-making StyleGAN [22] on this basis, which untangles latent code through a nonlinear mapping network to control high-level attributes of generated images. Aiming at the “water droplets” in images generated by StyleGAN, Karras et al. redesigned the normalization scheme used in the generator and put forward StyleGAN2

[23], solving the artifact problem of generated images. StyleGAN2 can generate high-quality virtual face images but does not achieve good equivariance. Shen et al. proposed InterFaceGAN [24], analyzed semantic characteristics of latent code, and constructed the theory of facial attribute editing through latent code. Grounded on virtual face generation technology, Wu et al. [12] presented PP-GAN for deidentification. It could generate private images of faces with a Generative Adversarial Network (GAN) to avoid its identification by face image recognition systems. Besides, a new validator and modulator were adopted to ensure the quality of private facial images but only experimented on black and white datasets. Based on Generative Adversarial Network and U-NET, He et al. [25] added tiny perturbation to each face image to make deidentified faces wrongly classified by face recognition network, but the “checkerboard effect” arose in deidentified faces. Yang et al. [13] proposed that principal component analysis of faces should be carried out to reduce data redundancy. Then the principal component of face images would be disturbed by adversarial samples and transformed into face images through PCA inversion. However, the quality of generated deidentified face images still needs to be improved. Proenca put forward UU-Net [26], which used Conditional Generative Adversarial Network to create synthetic face privacy images that retain the original posture, lighting, background information, and facial expressions. Lin et al. [27] proposed FPGAN (face deidentification method with generative adversarial networks for social robots). The pixel loss and content loss functions are designed to retain part of the link between the deidentified image and the original image, and U-Net is improved as a generator and applied to the deidentification of social robots. So far, there has been no research on deidentification aiming at private face image generation in low light yet. The method in this paper can break through the flaw of existing technologies which can only be applied under sufficient light and realize deidentification of face image data under low-light conditions, extending application scenarios of the face image deidentification method based on privacy protection.

THE PROPOSED METHOD

Definition of the Problem

Suppose there is a low-light face data set from IoT devices, $X = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$. For any low-light face image x_i and $x_i \in R^{m \times n}$, the

corresponding identity tag is y_i . The algorithm of Q generates corresponding underexposed face privacy images. Then for random face image recognition algorithm, there is

$$\begin{aligned} & \min_t \|\delta\|_2, \\ \text{s.t. } & \log \frac{\Pr[f(x'_i) = y_i]}{\Pr[f(x_i) = y_i]} < \epsilon \end{aligned} \quad (1)$$

Among them, δ represents the change amplitude of lowlight face image x_i and low-light private face image x'_i . To ensure the high practicability of the private face image x'_i , δ should be as small as possible. The face should be as real and natural as possible, and the brightness, contrast, and other indicators should be as similar as possible to low-light face images x_i . ϵ is the index of privacy protection degree. For the random face image recognition algorithm f, the probability of recognizing the real identity tag y_i , corresponding to lowlight face privacy images, should be minimized to realize privacy protection. (ϵ smaller the ϵ is, the better the privacy protection will be. The purpose of this paper is to generate low-light private face images x'_i on the premise of minimizing x_i and δ .

The Framework of the Proposed Method

The Overall Framework of the Proposed Method

The overall framework of the low-light private face image generation method based on deidentification is shown in Figure 2. To ensure the high practicability of the generated private face image x'_i , the low-light face images are firstly enhanced through a light enhancement network, and the face area is captured through a face cropping network. Then an enhanced face image \dot{x}_i is created. Then, a private face image \ddot{x}_i is input into the latent code generation net for latent code generation, and the latent code feature is disentangled through the mapping network of the latent code generation net. Then tiny noise is added to the enhanced face image \dot{x}_i with a synthesis network. Next, Pixel-Level Similarity Loss is adapted to constrain the similarity between the generated face and the enhanced face image \dot{x}_i , to create a deidentified face image \ddot{x}_i similar to the enhanced face image \dot{x}_i . Deidentified face image \ddot{x}_i is input in the light attenuation network to generate private face image x'_i in low-light style. To ensure that private

face image x'_i can successfully mislead face image recognition networks in face identification, it is input in the face image recognition network and Similarity Judgment Loss is set. If the face image recognition network can successfully recognize the face, the noise will be added to the latent code to generate a new deidentified face image \ddot{x}'_i . This step will be repeated until the face image recognition network fails in face recognition. Then the private face image \ddot{x}'_i will be output. Among them, the light enhancement and attenuation network are pretrained with paired low-light face images and normal face images. In order to ensure the privacy of the generated face images, the face recognition network adopts a well-trained model with high accuracy.

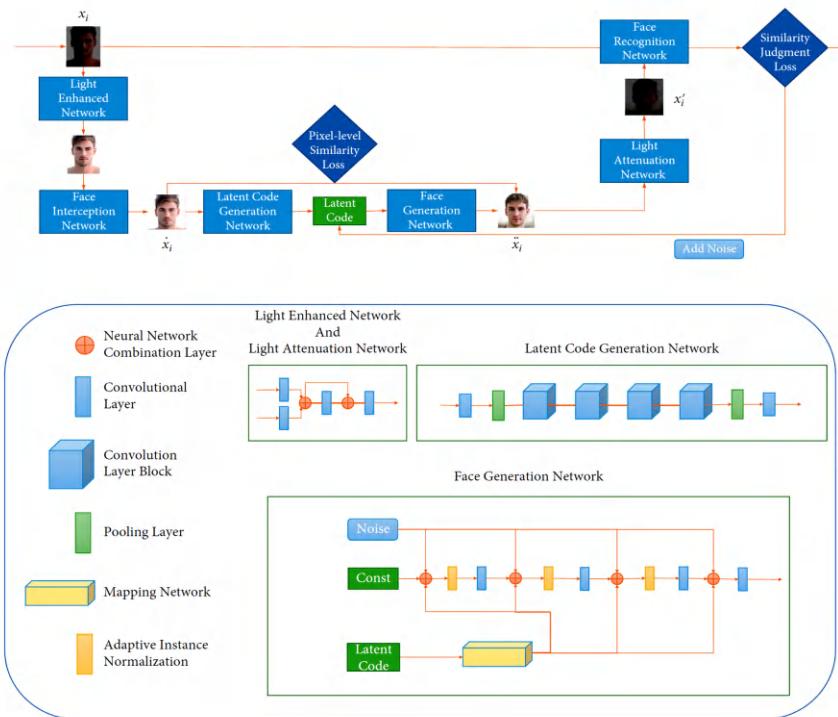


Figure 2. The overall framework of the proposed method.

The Light Enhancement and Attenuation Networks

The training principle of light enhancement and attenuation networks is shown in Figure 3. Use a specific method to perform low-light processing

on the normal-light image to obtain a low-light image paired with it. A low-light face image and a normal-light face image are paired as $\{x_j, \check{x}_j\}$. The low-light facial image is input into the light enhancement network. The face image \check{x}_j , output by Mean Squared Error, namely, equation (2), is close to its Euclidean distance with normal-light face image. The opposite strategy will be applied when training the light attenuation network. The normal-light face image is adopted as the input of the light attenuation network, and the face image \check{x}_j , output through loss function (2), approaches the Euclidean distance with a low-light face image. Noise is generated randomly. If the similarity between the deidentified face image and the original face image is too high, it is necessary to use noise to interfere with the latent code of the deidentified face. The latent code of the face image is multiplied by random noise to change the face generated by the target latent code. Through multiple trainings, the similarity between the generated face image and the original face image can be reduced.

$$L_J = \frac{1}{N} \sum_{i=1}^N (\check{x}_j - h_\theta(x_j))^2. \quad (2)$$

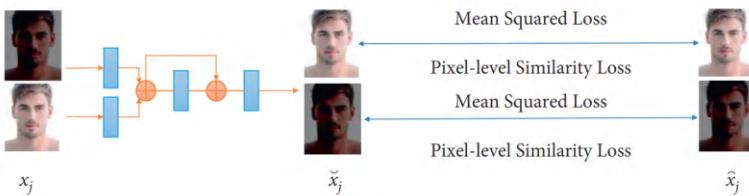


Figure 3. Training schematic diagram of light enhancement and attenuation networks.

In the loss function (2), h_θ represents the fitting function of light enhancement and attenuation networks. That is, $h_\theta(x_j)$ is the output of the fitting function \check{x}_j , and N represents the size of the training data set. To ensure the training quality of light enhancement and attenuation networks, SSIM (Structural Similarity) indexes [12, 28], shown in loss function (3), are also adopted to drive the output face image \check{x}_j to be structurally close to the training images \check{x}_j .

$$L_s = \frac{1}{2} (1 - \text{SSIM}(x, y)). \quad (3)$$

In the loss function (3), x, y represent two face images to be compared for structural similarity. The function of the loss functions L_E and L_W is to enable the deep network to achieve the effect of enhancing or attenuating the light of the face image. For the loss function L_E of the light enhancement network and the loss function L_W of the light attenuation network, there are

$$\begin{aligned} L_E &= L_W \\ &= \chi_1 L_J + \chi_2 L_s. \end{aligned} \tag{4}$$

The Generation of Private Face Images

The classical network model Senet50 is selected as the model of latent code generation network, taking the enhanced face image \dot{x}_i as the input of Senet50 and connecting the mapping network at the end of Senet50 network to transform the enhanced face images into latent space code. That is, the disentangled feature is latent code. The latent code can be used to control the style of the generated image. The mapping network consists of six full connection layers. Generated latent code is input into face generation network, and Mean Squared Error, namely, equation (5), is used as the loss function of latent code generation network, making the output of face generation network, namely, deidentified face image \ddot{x}_i , approximate to enhanced face images, to drive the latent code generation network to create latent code of enhanced face images in the initial domain.

$$L_C = \frac{1}{N} \sum_{i=1}^N (\dot{x}_i - \ddot{x}_i)^2. \tag{5}$$

The role of the synthesis network is to generate face images. The synthesis network of face generation network adopts the structure of StyleGAN2 and the loss function of logistic with single gradient penalty, as shown in equation (6), where D represents a discriminator, G stands for a generator, $\nabla_{T_{\text{real}}}^2$ serves as the gradient penalty of real samples, and $r1_{\text{gamma}}$ is the hyperparameter.

$$\begin{aligned} L_D &= \log(\exp(D(G(z))) + 1) + \log(\exp(-D(x)) + 1) + r1_{\text{gamma}} * 0.5 * \sum \nabla_{T_{\text{real}}}^2, \\ L_G &= -\log(\exp(D(G(z))) + 1). \end{aligned} \tag{6}$$

To ensure that generated private face images can successfully mislead face image recognition networks in face identification, as shown in formula (7),

Similarity Judgment Loss is also set to ensure that the generated deidentified face images \ddot{x}_i can lead to the failure of the face image recognition network.

$$L_A = \mathbb{E}_{x_i} \ell_f(\ddot{x}_i, y, y'). \quad (7)$$

Thereinto ℓ_f represents the fitting function of the face image recognition network. When it identifies the deidentity tag of the deidentified face image as the real label, the loss function will return to a higher value. Then the face generation network will add tiny noise to the latent code and repeat the above generation process until ℓ_f identifies the forged identity tag of the deidentified face image. Similarly, to ensure the quality of generated private face images, the synthesis network also adds SSIM (Structural Similarity) index loss function, so the loss function L_F of the face generation network is shown in equation (8), where χ_1 , χ_2 , χ_3 , and χ_4 are hyperparameters.

$$L_F = L_G + \chi_1 L_D + \chi_2 L_s + \chi_3 L_A + \chi_4 L_C. \quad (8)$$

EXPERIMENTS AND ANALYSIS

Experimental Settings

The hardware configuration used in the experiment is Intel 8700K CPU, 16G DDR4 memory, and 2070Ti graphics card. The implementation of the algorithm uses Python as the programming language and TensorFlow as the deep learning framework.

VGGFACE2 [29] covers a wide range of poses, ages, and races. It is a large-scale face recognition data containing 3.31 million pictures and 9131 IDs. The average number of pictures per ID is 362.6. Now the structure and model parameters of the trained VGG16, Resnet50, and other networks have been open sourced. The experimental data set adopted the public face data set VGGFACE2 from which 300,000 face images were randomly selected. All face images were converted into low-light face images through a new training method of the low-light environment data set [30]. The data set was divided into a training set, validation set, and test set according to the ratio of 98:1:1. The classical networks VGG16, Resnet50, MobileNet V3, and Senet50 were trained, respectively, to serve as face image recognition networks in the loss function by using transfer learning. All these face image recognition networks adopt triples to construct loss functions, so they all set the threshold of 0.3 to determine whether the input face images

belong to the category, as shown in Table 1. To prove the advancement of the method, threshold settings were all equal to or less than the common threshold (0.7–0.9) set by face image recognition networks. Four face image recognition networks all achieved a high recognition rate. Thereinto, True Positive Rate and False Positive Rate are two commonly used indicators in face recognition, and the calculation method is shown in formula (9). TP is correctly classified by the classifier as a positive example; TN is correctly classified by the classifier as a negative example, FP is wrongly classified by the classifier and it should be a negative example, and FN is wrongly classified by the classifier as it is a positive example. Therefore, TPR represents the rate of correctly judged positive among all positive samples. FPR represents the rate of false positives among all negative samples.

Table 1. Accuracy and threshold of four face image recognition networks

| Face recognition model | Training accuracy | Test accuracy | TPR | FPR |
|------------------------|-------------------|---------------|-------|-------|
| VGG16 | 0.982 | 0.951 | 0.938 | 0.081 |
| Resnet50 | 0.990 | 0.979 | 0.973 | 0.062 |
| MobileNet V3 | 0.999 | 0.987 | 0.997 | 0.031 |
| Senet50 | 0.975 | 0.960 | 0.917 | 0.172 |

$$\left\{ \begin{array}{l} \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{array} \right. \quad (9)$$

In recent years, deep learning has been widely used in research related to light enhancement such as dehazing and harsh environments [31–33]. Inspired by this, the method in this paper designs a deep neural network for light enhancement and attenuation of face images. The structure of light enhancement and attenuation networks is shown in Figure 4. About the Retinex theory [34], the network was designed as a cascade structure to decompose images into reflection components and illumination components. Among them, the illumination component reflects the slow illumination information of the overall face image. The reflection component reflects the authentic attributes of the face image. After the steps in Section 3.2.2, reconstructed images can be converted into light enhancement images and low illumination images. We paired face images in the training set with low-light face images, selected 100,000 pairs as the training set, and made

pretraining of illumination enhancement and attenuation networks. The loss function L_E of the light enhancement network reached 0.091 and that of the light attenuation network hit 0.121. Both latent code generation network and face generation network adopted pretraining model. We also compared the method with StyleGAN1 [22] and StyleGAN2 [23] to demonstrate its state of the art. We set the hyperparameter $\chi_1 = 1.0$, $\chi_2 = 0.5$, $\chi_3 = 0.3$, and $\chi_4 = 0.1$.

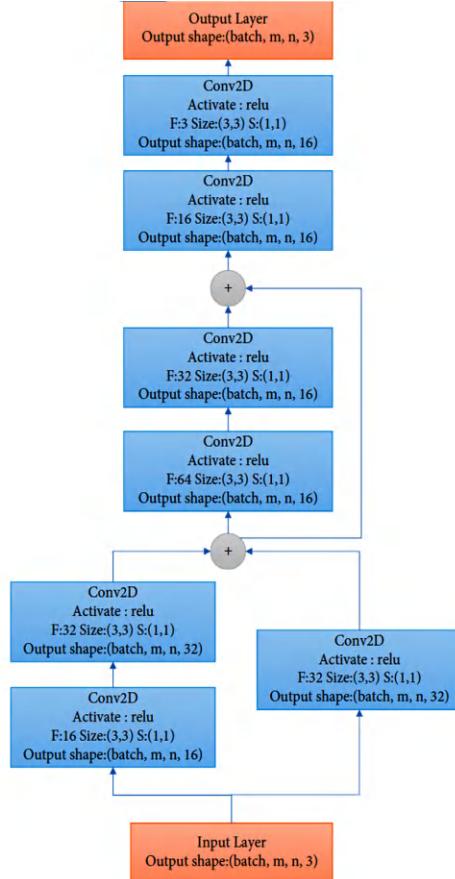


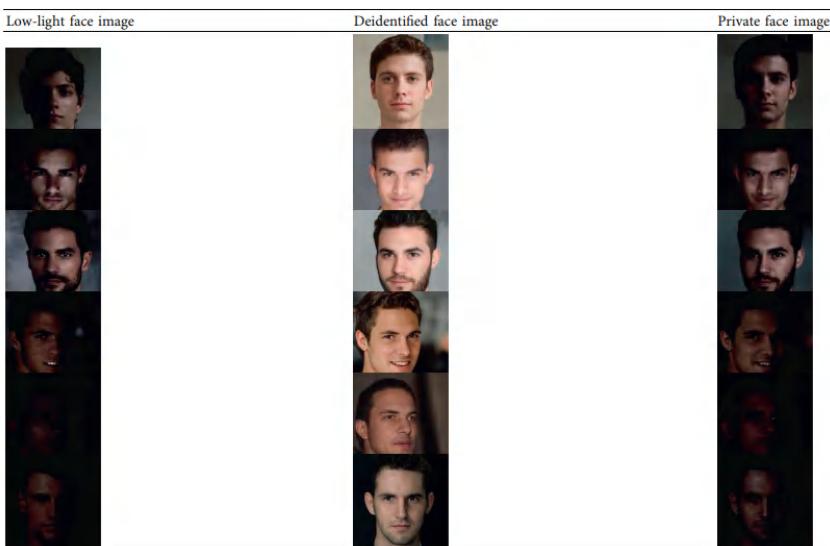
Figure 4. The structure of light enhancement and attenuation networks.

Experimental Results and Analysis

The private face images x'_i generated by the method are shown in Table 2, and the deidentified face image \ddot{x}_i was generated from a low-light face image x_i with the joint efforts of light enhancement network, latent code

generation network, and face generation network. Although the facial characters had become visibly different from that of the original face image, they still maintained the basic appearance of the original face. Inputting deidentified face image into the light attenuation network, a low-light private face image was obtained. Attaching it to the original video or image, we found the styles of the two images are unified, which improved the user experience.

Table 2. Private face images generated by the method



To test the privacy protection degree of the face privacy image x'_i generated by the method on face features, we applied the four pretrained face recognition networks in Table 1 respectively to recognize private face images. TPR and FPR were measured, as shown in Table 3.

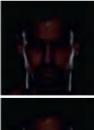
Table 3. TPR and FPR of four face recognition models to recognize private face images

| Face recognition model | TPR | FPR |
|------------------------|-------|-------|
| VGG16 | 0.148 | 0.850 |
| Resnet50 | 0.145 | 0.887 |
| MobileNet V3 | 0.119 | 0.815 |
| Senet50 | 0.103 | 0.854 |

It can be seen from Table 3 that if TPR declines while FPR rises, the proportion of correctly identified positive examples in total positive examples drops, and the percentage of negative examples predicted as positive one's increases. As such, private face image x'_i generated by this method can successfully mislead face recognition networks, thus protecting the privacy of users.

Low-light private face images generated by this method were compared with that by PGGAN [21], StyleGAN1 [15], and StyleGAN2 [16], as shown in Table 4. Because the method in this paper uses light enhancement and attenuation networks, the light of the face image generated by the method in this paper is more in line with the original image. It can be seen that the private face images generated by this method are more consistent with the original image in structural similarity than those generated by PGGAN, StyleGAN1, and StyleGAN2.

Table 4. Comparisons of low-light private face images generated by different methods

| Method | Low-light face image | Private face image | Low-light face image | Private face image |
|---------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| The proposed method |  |  |  |  |
| PGGAN |  |  |  |  |
| StyleGAN1 |  |  |  |  |
| StyleGAN2 | | | | |

The similarity comparison between private face images generated by this method and by PGGAN, StyleGAN1, and StyleGAN2 and the original face images is shown in Figure 5. In Figure 5, PSNR (Peak Signal to Noise Ratio) is the most common and widely used objective image evaluation index, which is the ratio of the energy of the peak signal to the average energy of the noise. SSIM (Structural Similarity) is a full-reference image quality evaluation index, which measures image similarity from three aspects: brightness, contrast, and structure. CS (Cosine similarity) calculates the angle between two vectors, which can be used to measure the direct

similarity of images. The private face images generated by this method have higher SSIM, CS, and PSNR than those created by the other three methods. It indicates that the low-light face images generated by the proposed method are closer to the style of the original images and more real and natural, thus greatly improving the user experience.

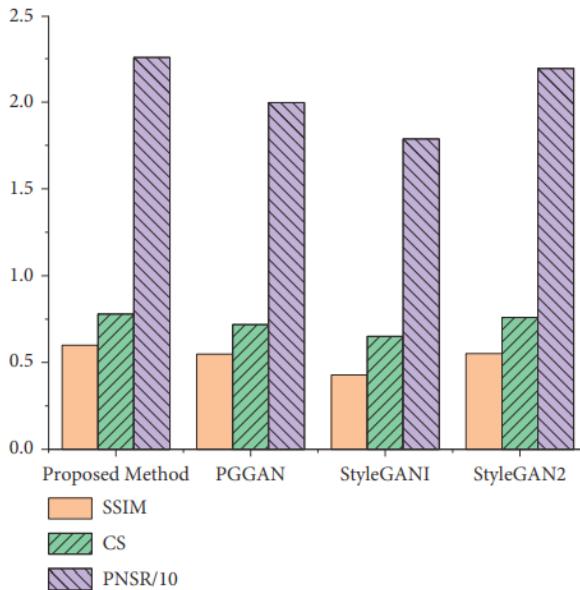


Figure 5. Comparisons of similarity between original face images and private face images generated by different methods.

The comparison of the success rate of the proposed method and PGGAN, StyleGAN1, and StyleGAN2 in generating low-light face deidentification images is shown in Figure 6. The success rate of the proposed method in generating low-light private face images is 100 percent, while that of PGGAN is 42.9%, StyleGAN1 is 46.2%, and StyleGAN2 is 68.3%. The front-facing face cropping network of PGGAN, StyleGAN1, and StyleGAN2 cannot completely detect the face area under low light, failing the generation of low-light images sometimes. Or the generated face images are completely black and cannot be recognized, resulting in generation failure. In our method, since the light enhancement network has enhanced the illumination of low-light photos, the front face region interception network is not affected by low light, and the success rate of detecting the face region is 100%. Moreover, under the effect of the light attenuation network, the style of the generated low-light face image is closer to the original face image.

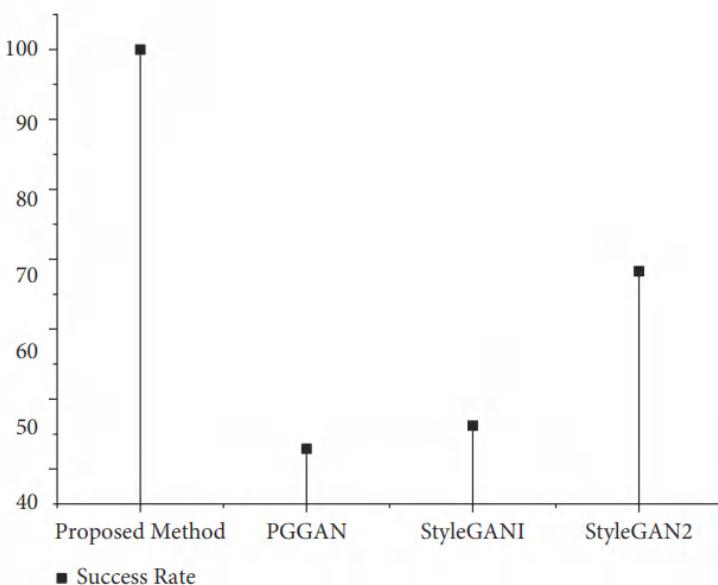


Figure 6. Comparisons of the success rate of low-light private face images generated by different methods.

CONCLUSION

This paper puts forward a low-light face image generation method based on deidentification. It overcomes the adverse effects of low light and generates face images of a low-light style, making private face images real and natural and thereby improving the user experience. Meanwhile, it reduces the accuracy of the face recognition network to protect the privacy of users. When IoT devices collect face images for internal storage, or IoT applications transmit face images through an external communication network, even if there is a storage data leakage or a man-in-the-middle attack, the method proposed in this article can effectively prevent the leakage of user privacy. The method in this paper can be applied to various application scenarios of face image collection, and it is an effective supplement to the existing face privacy image methods. In the future, we will do more research on lightweight models in private face image generation and optimize the operating speed to make its application more efficient in edge computing. In addition, there are deidentification methods in many special scenes, such as profile and occlusion, which need to be studied.

AUTHORS' CONTRIBUTIONS

All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication. Beibei Dong and Zhenyu Wang contributed equally to this work.

ACKNOWLEDGMENTS

This research was supported by the Natural Science Foundation of Hebei Province under Grant no. F2021405001, Basic Scientific Research Funds Projects of Hebei North University under Grant no. JYT2020029, the Three Three Talent Project Funding Project in Hebei Province under Grant no. A202001017, the General Project of Hebei North University under Grant no. XJ2021005, and Innovation and Entrepreneurship Training Program for College Students in Hebei Province under Grant no. S202110092017.

REFERENCES

1. G. Lou and H. Shi, “Face image recognition based on convolutional neural network,” *China Communications*, vol. 17, no. 2, pp. 117–124, 2020.
2. P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, R. Jidhesh, and T. Vadivel, “Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things,” *Cluster Computing*, vol. 22, no. S4, pp. 7733–7744, 2019.
3. P. Hu, H. Ning, T. Qiu, Y. Xu, X. Luo, and A. K. Sangaiah, “A unified face identification and resolution scheme using cloud computing in Internet of Things,” *Future Generation Computer Systems*, vol. 81, pp. 582–592, 2018.
4. J. Li, Y. Wang, G. Fang, and Z. Zeng, “Real-time detection tracking and recognition algorithm based on multi-target faces,” *Multimedia Tools and Applications*, vol. 80, Article ID 17238, 2021.
5. Y.-S. Su and C.-F. Lai, “Applying educational data mining to explore viewing behaviors and performance with flipped classrooms on the social media platform Facebook,” *Frontiers in Psychology*, vol. 12, 2021.
6. Y. Cai, T. Luan, H. Gao et al., “YOLOv4-5D: an effective and efficient object detector for autonomous driving,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.
7. D. Ma, X. Song, and P. Li, “Daily traffic flow forecasting through a contextual convolutional recurrent neural network modeling inter- and intra-day traffic patterns,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2627–2636, 2021.
8. F. Li, Z. Sun, A. Li, B. Niu, H. Li, and G. Cao, “HideMe: privacy-preserving photo sharing on social networks,” in *Proceedings of the IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 154–162, Paris, France, April 2019.
9. J. Yang, J. Liu, R. Han, and J. Wu, “Transferable face image privacy protection based on federated learning and ensemble models,” *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2299–2315, 2021.
10. J. Yang, J. Liu, R. Han, and J. Wu, “Generating and restoring private face images for Internet of vehicles based on semantic features and adversarial examples,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.

11. K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, “I know that person: generative full body and face de-identification of people in images,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1319–1328, Honolulu, HI, USA, July 2017.
12. Y. Wu, F. Yang, Y. Xu, and H. Ling, “Privacy-protective-GAN for privacy preserving face de-identification,” *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, 2019.
13. J. Yang, J. Liu, and J. Wu, “Facial image privacy protection based on principal components of adversarial segmented image blocks,” *IEEE Access*, vol. 8, Article ID 103394, 2020.
14. C. Shi, C. Wu, and Y. Gao, “Research on image adaptive enhancement algorithm under low light in license plate recognition system,” *Symmetry*, vol. 12, no. 9, p. 1552, 2020.
15. M. O. Oloyede, G. P. Hancke, and H. C. Myburgh, “A review on face recognition systems: recent approaches and challenges,” *Multimedia Tools and Applications*, vol. 79, Article ID 27922, 2020.
16. S. Çiftçi, A. O. Akyüz, and T. Ebrahimi, “A reliable and reversible image privacy protection based on false colors,” *IEEE Transactions on Multimedia*, vol. 20, pp. 68–81, 2018.
17. L. Meng and Z. Sun, “Face De-identification with perfect privacy protection,” in *Proceedings of the 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2014.
18. I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
19. M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, <https://arxiv.org/abs/1701.07875>.
20. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2016.
21. T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” 2017, <https://arxiv.org/abs/1511.06434>.
22. T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, Long Beach, CA, USA, June 2019.
- 23. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, Seattle, WA, USA, June 2020.
 - 24. Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, Seattle, WA, USA, June 2020.
 - 25. Y. He, C. Zhang, X. Zhu, and Y. Ji, “Generative adversarial network based image privacy protection algorithm,” in *Proceedings of the Tenth International Conference on Graphics and Image Processing, SPIE*, Chengdu, China, December 2019.
 - 26. H. Proen  , “The UU-net: reversible face de-identification for visual surveillance video footage,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 496–509, 2021.
 - 27. J. Lin, Y. Li, and G. Yang, “FPGAN: face de-identification method with generative adversarial networks for social robots,” *Neural Networks*, vol. 133, pp. 132–147, 2021.
 - 28. U. Sara, M. Akter, and M. S. Uddin, “Image quality assessment through FSIM, SSIM, MSE and PSNR-A comparative study,” *Journal of Computer and Communications*, pp. 8–18, 2019.
 - 29. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: a dataset for recognising faces across pose and age,” in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, IEEE, Xi'an, China, May 2018.
 - 30. K. G. Lore, A. Akintayo, and S. Sarkar, “LLNet: a deep autoencoder approach to natural low-light image enhancement,” *Pattern Recognition*, vol. 61, pp. 650–662, 2017.
 - 31. Z. Zhu, H. Wei, G. Hu, Y. Li, G. Qi, and N. Mazur, “A novel fast single image dehazing algorithm based on artificial multiexposure image fusion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–23, 2021.
 - 32. H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, “Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain

- adaptation person Re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1, 2021.
- 33. M. Huang, B. Zhang, W. Lou, and A. Kareem, “A deep learning augmented vision-based method for measuring dynamic displacements of structures in harsh environments,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 217, Article ID 104758, 2021.
 - 34. R. R. Hussein, Y. I. Hamodi, and R. A. Sabri, “Retinex theory for color image enhancement: a systematic review,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 2088–8708, 2019.

CHAPTER 5

Application of Remote Sensing Image Data Scene Generation Method in Smart City

Yuanjin Xu

Institute of Mathematical Geology and Remote Sensing Geology, School of Earth Resources, China University of Geosciences, 388 Lumo Road, Wuhan 430074, China

ABSTRACT

Remote sensing image simulation is a very effective method to verify the feasibility of sensor devices for ground observation. The key to remote sensing image application is that simultaneous interpreting of remote sensing images can make use of the different characteristics of different data, eliminate the redundancy and contradiction between different sensors, and improve the timeliness and reliability of remote sensing information

Citation: Yuanjin Xu, “Application of Remote Sensing Image Data Scene Generation Method in Smart City”, Complexity, vol. 2021, Article ID 6653841, 13 pages, 2021.
[https://doi.org/10.1155/2021/6653841..](https://doi.org/10.1155/2021/6653841)

Copyright: © 2021 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

extraction. The hotspots and difficulties in this direction are based on remote sensing image simulation of 3D scenes on the ground. Therefore, constructing the 3D scene model on the ground rapidly and accurately is the focus of current research. Because different scenes have different radiation characteristics, therefore, when using MATLAB to write a program generated by 3D scenes, 3D scenes must be saved as different text files according to different scene types, and then extension program of the scene is written to solve the defect that the calculation efficiency is not ideal due to the huge amount of data. This paper uses POV ray photon reverse tracking software to simulate the imaging process of remote sensing sensors, coordinate transformation is used to convert a triangle text file to POV ray readable information and input the RGB value of the base color based on the colorimetry principle, and the final 3D scene is visualized. This paper analyzes the thermal radiation characteristics of the scene and proves the rationality of the scene simulation. The experimental results show that introducing the chroma in the visualization of the scene model makes the whole scene have not only fidelity, but also radiation characteristics in shape and color. This is indispensable in existing 3D modeling and visualization studies. Compared with the complex radiation transmission method, using the multiple angle two-dimensional image generated by POV rays to analyze the radiation characteristics of the scene, the result is intuitive and easy to understand.

INTRODUCTION

With the development of science and technology, digital cities have received more and more attention [1]. The concept of digital city originates from the strategic concept of digital earth, also known as network city or smart city, or more precisely information city. It refers to the comprehensive use of computer tools (GIS, remote sensing, telemetry, network, multimedia, and virtual simulation technology). The use of digital technology to collect and process the city's infrastructure and functional mechanisms to enable it to have digital functions, which is conducive to optimizing and improving the city's ecological environment and resources, economy, population, and other complex fields, effectively predicts the future of the city. [2, 3]. The essence or core of a digital city is the fusion of massive urban spatial data with three-dimensional urban geographic information systems and time series urban geographic information systems [4]. The outstanding feature of digital cities is the ability of applying digital information to grasp the

changing process of urban regional structure in time and space. The application research of 3D urban geographic information system and time series urban geographic information system will be an important part of recent digital city theory research. Digital city construction will provide an information security system for the city's sustainable development strategy, meet the government's decision-making, macrocontrol, scientific and technological innovation, natural resources and environmental monitoring, intelligent transportation and urban management, and various social welfare undertakings, and further provide solutions for the sustainable development of cities. [5, 6]. To build a digital city, we must first apply high-tech means such as computer technology to model the urban environment [7].

Remote sensing image fusion is a process of comprehensive processing the image data obtained by multiple remote sensing sensors or the same kind of sensor for the same target at different times. The image is processed by using certain rules or algorithms, and the useful information contained in the image is fused into a new image. The image contains more accurate and abundant information than any single image, in order to achieve a comprehensive description of the target and ground objects.

The problem of three-dimensional reconstruction of urban buildings has been studied by experts and scholars from various countries for many years and has achieved a series of results. The most representative ones are Google Earth and Microsoft Virtual Earth, which use satellite remote sensing images to generate virtual ground scenes, which have been successfully commercialized on the Internet [8, 9]. Image-based three-dimensional reconstruction of urban buildings is mainly divided into three categories according to the different data sources used: (1) based on remote sensing images, this method uses three-dimensional reconstruction of urban buildings using approximately vertical satellite remote sensing images or aerial images [10]. According to the characteristics of remote sensing imaging, based on the reconstruction of remote sensing images, the reconstruction space is large, and the roof information of the building can also be obtained and the accumulation of errors can be effectively reduced, but the reconstructed buildings have poor fidelity [11]. (2) Based on ground image, ground image-based reconstruction is a three-dimensional reconstruction of urban buildings using images acquired by various ground shooting techniques. According to the characteristics of ground imaging, the reconstruction of this method is better, and the wall texture of the building can be obtained, but the roof information of the building is not obtained, the reconstruction scale is small, and the error accumulation is large [12].

(3) Regarding combination of remote sensing image and ground image, there are advantages and disadvantages in remote sensing and ground-based imaging reconstruction. In fact, remote sensing image and ground image are two important complementary source data. Combining the two for reconstruction is expected to be obtained. This resulted in a reconstruction method combining remote sensing images with ground images. Generally, there are insufficient data acquisition costs, large data volume, complicated calculation, and low automation [13, 14].

Based on the second generation bending wave transform and Dempster-Shafer (DS) evidence theory, Huang C proposed a new remote sensing image fusion method. Huang C uses the bending wave transform to decompose the remote sensing image to obtain the coefficients and uses DS evidence theory to optimize the high coefficients [15, 16]. First, the high-resolution and multiple spectral remote sensing images are decomposed by bending wave transform to obtain bending wave transform coefficients (coarse, detailed, and fine scale layers) of all layers. Second, the coarse scale layer uses the maximum fusion rule. The detailed scale layer is used by the weighted average fusion rule. The fine scale layer is optimized by DS evidence theory. Three features of the fine scale layer coefficients are obtained. These three characteristics are variance, information entropy, and energy. The use of these features is some parametric belief function and rationality function. The mass function is then combined and a new fusion factor is obtained. Finally, the scene image is obtained by inverse bending wave transform. Rhee et al. attempt to apply two types of image matching, object space based matching techniques and image space based matching techniques, and compare the performance of the two techniques [17, 18]. The object space based match used sets a list of candidate height values for a fixed horizontal position in the object space. For each height, its corresponding image points are calculated and similarity is measured by gray level correlation. The image space based matching used is a modified slack match. Rhee and Kim designed a global optimization scheme for finding the best pair (or group) to apply image matching, defining local matching regions in the image or object space, and merging local point clouds into global point clouds. For optimal pairing selection, the connection points between the images are extracted and a stereoscopic overlay network is defined by using the connection points to form a maximum spanning tree. Qin built the core technology and method related to 3D model reconstruction, focusing on matching of point cloud data registration to simplify denoising, 2D contour extraction, and finally achieving the high complexity of 3D geometric model of farmland site,

using advanced 3D printing. Technology produces small 3D printed point cloud data [19, 20].

The innovations of this paper are as follows: (1) introducing the principle of colorimetry method into the visualization of the image simulation scene model and replacing the complex texture with color can reflect the spectral radiation characteristics of the object to a certain extent. Through investigation and research, it is found that the only one that can correspond to the spectral characteristics is its chroma characteristic. Therefore, the chroma is introduced into the scene model visualization so that the whole scene not only is in shape and color, but also has radiation characteristics. This is not available in existing 3D modeling and visualization studies. (2) The remote sensing imaging process can be simulated using the POV ray visual ray tracing software package. It has a convenient and fast programming language, high computational efficiency, and intuitive output. Compared with the complex radiation transfer equation, the multiple angle two-dimensional image generated by POV ray is used to analyze the radiation characteristics of the scene, and the result is intuitive and easy to understand. POV ray can simulate the remote sensing imaging process, mainly because it first defines the position of the light source (sun) and camera (sensor), zenith angle and azimuth, and also sets the camera's field of view, which is another 3D visualization software. And the higher computational efficiency is also relatively advanced in the field of visualization. (3) Considering the differences of different scenes, the design of remote sensing images for different scenes is different, so that the experimental structure can reflect the diversity and rationality.

THE PROPOSED METHOD

Preparation of Remote Sensing Images

Image Preprocessing

When the obtained source image is blurred, the contrast is not strong, and the noise interference is large; the corresponding method needs to be used for some processing, so that the subsequent work can be carried out more effectively. Common methods include image enhancement, filtering, histogram correction, and gradation transformation. However, when the quality of the source image is good, it is not necessary to perform these processes. Therefore, the preprocessing of the source image is an optional

link, and the processing method is different for different images. The omnidirectional image is used in the experiment in this paper; because the camera has been calibrated before shooting, the lighting conditions are better at the time of shooting, so the obtained image quality is better and generally does not have to be preprocessed.

Registration

Registration is to find the mutual correspondence between the omnidirectional map and the remote sensing map and achieve the purpose of reconstruction service through information fusion. Information fusion is the foundation of all subsequent work, so registration is a core component of the entire reconstruction. Conventional methods generally solve the problem of registration between images from homologous images or from the same type of sensor. The omnidirectional and remote sensing images are images formed by heterogeneous sensors. The general reconstruction method of the source image requires registration of the source image, and more registration methods are available at present. However, the existing conventional methods cannot solve the registration problem of remote sensing maps well. The registration of remote sensing maps belongs to the specific key technical problems of 3D reconstruction of remote sensing images.

Height Extraction

Height extraction is to obtain the height value of the target building in real space, which is an important information of the space structure of the building. The height of the building can be combined with the top view of the building available in the remote sensing map to obtain the approximate spatial structure of the building. Therefore, the height of the building plays a crucial role in the reconstruction of the shape and contour. There are two conventional solutions; one is direct measurement using instruments and equipment, such as laser range finder. The second is to use computer vision principles for estimation. Direct measurement with related instruments is costly. With computer vision estimation, accurate absolute heights are generally not obtained without accurate scale reference objects.

Shape Modeling

The goal of shape modeling is to get the outer shape of the entire building. It is also an important part of reconstruction. It is generally modeled based on certain assumptions or prior knowledge (such as a box in the shape of

a box and a flat roof) using the obtained height and roof shape information obtained from the remote sensing map. The outer shape of the building is more complex and precise than the approximate spatial structure. The key to shape modeling is to extract the outline of the building. At present, the contour detection algorithms generally have shortcomings such as low detection rate and inability of fully automating. The method of semiautomatic human-computer interaction can be adopted; that is, the existing detection algorithm is combined with the manual correction method for detection. Due to limited time and energy, image-based 3D reconstruction mostly requires shape modeling, and shape modeling methods are almost universal.

Building Outline Segmentation and 3D Model Extraction

How to obtain the singular model of the building from the three-dimensional model of the scene generated by oblique photography is the goal of this paper. DOM can be seen from the high-resolution scene, which has obvious image difference between the building and other features. The image analysis method can be used to extract the outline of the building from the scene DOM and obtain the position information of the building outline, thereby realizing the positioning and segmentation of the building model in the three-dimensional scene model.

High-Resolution Image E Building Feature Analysis Method

In an image, the edge information is the most important and basic feature, and the edge feature is the most direct expression and embodiment of the image geometric information. The low-altitude drone oblique photography obtains a higher resolution of the image, and the texture of the scene DOM obtained after the correction is clear. The difference in shape, size, and texture patterns is the basic basis for distinguishing between different features. Through vision, the color of the scene DOM is very realistic, the texture information is very rich, the geometrical structure of the object is more refined, and the recognition of different target objects in the image is more accurate; from a local perspective, a single feature, especially the boundary between the edge of the building and its surrounding environment, is obvious, and the details inside the target object are richly expressed. These features are very advantageous for identifying and extracting the target individual of the building. However, because of the rich information contained in high-resolution images, the phenomenon of “homologous” and “homogeneous foreign matter” appears, resulting in increased noise interference.

In order to extract buildings (houses) in high-resolution images, the characteristics of the buildings are analyzed to establish a good basis for building identification. From spectral and texture characteristics, usually the gray distribution of buildings is relatively uniform. The gray value of the top is higher than the gray value of other surrounding objects and the texture mode is relatively regular. Generally, the texture performance of buildings has the outline direction of the building that is approximately uniform or orthogonal. For the shape feature, usually the building has geometrically regular edges and corners, and the whole is presented as a regular polygon. From the spatial distribution characteristics, usually the roads in the urban area will be divided into several blocks like chessboards. The buildings are regularly distributed in the block, and the surrounding objects mainly include roads and tree vegetation, so the roads and buildings have strong spatial associations, and usually densely populated buildings will also be arranged regularly and have similar configurations. When considering the geometric features, spectral features, and spatial distribution characteristics of features in high-resolution images, the algorithm is not mature enough and complicated, having low efficiency, which needs further study. According to the research objectives and needs of this paper, the rough positioning of the target of the suspected building does not require accurate and complete extraction of the building. Therefore, this paper only analyzes the geometric features of the building from the geometric shape. In the scene DOM, the geometry of buildings and other features is significantly different. From a straight-down perspective, no matter which building is an individual wrapped by its outer contour, in the high-resolution scene DOM, the outer contour of the building is a connected area with a certain length and area. As a basis for identifying buildings, the internal structure of vegetation such as trees is disorderly, and the boundaries of the whole forest are not clear enough and there are no rules: there are often no clear and regular boundaries on both sides of the road, and the roads are lacking in certain areas. From the perspective of individual buildings, there are also separate topological relationships between different buildings. In the scene DOM, the target features that can be approximated as connected areas tend not to have only buildings, but the buildings also contain other distinguishable geometric features, and the more prominent common features are the length of the outer contour of the building. The outer contour of the building must contain at least a certain number of linear features, according to which the connected areas surrounded by the outline of the building can be further identified and screened.

Therefore, by analyzing the features of the building on the image, the difference and segmentation between the building and other features and buildings can be realized, and the outline of a single suspected building can be extracted and used as a basis for the three-dimensional scene model. A rough monomer model was extracted.

Building Edge Feature Detection Method Based on Canny Operator

Edge detection is to obtain information about shape and reflection or transmittance in an image. It is a basic step in image processing, analysis, understanding, pattern recognition, and computer and human vision, being a very important technology. There are many edge detection methods, such as Roberts operator, Pruitt operator, Sobel operator, and Laplace operator. These are the operators that detect features through local window and are sensitive to noise. Canny proposed the best edge detection operator Canny operator. The operator determines the edge pixels by the maximum value of the image signal function, and the detection performance is good, which has been widely used. Therefore, this paper uses Canny operator to perform edge detection on scene digital image. The scene DOM is a true color image, which needs to be grayed out. The color image can be converted into a grayscale image by using the following formula:

$$P(x, y) = 0.3^*R + 0.59^*G + 0.11^*B, \quad (1)$$

where $P(x, y)$ is the gray value of the pixel at the (x, y) coordinate and R, G, and B are the values in the red, green, and blue primary color channels in the pixel at the (x, y) coordinate, respectively. There are four main steps in detecting the edge features of grayscale images using the Canny operator:

(1) *Eliminate Noise.* The differential algorithm is highly sensitive to noise, and the Gaussian smoothing filter is used to convolve the image before edge detection to reduce noise interference. The first following formula is a two-dimensional Gaussian function. The principle of Gaussian smoothing is the discrete Gaussian function. The Gaussian function value on the discrete point is used as the weight. For each pixel in the gray image, it is within the window neighborhood of a certain size. Considering pixel weighting to eliminate Gaussian noise, the second following formula is a discrete Gaussian function weight window template with a window size of 5×5 pixels, the third following formula is a convolution formula for Gaussian filtering of image J , and g is the result of convolution:

$$f(x, y) = \frac{1}{2\pi\sigma^2} e^{(-(x^2+y^2)/2\sigma^2)}, \quad (2)$$

$$K = \frac{1}{159} \begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix}, \quad (3)$$

$$g = I^* K. \quad (4)$$

(2) *Calculate the Image Gradient Amplitude Value and the Gradient Direction.* The first-order finite difference is used to approximate the gray gradient of the image. In the Canny operator, the first following formula is used to convolve in x and y directions of the image. As shown in the second and third following formulas, the Sobel template is shown, wherein S_x is a convolution template in the horizontal direction, and S_y is a convolution template in the vertical direction:

$$\begin{aligned} G_x &= g^* S_x, \\ G_y &= g^* S_y, \end{aligned} \quad (5)$$

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad (6)$$

$$S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \quad (7)$$

According to the following formulas, the gradient magnitude value G and the direction θ of the image can be, respectively, calculated and the gradient direction is approximated to be generally 0° , 45° , 90° , and 135° :

$$G = \sqrt{G_x^2 + G_y^2}, \quad (8)$$

$$\theta = \arctan\left(\frac{G_y}{G_x}\right). \quad (9)$$

(3) *Nonmaximum Suppression.* The pixel corresponding to the local maximum point is found and retained or marked as an edge pixel, and the gray value of the pixel of the nonmaximum point is suppressed and set as the background. This step is mainly to discriminate and remove nonedge pixels, leaving only the candidate image edges.

$$A = \sum \left[S_i^* \log\left(\frac{S_i}{Q_i}\right) \right]. \quad (10)$$

The image edge is solved using a hysteresis threshold algorithm. The method of lag threshold is to use two thresholds (high and low). The following discriminant basis is used when determining the true edge and removing the false edge: if the gradient amplitude value of a pixel is greater than the high threshold, it is determined that the pixel is a true edge pixel and is retained; if the gradient amplitude value of a pixel is less than the low threshold, it is determined that the pixel is not a true edge pixel and is excluded; if the gradient amplitude value of a pixel is between the high and low thresholds, the pixel has only one gradient amplitude value. When pixels larger than the high threshold are connected, it is determined that the pixel is reserved for the real edge pixel. The scene digital orthophoto is rich in texture, and the geometric features of the features are complex. The edge obtained by the Canny operator for edge detection usually contains a lot of noise. Therefore, the DOM edge detection result needs to be Gaussian. In order to facilitate the identification of the building outline, it is necessary to highlight the contour edge features of the target such as a building from the numerous edge feature pieces of information in the scene and further suppress the edge pixels whose edge features are not obvious into the background. Therefore, it is necessary to perform Gaussian smoothing on the scene edge detection result and then perform binary processing. Treated by formula (10), we get

$$g_{\text{new}}(x, y) = \begin{cases} 255 & (g_{\text{old}}(x, y) > T), \\ 0 & (g_{\text{old}}(x, y) \leq T), \end{cases} \quad (11)$$

In the formula, $g_{\text{new}}(x, y)$ is the new gray value of the pixel at the (x, y) coordinate; $g_{\text{old}}(x, y)$ is the previous gray value of the pixel; T is the gray threshold set according to experimental experience and belongs to 0–255. If the gray value of a pixel is greater than the threshold, the gray value of the pixel is set to 255. If it is not greater than the threshold, the gray value is set to 0, and the background is suppressed, thereby obtaining a scene edge with a distinct target edge feature.

Construction of the Original Features of the Front Image

According to the previous preset, the output of the CNN convolutional layer is more than fully connected:

$$a_k = \sum_{y=1}^H \sum_{x=1}^W w_{x,y} b_{x,y} \cdot f_{x,y,k}. \quad (12)$$

Among them, $w_{x,y}$ is the response weight at the point (x, y) in the feature map, and $b_{x,y}$ is the depth weight of the point. For the response weight matrix $W \in \mathbb{R}^{W \times H}$, we use the feature map f_k to construct

$$W = \sum_{k=1}^N f_k. \quad (13)$$

For the depth weight, we use the depth information of the image to assign the weight. We first scale the depth map of the input image to $W \times H$; then

$$d_{x,y} = \left(\frac{d_{\max} - d(x, y)}{d_{\max} - d_{\min}} \right) + \gamma. \quad (14)$$

Among them, d_{\max} is the maximum depth, d_{\min} is the minimum depth, and $d(x, y)$ is the depth information at (x, y) . γ is a very small amount to ensure that the depth estimation process of the monocular image is very far away (the misjudgment). The weight value tilts the image feature to the close range.

For the obtained weight matrix, we use the $L2$ norm to normalize; namely,

$$V = \frac{W}{\|W\|_2}. \quad (15)$$

We perform the above sum pooling calculation on all N feature maps output by the convolutional layer L to obtain an N -dimensional feature vector φ of the convolutional layer, and use the same-dimensional PCA whitening, and then perform the whitening features obtained. The $L2$ norm is normalized, and finally N -dimensional image features are obtained.

For comparison, we extract the SPOC algorithm as follows:

$$a_k = \sum_{y=1}^H \sum_{x=1}^W c_{x,y} f_{x,y,k}. \quad (16)$$

Among them, $c_{x,y}$ is the Gaussian center prior, and its weights are set as follows:

$$c_{x,y} = \exp \left(\frac{(y - H/2)^2 + (x - W/2)^2}{2\sigma^2} \right). \quad (17)$$

Among them, ∂ is the distribution covariance, which is set to one-third of the length of the feature map center from the nearest boundary. It can be seen that the SPOC algorithm adds a Gaussian center prior on the basis of sum pooling and does not effectively reflect the icon in the image. Objects cannot reflect the characteristics of close-up shots.

EXPERIMENTS

Data Acquisition

Remote sensing is a means of collecting electromagnetic fields, force fields. Remote sensing maps record this information as an image. The classification of remote sensing maps is more complicated due to differences in sensors, imaging conditions, and types of information collected. They are also treated differently. The proposed method is suitable for visible light imaging and satisfies the remote sensing map of the vertical parallel projection imaging model. Considering the cost and simplicity of acquisition, this paper uses a satellite remote sensing map downloaded from the “satellite” mode of Google Maps. Its resolution accuracy is acceptable, it can be downloaded as long as it can be connected to the Internet, and it is completely free, so it is not only simple and practical but also low in cost. Its data components are more complex, mainly from Digital Globe and MDA Federal. Its imaging feature is a high-altitude bird’s-eye view, which provides information on the roof of the building and covers a large area. In addition, it is visible light imaging and approximate vertical shooting; it can be assumed that the downloaded remote sensing image conforms to the vertical parallel projection imaging model; that is, it can meet the requirements of the algorithm.

Scene Visualization

Basic Steps of Reverse Tracking

According to the set image size, the number of rays is determined to be slightly higher than the total number of pixels in the image. If the image size is 160*120, the total number of pixels is 19200, and the light is 22630; if the image size is 640*480, the number of pixels is 307200, and the light is 363388; when the image size is 1024*1280, the number of pixels is 1310720, the light is 1550877, and the number of rays is more than 18% of the number of pixels, so that each cell has a light, and the extra light

can be used to verify the correctness of each cell calculation. When the number of rays is determined, tracking begins. The specific tracking process is as follows: Step 1: Determine the position of the sensor and the viewing plane, and the light is directed to the scene through the viewing plane; Step 2: When the light reaches the set opaque surface, the light source is tracked according to the surface reflection principle of the object. At this time, no other object is blocked between the light sources, and the reflection portion is a bright portion; Step 3: When the light is reflected by the opaque surface, the scene entity that is set when tracking the light source is occluded (the sphere in the figure), and the reflective surface is dark; Step 4: When the light hits the solid in the scene (the sphere in the figure), it is reflected to the opaque surface, and then the reflection can be traced to the light source. The reflection of the object is bright and there are some optics of the opaque surface characteristic; the light reverse tracking step is shown in Figure 1.

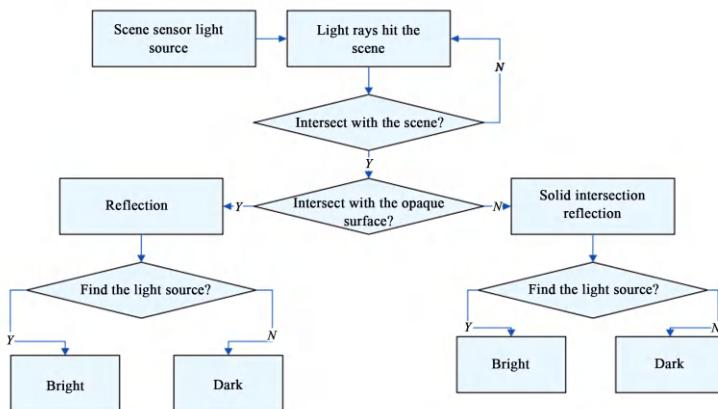


Figure 1. Light reverse tracking step.

Acceleration Algorithm

In the process of ray tracing by POV ray, there are a large number of rays intersecting the object. In order to improve the judgment efficiency of the intersection of light and object, POV ray uses a variety of acceleration algorithms, including multiple layers nested bounding box algorithm. The most important multiple levels nested bounding box algorithm is introduced here, depending on the buffer algorithm and the ray buffer algorithm. The bounding box algorithm is widely used in ray tracing. The traditional bounding box algorithm divides the scene into virtual cubes (such as

bounding boxes). First, it is determined which intersecting box the light intersects. If it intersects, it is judged whether the light and the entity in the bounding box are intersecting; compared with the original algorithm that uses image space as an order, this can greatly reduce the number of judgments and improve efficiency. The traditional bounding box algorithm steps are shown in Figure 2.

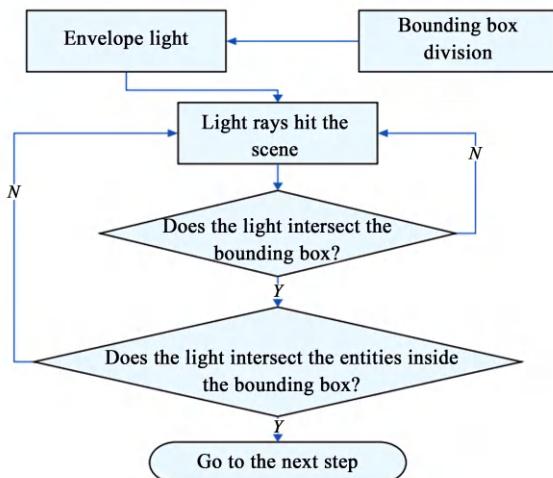


Figure 2. Traditional bounding box algorithm steps.

Since not every bounding box contains entities, POV ray uses a multiple layers nested bounding box algorithm. The multiple levels nested bounding box is similar to the structure of a tree, and the whole scene is first divided into larger bounding boxes.

It is judged whether the light intersects with the bounding box. If they intersect, the intersecting bounding box is decomposed into smaller bounding boxes, and the judgment is performed again, and the layers are subdivided into multiple layers in order. In this way, when the entities are discretely distributed in the scene, the computational efficiency can be greatly improved; however, when the entities are continuously distributed, the efficiency is not as good as the traditional bounding box algorithm. Therefore, when using the multiple layers nested bounding box algorithm, the key is to start the bounding box. In POV ray, Bounding=on/off controls whether to use the bounding box, and Bounding Threshold= n controls the starting layer number n of the bounding box. POV ray defaults to 3 layers.

Generation of House Scenes

The three-dimensional model of the house can be divided into two parts: the main body of the house and the roof. Therefore, the generation of the house scene can be divided into two parts: the main body of the house and the roof. Considering the complexity of the actual building, this article has been simplified accordingly: the complex building is broken down into a simple four-corner house model.

For four-corner houses, there are roughly two categories, one is the most common flat-top houses, and the other is non-flat-top houses. To build a three-dimensional model of a house, you must first obtain some information about the house from some way. Undoubtedly, the information related to the house is mainly the corner coordinates of the house, the elevation of the bottom, and the height of the house; for nonflat houses, the information about the height of the roof of the house is also known.

Acquisition of Corner Information of Houses

The house generally appears as a relatively regular shape in the two-dimensional image, so this paper decomposes it into a rectangle or a square, so that, for the house in the south of the south, it is only necessary to extract the coordinates of the two corners of the diagonal of the house, but for nonpositive south and north houses, the coordinates of the four corners of the house do not have a mutual relationship and must be extracted. In order to be able to generate the above two types of house models at the same time, this paper extracts all the corner coordinates of the house.

So, how to extract corner coordinates has become a focus of this article. There are two ways to extract the corner information. One is to extract the boundary information of the house as a straight line segment and then find the intersection point of the straight line segment as the coordinates of the corner of the house; the other is to extract directly according to the gray information of the image (corner coordinates). By comparison, the latter method is found to be simpler and more accurate than the first method. When directly using the image gray scale to extract the angular coordinates of the house, a corner feature extraction algorithm such as a Harris operator and a Moravec operator is often used, which are all difficult to meet the needs of this study to extract the corner coordinates of the house. But with MATLAB's powerful matrix processing capabilities, you can easily solve the problem.

Acquisition of the Elevation of the Bottom of the House

Using the classification map, we obtained the coordinate information of the four corner points on the bottom of the house. The acquisition of the elevation of the bottom of the house is simpler. Since the bottom of the house in the study area is generally flat, the elevation of the bottom surface is uniform. The DEM elevation value of the selected area has been assumed to be 0 before, so the bottom elevation value is 0 from the DEM.

DISCUSSION

Remote Sensing Image Registration Analysis

Registration Error

The registration error is shown in Table 1 and Figure 3. It can be seen from the table that the camera optical axis position error is below 1.3 meters. The result of this accuracy is acceptable. The result has reached the practical standard and is large. In some cases, the resolution of the satellite remote sensing map is also meter level or lower.

Table 1. Registration error

| | Isometric search registration method (m) | Main point registration method (m) |
|--------|------------------------------------------|------------------------------------|
| Test 1 | 1.2955 | 0.6411 |
| Test 2 | 0.4324 | 0.6575 |

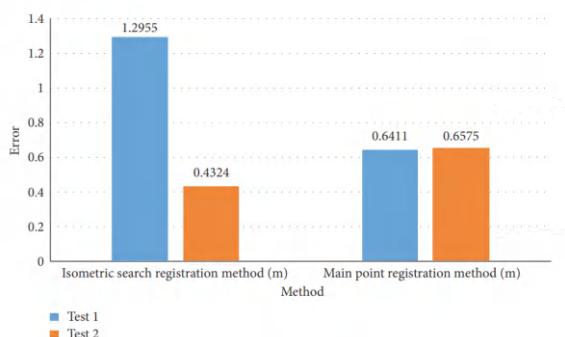


Figure 3. Registration error.

The error of the registration algorithm mainly comes from two aspects, one is the error of the experiment itself, and the other is the algorithm error. For the omnidirectional diagram, errors in the relative positional mounting of the components of the camera, errors in the attitude of the device at the time of shooting (such as the level of the device), errors in the imaging device itself, etc. can cause errors in the imaging of the omnidirectional image. Remote sensing maps also have corresponding errors. The error of the experiment itself can be reduced by precisely adjusting the installation of the imaging device and correcting the obtained image. This part will not be discussed in this paper. The registration algorithm error mainly comes from feature extraction, because, from the principle of the algorithm, as long as the extracted features are absolutely accurate, the calculated registration result error is very small. Therefore, the accuracy of feature extraction is the bottleneck of the accuracy of the registration algorithm, especially the principal point registration method, because the features used are basically no information redundancy.

Registration Time Consuming

The registration time consumption is shown in Table 2 and Figure 4. It can be seen from the table that both methods are relatively fast and can be controlled within 6 seconds. In comparison, the primary point registration method is faster.

Moreover, the time-consuming search registration method will increase with the increase of scene buildings, and the main point registration method will be more time-consuming.

This is consistent with theoretical analysis because the equal angle search registration method requires searching for a set of points. And the calculation amount used to judge the feasibility of any point in the point set increases with the increase of the number of linear features used and the principal point registration method only. One or several points need to be calculated, and the amount of calculation for calculating any point is fixed.

Table 2. Registration time

| | Isometric search registration method (m) | Main point registration method (m) |
|--------|------------------------------------------|------------------------------------|
| Test 1 | 5.058 | 0.162 |
| Test 2 | 1.813 | 0.114 |

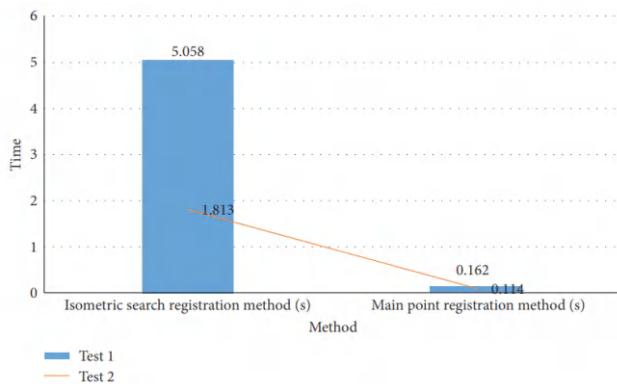


Figure 4. Registration time.

Scene Generation Results

Acquisition of Corners of Houses

Therefore, based on the classification map obtained from the remote sensing image, the image is read into MATLAB using the `imread` function to generate a matrix, and then the matrix gray information of the house area is given to another matrix of the same size as the original image, so that these two matrices coincide.

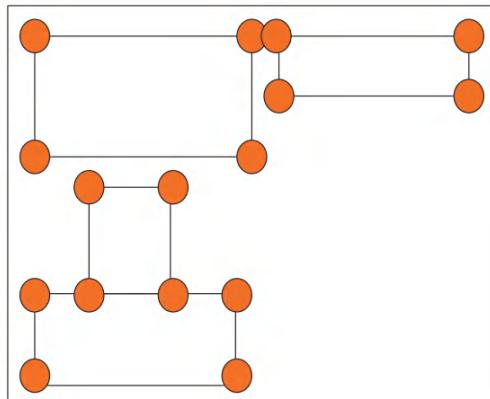


Figure 5. House corner extraction effect map.

There are only two values: the gray value of the house and the background gray value. Each house is then marked or given a different color using the

bwlabel function. Finally, the regionprops function can be used to extract the coordinates of the corner points of the region boundary. According to a certain rule, the coordinates of the four corner points of each house can be obtained. The corner points of the house extracted from the classification map are shown in Figure 5.

Generation of House Scenes

When obtaining the building model construction information, the various types of information should be stored in the form of a matrix or an array, and then the patch function is used to draw the planes. The color of the roof and the color of the surrounding walls are separated, and the storage of the coordinate files is also required. The results of using MATLAB 3D building modeling are shown in Figure 6.

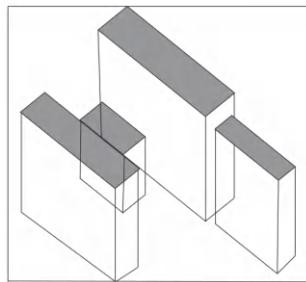


Figure 6. Housing scenario generation results.

The modeling method of the three-dimensional house is relatively simple, and the required three-dimensional data is small, but the distortion is large, and there is a defect in the aesthetic angle when used for three-dimensional visualization. However, the modeling required in this paper can reflect the radiation characteristics of the ground object. Because the radiation characteristics are only related to the material of the ground under the conditions determined by the lighting conditions and the external environmental conditions, the radiation characteristics of the house are mainly the study of the radiation characteristics of both the exterior wall and the roof of the house. For this purpose, this modeling approach is feasible.

Clustering Algorithm on Remote Sensing Images

Through experiments, it can be concluded that the clustering collective scale of ECUNGA algorithm is set to 5, 10, 20, 30, 40, and 50, the initial random

parameter is 20, the maximum allowable algebra of GA is 500, and the maximum allowable stagnant algebra is 50. Perform clustering on the three data sets of Iris, Wine, and Glass, respectively, and compare the best correct rate, average correct rate of 20 times, and worst correct rate of the clustering results under the collective scale of each cluster. The experimental results are shown in Table 3 and shown in Figure 7.

Table 3. The influence of remote sensing image clustering algorithms.

| Data set | Scale | Best | Average | Worst |
|----------|-------|------|---------|-------|
| Iris | All | 0.89 | 0.89 | 0.89 |
| Wine | All | 0.7 | 0.7 | 0.7 |
| Glass | 5 | 0.56 | 0.529 | 0.46 |
| | 10 | 0.54 | 0.48 | 0.46 |
| | 20 | 0.54 | 0.52 | 0.52 |
| | 30 | 0.54 | 0.53 | 0.52 |
| | 40 | 0.54 | 0.51 | 0.52 |
| | 50 | 0.54 | 0.53 | 0.52 |

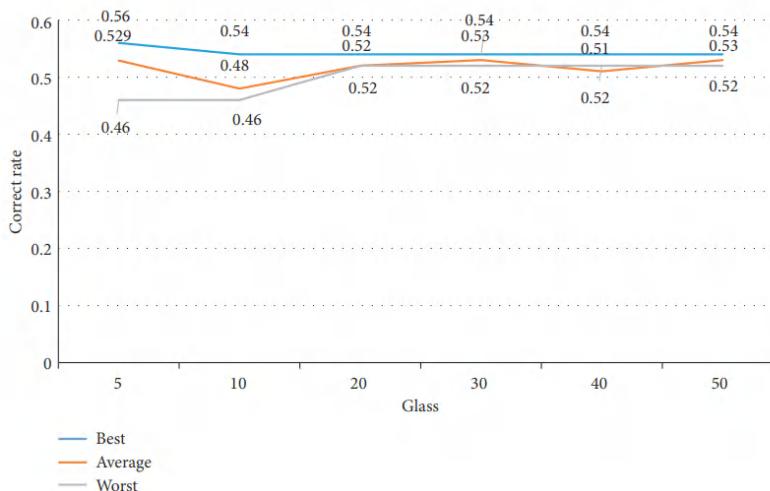


Figure 7. The influence of remote sensing image clustering algorithm.

From the experimental results, it can be found that the algorithm is not sensitive to the size of the clustering collective on the data sets Iris and Wine,

while the clustering collective scale on the Glass data set is not very stable at 5 and 10, the best and worst results. The difference is relatively large.

Remote Sensing Image Registration

Analyze the experimental part and get Tables 4 and 5 and Figure 8. Figure 9 shows the relationship between the number of sampling times of the RANSAC algorithm and the number of last correct matching point pairs and the execution time of the RANSAC algorithm. Red, green, and blue curves represent the experimental results on the test image pairs PA, PB, and PC, and the time unit is seconds. It can be clearly seen from the figure that the number of correctly matched feature points remains unchanged after the number of iterative samples is greater than 80, so the threshold value is set to 100 in the RANSAC algorithm based on the similarity transformation model. RANSAC algorithm execution time and sampling times meet a linear relationship, so the RANSAC algorithm has a higher execution efficiency. It is worth noting that the RANSAC algorithm based on affine transformation usually requires more iterations. Since SIFT is a classic image registration algorithm based on scale space and point features, the SIFT algorithm is still compared here.

Table 4. The relationship between algorithm RANSAC sampling times and correct matching point pairs

| Sampling times | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|----------------|----|----|----|----|----|----|----|----|----|-----|
| PA | 2 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| PB | 13 | 20 | 27 | 27 | 27 | 27 | 32 | 32 | 32 | 32 |
| PC | 76 | 76 | 76 | 76 | 76 | 91 | 91 | 91 | 91 | 91 |

Table 5. The relationship between algorithm RANSAC sampling times and running time

| Sampling times | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| PA | 0.05 | 0.021 | 0.026 | 0.039 | 0.04 | 0.045 | 0.047 | 0.06 | 0.07 | 0.08 |
| PB | 0.054 | 0.007 | 0.017 | 0.021 | 0.023 | 0.028 | 0.034 | 0.042 | 0.05 | 0.056 |
| PC | 0.045 | 0.017 | 0.021 | 0.031 | 0.034 | 0.037 | 0.041 | 0.054 | 0.063 | 0.08 |

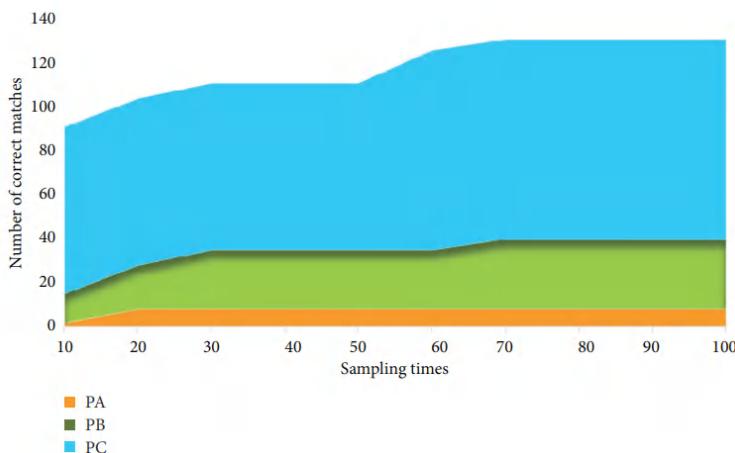


Figure 8. The relationship between the number of sampling times of the algorithm RANSAC and the correct matching point pair and running time.



Figure 9. The relationship between the number of samples of the algorithm RANSAC and the running time.

It shows the image after registration and fusion using the algorithm proposed in this chapter. It can be seen that the edges and regions overlap well. Therefore, it can be judged intuitively that the registration result is accurate, which again proves the effectiveness of the algorithm proposed in this chapter.

CONCLUSIONS

This paper analyzes and elaborates the concept of hyperspectral remote sensing system and the absorption and reflection of electromagnetic waves. The structure of the hyperspectral scene system is analyzed, and the influencing factors of solar radiation and atmospheric effects and ground reflectivity model are introduced. The imaging mode and imaging principle of the imaging spectrometer were studied and discussed, and the parameters of the remote sensing system of the two scenes were determined.

For the simple scene, the simulation principle and implementation method of spatial correlation and spectral correlation in the ground reflectivity model are studied. The research shows that the spectral and spatial correlation of the features makes the spectral reflection of the pixels belong to different locations of the same type of features. The rate curve fluctuates around its mean reflectance. At the same time, the influence of the atmosphere on solar irradiation intensity and atmospheric transmission coefficient is analyzed. The analysis shows that the worse the visibility of the atmosphere, the smaller the solar irradiation intensity and the atmospheric transmission coefficient.

Based on the remote sensing image, a three-dimensional scene model is constructed. Based on the three-dimensional scene model and the scene digital orthostatic image, a series of processing is extracted from the geometric and texture features of the scene orthostatic image and the contour features of the building. The minimum rule of the outline of the building is used to outsource the rectangle, and then these outsourcing rectangles are used as the constraint domain to segment the three-dimensional scene model to obtain the coarse single model. On this basis, the triangular and vertical classes of the triangular patches in the three-dimensional space are performed (classification). Calculate the roughness of the regional triangular patches. Based on this, combined with the height of the patch, the triangular patches that are not part of the building features in the coarse single model are purified, and then the adjacent patch growth method is used. The number of feature patches generates a more accurate building unit model and stores a single model from a single file, thus enabling automatic singular modeling of buildings in 3D scenes.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities, China University of Geosciences, Wuhan (no. CUGQY1911).

REFERENCES

1. A. Vanolo, "Smartmentality: the smart city as disciplinary strategy," *Urban Studies*, vol. 51, no. 5, pp. 883–898, 2014.
2. P. Neirotti, A. De Marco, A. C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in smart city initiatives: some stylised facts," *Cities*, vol. 38, no. 5, pp. 25–36, 2014.
3. L. Sanchez, L. Muñoz, J. A. Galache et al., "SmartSantander: IoT experimentation over a smart city testbed," *Computer Networks*, vol. 61, no. 6, pp. 217–238, 2014.
4. J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through Internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 112–121, 2014.
5. Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.
6. F. Leccese, M. Cagnetti, and D. Trinca, "A smart city application: a fully controlled street lighting isle based on raspberry-pi card, a ZigBee sensor network and WiMAX," *Sensors*, vol. 14, no. 12, pp. 24408–24424, 2014.
7. J. Gabrys, "Programming environments: environmentality and citizen sensing in the smart city," *Environment and Planning D: Society and Space*, vol. 32, no. 1, pp. 30–48, 2014.
8. K. Czynska, "Application of lidar data and 3D-city models in visual impact simulations of tall buildings," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, no. 7, pp. 1359–1366, 2015.
9. R. Heno and L. Chandelier, "3D modeling of buildings: outstanding sites," *IEEE Transactions on Ultrasonics Ferroelectrics & Frequency Control*, vol. 50, no. 11, pp. 1429–1435, 2014.
10. X. Zhang, Y. Lv, J. Tian, and Y. Pan, "An integrative approach for solar energy potential estimation through 3D modeling of buildings and trees," *Canadian Journal of Remote Sensing*, vol. 41, no. 2, pp. 126–134, 2015.
11. H. Son and C. Kim, "Semantic as-built 3D modeling of structural elements of buildings based on local concavity and convexity," *Advanced Engineering Informatics*, vol. 34, no. 1, pp. 114–124, 2017.

12. X. Zhang, L. Yang, and Y. Liu, “3D modeling of urban buildings and trees and its application in building-scale solar energy potential mapping,” *Journal of Basic Science & Engineering*, vol. 22, no. 3, pp. 415–425, 2014.
13. L. Díaz-Vilarino, K. Khoshelham, J. Martínez-Sánchez, and P. Arias, “3D modeling of building indoor spaces and closed doors from imagery and point clouds,” *Sensors*, vol. 15, no. 2, pp. 3491–3512, 2015.
14. M. Kedzierski and A. Fryskowska, “Terrestrial and aerial laser scanning data integration using wavelet analysis for the purpose of 3D building modeling,” *Sensors*, vol. 14, no. 7, p. 12070, 2014.
15. C. Huang and W. Bao, “A remote sensing image fusion algorithm based on the second generation curvelet transform and DS evidence theory,” *Journal of the Indian Society of Remote Sensing*, vol. 42, no. 3, pp. 645–650, 2014.
16. K. Sugihara and Z.-J. Shen, “Automatic generation of 3D building models by rectification of building polygons,” *Advanced Science Letters*, vol. 21, no. 12, pp. 3649–3654, 2015.
17. S. Rhee and T. Kim, “Dense 3d point cloud generation from uav images from image matching and global optimazation,” *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B1, no. 1, pp. 1005–1009, 2016.
18. S. G. Nebaba and A. A. Zakharova, “An algorithm for building deformable 3D human face models and justification of its applicability for recognition systems,” *SPIIRAS Proceedings*, vol. 3, no. 52, pp. 157–179, 2017.
19. J. Qin, C. Fang, Y. Wang, G. Li, and S. Wang, “Evaluation of three-dimensional urban expansion: a case study of Yangzhou City, Jiangsu Province, China,” *Chinese Geographical Science*, vol. 25, no. 2, pp. 224–236, 2015.
20. L. Ding, Y. Zhou, and B. Akinci, “Building Information Modeling (BIM) application framework: the process of expanding from 3D to computable nD,” *Automation in Construction*, vol. 46, no. 6, pp. 82–93, 2014.

SECTION 2

VIDEO GENERATION TECHNIQUES

CHAPTER 6

Realistic Speech-Driven Talking Video Generation with Personalized Pose

Xu Zhang and Liguo Weng

Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China

ABSTRACT

In this work, we propose a method to transform a speaker's speech information into a target character's talking video; the method could make the mouth shape synchronization, expression, and body posture more realistic in the synthesized speaker video. This is a challenging task because changes of mouth shape and posture are coupled with audio semantic information. The model training is difficult to converge, and the model effect is unstable in complex scenes. Existing speech-driven speaker methods cannot solve this problem well. The method proposed in this paper first generates the

Citation: Xu Zhang, Liguo Weng, "Realistic Speech-Driven Talking Video Generation with Personalized Pose", Complexity, vol. 2020, Article ID 6629634, 8 pages, 2020. <https://doi.org/10.1155/2020/6629634>.

Copyright: © 2020 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sequence of key points of the speaker's face and body postures from the audio signal in real time and then visualizes these key points as a series of two-dimensional skeleton images. Subsequently, we generate the final real speaker video through the video generation network. We take a random sampling of audio clips, encode audio contents and temporal correlations using a more effective network structure, and optimize and iterate network outputs using differential loss and attitude perception loss, so as to obtain a smoother pose key-point sequence and better performance. In addition, by inserting a specified action frame into the synthesized human pose sequence window, action poses of the synthesized speaker are enriched, making the synthesis effect more realistic and natural. Then, the final speaker video is generated by the obtained gesture key points through the video generation network. In order to generate realistic and high-resolution pose detail videos, we insert a local attention mechanism into the key point network of the generated pose sequence and give higher attention to the local details of the characters through spatial weight masks. In order to verify the effectiveness of the proposed method, we used the objective evaluation index NME and user subjective evaluation methods, respectively. Experiment results showed that our method could vividly use audio contents to generate corresponding speaker videos, and its lip-matching accuracy and expression postures are better than those of previous work. Compared with existing methods in the NME index and user subjective evaluation, our method showed better results.

INTRODUCTION

The task of a speech-driven speaker video refers to a technology that automatically generates a video of a corresponding character's speech through a computer-based audio information. The content of the talking must be consistent with the character's pose in the video. Traditional speech-driven talking video requires professional equipments and operators to perform character modeling, which is usually very expensive for custom use. In recent years, with the successful application of deep neural networks, data-driven speech and video synthesis methods have been proposed. These methods often require the use of a large amount of high-quality audio and video data, and the production process is complex, but the synthesized speaker's mouth posture matching effect is poor.

The current mainstream methods mainly focus on facial speaker synthesis and do less work on body postures and facial expressions. Specifically,

the existing methods [1, 2] input the speaker’s voice information into the recurrent neural network to obtain 3D face model parameters, then map the fitted 3D face model to 2D key points as inputs of the video synthesis module, and then output corresponding speaker pictures through the video synthesis model.

Due to the weak representation ability of the 3D face model parameter network, the key point error obtained from the 3D face model conversion is larger, the 3D face model needs to be used as an intermediate state for conversion, resulting in a complicated overall process. Eskimez et al. [3] converted the facial key points into the average face space in the dataset to remove ID features and simplified the task. Although the key point indicators obtained from the network output are relatively low, the posture expressions are very monotonous and rigid, and hence, the synthesized speaker video is not realistic enough.

As mentioned above, the matching effect of existing speech-driven speaker methods is not ideal, and the synthesized speaker video has a jitter phenomenon. In order to solve the above problems, this paper proposes a method to convert the speaker’s voice information into the target person’s talking video. We use the Dilated Depthwise Separable Residual (DDSR) unit to encode the audio features [4, 5], and then use the GRU network layer [6] to learn the temporal features and constrain the network outputs using content loss functions.

Through this network structure, the audio content and temporal correlation information are effectively encoded simultaneously, the facial key point index of the model output is lowered, and the mouth shapes and postures of the synthesized speaker video are matched with audio contents better, plus, the synthesized speaker video is more natural and realistic. In the process of training and testing, we insert the specified pose sequence frame into the pose sequence, which makes the audio conversion to the speaker’s mouth shape and posture more natural and vivid. In order to enrich the speaker’s detailed texture, we introduce a local attention mechanism in the key point network and add spatial weights to the face, fingers, and other parts of the character to get higher attentions.

Finally, in order to better evaluate our system, we used high-resolution and frame rate (FPS) cameras to create a dataset containing audio and video for multiple targets reading selected articles. Compared with the existing methods, our method produces better visual perception. In Figure 1, we show some images of our synthesized speaker video.

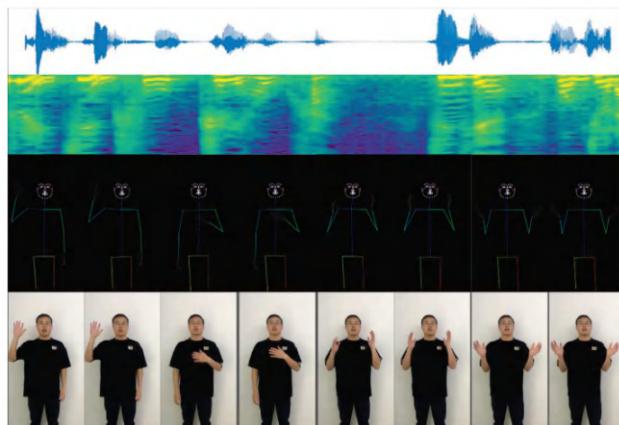


Figure 1. Speech-driven talking video: a given piece of audio/text can be used to drive the video of the specified speaker.

In summary, the contributions of our work are

(1) We use a novel Dilated Depthwise Separable Residual (DDSR) unit. This network structure can effectively represent the audio content and temporal correlation, and the facial key point index of the model output is lower. At the same time, the network model is used to model the key points of the face and human posture, respectively. After preprocessing, it uses the loss function to optimize iteratively. The results show that the face details and human postures are better.

(2) We use the first-order differential loss function and the pose perception loss function [7, 8] to optimize the model. Among them, the first-order differential loss function can smooth the pose of the front and rear frames, and the pose perception loss function uses the spatiotemporal graph to form a hierarchical representation of the pose sequence, so as to constrain the temporal-spatial information output from the network.

(3) We establish a pose keypoint map to add richer poses and expressions to the generated human poses. In addition, we also provide a method to convert the pose in the existing sequence window into the corresponding keyframe pose sequence.

RELATED WORK

Given a speaker’s audio information, the generation of the corresponding person speaking video has attracted many researchers’ interests. Earlier

works mainly used the Hidden Markov model (HMM) to generate corresponding relationships between speech and facial motions [9–14]. Among them, Brand [15] proposed voice puppetry as an HMM-based method for generating conversation faces driven only by voice signals. In another study, Cosker et al. [10, 11] proposed a hierarchical model that can animate the subregions of the face independently of speech and merge them into a complete face video.

In recent years, with successful applications of deep neural networks, the related work of speech-driven speaker based on deep learning method has been proposed. Among them, Suwajanakorn et al. [16] designed an LSTM network to directly generate the target identity talking face video from the audio. However, this method needs to record a large number of facial videos with specific target identities, it limits its application in many scenarios. Linsen et al. converted audio information into the 3D face model parameter space and then the fitted 3D face model to 2D facial key points. Their network uses several layers of recurrent neural networks as encoding, and the network feature learning ability is relatively weak. The facial key points obtained by the conversion of the 3D face model have a large error, and the 3D face model needs to be used as an intermediate state for conversion. This leads to the complexity of the overall process.

In addition, including the single-stage method of direct conversion of audio to speaker video space, many researchers divide the task of speech generation into two stages. Usually, the key point information only responds to the voice content information. Pham et al. [17] first used the LSTM network to map voice features to 3D deformable shapes and rotation parameters and finally generated 3D animated faces in real time based on the predicted parameters. In literature [18], they further improved this method, replacing speech features with original waveforms as inputs and the LSTM network with a convolutional structure. However, compared with the speech-generated gesture keypoint network in our method, their method is less intuitive in shapes and rotation parameters, and the mapping from these parameters to specific gestures or facial expressions is not clear. In another related work, the key points of the face that they generated are for a standardized average face, rather than for a specific target identity. Although this helps to eliminate factors that are not directly related to voice, the predicted sequence of key points for the posture is unnatural. [19] An extended complex human motion synthesis method based on autotuning recurrent network is proposed. They can simulate more complex movements, including dances or martial arts. In the second stage of work,

most methods use vid2vid [20] to enhance the time consistency between adjacent frames. Shysheya et al. [21] proposed a method to generate realistic videos from skeleton sequences without establishing a 3D model. Our method also uses the vid2vid network to synthesize the final speaker video from the posture skeleton picture and obtains better results. For the detailed texture information of the face and hands, we use separate discriminators to optimize these parts in vid2vid.

Our method expands the data of random audio samplings and uses a more effective network structure to learn audio contents and timing correlations. The loss function uses the first-order differential loss and poses perception loss to optimize output pose timing stability and matching accuracy. At the same time, the keyword wake-up technology is used to convert the generated sequence poses into specified action poses. A large number of experimental results show that our method generates a natural and realistic speaker video for talking audio, and its lip matching and expression posture are more expressive than those of the previous work.

METHODS

In this section, we mainly introduce different modules of the network. The overall network structure is shown in Figure 2. In our approach, the input information can be either audio or text.

When the audio information is used as the speaker synthesis network input, we convert the audio data into log-mel features; the aud2kps network is used to get the human body postures and facial key points.

Using the Dictionary Building and Key Pose Insertion method to insert a specified action frame into the generated key point sequence, the synthesis effect is more natural and realistic, and then the output key points of facial and human posture are visualized as a series of 2D skeleton images, and these 2D skeleton images are further fed into the Vid2vid generation network to generate the final talking images.

When the input is text information, it is necessary to use the acoustic model to convert the text information to obtain a unified log-mel feature as the input of the Aud2Kps network. The following steps are the same as the audio signal input process. The text-to-speech method (TTS) is currently very mature and commercialized, and we use the open source tactron2 [22] to complete the text conversion results which we want. In the following sections, we describe each module of our architecture.

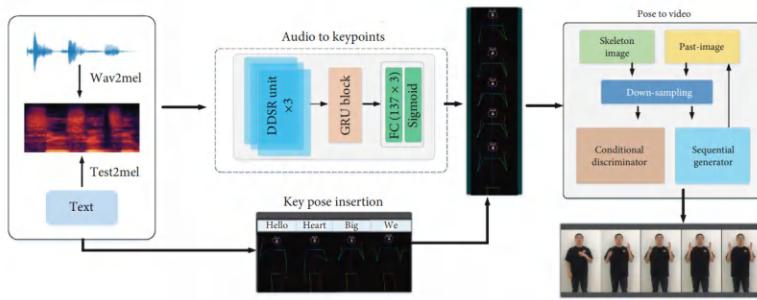


Figure 2. Pipeline of our method: the input information can be audio or text. When the audio information is used as the speaker synthesis network input, we convert the audio data into log-mel features and then input the Aud2Kps model to get the pose key points. When the input is text information, it is necessary to use the acoustic model to convert the text information to the log-mel feature as the input of the Aud2Kps network. The following steps are the same as the audio signal input process.

Pose Keypoints

In the process of audio-video conversion, we use the key points of human body posture as the intermediate state representation so that the span of the two spatial features will not be too large. Compared with using the 3D human body model as the intermediate state representation, it is more convenient and universal in the process of training and reasoning. We use the open source method OpenPose [23, 24] to obtain the key points of the human body posture. These key points include a total of 137 position coordinate information of the body, feet, hands, and faces. Firstly, we construct these 2D key points and audio information into a content sequence and then train the Aud2Kps network to generate 2D coordinates corresponding to the posture key points from the audio speech information.

Audio to Keypoints (Aud2Kps)

As shown in Figure 2, our Aud2Kps network takes log-mel spectrogram as the input. $[x_0, x_1, \dots, x_n]$ is the input vector of audio/text encoding and $[y_0, y_1, \dots, y_n]$ is the output open-pose key point vector. The log-mel spectrum feature extracted from audio [25] is a set of 80-dimensional vectors. We designed a DDSR unit to encode the semantic content of features, then input the GRU model to learn the timing features, and finally input the full

connection layer and sigmoid activation function to obtain the key point information of the face and human body posture. Our network structure effectively characterizes the audio content information and the correlations between the front and rear time series so that the NME index of the facial key points output by the model is lower. When Aud2Kps maps the audio sequence to the pose sequence, since different parts of the human body have different scales, we need to give them different weights. Therefore, for the body, hands, facial contours, and mouth positions, we set the attention weights as 1, 10, 50, and 100, respectively. We also use the first-order differential loss between two consecutive poses to ensure that the output pose key points are more smooth and natural.

The MSE loss function L_{MSE} is given by

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N \|y_i - \hat{y}_i\|_{l_2}. \quad (1)$$

The first-order temporal differential loss L is given by

$$L_{\text{First-order}} = \frac{1}{N} \sum_{i=1}^N \|\hat{y}_i - \hat{y}_{i-1}\|_{l_1}. \quad (2)$$

At the same time, we use a pose-perception loss function to calculate the content loss between the real and generated pose key points. In most content loss, the VGG network is used as the feature extractor [26, 27], the pose perception loss function uses ST-GCN as the feature extractor of the perception loss function, and the hierarchical representation of the skeleton sequence is formed by using the space-time graph and can be obtained from automatically learn spatial and temporal patterns in the data.

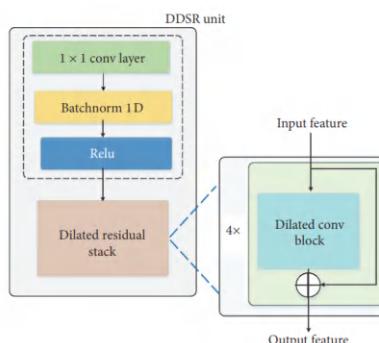


Figure 3. Dilated Depthwise Separable Residual (DDSR) unit network.

We use a dilated residual block in each DDSR unit [28] so that each subsequent layer has a long time span, and the receptive field of the convolutional layer after expansion increases exponentially with the number of layers. This method can effectively increase the sensing receptive field of each output time step and obtain a better long-range correlation. The implementation details of the DDSR unit are shown in Figure 3.

Given a pretrained GCNnetwork φ , we define a collection of layers φ as φ_l . For a training pair (P, M) , where P is the ground truth skeleton sequence and M is the corresponding piece of audio, our perceptual loss is

$$L_{\text{Perceptual}} = \sum_{l=1}^N \beta_l \|\varphi_l(P) - \varphi_l(G(M))\|_{l_1}. \quad (3)$$

Here, G is the first-stage Aud2Kps network in our framework. The hyperparameters β_l balance the contribution of each layer l to the loss.

Since the text input will not affect the model efficiency even there is difference in voice characteristics between people, the text input will make the network model more general. Similar to the process of using audio-training Aud2Kps, we convert the text segmentation into phonemes and then use the acoustic model through feature encoding to generate log-mel features as the input of the subsequent speaker synthesis model. We use the open source tacotron2 model to convert the text into a log-mel feature. The following process is the same as the process of audio-to-keypoint.

Key Pose Insertion

During the model training process, we found that although the Aud2Kps model can synchronize the audio and video content of the speaker very well, the generated character action sequence is too monotonous. This is mainly because the character action sequence is the same at most times in the training set, and the action sequence with posture change is very sparse in the whole training set [29]. In order to make the gesture actions in the synthesized speaker video more expressive and diverse, we designed a gesture sequence dictionary. When the specified keywords appear in the audio content, the corresponding window of the gesture sequence output by Aud2Kps is converted into the specified action, and the posture transformation here uses the posture transformation matrix stored in the posture sequence dictionary.

We select some posture action sequences from the recorded videos and then construct these posture sequences and the corresponding wake-up words into a posture sequence transformation dictionary (composed

of transformation matrix). Once the input audio content appears in the dictionary, we will transform the existing pose sequence with a certain probability. The probability between different words may be different. In order to maintain a smooth transition to this pose, we smooth the adjacent frames.

Pose to Video

We use the vid2vid generator network to convert our generated skeleton images into corresponding speaker videos. After the key points of the human body posture are obtained from the Aud2Kps network, they are visualized as a series of 2D skeleton images, and these 2D images are further fed into the Vid2vid generator network [20] to synthesize the final speaker video. In our network structure, different positions of the human body pay attention to different degrees of importance and people tend to pay more attention to the part of the face and hands. In order to make the vid2vid network pay more attention to the detail texture synthesis of face and hands, we use a separate discriminator network to train the models of face and hand regions to ensure that the discriminator pays more attention to the generated facial and hand details.

EXPERIMENTS

TalkingPose Dataset

Our audio and video data can be from related speeches or broadcast videos on websites. However, most of the video resources on websites are shot at different times with change of character decorations and clothing styles, which increases uncontrollable factors of samples and increases the difficulty of training. Therefore, we specify speakers to perform audio and video recording. Our speakers read different themes and scripts, and the entire recording time of audio and video is about 2 hours. The video resolution is 1920×1080 , and the speed is 30 frames per second.

After recording the video data, the audio data can be directly separated from the corresponding video data. We sample audio data with a sampling rate of 16 kHz and convert them into log-mel features as the network input. Since audio may have different volume levels, we first normalize its volume through RMS-based normalization [29]. Then, through sparse fast Fourier transform (sfft), the audio is converted from time-domain representation to

frequency-domain representation. The value on each frequency represents the energy of the frame of speech signal at the current frequency, and a set of multiple triangular filters are used. The linear spectrum after sfft is processed to obtain 80-dimensional low-dimensional features to simulate the suppression of high-frequency signals by human ears. This method is widely used in speech feature extraction. We use random sampling strategies to expand the dataset for the audio features in the same segment, and the log-mel feature and the posture key point sequence are 1 : 4 as the model input. Figure 2 is a partial example of our dataset.

Implementation Details

All the models are trained on 8 Nvidia GeForce GTX 1080 Ti GPUs. For the first stage of the Aud2Kps model in our framework, the model is implemented in PyTorch [24] and takes approximately one day to train for 500 epochs. For the hyperparameters, the dimensions of the output channels of the three DDSR units are set to [128, 256, 512], the number of hidden nodes in the GRU timing network is set to 256, and the number of nodes in the final fully connected layer of the network is set to the number of OpenPose parameters 137×3 . For the pretraining process of ST-GCN, ST-GCN achieves 49% precision on our TalkingPose dataset. By using the Adam optimizer [30] to minimize the L_2 norm loss of key points in Pytorch, we ensure that the audio features are effectively converted to the corresponding pose key points. The network training batch size is 64, and the learning rate is 0.001. For the second stage that transfers pose to video, the Vid2vid model takes approximately seven days to train for 20 epochs, and the hyperparameters of it adopts the same as [20]. During model training, the data preprocessing part will automatically crop the original video resolution to 1024×1024 . Therefore, our results are all 1024×1024 resolution.

Evaluation Metrics

The task of evaluating speech-driven talking videos is not simple because (1) there is no benchmark dataset to evaluate speech-to-human pose video; (2) the effect of people's speech-driven talking video performance is very subjective, so it is difficult to define model performance. We choose to compare our results with SoTA approaches using the user study. We compare LearningGesture [31], neural-voice-puppetry [32], EverybodyDance [33], and Personalized-bodyPose [29] in our user study. In the evaluation metrics of the user study, we refer to the Mean Opinion Score (MOS) [30] of the

evaluation index in the text-to-speech (TTS) method [34] to measure the effectiveness of different models. Table 1 shows the MOS of user study for all methods. We get the best overall quality score over the other 4 SOTA methods.

Table 1. Mean Opinion Score (MOS) of 100 participants on 4 questions. *Q1*: completeness of body. *Q2*: the face is clear. *Q3*: the body movement is correlated with audio. *Q4*:overall quality

| | <i>Q1</i> | <i>Q2</i> | <i>Q3</i> | <i>Q4</i> |
|---------------------------|-----------|-----------|-----------|-----------|
| Learning gesture [31] | 3.414 | 3.659 | 3.914 | 3.308 |
| Neural-voice-puppetry[32] | 3.202 | 3.840 | 3.180 | 3.542 |
| EverybodyDance [33] | 3.944 | 3.662 | 3.680 | 3.681 |
| Personalized-bodyPose[29] | 3.894 | 4.011 | 3.383 | 3.762 |
| Our method | 3.901 | 4.083 | 3.526 | 3.778 |

The quantitative model predicts the effect of speaking posture. Even if the people speak the same sentence, he will not perform the same gesture at different moments. It is difficult to judge whether the speech content is correctly converted to the human body posture. However, the facial and mouth shapes of the same sentence are almost the same. Therefore, we evaluate the performance of the model through facial key points. We use the NME indicator [35] to measure the deviation degree that the audio information is converted into corresponding real facial key points. NME is widely used in facial landmark detection to evaluate the quality of models. It is calculated by the average Euclidean distance between predicted and ground truth landmarks, and then it is normalized to eliminate the impact caused by the image size inconsistency. NME for each pose is defined as

$$\text{NME} = \frac{1}{L} \sum_{k=1}^L \frac{\|p_k - \hat{p}_k\|_2}{d}, \quad (4)$$

where L refers to the number of landmarks, p_k and \hat{p}_k refer to the predicted and ground truth coordinates of the k_{th} landmark, respectively, and d is the normalization factor, such as the distance of eye centers (interpupil normalization, IPN) or the distance of eye corners (interocular normalization, ION).

To evaluate the effect of pose to video, we use a subjective evaluation method, a user study. In order to evaluate the final output video, we invited 100 participants on the Internet to conduct a subjective test. We showed a

total of three videos to participants. Two of them are our synthetic videos, of which, one is a speaker video generated from real human audio, and the other one is a speaker video generated from TTS synthetic audio, and the remaining one is the original real speaker video. These 3 videos are randomly scrambled, and we did not tell the participants the tags behind the videos. Participants need to subjectively rate the quality of these videos, from 1 (strongly disagree) to 5 (strongly agree). The evaluation options include (1) the integrity of the human body; (2) the face of the speaker in the video is clear; (3) the posture of the person in the video looks natural and smooth; (4) the overall visual experience of the video is realistic.

As shown in Table 2, the overall score of our synthetic video four items is 3.795, and the real video is 4.365, which means that the overall effect of our proposed synthetic talking video reaches 86.94% of the real video. It is closer to the real speaker effect in terms of facial details and human body posture integrity. The video score generated by TTS is worse than the voice generation effect, and the reasons are the same as those in Table 3. The main reason is that the synthesized audio has information loss, and hence it is different from the original audio. This loss brings errors into the generated human body postures so that the visual score of the synthesized speaker video is low.

Table 2. Mean Opinion Score (MOS) of 100 participants on 4 questions. $Q1$: completeness of body. $Q2$: the face is clear. $Q3$: the body movement is correlated with audio. $Q4$: overall quality

| | $Q1$ | $Q2$ | $Q3$ | $Q4$ |
|--------|------|------|------|------|
| Synth. | 4.14 | 4.37 | 2.92 | 3.75 |
| TTS | 4.10 | 3.80 | 2.58 | 3.39 |
| Real | 4.31 | 4.42 | 4.33 | 4.40 |

Table 3. Evaluation metrics used NME (%) on facial landmarks (lower is better)

| | Orig. | Only-GRU | TTS-mel | Text |
|-----|-------|----------|---------|-------|
| 0.5 | 4.925 | 5.673 | 5.871 | 5.693 |
| 1.0 | 4.921 | 5.640 | 5.885 | 5.690 |
| 1.5 | 4.853 | 5.644 | 5.877 | 5.614 |
| 2.0 | 4.907 | 5.647 | 5.829 | 5.607 |

Ablation Study

We use the NME index to evaluate facial key points on the test set. As shown in Table 3, we use different time-length datasets (0.5 h, 1.0 h, 1.5 h, and 2.0 h, respectively) to train the model and observe the impact on the accuracy of pose prediction. In addition, we evaluate the audio data of text synthesis to observe the impact of sound changes on the results, use text to train and test the network, and compare the results with the audio results. Finally, we compare the training using only the GRU network with that using our network structure.

From Table 3, we can notice the following. (1) After the audio training set is increased to 1.5 h, the model benefit will not be great by increasing the dataset, but the model effect can also be improved by further increasing the amount of data on the text training set. (2) From the model indicators obtained from audio and text data, it can be seen that the effect of audio is worse than that of text, indicating that the audio conversion to the key points of the face is more accurate. (3) The audio data synthesized by text is tested on the model. The effect is not as good as the original audio mainly because the synthesized audio has information loss, and hence it is different from the original audio. (4) Using the DDSR unit network model is better than only using the GRU network structure as feature extractor. Although only using the GRU network can capture the correlation between the front and rear frames, the feature representation ability is weak. The combination of the DDSR unit and the GRU can make up for this shortcoming.

To prove the effectiveness of our key pose insertion method, we conducted another user study. In this study, we simply presented a pair of composite videos with and without inserting key poses. Participants only need to evaluate which of the two videos is more natural and realistic. From the final user rating, it is shown that the synthesized video with gesture actions being inserted into its existing posture sequence scored 81.3% and the synthesis video without the key frame poses only received 18.7% of votes. This illustrates the effectiveness of inserting pose key points to enrich speech-driven talking video synthesis.

CONCLUSION AND FUTURE WORK

In this work, we propose a new method to generate realistic talking video from audio information. We sample the audio data randomly and use a more effective network structure to learn the audio content and timing correlation.

We use first-order differential loss and pose perception loss to optimize the network output so that the face and pose key points obtained by audio conversion are smoother and the index performance is better. At the same time, by inserting a specified action frame into the synthesized human pose sequence window, the synthesized speaker's action posture is more natural and realistic. Our objective and subjective evaluation comparison results are very competitive over the existing methods. Our current method has good results in single-speaker scenarios. In multispeaker audio-video conversion tasks, we use TTS technology to convert speech to text to eliminate the inconvenience caused by voice ID information. In the future, we will further explore the work related to multispeaker to multitarget character video synthesis.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of PR China (42075130).

REFERENCES

1. T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, “Audio-driven facial animation by joint end-to-end learning of pose and emotion,” *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–12, 2017.
2. Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, “VisemeNet: audio-driven animator-centric speech animation,” *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–10, 2018.
3. S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “Generating talking face landmarks from speech,” in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, pp. 372–381, Guildford, UK, June 2018.
4. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
5. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.
6. K. Cho, B. Van Merriënboer et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014.
7. S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the Association for the Advancement of Artificial Intelligence*, pp. 1–9, New Orleans, LA, USA, February 2018.
8. J. B. Estrach, P. Sprechmann, and Y. LeCun, “Super-resolution with deep convolutional sufficient statistics,” in *Proceedings of the 4th International Conference on Learning Representations, ICLR*, San Juan, Puerto Rico, May 2016.
9. K. Choi, Y. Luo, and J.-N. Hwang, “Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system,” *The Journal of VLSI Signal Processing*, vol. 29, no. 1/2, pp. 51–61, 2001.

10. D. Cosker, D. Marshall, P.L. Rosin, and Y. Hicks, “Speech driven facial animation using a hidden markov coarticulation model,” in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 128–131, Cambridge, UK, August 2004.
11. D. Cosker, D. Marshall, P. Rosin, and Y. Hicks, “Video realistic talking heads using hierarchical non-linear speech-appearance models,” in *Proceedings of the Mirage 2003*, Le Chesnay-Rocquencourt, France, March 2003.
12. L. D. Terissi and J. C. Gómez, “Audio-to-visual conversion via HMM inversion for speech-driven facial animation,” in *Proceedings of the Brazilian Symposium on Artificial Intelligence*, pp. 33–42, Salvador, Brazil, October 2008.
13. L. Xie and Z.-Q. Liu, “A coupled HMM approach to video-realistic speech animation,” *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.
14. X. Zhang, L. Wang, G. Li, F. Seide, and F.K. Soong, “A new language independent, photo-realistic talking head driven by voice only,” in *Interspeech*, pp. 2743–2747, Springer, Berlin, Germany, 2013.
15. M. Brand, “Voice puppetry,” in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 21–28, Los Angeles, CA, USA, August 1999.
16. S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: learning lip sync from audio,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.
17. H.X. Pham, S. Cheung, and V. Pavlovic, “Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach,” in *Proceedings of the 1st DALCOM Workshop, CVPR*, Guildford, UK, 2017.
18. H. X. Pham, Y. Wang, and V. Pavlovic, “End-to-end learning for 3D facial animation from speech,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 361–365, Boulder, CO, USA, October 2018.
19. Y. Zhou, Z. Li, and S. Xiao, “Auto-conditioned recurrent networks for extended complex human motion synthesis,” in *Proceedings of the International Conference on Learning Representations*, Vancouver, Canada, May 2018.

20. T. C. Wang, M. Y. Liu, and J. Y. Zhu, “Video-to-video synthesis,” in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, December 2018.
21. A. Shysheya, E. Zakharov et al., “Textured neural avatars,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2397, San Juan, PR, USA, June 2019.
22. J. Shen, R. Pang, R. J. Weiss et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, Calgary, Canada, April 2018.
23. Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, p. 1, 2019.
24. S. E. Wei, V. Ramakrishna, and T. Kanade, “Convolutional pose machines,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, Las Vegas, NV, USA, June 2016.
25. K. Kumar, R. Kumar, and T. De Boissiere, “Melgan: generative adversarial networks for conditional waveform synthesis,” in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 14910–14921, Vancouver, Canada, 2019.
26. Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1520, Venice, Italy, October 2017.
27. J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proceedings of the European Conference on Computer Vision*, pp. 694–711, Amsterdam, Netherlands, October 2016.
28. S. Mehta, M. Rastegari, and A. Caspi, “Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 552–568, Munich, Germany, September 2018.
29. M. Liao, S. Zhang, and P. Wang, “Speech2video synthesis with 3D skeleton regularization and expressive body poses,” in *Proceedings of the Asian Conference on Computer Vision*, Kyoto, Japan, December

2020.

30. R. Skerry-Ryan and E. Battenberg, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *Proceedings of the 35th International Conference on Machine Learning*, pp. 4693–4702, Stockholm Sweden, July 2018.
31. S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, “Learning individual styles of conversational gesture,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506, Long Beach, CA, USA, June 2019.
32. J. Thies, M. Elgharib, and A. Tewari, “Neural voice puppetry: audio-driven facial reenactment,” in *Proceedings of the European Conference on Computer Vision*, pp. 716–731, Glasgow, UK, August 2020.
33. C. Chan, S. Ginosar, T. Zhou, and A.A. Efros, “Everybody dance now,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5933–5942, Seoul, Republic of Korea, October 2019.
34. J. M. Valin and J. Skoglund, “LPCNet: improving neural speech synthesis through linear prediction,” in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, Brighton, UK, May 2019.
35. A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D& 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks),” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1021–1030, Venice, Italy, October 2017.

CHAPTER 7

Video Transformation in Big Video Era and its Impact on Content Editing

Mingzhi Yin

School of Foreign Languages, North China Electric Power University,
Beijing, China.

ABSTRACT

In the “Big Video Era”, the amount of video has increased dramatically and the presentation mode of videos has also changed fundamentally with the development of technology. As huge changes will bring many new opportunities, it is urgent to explore the changes of video in the big video era and its impact on content editing for the development of the video industry. By studying and analyzing the relevant literature, combining the current development of the video industry and specific practical cases, this thesis explores the transformation of video content in the “Big Video Era”, analyzes

Citation: Yin, M. (2021), “Video Transformation in Big Video Era and Its Impact on Content Editing. Open Journal of Social Sciences”, 9, 116-124. doi: 10.4236/jss.2021.911010.

Copyright: © 2021 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

the positive and negative impacts, and discusses the new requirements in video content editing.

Keywords:- Big Video Era, Video Transformation, Content Editing

INTRODUCTION

Now we have entered a new era of video, which is the Big Video Era. First of all, what is the Big Video Era? The definition of the Big Video Era is the comprehensive expression of video. In the era of mobile and socialization, there are many different media forms, so it presents the pattern of “big video”. The beginning of the Big Video Era is the network video and sharing video, which is also the product of media integration. In the era of big video, technology is being upgraded constantly, and now we have 5G technology.

The Big Video Era has already had a huge impact on video content with the explosion of new videos appearing every day on various platforms and software. John Hoffman, CEO and Director for GSMA Ltd., said at WAIC 2019 that, “We have a huge unexplored frontier, a big opportunity for growth, and unlimited potential for growth. Because video is the biggest driver of data flow in the world today, video will bring endless opportunities for all players in the industry” (Embrace the Big Video Era, 2017). Short videos are becoming the mainstream of the day, creating an interceptable, reconfigurable, and embeddable TV stream and video textualization dynamic.

Therefore, in big video era, changes in the way media converge have led to changes in video, which also have a huge impact on content editing. In order to grasp these new opportunities and seize the opportunities created by the times, we need to explore the new requirements for video editing and video content in the big video era in order to better create videos that meet the progress of the times and human aesthetics.

LITERATURE REVIEW

If you search “big video era” in JSTOR, you can get 5059 results. Of academic content, you can get 424 Journals, 1001 Book Chapters and 576 Research Reports. Of primary source content, you can get 2866 serials, 106 books and 86 documents.

In China, if you search for “big video era” or its related keywords, such as “short video”, “big data era”, “video content”, etc. on CNKI, you can get 2W+ results. In terms of years, the number of papers shows a gradual

increase, and reaches the peak in 2020. Its research content is mainly about short videos, Internet, new media, various short video platforms, media integration, communication strategies, etc.

Its research is mainly based on several directions. Firstly, it is a specific analysis study of a particular short video. Based on a specific short video software to carry out research, to carry out analysis of its data and audience, to carry out case studies. For example, the analysis of Li Ziqi's video in TikTok platform accounts for more than 200 articles. Secondly, it is a kind of research which combines short video and other forms of media together. As an emerging media means, the integration of short video and other media has been attracting much attention, and some researchers are also very interested in this field, and they summarize and explore the integration practice of short video and other media from the current situation of the industry, and summarize their experiences. Articles such as "The road to break the siege of mainstream media in the era of short video" and "How local traditional media can do well in short video to enhance their influence" have been carefully analyzed and studied through the perspective of the mutual influence of short video and traditional media and the drastic impact of short video on traditional media. We can also search a lot of research on the change of traditional media, which is always closely related to the new media and the big video era. When you search for traditional media in the search bar, you will find a large number of papers on traditional media development and innovation. This also coincides with the main tone of this thesis. Changes corresponding to the new media era must occur and new requirements must be put forward in order to comply with the new trend. Fourth, the future development trend of short video has also been the mainstream of popular research. For example, Nirobaier Elti and Zheng Liang mentioned in "Characteristics, Trends and Dilemmas of Short Video Content Production in the New Media Era" (Alti & Liang, 2021) that the content production of short video in the new media era has formed socially driven, emotionally stimulating and multi-valued production modes while developing rapidly, and has shown the integration into more complex In addition. Some scholars have also explored the development of short video from other perspectives, and the prospects for the development of the short video industry from a technological perspective. Some scholars believe that the main development trend of the short video industry should focus on technology upgrading, such as AR, VR, data visualization and other high-tech. These will become an important part of the promotion of short video field updates and iterations.

ANALYSIS OF THE POSITIVE IMPACT OF THE BIG VIDEO ERA ON VIDEO ORIENTATION

The Increase of Public Participation

The public is involved in the process of video production and grassroots culture is becoming popular. In the past, only authoritative media have the credential to produce and publish videos on public channels. However, nowadays, it is the era of universal video. Around 2010, as media competition intensified and new media, especially online media, cell phones and other electronic media, diverted TV audiences, leading to the decline of TV viewership (Yu, 2021). Thanks to the development of technology, all it takes is a phone for the public to become a video producer. The number of video-based software is also growing. On October 20, 2021, at the launch of Snack Video Short Drama Starmount, the person in charge, Yu Ke, announced the scale of 230 million daily active users and 770 billion total plays of Snack Video Short Drama (Sun, 2021). The prevalence of short videos on TikTok has made it easier to get started with video production. Everyone can use the function of video shooting, so anyone with a cell phone and WiFi can share their life on the platform. The number of online video releases is growing in spurts. C. Stokel-Walker analyzes the current state of Jitterbug and concludes that Jitterbug has seen a dramatic rise in popularity since its merger with Musical.ly. The users of TikTok open the app an average of 121 times per week and watching nearly 27 minutes per day, showing that the rise of TikTok has a little sign of slowing down (Stokel-Walker, 2020). In this case, the public will prefer videos made by ordinary people to those released by official media, because those videos are more relevant to the subject matter of life and are more interesting.

Changes in the Function of Video Discourse

The discourse function of video has been reinvented. Since video has the function of spoken language, it has become our social language. As humans are social-loving animals, they are willing to share those videos which make them have emotional resonance further to their friends or their followers, which further triggers the generation of retweeting behavior. For example, if a cute pet blogger uploads a video of a cat, the cute behavior or beautiful appearance of the cat will make the viewers who like cats or cute things see it and have emotional resonance, thus leading to the generation of retweeting

behavior. For example, during the parade of China's National Day, a video called *the armed police squad kicking the march blindfolded* generated 35,442,000 likes, 109,000 comments and 419,000 retweets due to its passionate soundtrack and content that touched everyone's love for the country (Zhang, 2021a).

Reconfiguration of the Video Communication Context and Expansion of the Consumption Scenario

In the era of big video, the video communication context has been reconstructed, and the consumption scenario has been expanded. The original way and place of video communication may be relatively single, but in the era of big video, the communication context has undergone a radical change. The boom in immersive experiences was brought about by the changing needs of the public at the aesthetic and experiential levels (Huang, 2021a). First, the video itself has become a new communication scenario. With the rise of live video, we can watch the lives of the anchors, getting closer to the scenes of others' lives, and satisfy our own voyeurism. On common video platforms, people are able to send pop-ups, making timely comments on videos, and sharing their own feeling after watching the videos.

In addition, some special technologies have been applied to different scenarios, creating many new experiences. For example, in the opening show of League of Legends, the audiences are completely shocked by the exciting sensory experience of the live song performances with VR character interaction and the dazzling VR special effects. As the communication context of video has changed, the science and technology applied to it has further expanded our consumption scenarios. Audition content is closely connected to viewing scenarios, consumption behaviors, and usage contexts. Today we can see audition content entering our lives. When we buy cosmetics such as lipsticks and foundations, we can try them out online and complete the whole shopping behavior without leaving home. For example, Maybelline premiered the world's first AI foundation adapter "360° Intelligent Color Selection", using the AI foundation adapter to scan the face and surrounding light environment 360 degrees. Through AI analysis, it can ensure that when consumers buy the new Superstay Makeup Foundation and the classic Fit Me custom water-based foundation online, they can enjoy convenient and intelligent personalized color selection recommendation service to accurately locate the right foundation shade for you.

Change in the Form of Video Interaction

In the era of big video, the interactive form of video is upgraded. In the past, we could only sit in front of the TV or computer, watching the characters in TV. There was no way for us to interact with them. However, interactive narrative, immersive experience, multi-interaction, and other diverse video forms have emerged and come into use nowadays. Artificial intelligence, big data, VR, and other technologies have gradually come into our lives. This has greatly enriched our entertainment life. According to the authoritative industry report *VR and AR: Decoding the Next Universal Computing Platform*, gaming, live streaming and video entertainment will account for 60 percent of overall VR/AR revenue expectations, and statistics from Goldman Sachs and Sadie Consulting show that the global VR live streaming market will increase in revenue from \$1.161 billion to \$4.113 billion from 2021-2025. 4.113 billion (Song, 2021). Cinema has also been updated; IMAX and 4D movies are gradually more and more common, ensuring us to have a better screen viewing experience. Those cool future times in technology movies seem to be no longer far away from us. Google Glass has shone a ray of success into reality. It combined AR, VR with the glasses, equipped them with a variety of intelligent functions, so that the flattened glasses can bring a brand new audition experience. Although this technology is still immature, I believe that human beings will be able to make a breakthrough in this technology with the fast-developing technology.

ANALYSIS OF THE NEGATIVE IMPACT OF THE BIG VIDEO ERA ON VIDEO ORIENTATION

Lack of Control over Error Messages

Due to the technological development in the era of big video, the information flow is accelerated. While the efficiency of the delivery of correct information is improved, wrong information is also being delivered faster. Many media, as well as marketing numbers, often publish misguided statements in order to gain attention, inciting public sentiment and misleading audiences, resulting in bad consequences. Although many platforms have increased the number of machine and manual audits, we still haven't found a completely effective way to thoroughly screen the authenticity of information today. In order to prevent public riot, confusion, and panic caused by wrong information, it is extremely essential not only to improve the quality of video and content

output from ourselves, but also to generate video screening means that match with the Big Video Era.

The Serious Problem of Video Quality and Content Homogenization

The diversification of the media and the simplification of blogger entrance has resulted the video quality to vary. Some data show that more than 30% of consumer users believe that there is too much homogenized content in TikTok, and the phenomenon of following the trend of shooting has led to a decrease in goodwill of some users, resulting in the loss of a large number of short video users (Deng, 2020). In order to be popular, some bloggers may be unscrupulous and choose to copy other people's videos. However, for now, there is neither a very effective mechanism to determine whether a video is plagiarized, nor strict penalties to punish plagiarism, which has allowed plagiarism to run even more rampant. *The 2020 China Internet Short Video Copyright Monitoring Report* released by 12,426 Copyright Monitoring Center shows that between 2019 and October 202, a total of 16,026,900 suspected infringement links were monitored, and the rate of exclusive original creators being infringed was as high as 92.9% (Zhang, 2021b). This is a great disrespect to the hard-working original video makers. As a positive example, YouTube has stricter requirements on whether a video is original or not, even the use of background music without consent is considered as a severe violation. When multiple unoriginal acts are found, the account of this video producer will be directly banned forever, which greatly reduces the number of plagiarism. In the past, the delivery of information was point-to-point and the media was able to control the quality of video content as much as possible. But considering the universal use of video software today, it is difficult for us to make strict requirements for the quality of the video, which has led to the quality issue of the video, and even some vulgar videos and videos with incorrect values are still being delivered in the network. Improving the quality of original user content is imminent. *In The Research of Short Video App: The Case of TikTok*, Janssen Richardson examines the factors that make TikTok short videos successful in terms of platform marketing strategy and user use and satisfaction theory, and argues that in a short video space with intense content homogenization, only content of sufficient quality can make In the short video space with intense content homogeneity, only high quality content can lead to higher user survival rate (Richardson, 2019). In the Big Video Era, in order to better

restrain video content and distribution channels, further laws and a better system are required urgently.

Lack of Guidance on Correct Value Orientation for Teenagers

In the Big Video Era, children are exposed to the Internet at an early age. Until June 2020, the teenage group aged 10 to 19 years old accounted for 14.8% of the short video users. Students have a high share of 23.7% in the occupational dimension (Huang, 2021b). However, some of the videos on the Internet don't have good qualities, and some of them may be misleading in their translated values. Some of the videos are made by people who overemphasize the role of love in a person's life, or who are overly interested in money and material life. These videos along with the bad Internet environment will have a side effect for teenagers who are at the age of forming their own view towards the world. Nowadays, we have started to pay attention to applying stricter control of video quality, but Rome is not built in one day. Current major video platforms have launched a youth model that can effectively block some of the videos with improper values, but still cannot solve the problem fundamentally. Relevant policies and programs should be introduced as soon as possible to positively guide young people searching the Internet and viewing online videos, growing up healthily in the Big Video Era.

NEW REQUIREMENTS OF VIDEO CONTENT EDITING IN THE BIG VIDEO ERA

To Tell Good Stories

Video editors need to be able to tell good stories. With the explosion of video information in the Big Video Era and the accelerated pace of people's lives, people usually have only fragmented time. As a result, their demand for stories has become higher. Many stories today lack vividness and a sense of immersion, and are just flat and straightforward, so we need to help the public distill effective information and tell the best stories in the shortest time. Nowadays, a short video format for narrating stories is quietly gaining popularity on various platforms. As people don't have time to watch a whole TV series by themselves, this narrated form of short video will edit a good TV series, leaving the best bits and presenting them to the audience with a narration. The editing in this way is a case of telling a good story. Only those

quality contents will be recognized by the public. Creating a good story from this requires more of a blend of various elements and eventually create work with soul.

To Tell Right Stories

The story after video editing should have been positive in its nature. Nowadays, people are enjoying various ways of entertainment, TV and online entertainment programs are no longer the only way for people to find fun. So some programs do whatever they can for the sake of attracting followers, using editing to constantly intensify conflicts and triggering online exposure of someone or something to contribute to the explosion of the program. This is not something we want. When editing, we should pursue more high-quality content output and try to establish the guidance of the right values,, rather than deliberately create intensified conflicts.

To Tell Creative Stories

As mentioned in the previous article, the era of big video is full of opportunities. I think whoever has more creativity and innovation is blessed with bigger chance of success. Originality and uniqueness of content are advocated by the cultural and creative industry and the media industry (Zhao, 2021). The core of the “traffic code” is creativity. As today’s videos are mostly the same, people can reap a better response if the output of the video can be creative. On the short video platform, there is always creative editorial content output. For example, launching new challenges, using filters and face effects to shoot videos, technical streaming videos, jamming videos, including the recent use of artificial intelligence effects to complete the video, and so on. All these ways combine music, special effects, VR, and other elements with video in a good way, bringing people a new audio-visual experience.

Creativity is important not only for content editing in short videos, but is also indispensable for long videos. The Korean variety show “Heart Signal” series exploded all over the internet because it is a very creative reality show. With a new theme, excellent characterization, and perfect music, it created a brand new relationship variety show. The content explored and the events that really happened in the show also resonated with the audience.

CONCLUSION

After studying the two aspects above in general, we can finally give out a conclusion. Big video era has put forward different requirements for video. Firstly, we should combine some of the existing advantages of the big video era, make full use of audience participation, leverage the dividends of multiple platforms, develop the technologies of AR, VR and further combine these technologies with the video industry. At the same time, we should also further improve the existing loopholes and introduce relevant regulations and laws to regulate various negative behaviors on the Internet as soon as possible, so as to create a healthy and positive network platform for the teenagers and let video communication become the main body of information delivery. In the process of video content output, we also need to combine the new requirements of the big video era and pay further attention to the content of video. While outputting high-quality original internal stories, we should also focus on its form. We have to add innovative content and ideas to the content so as to further attract the attention of all the viewers.

“Let content creation present maximum value”, says the editor of the Audiovisual Department. In contrast to equal opportunities for content producers, users have a very limited attention span. “As a result, users’ choices become quite valuable. In this case, quality content can stand out and realize its headline value. Vertical and distinctive niche and niche programs can also reap their segmentation value and even long-tail value from them” (Zhao, 2018).

In the Big Video Era, the transformation of videos will undoubtedly be a new challenge for all media people, and video-based communication innovation has infinite possibilities. For us to create works that are valuable, worth watching, and worth repeating, we still need further in-depth learning, thinking, and practice!

REFERENCES

1. Alti, N. & Liang, Z. (2021). Characteristics, Trends and Dilemmas of Short Video Content Production in the New Media Era. *China Editorial*, No. 3, 81-85.
2. Deng, Y.-S. (2020). Research on the Communication Development Strategy of Jitterbug Short Video in the Context of the Internet. *Journalism and Culture Construction*, No. 17, 167-168.
3. Embrace the Big Video Era (2017). <https://www.huawei.com/cn/events/ubbf2017/big-video-era>
4. Huang, C. (2021a). Analysis of the Application of Artificial Intelligence in Multimedia Art. *Art Pinnacle*, No. 27, 129-130.
5. Huang, W. W. (2021b). The Influence of Jitterbug on Teenagers and Guidance Strategies. *Journalism Research Guide*, No. 6, 79-80.
6. Richardson, J. (2019). The Research of Short Video App: The Case of TikTok. *IEEE Transactions on Multimudha*, 3, 91-100.
7. Song, J. (2021). VR Live Streaming: 5G Helps Usher in the Market Inflection Point. *China Electronics News*, No. 4, p. 2.
8. Stokel-Walker, C. (2020). Tik Tok's Global Surge. *New Scientist*, No. 245, 31.
9. Sun, L. (2021). Racer Sunshine Short Drama Business Report Card: 850 Short Dramas Played over 100 Million. *Chongqing Business Daily*, No. 4, p. 1.
10. Yu, W. D. (2021). A Review of Twenty Years of Viewership Application in Shanghai. *Shanghai Radio and Television Research*, No. 2, 114-121.
11. Zhang, C.-C. (2021b). Short Video Clips: A Strong Demand That Can't Be Suppressed in the Legal Crevice. *Youth Journalist*, No. 9, 111.
12. Zhang, Y. L. (2021a). Short Video Communication and National Identity Construction of the 70th Anniversary National Day Parade. Master's Thesis, Lanzhou: Lanzhou University.
13. Zhao, G. H.. (2021). Research on Short Video Production and Dissemination Mechanism in the Era of Big Data. *Media*, 13, 50-52.
14. Zhao, Q. J. (2018). Good Programming in the Era of Big Video. *Audiovisual World*, No. 1, p. 1.

CHAPTER 8

A Fast Depth-Map Generation Algorithm Based on Motion Search from 2D Video Contents

Weiwei Wang¹, Yuesheng Zhu²

¹Communication and information Security Lab

² Shenzhen Graduate School, Peking University, China

ABSTRACT

Generation of a depth-map from 2D video is the kernel of DIBR (Depth Image Based Rendering) in 2D-3D video conversion systems. However it occupies over most of the system resource where the motion search module takes up 90% time-consuming in typical motion estimation-based depth-map generation algorithms. In order to reduce the computational complexity, in this paper a new fast depth-map generation algorithm based on motion search is developed, in which a fast diamond search algorithm is adopted to decide whether a 16x16 or 4x4 block size is used based on Sobel operator in

Citation: W. Wang and Y. Zhu, "A Fast Depth-Map Generation Algorithm based on Motion Search from 2D Video Contents," Journal of Software Engineering and Applications, Vol. 5 No. 12B, 2012, pp. 144-148. doi: 10.4236/jsea.2012.512B028.

Copyright: © 2012 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

the motion search module to obtain a sub-depth-map. Then the sub-depth-map will be fused with the sub-depth-maps gotten from depth from color component Cr and depth from linear perspective modules to compensate and refine detail of the depth-map, finally obtain a better depth-map. The simulation results demonstrate that the new approach can greatly reduce over 50% computational complexity compared to other existing methods.

Keywords: Block-Matching; Depth-Map; Motion Search; DIBR

INTRODUCTION

Commercialization and industrialization of three-dimension televisions (3D TV) [1]not only depend on the development of 3D display as well as standardized technology, but also rely on a large amount of 3D video contents. Although 3D movie are on its way to develop, currently the 3D video contents are still not rich enough to satisfy the 3D-video market needs. In fact, the market is overwhelmed with 2D video. Converting 2D video into 3D video automatically and enabling the existing movies to be played on 3D displays becomes an important way to alleviate the shortage of 3D program. Also the 2D to 3D conversion technique can deliver the 3D videos effectively and efficiently. Therefore, the transition from 2D to 3D video is a low cost solution for the 3D industry compared with that captures 3D video directly.

There are some approaches for converting 2D video into 3D video [2-10]. Depth-map contains information relating to the distance of the scene objects from a viewpoint in a video content, and generating a depth-map effectively from 2D video is the kernel of DIBR in 2D-3D video conversion systems. The basic principle of DIBR [2,11] is to obtain a depth-map from 2D video and then synthesize the left and right views. The depth from motion (DFM) [5] is a kind of depth-map generation algorithm in which video is segmented first and the frame disparity is estimated to obtain the depth-map. But the DFM requires that moving objects must exist in successive frames. Fusion with color information can improve the depth-map quality [12] [13]. In the literature [12], the depth-map generated from motion-parallax is fused with color segmentation to obtain a clear and reliable depth-map, but its color segmentation algorithm introduces high computational complexity. In the literature [13] motion estimation is performed by using luminance and chrominance information in motion search module to yield a reliable depth-map and reduce the computation complexity. However, the computational complexity of the color information in motion search module is still high.

In this paper, a new fast depth-map generation algorithm based on motion search is developed, in which a fast diamond search algorithm is adopted to decide whether a 16 x 16 or 4x4 block size is used based on Sobel operator[14,15] in the motion search module without using color information to obtain a main sub-depth- map, and then the depth from color component Cr [4] and the depth from linear perspective [6] are used as auxiliary sub-depth-maps to fuse the main sub-depth- map. Finally the bilateral filter is adopted to eliminate the block effect and the staircase edges that remained in the fused depth-maps. The results show that with the proposed algorithm a smooth and reliable depth-map and a better visual 3D video can be obtained with low computational complexity compared to the methods in [12] and [13]. The remainder of the paper is organized as follows. The proposed depth-map generation algorithm is presented in section II. Experimental results are provided in section III. Finally, a conclusion is given in section IV.

THE PROPOSED DEPTH-MAP GENERATION ALGORITHM

The block diagram of proposed algorithm is shown in Figure 1. In Figure 1, the final depth-map is fused with three sub-depth-maps, that is, depth from improved motion estimation, depth from color component Cr, and depth from linear perspective. In depth-map fusion, depth from color component Cr and depth from linear perspective are used as auxiliary sub-depth-map to compensate the main sub- depth-map gotten from improved motion estimation. And then a bilateral filter is adopted to eliminate the block effect and the staircase edges. In this section, the module of improved block-matching based depth from motion estimation is developed and described. The approach and the corresponding algorithms are described in detail as follows.

Depth from Improved Motion Estimation

In paper [7], the motion estimation is performed by using luminance information, which may cause mismatch in the areas where the luminance components tend to distribute uniformly. In paper [10], luminance and chrominance information is adopted in the motion vector processing which uses Y (luminance component), C_r (red component) and C_b (green component) to calculate the motion vectors. Our comparison results of time consumption in motion search module for four cases:(1) Y, (2) Y and C_b ,(3) Y and C_r , (4) Y, C_r and C_b are shown in Figure 2.

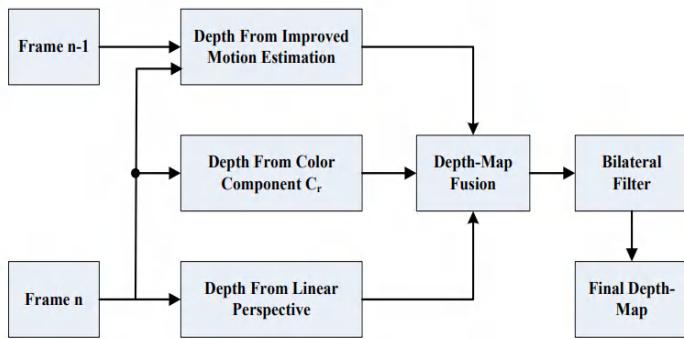


Figure 1. Block diagram of proposed algorithm.

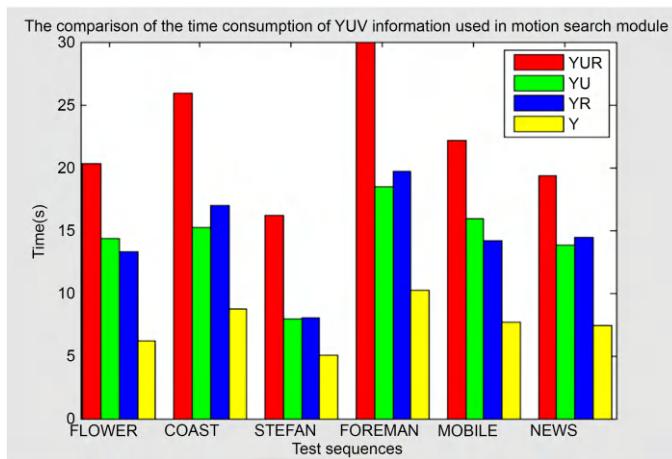


Figure 2. The time consumption in motion search module in four cases.

The results indicate that using color information in motion search module would increase the computational complexity. So, in the proposed module, the motion estimation is performed by using luminance information only in motion search module without using color information. But the depth from color component Cr is used as auxiliary sub-depth-map to fuse the sub-depth-map obtained from improved motion estimation.

In motion search module current frame is divided into small blocks for depth assignment while the corresponding small blocks in reference frame are used as center and expanded to bigger blocks for matching. Then block-matching based motion estimation is performed to find the best match block and the motion vectors generated are used to assign depth for small blocks. The block-matching algorithm based motion estimation [7] utilizes

the fact that objects with different motions usually have different depths. Near objects move faster than far objects and the relative motions are used to estimate the depth-map. The depth value $D(i,j,k)$ are estimated by the magnitude of the motion vectors as follows:

$$D(i, j, k) = C \sqrt{MV(i, j, k)_x^2 + MV(i, j, k)_y^2} \quad (1)$$

where $MV(i,j,k)x$ and $MV(i,j,k)y$ are horizontal and vertical components of the motion vectors and C is a predefined constant. It is noted that the motion searching module takes as much as 40 percentage of the total time consumption. In order to reduce the computational complexity a fast diamond search algorithm is adopted in the new motion search module.

The common algorithms generate depth-map based on motion estimation in 4×4 block size. While, we observe that if the frame picture is homogenous enough, the depth value of blocks is close to their neighbors. We find it will save much computational complexity if we use 16×16 block size instead of 4×4 block size in homogeneous area. To evaluate the smoothness of a picture, statistical measurement such as standard deviation, variance, skewness and kurtosis [16] are used. Paper [14,15] use Sobel operator to create the edge maps of pictures for high efficiency of video coding process. In order to classify the homogeneity of a block, the amplitude of edge vector is defined by formula (2).

The vertical and horizontal directions are defined according to a luminance or chrominance pixel at position (i,j) with value V_{ij} by formula (2)(3). Then the block homogeneity measurement H can be set by formula (5).

$$Amp(\bar{E}_{i,j}) = |Ex_{i,j}| + |Ey_{i,j}| \quad (2)$$

$$Ex_{i,j} = v_{i-1,j+1} + 2 \times v_{i,j+1} + v_{i+1,j+1} - v_{i-1,j-1} - 2 \times v_{i,j-1} - v_{i+1,j-1} \quad (3)$$

$$Ey_{i,j} = v_{i+1,j-1} + 2 \times v_{i+1,j} + v_{i+1,j+1} - v_{i-1,j-1} - 2 \times v_{i-1,j} - v_{i-1,j+1} \quad (4)$$

where $i \in 1, 2, \dots, R, j \in 1, 2, \dots, C$,

$$H_{r,c} = \begin{cases} 1, & \sum_{0 \leq i,j < N} Amp(\bar{E}_{i,j}) < Thd_H \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

After the block size is assigned, the fast diamond search algorithm makes two round iterations to calculate the motion vectors [17]. In the first round, it takes a large diamond search pattern (Fig.3a) and 9 test points are compared to find the best points. The iteration process breaks out when the

center point is just the most matching point. The second round aims to find the best point in a small diamond search pattern (Fig.3b). The first round takes at least 90% time according to related experiment. So we design a quick quit scheme to exit the iterative loop ahead of time if the SAD (Sum of Absolute Difference) of current point is less than the predefined threshold. A fast depth-map generation algorithm based on motion search is as follows.

Step1. Read the current macro block data (16x16 block size) and compute the homogeneity measurement H of the block by formulas (2)(3)(4)(5). If H equals to 0, the block is divided into 16 little 4 x 4 block size, and the motion search module will be based on the 4 x 4 block size, otherwise the large 16 x 16 block size will be used.

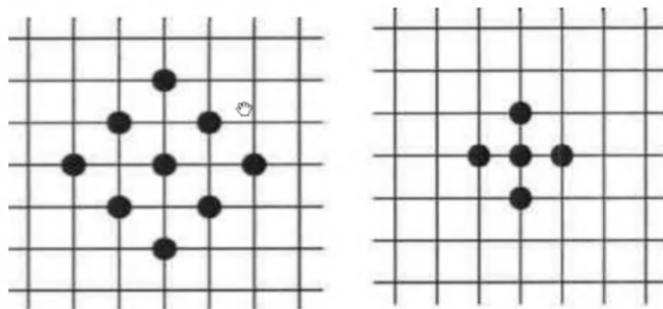


Figure 3. Diamond search pattern. (a) Large diamond search pattern. (b) Small diamond search pattern.

Step2. Compute the SAD of the 9 neighbor points of center point in large diamond pattern (Fig.3a). If the SAD value of current point is less than the given threshold T1, go to step 4; if it is larger than T1 but less than T2, go to step 3; otherwise, compute the minimum SAD value and loop over the 9 points.

Step3. Compute the SAD of the 5 neighbor points of center point in small diamond pattern (Fig.3b). The best point will be chosen as the matching point.

Step4. Compute the depth value of the block with formula (1).

Depth from Color Component Cr

In the research of [4], it has been proved that different objects have different hues in the 2D color video sequences, and each of the hues has its own associated grey level intensities in the Cr color component images. If we take

the gray level intensities as indexes of depth, the depth of the boundaries of each object is different from that of its immediate surroundings. Therefore gray intensity images associated with the color component Cr of standard 2D-colour video sequences can be used as proxy depth-map. In our situation, the depth from color component Cr which is derived directly from the current frame of the 2D images is used as auxiliary sub-depthmap to fuse the sub-depth-map from improved motion estimation. The fused depth-map can increase the accuracy and detail of sub-depth-map. Others, for static scene we can get the different depth values using color component Cr while the other methods [7] obtain the same depth values. So depth from color component Cr can strengthen the layer of the stereo videos. The depth from color component Cr is shown in Fig. 4(b).

Depth from Linear Perspective

Research in [6] shows that depth from linear perspective can make the stereoscopic video more comfortable for human to watch. In our method, near to far global scene depth gradient is applied as the auxiliary depth map as human visual perception tends to interpret most of the images represent scenes in which the bottom part is related to the ground and consequently close to us and the upper part represents the sky and consequently far from us.



Figure 4. (a) The depth-map from color component Cr.(a) The original “cheerleader” video image.(b)The depth from color component Cr.

Depth-Map Fusion

The final depth-map is fused with three sub-depth-maps, that is, depth from improved motion estimation, depth from color component Cr, and depth from linear perspective. In this paper a simple linear module is used to fuse

the three kinds of sub-depth-maps. The fused depth-map can be described in the following equation:

$$D_{all} = D_m \times W_m + D_c \times W_c + D_l \times W_l \quad (6)$$

$$W_m + W_c + W_l = 1 \quad (7)$$

Where D_m , D_c and D_l are the values of sub-depth-maps estimated by motion estimation, color component Cr and linear perspective respectively. D_{all} is the values of the fused depth-map while W_m , W_c and W_l are the weights of them.

Because in depth-map fusion, depth from improved motion estimation is used as the main sub-depth-map, and depth from color component Cr and depth from linear perspective are used as the auxiliary sub-depth-maps, selecting the values of W_m , W_c and W_l we follow the principle that the value of W_m is larger than W_c and W_l .

EXPERIMENTAL RESULTS

To evaluate our proposed algorithm, several test sequences are used to perform the [12], [13] and our method to make some comparisons on the efficiency. We set the threshold of Sobel to 5000 while the thresholds of motion search modules are 0.15 and 0.3. The test results are shown in Table 1.

Table 1. The test results of several sequence

| Sequence | Time(s) | | |
|-------------|---------|-----------|------------|
| | Po [12] | Chen [13] | Our method |
| FLOWER | 4.12 | 8.12 | 2.92 |
| COAST GUARD | 3.45 | 8.23 | 1.98 |
| CARPHONE | 4.23 | 7.98 | 2.03 |
| FOREMAN | 4.47 | 8.43 | 2.65 |
| MOBILE | 5.01 | 8.90 | 3.21 |
| CALENDAR | 5.43 | 8.79 | 3.45 |

According to Table 1, we can easily find that the efficiency of our method is better than the algorithms in [12] and [13], about 31% of time saving than [12] and 65% to [13]. The promotion is especially remarkably in CAR PHONE and COASTGUARD for the large scale of smooth area in these two sequences. It is the contribution of Sobel operator which decides whether a 16x16 or 4x4 block size is used, and the quick quit scheme to

exit the iterative loop ahead of time in the search module. The sub-depth-map obtained from the improved motion estimation module is shown in Fig.5.(b). From the Fig.5.(b), we can see that there are many isolated points in this sub-depth-map. So this sub-depth-map will be fused with the sub-depth-maps gotten from depth from color component Cr and depth from linear perspective modules to eliminate these isolated points and compensate this sub-depth-map and get a better fused depth-map as shown in Fig.5.(c). Finally, a smooth and reliable depth-map shown in Fig.5.(d) is obtained by passing the bilinear filter.

CONCLUSIONS

In this paper, a fast depth-map generation algorithm based on motion search is proposed to enhance the efficiency of the generation of depth-map. In the new proposed modules, a fast diamond search algorithm is adopted to decide whether a 16x16 or 4x4 block size is used based on Sobel operator in the motion search module without using color information to obtain a sub-depth-map, and this sub-depth-map will be fused with the sub-depth-maps gotten from depth from color component Cr and depth from linear perspective modules respectively to compensate this sub-depth-map and obtain a improved fused depth-map. Finally, the bilateral filter is adopted to eliminate the block effect and the staircase edges that remained in the fused depth-map. The results show that with the proposed algorithm a smooth and reliable depth-map and a better visual 3D video can be obtained with over 50% reduction of computational complexity compared to the other methods.

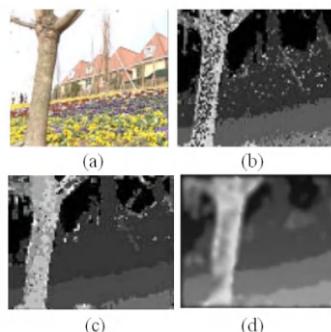


Figure 2. Depth-map from our proposed algorithm (a) The original “flower” video image. (b) The sub-depth-map estimate by improved motion estimation module. (c) The fused depth-maps after fused with three sub-depth-maps. (d) The final depth-map by passing bilinear filter.

REFERENCES

1. M. Op de Beeck, and A. Redert, “Three Dimensional video for the Home,” Proceedings of International Conference on Augmented, Virtual Environments and Three-Dimensional Image, May-June 2001, pp. 188-191.
2. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. SPIE, vol.5291, no.2, pp. 93-104 2004.
3. P. Harman, J. Flack, S. Fox, M. Dowley.: Rapid 2D to 3D Conversion. Proceedings of SPIE, vol. 4660, pp. 78-86, 2002
4. Wa James Tam, Carlos Vázquez, Filippo Speranza.: Three-dimensional TV: A novel method for generating surrogate depth maps using color information. Proc. SPIE Electronics Imaging-stereoscopic Displays and Applications XX, 2009.
5. D. Kim, D. Min, K. Sohn.: Stereoscopic video generation method using motion analysis. in Proceedings of the 3DTV Conference, pp. 1–4, Kos Island, May 2007.
6. Sung-Fang Tsai, Chao-Chung Cheng, Chung-Te Li, Liang-Gee Chen. □A Real-Time 1080p 2D-to-3D Video Conversion System. IEEE Transactions on Consumer Electronics, Vol. 57, No. 2, pp. 803–804, May 2011.
7. I. Ideses, L. P. Ya-roslavsky, B. Fishbain.: Real-time 2D to 3D video conversion. Journal of Real-Time Image Processing, vol. 2, no.1, pp. 3–9, 2007.
8. C. Tomasi, R. Manduchi.: Bilateral Filtering for Gray and Color Images. Proceedings of the IEEE International Conference on Computer Vision, Bombay, India, pp.839-846, Bombay January 1998.
9. M. T. Pourazad, P. Nasiopoulos, and R. K. Ward.: An H.264-based scheme for 2D to 3D video conversion. IEEE Transactions on Consumer Electronics, vol.55, no.2, pp. 742–748, 2009.
10. A.-M. Huang, T. Nguyen.: Motion vector processing using the color information. IEEE International Conference on Image Processing, ICIP, pp. 1605-1608, Cairo, 2009.
11. W. J. Tam, f. Speranza, L. Zhang, R. Renaud, J. Chan, C. Vazquez.: Depth Image Based Rendering for Multiview Stereoscopic Displays: Role of Information at Object Boundaries. Three-Dimensional TV, Video, and Display IV, vol. 6016, pp. 75-85, 2005.

12. L. Po, X. Xu, Y. Zhu, S. Zhang: Automatic 2D-to-3D video conversion technique based on depth-from-motion and color segmentation. IEEE International Conference on Signal Processing, ICSP, pp.1000-1003, 2010.
13. J. Chen, Yuesheng Zhu and X. Liu.: A New Block-Matching Based Approach for Automatic 2D to 3D Conversion. The 4th International Conference on Computer Engineering and Technology, ICCET, pp. 109-113, Thailand, 2012.
14. D. Wu, S. Wu, K. Lim, F. Pan, Z. Li, X. Lin: Block INTER mode decision for fast encoding of H.264. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). vol.3, 2004 pp. iii- 181.
15. F. Pan, X. Lin, R. Susanto, K. P. Lim, Z. G. Li, G. N. Feng,D. J. Wu, and S. Wu.: Fast Mode Decision Algorithm for Intra Prediction in JVT. 7th JVT meeting, JVT-G013, Thailand, March 2003.
16. K. R. Castleman.: Digital Image Processing. Prentice Hall Inc, 1996.
17. Shan Zhu, Kai-Kuang Ma.: A new diamond search algorithm for fast block-matching motion estimation. IEEE Transactions on Image Processing, vol.9, no.2, pp.287-290, Feb 2000.

CHAPTER 9

Adaptive Content Management for UGC Video Delivery in Mobile Internet Era

Qilin Fan¹, Hao Yin¹, Zexun Jiang¹, Haojun Huang², Yan Luo³, and Xu Zhang¹

¹National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, China

²Department of Communication Engineering, Wuhan University, China

³Department of Electrical and Computer Engineering, University of Massachusetts Lowell, USA

ABSTRACT

The demand of storing and transferring user generated content (UGC) has been rapidly growing with the popularization of mobile devices equipped

Citation: Q. Fan, H. Yin, Z. Jiang, H. Huang, Y. Luo, X. Zhang, “Adaptive Content Management for UGC Video Delivery in Mobile Internet Era”, *Mobile Information Systems*, vol. 2016, Article ID 3624860, 9 pages, 2016. <https://doi.org/10.1155/2016/3624860>.

Copyright: © 2016 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

with video recording and playback capabilities. As a typical application of software-defined networks/network functions virtualization-based pervasive communications infrastructure, content delivery networks (CDNs) have been widely leveraged to distribute contents across different geographical locations. Nevertheless, the content delivery for UGC is inefficient with the existing “pull-based” caching mechanism in traditional CDNs, because there exists a huge volume of lukewarm or cold UGC which results in a low cache hit ratio. In this paper, we propose a “push-based” caching mechanism to efficiently and economically deliver UGC videos. Different from traditional CDNs which separate the original content storage and caching, we directly store UGC videos into selective servers which serve as both reliable storages and user-facing uploading servers. By carefully and dynamically selecting the storage locations of each UGC object based on its popularity and locality, we not only guarantee the data availability but also remarkably improve the content distribution performance and reduce the distribution cost.

INTRODUCTION

Videos in video on demand (VOD) systems have historically been created and supplied by a limited number of media producers. The emergence of mobile devices such as high quality smart phones and tablets equipped with video recording capabilities has enabled the general public to record events, generate videos, and upload them to video-sharing sites such as YouTube. Nowadays, Internet users are not only content consumers, but also content publishers as well. Besides, users could access contents via mobile devices at any time and anywhere. Such advent of user generated content (UGC) in mobile Internet era has remarkably reshaped the online video industry.

As a typical application of software-defined networks/network functions virtualization-based pervasive communications infrastructure, content delivery networks (CDNs) have been playing a critical role in offering fast and reliable communication services by distributing content to cache or edge servers located close to users. Today, video providers rely on overlay CDNs like Akamai, Limelight to leverage their presence across different geographical locations to serve video contents. However, the explosive video consumption paradigm shift in the mobile Internet environment has introduced new challenges in distributing UGC videos for CDNs.

(1) The conventional caching schemes utilized in traditional CDNs are ineffective with UGC. We crawl the request logs from Youku (<http://www.youku.com/>), the largest video-sharing website in China, to simulate the

impact of UGC on traditional CDN and compare it with provider generated content (PGC). We sample 20,000 UGC videos and 2000 PGC videos in Youku. Figure 1 shows how the hit rate of UGC and PGC evolves with cache size using most common caching technique, least recently used (LRU). After cache size achieving 10% of the total video volume, the hit rate gain becomes slower while expanding cache size in UGC. It suggests that the long tail of lukewarm videos in UGC exacerbates the efficacy of cache deployment.

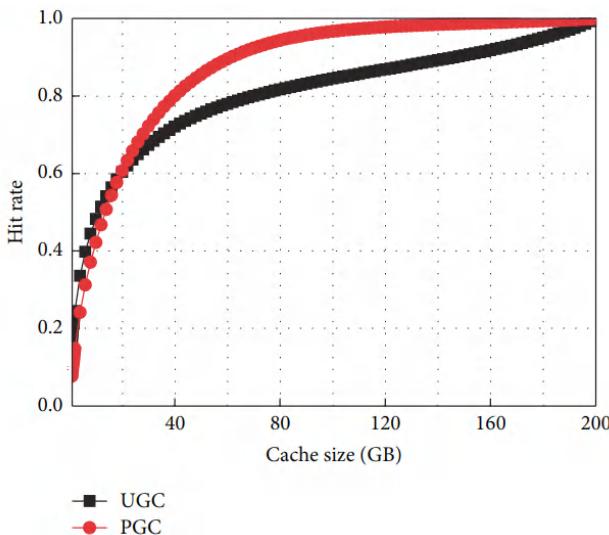


Figure 1. Hit rate as function of cache size.

(2) The passive content management does not apply to cost control in UGC. Supposing that the UGC and PGC video number is 2×10^8 and 2×10^5 , we calculate the total cost including storage and bandwidth as cache proportion increases according to current pricing norms [1] (<http://www.bizety.com/2014/08/24/cdn-storage-selling-feature/>). Figure 2 illustrates that caching 50% content in edge servers passively is sufficient for PGC to achieve low cost, while partial passive caching in traditional CDNs does not contribute to cost reduction in UGC. Meanwhile, the content volume is still increasing at a speed of 53% (<http://blog.performics.com/performics-weekly-digital-digest-5-23-13/>) per year, faster than the decreasing speed 28% (<http://www.dostor.com/article/2013-08-29/3649023.shtml>) of storage cost, thus leading to ever-increasing cost.

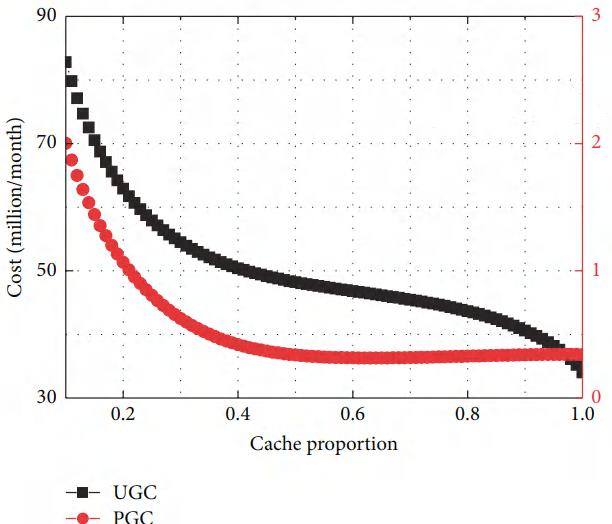


Figure 2. Cost as function of cache proportion.

(3) The usefulness of the most widely used DNS name resolution approach based on URL in traditional CDNs is declining for massive content. Authoritative DNS maps the URL to an IP address in TTL period. However, the explosive growth of the UGC namespace has decreased the effectiveness of DNS caching. Further, the timeout-based invalidation of stale mapping cannot guarantee cache coherency [2]. Nowadays, many UGC video portals build their own naming system. However, as far as we know, the study has stayed in measurement and analysis [3].

In this paper, we firstly propose a general framework for UGC video delivery, namely, adaptive content management- (ACM-) based CDN. By introducing proactive content management, joint content replication, and request routing design principles into system design, it could achieve high scalability, flexibility, and performance goals and reduce cost as well.

Second, after conducting extensive measurement on Youku, we analyze temporal popularity evolution and geographic location distribution for UGC videos. Based on the popularity predictability and geographic locality characteristics, we present data-driven content replication and request routing algorithms so that videos are replicated at “cost-effective” locations and server selection is content-aware. In order to evaluate our system, we build an experiment platform with realistic UGC traces. Our trace-driven simulation clearly demonstrates the quantitative benefits of our ACM-

based CDN. In particular, our ACM-based CDN reduces latency by 14.7%, network load by 63%, and server load by 50% at 95th percentile.

FRAMEWORK DESIGN

In this section, we begin with an overview of our goals and principles for guiding the design of framework for UGC video delivery. Then we describe the key components to satisfy the design goals and philosophies.

Design Goals

To design UGC video delivery framework, we first illustrate four significant design goals.

- (1) *High Scalability.* At the fundamental level, scalability for UGC video means handling more clients, content, and traffic (e.g., over 1 billion unique users have visited YouTube each month, over 100 hours of video has been uploaded every minute, and over 6 billion hours of video has been watched each month on YouTube (<https://www.youtube.com/yt/press/statistics.html>)). This also means that a name resolution system should support ever-growing number of content and distributed servers.
- (2) *Flexibility.* The video popularity and geographical access change dynamically as time goes by under UGC environment [4]. This requires the system to support changes in name-address mapping to rapidly propagate new mappings to users.
- (3) *High Performance.* The multimedia streaming service requires higher QoS, such as lower startup delay, lower transmission delay, and higher continuity. Any degradation in any of these factors may impact users' experience. Balachandran et al. [5] observed that an increase of the buffering ratio of only 1% can lead to more than three minutes of reduction in the user engagement.
- (4) *Controllable Cost.* The massive number of user generated videos and visits consume a huge volume of storage and network resource (e.g., Tudou (<http://www.tudou.com/>) consumed 1 PB bandwidth each day for transferring videos in 2012). The system should be carefully designed to reduce unnecessary resource consumption.

Design Principles

Then, we introduce two design principles into framework design to satisfy the above requirements. The first principle is to *bring proactive self-adapting content management into content distribution*. Content management is strategically vital to a CDN for efficient content delivery and for overall performance. The UGC video distribution and propagation in mobile Internet environment have brought new features with respect to highly dynamic access, flattening popularity distribution [4], and marginalized content delivery. Therefore, it is necessary to develop a proactive self-adapting content management mechanism into UGC content delivery and guarantee favourable users' experience. The second principle is to *merge content replication and request routing together*. A CDN must decide on how and where to replicate the content in an intelligent way, referred to as content replication problem. Also, it is challenging for a CDN to select the best server to respond to the user, known as request routing problem. Both problems are interdependent and thus should be considered together to operate in an efficient manner.

Architecture

According to the above design goals and philosophies, we propose a general architecture for UGC video delivery. As shown in Figure 3, it has three components.

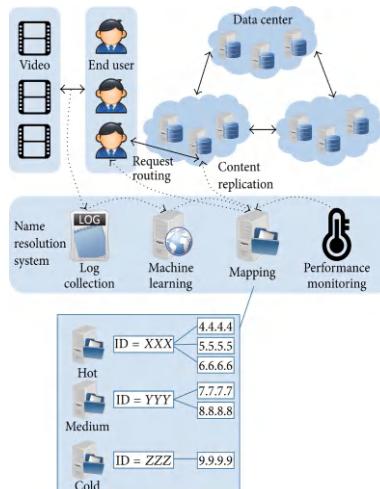


Figure 3. Architecture for UGC video delivery.

(1) *Video ID Space.* Each video is uniquely identified by a “flat” name instead of the Internet’s current host-centric naming for the reason that DNS overloads the names and rigidly associates them with specific network locations, making it inconvenient to migrate data. Flat namespace’s scalability issue could be solved by distributed hashing table (DHT).

(2) *Name Resolution System.* It comprises the log collection, machine learning, mapping, and performance monitoring modules. Log collection module keeps track of users’ access behaviour. Machine learning module analyzes collected data and predicts video access characteristics with respect to video’s popularity evolution and geographic distribution to guide the content replication. Performance monitoring module gathers information about performance of network and servers and maintains an up-to-date view of network resources. The mapping module establishes video ID-to-IP address index and provides content-aware request routing service.

(3) *Distributed-Tiered Hybrid Storage System.* Traditional CDN consists of a centralized storage data center which stores all the content and multiple delivery servers responsible for handling users requests. However, it is challenging for a centralized data center to store massive and rapidly growing content. Different from traditional CDN which separates the original content storage and caching, we directly store UGC into selective servers which serve as both reliable storages and user-facing uploading servers. We adopt distributed storage to manage video collection and guarantee dataset integrity. While videos with different popularity have varying access characteristics, they should not be considered collectively. We sort videos into three logical layers according to their popularity: hot, medium, and cold. Each layer determines how many replicas are kept on servers, where to place these replicas, and how often to update them.

The name resolution space combined with flat video ID space achieves qualitative goals. (1) *Scalability.* In our design, mapping module could provide IP addresses in content-granularity to users directly. This solves the inefficient caching and slow update problems [2] existing in traditional DNS resolution mechanism. Further, due to loose coupling between mapping module and physical servers, it is straightforward to expand service capability of mapping servers or storage servers individually, without mutual interference. (2) *Flexibility.* Data becomes the first-class entity; it can be freely migrated or replicated across hosts and administrative boundaries. We can easily decouple content replication and request routing from sticky dependency via name resolution space and flexibly customize policies

separately. Further, mobility and multihoming can be elegantly supported. The name resolution space combined with hybrid storage system achieves quantitative goals. (1) *High Performance*. Using data-driven technology to guide the videos replicating at most appropriate locations, users could quickly find a nearby replica. The periodical push-based approach reduces frequent video fetch and replacement, thus avoiding network congestion. Further, each server only deals with a small set of videos, refraining from overloading a server. (2) *Controllable Cost*. Our design could save storage cost as videos are replicated on demand and change as time goes by. We could even find a tradeoff between the storage cost and performance in our future work. Meanwhile, the content-aware request routing can eliminate the bandwidth waste due to frequent content migration in a conventional network-aware request routing.

KEY ALGORITHMS

In this section, we first investigate the popularity and geographical distribution of UGC and then explore how these characteristics can be used for guiding content replication and request routing, two important algorithms in the system design.

UGC Measurement and Analysis

In our measurement, we have crawled our dataset from Youku during the first two weeks of August 2015, using snowball sampling with initial set consisting of 10 random videos. We processed our collected datasets to remove (1) videos with missing or inconsistent information and (2) non-UGC videos according to category. The total number of samples was 200,000. For each video, we collected the following attributes: (1) its total number of views; (2) its views per day over time since it is uploaded; (3) its geographic distribution which represents how many views it received from each province; and (4) its top ten list of cities with the most traffic.

Temporal Popularity

Szabo and Huberman [6] first observed that the log-transformed popularity exhibits strong correlations between early and later periods. In this paper, we use Multivariate Linear (ML) model [7] to predict the popularity $\hat{x}(V)$ of a video V on target day t . Given the number of views $x(V)$ on each day before the target day t , we can define the feature vector $X_{t,n}(V)$ of video V as

$$X_{t,n}(v) = (x_{t-1}(v), x_{t-2}(v), \dots, x_{t-n}(v))^T. \quad (1)$$

Then we can estimate the number of views video V can get on target day t as

$$\hat{x}_t(v) = \Theta_n \cdot X_{t,n}(v), \quad (2)$$

where $\Theta_n = (\theta_{t-1}, \theta_{t-2}, \dots, \theta_{t-n})$ is the vector of model parameters. Intuitively, each parameter value represents the importance of each day for estimating views on target day. To train model parameters, we use mean Relative Square Error (mRSE) as cost function on training videos set V . We define cost function (Θ_n) as follows:

$$J(\Theta_n) = \frac{1}{|V|} \sum_{v \in V} \left(\frac{\hat{x}_t(v)}{x_t(v)} - 1 \right)^2, \quad (3)$$

where $x_t(V)$ is the actual number of views of V on target day t .

The global optimal solution is to find the best parameter vector which minimizes the cost function:

$$\begin{aligned} \Theta_n^* &= \arg \min \frac{1}{|V|} \sum_{v \in V} \left(\frac{\hat{x}_{t,v}}{x_{t,v}} - 1 \right)^2, \\ &= \arg \min \frac{1}{|V|} \sum_{v \in V} \left(\Theta_n \cdot \frac{X_{t,n}(v)}{x_t(v)} - 1 \right)^2. \end{aligned} \quad (4)$$

The optimization problem can be solved by gradient descent algorithm which starts with some initial Θ_n and repeatedly performs the update:

$$\Theta_{n,j} = \Theta_{n,j} - \alpha \frac{\partial}{\partial \Theta_{n,j}} J(\Theta_n), \quad (5)$$

where α is the learning rate and $\Theta_{n,j}$ is the j th weight of Θ_n .

To validate our model training process, we randomly extract three categories: music, game, and entertainment. For each category, along with all videos, we use 10-fold cross validation to calculate predicted popularity. During training process, given different number n , we can get different model parameter vector Θ_n , and then we use Θ_n^* to estimate $\hat{x}(V)$ on validation data based on (2). Figure 4 shows the correlation between mRSE and latest n days for different categories. We can draw two conclusions that are helpful for guiding content replication from temporal perspective. (1) The prediction of popularity on target day based on historical data is highly accurate. In fact, we only need views data on latest three days, yielding

18% to 25% prediction error to predict video’s popularity.(2)Prediction on subsamples of the dataset extracted by video category reduces 5% mRSE on average compared to the whole dataset. Prediction based on category is more accurate.

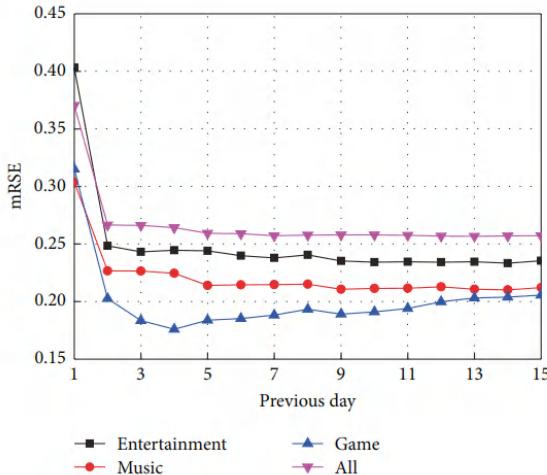


Figure 4. mRSE as various previous n days.

Geographic Location

We examine the geographic distribution of views for UGC videos on province granularity and city granularity. We divide our dataset into four categories according to view numbers (Table 1).

Table 1. Popularity category

| Category | # of views | % of videos | Average views |
|----------|------------------|-------------|---------------|
| C1 | $[0, 10^4)$ | 45.0 | 2792 |
| C2 | $[10^4, 10^5)$ | 35.1 | 38710 |
| C3 | $[10^5, 10^6)$ | 16.8 | 281105 |
| C4 | $[10^6, \infty)$ | 3.1 | 2427914 |

Province Granularity. For each video, geographic views from each province are sorted in decreasing order. We then compute the cumulative distribution of views of each video and plot the average over each popularity category. Figure 5 shows that 30% of provinces could cover 70% of video views for all categories.

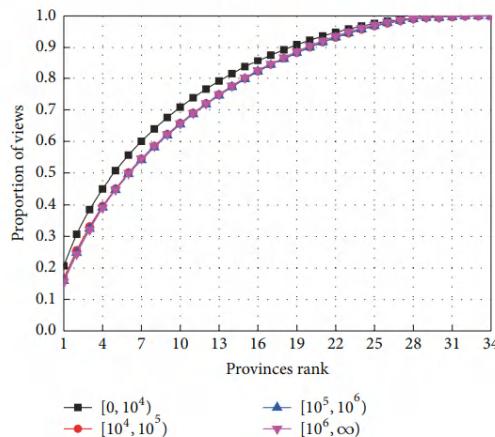


Figure 5. Geographic cumulative distribution of views.

City Granularity. Due to the limitation of dataset, we use linear fitting approach for ten most traffic cities in log-log coordinate to establish geographic distribution on city granularity (Figure 6). The distribution function can be expressed as follows:

$$\log y_n = k \log n + b, \quad (6)$$

where n is the city traffic rank and y_n is the corresponding views. From Figure 6, we observe that cities distribution satisfies power law profile. China has 287 cities. Top 10 cities (3.5% of cities) could cover 30% of all traffic. Given (6), we can calculate that 60 cities (20% of all cities) could hold 70% of the overall traffic.

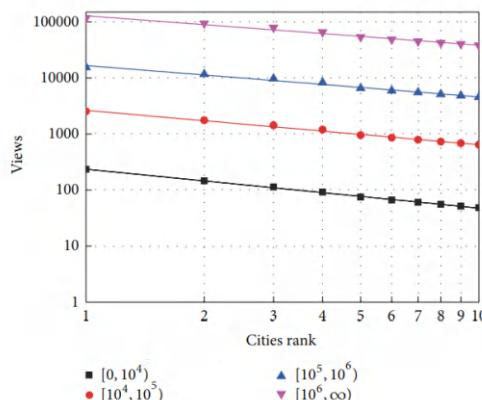


Figure 6. Views distribution of top 10 cities.

The above measurement illustrates that locality of geographic access is universal, whether from coarse or fine granularity. It is feasible to direct content replicating in small-scale locations to achieve most of the traffic.

Content Replication

Initial Replication. When video is first uploaded by a user, it will be stored by a server which is closest to uploader in cold level.

Replication Update. Assume time advances in time slots. First, we estimate the number of views $\hat{x}(V)$ video V may achieve at time t using (2). The thresholds of medium level and hot level are denoted as T_M and T_H . Only if $\hat{x}(V)$ exceeds the corresponding threshold, can the video migrate to higher level. The cold level is responsible for permanent storage. The replication location sets for cold, medium, and hot level are denoted as L_C , L_M , and L_H . We define $\lambda_l(V)$ as the views of video V generated from location l at time slot t . α_M and α_H represent geographic access locality in medium level and hot level. For example, if China's provinces ($|L_H| = 34$) reach the hot level, then α_H can be set 70%/30%, denoting 70% of the views divided by corresponding location numbers ($34 \times 30\%$) as the geographic propagation threshold. This threshold guides video replicating at desirable locations and these parameters can be tuned for balancing between user performance and storage cost. Our content replication algorithm is illustrated in Algorithm 1.

Algorithm 1. Content replication algorithm.

```

(1) if v is newly updated then
(2)   replicate v at closest location l ∈ LC
(3) else
(4)   estimate  $\hat{x}_t(v)$ 
(5)   if  $\hat{x}_t(v) < T_M$  then
(6)     if v is in medium or hot level at time t - 1 then
(7)       delete v from medium or hot level
(8)     end if
(9)   else if  $T_M \leq \hat{x}_t(v) < T_H$  then
(10)    for l ∈ LM do
(11)      if  $\lambda_{l,t-1}(v) \geq x_{t-1}(v) \times \alpha_M / |L_M|$  then
(12)        replicate v at location l
(13)      end if
(14)    end for
(15)    delete extra replicas of v in medium or hot level
(16)  else
(17)    for l ∈ LH do
(18)      if  $\lambda_{l,t-1}(v) \geq x_{t-1}(v) \times \alpha_H / |L_H|$  then
(19)        replicate v at location l
(20)      end if
(21)    end for
(22)    delete extra replicas of v in medium or hot level
(23)  end if
(24) end if

```

Request Routing

We use an abstract function $\text{QoS}(u, r)$ to quantify the quality of service between user u and replica r . The QoS metric can be related to many factors such as latency, network congestion, and server load. This provides content provider the flexibility to define its own QoS metric. Further, multihoming is naturally supported in our contentaware request routing mechanism. Many specific server selection or scheduling algorithms [1, 8] based on multihoming can also be applied in our video delivery framework. For intuitive comparison with traditional CDNs, we adopt a simple server selection mechanism: choose one replica r^* with best QoS serving the user u by using criterion $r^* \leftarrow \operatorname{argmax}_{r \in R_V} \text{QoS}(u, r)$, where R_V is the replica list for video V.

SIMULATION AND EVALUATION

In this section, we mark our proposed framework as ACM-based CDN. We collected 5,000 pieces of real request data for a week from September 1 to September 7 in 2015 from Youku portal. We conduct the experiment based on the data to analyze ACM-based CDN's performance from the following perspectives: (1) latency; (2) network congestion; (3) server load.

Experiment Setup

Our experiment is conducted using an event-based simulator implemented in Java. Figure 7 shows our experiment map with locators indicating the replication locations. The locations in medium level or cold level are chosen by k -means clustering. To keep the experiment simple and generic, we select server according to the geographical distance. More QoS definitions of server selection will be planned in our future work. We set the time slot to 1 day due to the granularity of access data we could achieve. According to the result illustrated in Figure 4, we set $n=4$ which is accurate enough for videos which are uploaded more than 4 days. That is, we only need to persist latest 4 day's historical access data for prediction. In order to reduce the perturbations on system performance in new time slot, videos will be replicated incrementally during idle time. Vectors of model parameters Θ_1 to Θ_4 for each category and replication strategy are calculated offline. The simulator uses extra previous 4 days of the logs to build request histories and does not report the performance of the caching algorithm during those days. This is because our algorithm needs the data of previous days to replicate

the initial videos in the cache. Table 2 summarizes more detailed experiment parameters.

Table 2. Experiment parameters

| Parameter | Definition | Value |
|------------------|-------------------------------------------------|-------|
| T | Duration of time | 7 |
| N_{old} | Initial number of old views | 5000 |
| N_{new} | Number of videos updated per time slot | 10 |
| T_M | Threshold for medium level | 50 |
| T_H | Threshold for hot level | 500 |
| L_C | Number of replication locations in cold level | 4 |
| L_M | Number of replication locations in medium level | 12 |
| L_H | Number of replication locations in hot level | 34 |
| α_M | Geographic locality indicator in medium level | 2 |
| α_H | Geographic locality indicator in hot level | 2 |

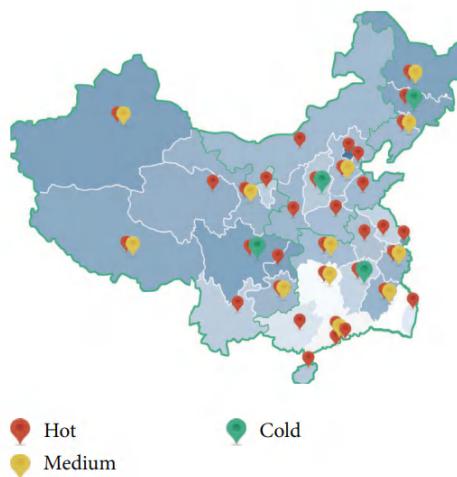


Figure 7. Experiment map of ACM-based CDN for UGC video delivery.

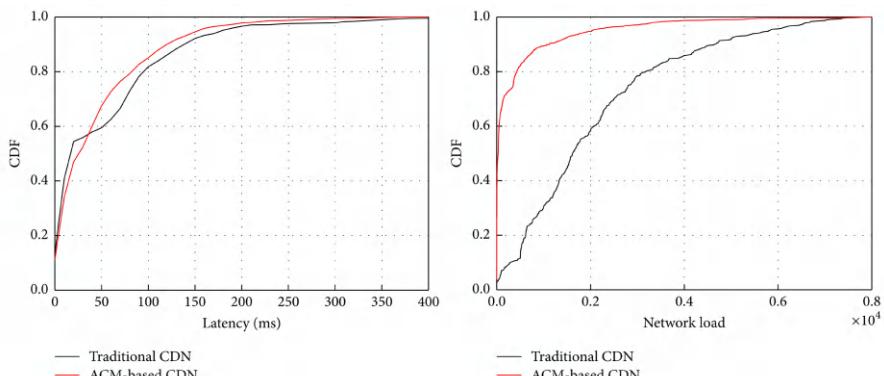
For comparison, we take traditional content delivery network as the baseline. We assume that it is composed of origin server containing all the videos which is located in the map's center, L_M parent servers, and L_H edge servers which are in the same locations illustrated in Figure 7. The origin server, parent servers, and edge servers form a tree-shaped network. A request arrives at the closest edge server and is routed along the closest parent server towards the origin server until it finds the server with the

requested video. A traditional CDN adopts LRU cache replacement policy when request misses. In order to compare fairly, we try to allocate the same storage capacity for different CDNs. We first compute the average cache percentage consumed in hot and medium level in our framework. Depending on the result, 7%, the cache percentage for parent servers and edge servers in traditional CDN is set 10%, considering that our framework requires extra name resolution system.

Results

For comparison, we conduct experiments of two CDN systems based on the same real request data. The experiment shows that our ACM-based CDN outperforms traditional CDN for each performance metric.

Latency. We first present the latency which indicates the transmission time between the request and the location from which it was served. In our experiment, we assume the latency is proportional to geographical distance which is expressed as tenfold geographic distance between the user and selected server in longitude and latitude coordinates. Figure 8(a) shows the latency gap between our ACM-based CDN and traditional CDN is 19 ms at 90th percentile and 27 ms at 95th percentile, that is, 13.2% and 14.7% improvement, respectively. The reason is that, by inferring videos access temporal pattern and geographical locality, better prediction of videos access times can be utilized to select closer replication positions to client, especially for a majority of lukewarm videos compared to traditional “pull-based” caching approach.



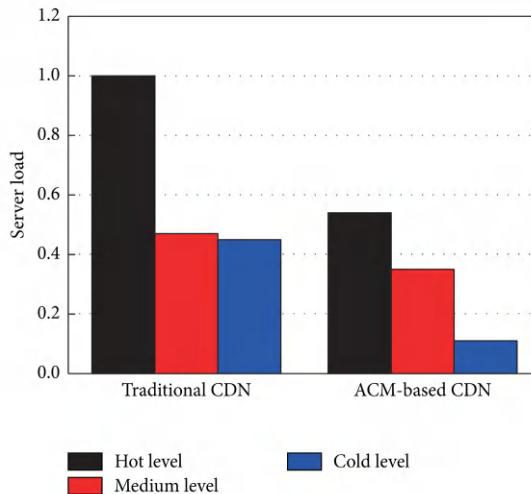


Figure 8. Cost and performance comparison.

Network Load. Next, we investigate the network congestion under two different CDNs. The network load is calculated as the number of videos transferring over the links. Figure 8(b) shows that the network load in our ACM-based CDN is only 37% of that in traditional CDN at 95th percentile.

Server Load. Finally, we study the load on the servers in Figure 8(c). The metrics is the requests served by the servers. We normalize the total server load in our ACM-based CDN. Servers for each level in the traditional CDN bear more load than servers in our ACM-based CDN: 90.4%, 37.1%, and 253.8%, respectively. On the condition that the storage capacity is almost equal, the total server load in the traditional CDN doubles that in the ACM-based CDN. The comparison experiment of network and server load illustrates that our content-aware request routing improves the efficiency to find an appropriate server for serving client, instead of wasting bandwidth on connecting to the upper server to find and transfer the content.

From the result of experiment, the “push-based” adaptive content replication algorithm, together with content-aware request routing mechanism, supports faster video delivery and imposes less traffic burden in network level. Furthermore, CDNs pay for bandwidth based on how many bits exit their servers which can be reflected by our server load experiment. Therefore, our framework together with our algorithm could distribute UGC videos efficiently and economically as well. But it is necessary to point out that compared to traditional CDN which utilizes DNS for request

routing, our framework builds its own name resolution system to implement refined content management. As video scale grows, we will partition the mappings into different index servers using consistent hashing. To look for the accurate index server for a particular video ID, each index server needs to be cooperative with the previous and next index servers to establish route. Our ACM-based CDN will take a longer time to find the video ID-to-IP address mapping but achieving higher performance and lower cost in video transferring phase is worth the sacrifice.

RELATED WORK

Caching mechanisms exploit storage capacity to absorb traffic by replicating content closer to the network edge rather than storing it in a central location which requires high processing power. Most caching schemes utilized in wide-area, distributed systems are initiated by clients (pull-based). The problem of pull-based caching and eviction has received many research efforts. For example, [9, 10] focused on online eviction algorithms (LRU, FIFO, and LFU) and their variants such as greedy and randomized versions. The drawback of these pull-based approaches is that an optimal server is not always chosen to serve content request.

To further improve the Web performance, several works [11, 12] proposed push-based caching as a complementary technique. It is formulated as an optimization problem under a given traffic pattern and a set of resource constraints.

Many studies have examined the characteristics of user generated videos. Cha et al. [4] observed the skewed distribution with long tails. Huguenin et al. [13] showed the correlation between the content locality and geographical locality. Cha et al. [14] observed the correlation between a video's history and its future demand. In this paper, we utilize these characteristics to derive models for prediction and guide intelligent “push-based” caching, rather than assuming idealistic traffic pattern [11, 12]. In addition, compared to the unique solution given by previous works [11, 12], server capacity in our framework can be scaled up and down by adjusting the thresholds in hot and medium level. It provides much more flexibility in cost management.

CONCLUSIONS

We address in this paper the challenges in distributing UGC videos, resulting from the gap between the new features of UGC in mobile Internet

environment and rigid architecture of traditional CDN. We propose a new framework for UGC video delivery that takes proactive content management, joint content replication, and request routing into consideration. Based on the UGC trace analysis for temporal predictability and geographic locality, we present a data-driven content replication algorithm, which distributes content into “cost-effective” locations, and corresponding content-aware request routing algorithm. Extensive experiments driven by the real-world traces demonstrate the high performance and low cost of our design.

Competing Interests

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant no. 2012CB315801, in part by the independent research project of Tsinghua University under Grant no. 20131089304, in part by the projects of Tsinghua National Laboratory for Information Science and Technology(TNLList), in part by the European Seventh Framework Programme (FP7) under Grant no. PIRSES-GA-2012-318939, in part by the National Natural Science Foundation of China under Grant no. 61402343, and in part by Jiangsu International Cooperation Program of Science and Technology under Grant no. BZ2013018.

REFERENCES

1. H. H. Liu, Y. Wang, Y. R. Yang, H. Wang, and C. Tian, “Optimizing cost and performance for content multihoming,” in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM ’12)*, pp. 371–382, Helsinki, Finland, August 2012.
2. V. Ramasubramanian and E. G. Sizer, “The design and implementation of a next generation name service for the internet,” *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 331–342, 2004.
3. V. K. Adhikari, S. Jain, Y. Chen, and Z. L. Zhang, “Vivisecting youtube: an active measurement study,” in *Proceedings of the IEEE Computer and Communications Societies, IEEE Annual Joint Conference (INFOCOM ’12)*, pp. 2521–2525, Orlando, Fla, USA, March 2012.
4. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, “Analyzing the video popularity characteristics of large-scale user generated content systems,” *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357–1370, 2009.
5. A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, “Developing a predictive model of quality of experience for internet video,” in *Proceedings of the ACM Conference on SIGCOMM*, pp. 339–350, ACM, 2013.
6. G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.
7. H. Pinto, J. M. Almeida, and M. A. Gonçalves, “Using early view patterns to predict the popularity of YouTube videos,” in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM ’13)*, pp. 365–374, Rome, Italy, February 2013.
8. P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, “Donar: decentralized server selection for cloud services,” *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4, pp. 231–242, 2010.
9. P. Cao and S. Irani, “Cost-aware www proxy caching algorithms,” in *Proceedings of the Usenix Symposium on Internet Technologies and Systems*, vol. 12, pp. 193–206, 1997.
10. K. Psounis and B. Prabhakar, “Efficient randomized web-cache replacement schemes using samples from past eviction times,” *IEEE/ACM Transactions on Networking*, vol. 10, no. 4, pp. 441–454, 2002.

11. J. Kangasharju, J. Roberts, and K. W. Ross, “Object replication strategies in content distribution networks,” *Computer Communications*, vol. 25, no. 4, pp. 376–383, 2002.
12. T. Bektaş, J.-F. Cordeau, E. Erkut, and G. Laporte, “Exact algorithms for the joint object placement and request routing problem in content distribution networks,” *Computers & Operations Research*, vol. 35, no. 12, pp. 3860–3884, 2008.
13. K. Huguenin, A. Kermarrec, K. Kloudas, and F. Taïani, “Content and geographical locality in user-generated content sharing systems,” in *Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '12)*, pp. 77–82, Toronto, Canada, June 2012.
14. M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, “I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system,” in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '12)*, pp. 1–14, San Diego, Calif, USA, October 2007.

SECTION 3

VOICE AND SPEECH GENERATION

CHAPTER 10

Generating the Voice of the Interactive Virtual Assistant

Adriana Stan¹ and Beáta Lőrincz^{1,2}

¹ Technical University of Cluj-Napoca, Cluj-Napoca, Romania

² “Babeş-Bolyai” University, Cluj-Napoca, Romania

ABSTRACT

This chapter introduces an overview of the current approaches for generating spoken content using text-to-speech synthesis (TTS) systems, and thus the voice of an Interactive Virtual Assistant (IVA). The overview builds upon the issues which make spoken content generation a non-trivial task, and introduces the two main components of a TTS system: text processing and acoustic modelling. It then focuses on providing the reader with the minimally required scientific details of the terminology and methods involved in speech synthesis, yet with sufficient knowledge so as to be able to make the initial decisions regarding the choice of technology for the vocal identity of the

Citation: Stan, A., & Lőrincz, B. (2021). “Generating the Voice of the Interactive Virtual Assistant”. IntechOpen. doi: 10.5772/intechopen.95510.

Copyright: © 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IVA. The speech synthesis methodologies' description begins with the basic, easy to run, low-requirement rule-based synthesis, and ends up within the state-of-the-art deep learning landscape. To bring this extremely complex and extensive research field closer to commercial deployment, an extensive indexing of the readily and freely available resources and tools required to build a TTS system is provided. Quality evaluation methods and open research problems are, as well, highlighted at end of the chapter.

Keywords: text-to-speech synthesis, text processing, deep learning, interactive virtual assistant

INTRODUCTION

Generating the voice of an interactive virtual assistant (IVA) is performed by the so called *text-to-speech synthesis (TTS)* systems. A TTS system takes raw text as input and converts it into an acoustic signal or waveform, through a series of intermediate steps. The synthesised speech commonly pertains to a single, pre-defined speaker, and should be as natural and as intelligible as human speech. An overview of the main components of a TTS system is shown in Figure 1.

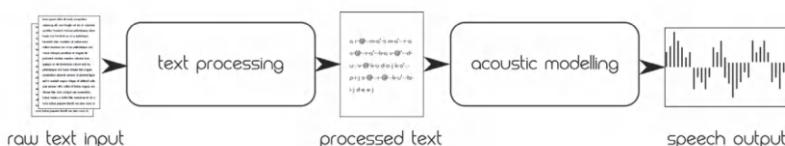


Figure 1. Overview of a text-to-speech synthesis system's main components.

At first sight this seems like a straightforward mapping of each character in the input text to its acoustic realisation. However, there are numerous technical issues which make natural speech synthesis an extremely complex problem, with some of the most important ones being indexed below:

the written language is a discrete, compressed representation of the spoken language aimed at transferring a message, irrespective of other factors pertaining to the speaker's identity, emotional state, etc. Also, in almost any language, the written symbols are not truly informative of their pronunciation, with the most notable example being English. The pronunciation of a letter or sequence of letters which yield a single sound is called a *phone*. One exception here is the Korean alphabet for which the symbols approximate the position of the articulator organs, and was introduced in 1443 by King Sejong the Great to increase the literacy among the Korean population. But for most languages, the so called orthographic transparency is rather opaque;

the human ear is highly adapted to the frequency regions in which the relevant information from speech resides (i.e. 50–8000 Hz). Any slight changes to what is considered to be natural speech, any artefacts, or unnatural sequences present in a waveform deemed to contain spoken content, will be immediately detected by the listener;

speaker and speech variability is a result of the uniqueness of each individual. This means that there are no two persons having the same voice timbre or pronouncing the same word in a similar manner. Even more so, one person will never utter a word or a fixed message in an exactly identical manner even when the repetitions are consecutive;

co-articulation effects derive from the articulator organs' inertial movement. There are no abrupt transitions between sounds and, with very few exceptions, it is very hard to determine the exact boundary of each sound. Another result of the co-articulation is the presence of reductions or modifications in the spoken form of a word or sequence of words, derived from the impossibility or hardship of uttering a smooth transition between some particular phone pairs;

prosody is defined as the rhythm and melody or intonation of an utterance. The prosody is again related to the speaker's individuality, cultural heritage, education and emotional state. There are no clear systems which describe the prosody of a spoken message, and one's person understanding of, for example, portraying an angry state of mind is completely different from another;

no fixed set of measurable factors define a speaker's identity and speaking characteristics. Therefore, when wanting to reproduce one's voice the only way to do this for now is to record that person and extract statistical information from the acoustic signal;

no objective measure correlates the physical representation of a speech signal with the perceptual evaluation of a synthesised speech's quality and/or appropriateness.

The problems listed above have been solved, to some extent, in TTS systems by employing high-level machine learning algorithms, developing large expert resources or by limiting the applicability and use-case scenarios for the synthesised speech. In the following sections we describe each of the main components of a TTS system, with an emphasis on the acoustic modelling part which poses the greatest problems as of yet. We also index some of the freely available resources and tools which can aid a fast development of a synthesis system for commercial IVAs in a dedicated section of the chapter, and conclude with the discussion of some open problems in the final section.

SPEECH PROCESSING FUNDAMENTALS

Before diving into the text-to-speech synthesis components, it is important to define a basic set of terms related to digital speech processing. A complete

overview of this domain is beyond the scope of this chapter, and we shall only refer to the terms used to describe the systems in the following sections.

Speech is the result of the air exhaled from the lungs modulated by the articulator organs and their instantaneous or transitioning position: vocal cords, larynx, pharynx, oral cavity, palate, tongue, teeth, jaw, lips and nasal cavity. By modulation we refer to the changes suffered by the air stream as it encounters these organs. One of the most important organs in speech are the vocal cords, as they determine the periodicity of the speech signal by quickly opening and closing as the air passes through. The vocal cords are used in the generation of vowels and voiced consonant sounds [1]. The perceived result of this periodicity is called the *pitch*, and its objective measure is called *fundamental frequency*, commonly abbreviated F_0 [2]. The slight difference between pitch and F_0 is better explained by the auditory illusion of the *missing fundamental* [3] where the measured fundamental frequency differs from the perceived pitch. Commonly, the terms are used interchangeably, but readers should be aware of this small difference. The pitch variation over time in the speech signal gives the melody or intonation of the spoken content. Another important definition is that of *vocal tract* which refers to all articulators positioned above the vocal cords. The resonance frequencies of the vocal tract are called *formant frequencies*. Three formants are commonly measured and noted as F_1 , F_2 and F_3 .

Looking into the time domain, as a result of the articulator movement, the speech signal is not stationary, and its characteristics evolve through time. The smallest time interval in which the speech signal is considered to be *quasi-stationary* is 20–40 msec. This interval determines the so-called *frame-level analysis* or *windowing* of the speech signal, in which the signal is segmented and analysed at more granular time scales for the resulting analysis to adhere to the digital signal processing theorems and fundamentals [4].

The *spectrum* or *instantaneous spectrum* is the result of decomposing the speech signal into its frequency components through Fourier analysis [5] on a frame-by-frame basis. Visualising the evolution of the spectrum through time yields the *spectrogram*. Because the human ear has a non-linear frequency response, the linear spectrum is commonly transformed into the *Mel spectrum*, where the Mel frequencies are a non-linear transformation of the frequency domain pertaining to the pitches judged by listeners to be equal in distance one from another. Frequency domain analysis is omnipresent in all speech related applications, and Mel spectrograms are the

most common representations of the speech signal in the neural network-based synthesis.

One other frequency-derived representation of the speech is the *cepstral* [6] representation which is a transform of the spectrum aimed at separating the vocal tract and the vocal cord (or glottal) contributions from the speech signal. It is based on homomorphic and decorrelation operations.

TEXT PROCESSING

Text processing or *front-end processing* represents the mechanism of generating supplemental information from the raw input text. This information should yield a representation which is hypothetically closer and more relevant to the acoustic realisation of the text, and therefore tightens the gap between the two domains. Depending on the targeted language, this task is more or less complex [2]. A list of the common front-end processing steps is given below:

text tokenisation splits the input text into syntactically meaningful chunks i.e. phrases sentences and words. Languages which do not have a word separator such as Chinese or Japanese pose additional complexity for this task [7];

diacritic restoration - in languages with diacritic symbols it might be the case that the user does not type these symbols and this leads to an incorrect spoken sequence [8]. The diacritic restoration refers to adding the diacritic symbols back into the text so that the intended meaning is preserved;

text normalisation converts written expressions into their ““spoken”” forms e.g. \$3.16 is converted into “three dollars sixteen cents.” or 911 is converted into “nine one one” and not “nine hundred eleven” [9]. An additional problem is caused by languages which have genders assigned to nouns e.g. in Romanian “21 oi = douăzeci și una de oi” (en. twenty one sheep–feminine) versus “21 cai = douăzeci și unu de cai” (en. twenty one horses-masculine);

part-of-speech tagging (POS) assigns a part-of-speech (i.e. noun, verb, adverb, adjective, etc.) to each word in the input sequence. The POS is important to disambiguate non-homophone homographs. These are words which are spelled the same but pronounced differently based on their POS (e.g. *bow* - to bend down/the front of a boat/tied loops). POS are also essential for placing the accent or focus of an utterance on the correct word or word sequence [10];

lexical stress marking - the lexical stress pertains to the syllable within a word which is more prominent [11]. There are however languages for which this notion is quite elusive such as French or Spanish. Yet in English a stress-timed language assigning the correct stress to each word is essential for conveying the correct message. Along with the POS the lexical stress also helps disambiguate non-homophone homographs in the spoken content. There are also phoneticians who would mark a secondary and tertiary stress but for speech synthesis the primary stress should be enough as the secondary does not affect the meaning but rather the naturalness or emphasis of the speech;

syllabification - syllables represent the base unit of co-articulation and determine the rhythm of speech [12]. Again different languages pose different problems and languages such as Japanese rely on syllables for their alphabetic inventory. As a general rule every syllable has only one vowel sound but can be accompanied by semi-vowels. Compound words generally do not follow the general rules such that prefixes and suffixes will be pronounced as a single syllable;

phonetic transcription is the final result of all the steps above. Meaning that by knowing the POS the lexical stress and syllabification of a word the exact pronunciation can be derived [13]. The phones are a set of symbols corresponding to an individual articulatory target position in a language or otherwise put it is the fixed sound alphabet of a language. This alphabet determines how each sequence of letters should be pronounced. Yet this is not always the case and the concept of orthographic transparency determines the ease with which a reader can utter a written text in a particular language;

prosodic labels, phrase breaks - with all the lexical information in place there is still the issue of emphasising the correct words as per intent of the writer. The accent and pauses in speech are very important and can make the message decoding a very complex task or an easier one with the information being able to be faster assimilated by the listener. There is quite a lot of debate on how the prosody should be marked in text and if it should be [14]. There is definitely some markings in the form of punctuation signs yet there is a huge gap between the text and the spoken output. However public speaking coaching puts a large weight on the prosodic aspect of the speech and therefore captivating the listeners attention through non-verbal queues;

word/character embeddings - are the result of converting the words or characters in the text into a numeric representation which should encompass more information about their identity pronunciation syntax or meaning than the surface form does. Embeddings are learnt from large text corpora and are language dependent. Some of the algorithms used to build such representations are: Word2Vec [15] GloVe [16] ELMo [17] and BERT [18].

ACOUSTIC MODELLING

The acoustic modelling or *back-end processing* part refers to the methods which convert the desired input text sequence into a speech waveform. Some of the earliest proofs of so-called talking heads are mentioned by Aurrilac (1003 A.D.), Albert Magnus (1198–1280) or Roger Bacon (1214–1294). The first electronic synthesiser was the VODER (Voice Operation DEMonstratoR) created by Homer Dudley at Bell Laboratories in 1939. The VODER was able to generate speech by tediously operating a keyboard and foot pedals to control a series of digital filters.

Coming to the more recent developments, and based on the main method of generating the speech signal, speech synthesis systems can be classified into rule-based and corpus-based methods. In rule-based methods, similar to the VODER, the sound is generated by a fixed, pre-computed set

of parameters. Corpus-based methods, on the other hand, use a set of speech recordings to generate the synthetic output or to derive statistical parameters from the analysis of the spoken content. It can be argued that using pre-recorded samples is not in itself synthesis, but rather a speech collage. In this sense Taylor gives a different definition of speech synthesis: “*the output of a spoken utterance from a resource in which it has not been prior spoken*” [2].

Rule-based Synthesis

Formant synthesis is one of the first digital methods of speech generation. It is still used today, especially by phoneticians who study various spoken language phenomena.

The method uses the approximation of several speech parameters (commonly the F_0 and formant frequencies) for each phone in a language, and also how these parameters vary when transitioning from one phone to the next one [19]. The most representative model of formant synthesis is the one described by [20], which later evolved into the commercial system of MITalk [21]. There are around 40 parameters which describe the formants and their respective bandwidths, and also a series of frequencies for nasals or glottal resonators.

The advantages of formant synthesis are related to the good intelligibility even at high speeds, and its very low computation and memory requirements, making it easy to deploy on limited resource devices. The major drawback of this type of synthesis is, of course, its low quality and robotic sound, and also the fact that for high-pitched outputs, the formant tracking mechanisms can fail to determine the correct values.

Articulatory synthesis uses mechanical and acoustic models of speech production [1]. The physiological effects such as the movement of the tongue, lips, jaw, and the dynamics of the vocal tract and glottis are modelled.

For example, [22] uses lip opening, glottal area, opening of nasal cavities, constriction of tongue, and rate between expansion and contraction of the vocal tract along with the first four formant frequencies. Magnetic resonance imaging offers some more insight into the muscle movement [23], yet the complexity of this type of synthesis makes it rather unfeasible for high naturalness and commercial deployment. One exception in the project GNUSpeech [24] but its results are still poor compared to what corpus-based synthesis is able to achieve nowadays.

Corpus-based Synthesis

Concatenative Synthesis

As the name entails, concatenative synthesis is a method of producing spoken content by concatenating pre-recorded speech samples. In its most basic form, a concatenative synthesis system contains recordings of all the words needed to be uttered, which are then combined in a very limited vocabulary scenario. For example, in a rudimentary IVA, it will combine the typed-in phone number of a customer by combining pre-recorded digits. Of course, in a large vocabulary, open-domain system, pre-recording all the words in a language is unfeasible. The solution to this problem is to find a smaller set of acoustic units which can be then combined into any spoken phrase. Based on the type of segment stored in the recorded database, the concatenative synthesis is either fixed inventory – segments in the database have the same length, or variable inventory or unit selection – segments have variable length. As the basic acoustic unit of any language is its phone set, a first open-domain fixed inventory concatenative synthesis made use of *diphones* [25, 26]. A diphone is the acoustic unit spanning from the middle of a phone to the middle of the next one in adjoining phone pairs. Although this yields a much larger acoustic inventory, the diphones are a better choice than phones because they can model the co-articulation effects. For a primitive diphone concatenation system, the recorded speech corpus would include a single repetition of all the diphones in a language. More elaborate systems use diphones in different context (e.g. beginning, middle or end of a word) and with different prosodic events (e.g. accent, variable duration etc.). Another type of fixed inventory system is based on the use of *syllables* as the concatenation unit [27, 28, 29]. Some theories state that the basic unit of speech is the syllable and, therefore, the co-articulation effects between them is minimum [30], but the speech database is hard to design. The average number of unique syllables in one language is in the order of thousands.

A natural evolution of the fixed inventory synthesis is the variable length inventory, or unit selection [31, 32]. In unit selection, the recorded corpus includes segments as small as half-phones and go up to short common phrases. The speech database is either stored as-is, or as a set of parameters describing the exact acoustic waveform. The speech corpus, therefore, needs to be very accurately annotated with information regarding the exact phonetic content and boundaries, lexical stress, syllabification, lexical focus

and prosodic trends or patterns (e.g. questions, exclamation, statements). The combination of the speech units into the output spoken phrase is done in an iterative manner, by selecting the best speech segments which minimise a global cost function [31] composed of: a *target cost* - measuring how well a sequence of units matches the desired output sequence, and a *concatenation cost* - measuring how well a sequence of units will be joined together and thus avoid the majority of the concatenation artefacts.

Although this type of synthesis is almost 30 years old, it is still present in many commercial applications. However, it poses some design problems, such as: the need for a very large manually segmented and annotated speech corpus; the control of prosody is hard to achieve if the corpus does not contain all the prosodic events needed to synthesise the desired output; changing the speaker identity requires the database recording and processing to be started from scratch; and there are quite a lot of concatenation artefacts present in the output speech making it unnatural, but which have, in some cases, been solved by using a hybrid approach [33].

Statistical-parametric Synthesis

Because concatenative synthesis is not very flexible in terms of prosody and speaker identity, in 1989 a first model of statistical-parametric synthesis based on Hidden Markov Models (HMMs) was introduced [34]. The model is parametric because it does not use individual stored speech samples, but rather parameterises the waveform. And it is statistical because it describes the extracted parameters using statistics averaged across the same phonetic identity in the training data [35]. However this first approach did not attract the attention of the specialists because of its highly unnatural output. But in 2005, the HMM-based Speech Synthesis System (HTS) [36] solved part of the initial problems, and the method became the main approach in the research community with most of its studies aiming at fast speaker adaptation [37] and expressivity [38]. In HTS, a 3 state HMM models the statistics of the acoustic parameters of the phones present in the training set. The phones are clustered based on their identity, but also on other contextual factors, such as the previous and next phone identity, the number of syllables in the current word, the part-of-speech of the current word, the number of words in the sentence, or the number of sentences in a phrase, etc. This context clustering is commonly performed with the help of decision trees and ensures that the statistics are extracted from a sufficient number of exemplars. At synthesis time, the text is converted in a context aware complex label and drives the selection of the HMM states and their

transitions. The modelled parameters are generally derived from the source-filter model of speech production [1]. One of the most common vocoders used in HTS is STRAIGHT [39] and it parameterises the speech waveform into F_0 , Mel cepstral and aperiodicity coefficients. A less performant, yet open vocoder is WORLD [40]. A comparison of several vocoders used for statistical parametric speech synthesis is presented in [41].

There are several advantages for the statistical-parametric synthesis, such as: the small footprint necessary to store speech information; automatic clustering of speech information—removes the problems of hand-written rules; generalisation—even if for a certain phoneme context there is not enough training data, the phone will be clustered along with similar parameter characteristics; flexibility—the trained models can be easily adapted to other speakers or voice characteristics with minimum amount of adaptation data. However, the parameter averaging yields the so-called *buzziness* and low speaker similarity of the output speech, and for this reason the HTS system has not truly made its way into the commercial applications.

Neural Synthesis

In 1943, McCulloch and Pitts [42] introduced the first computational model for artificial neural networks (ANN). And although the incipient ANNs have been successfully applied in multiple research areas, including TTS [43], their learning power comes from the ability to stack multiple neural layers between the input and output. However, it was not until 2006 that the hardware and algorithmic solutions enabled adding multiple layers and making the learning process stable. In 2006, Geoffrey Hinton and his team published a series of scientific papers [44, 45] showing how a many-layered neural network could be effectively pre-trained one layer at a time. These remarkable results set the trend for all automatic machine learning algorithms in the following years, and are the bases of the deep neural network (DNN) research field. Nowadays, there are very few machine learning applications which do not cite the DNNs as attaining the state-of-the-art results and performances.

In text-to-speech synthesis, the progression from HMMs to DNNs was gradual. Some of the first impacting studies are those of Ling et al. [46] and Zen et al. [47]. Both papers substitute parts of the HMM-based architecture, yet model the audio on a frame-by-frame basis, maintaining the statistical-parametric approach, and also use the same contextual factors in the text processing part. The first open source tool to implement the DNN-

based statistical-parametric synthesis is Merlin [48]. A comparison of the improvements achieved by the DNNs compared to HMMs is presented in [49]. However, these methods still rely on a time-aligned set of text features and their acoustic realisations, which requires a very good frame-level aligner systems, usually an HMM-based one. Also, the sequential nature of speech is only marginally modelled through the contextual factors and not within the model itself, while the text still needs to be processed with expert linguistic automated tools which are rarely available in non-mainstream languages.

An intermediate system which replaces all the components in a TTS pipeline with neural networks is that of [50], but it does not incorporate a single end-to-end network. The first study which removes the above dependencies, and models the speech synthesis process as a sequence-to-sequence recurrent network-based architecture is that of Wang et al. [51]. The architecture was able to “*synthesise fairly intelligible speech*” and was the precursor of the more elaborate Char2Wav [52] and Tacotron [53] systems. Both Char2Wav and Tacotron model the TTS generation as a two step process: the first one takes the input text string and converts it into a spectrogram, and the second one, also called the *vocoder*, takes the spectrogram and converts it into a waveform, either in a deterministic manner [54], or with the help of a different neural network [55]. These two synthesis systems were also the first to alleviate the need for more elaborate text representations, and derived them as an inherent learning process, setting the first stepping stones towards true end-to-end speech synthesis [56]. However, for phonetically rich languages it is common to train the models on phonetically transcribed text, and also to augment the input text with additional linguistic information such as part-of-speech tags which can enhance the naturalness of the output speech [57, 58].

Starting with the publication of Tacotron, the DNN-based speech synthesis research and development area has seen an enormous interest from both the academia and the commercial sides. Most focus has been granted on generating extremely high quality speech, but also to the reduction of the computational requirements and generation speed—which in the DNN domain is called *inference speed*. A major breakthrough was obtained by the second version of Tacotron, Tacotron 2 [59], which achieved naturalness scores very close to human speech. However, both systems’ architectures involve attention-based recurrent auto-regressive processes which make the inference step very slow and prone to instability issues, such as word skipping, deletions or repetitions. Also, the recurrent neural networks

(RNNs) are known to have high demands in terms of data availability and training time. So that, the next step in DNN-based TTS was the introduction of CNNs, in systems such as DC-TTS [60], DeepVoice 3 [61], ClariNet [62], or ParaNet [63]. The CNNs enable a much better data and training efficiency and also a much faster inference speed through parallel processing. And also, recently, the research community started to look into ways of replacing the auto-regressive attention-based generation, and incorporated duration prediction models which stabilise the output and enable a much faster parallel inference of the output speech [64, 65].

Inspired by the success of the Transformer network [66] in text processing, TTS systems have adopted this architecture as well. Transformer based models include Transformer-TTS [67], FastSpeech [68], FastSpeech 2 [69], AlignTTS [70], JDI-T [71], MultiSpeech [72], or Reformer-TTS [73]. Transformer-based architectures improve the training time requirements, and are capable of modelling longer term dependencies present in the text and speech data.

As the naturalness of the output synthetic speech became very high-quality, researchers started to look into ways of easily controlling the different factors of the synthetic speech, such as duration or style. The go-to solution for this are the Variational AutoEncoders (VAEs) and their variations, which enable the disentanglement of the latent representations, and thus a better control of the inferred features [74, 75, 76, 77, 78]. There were also a few approaches including Generative Adversarial Networks (GANs), such as GAN-TTS [79] or [80], but due to the fact that GANs are known to pose great training problems, this direction was not that much explored in the context of TTS.

A common problem in all generative modelling irrespective of deep learning methodologies, is the fact that the true probability distribution of the training data is not directly learned or accessible. In 2015, Rezende et al. [81] introduced the normalising flows (NFs) concept. NFs estimate the true probability distribution of the data by deriving it from a simple distribution through a series of invertible transforms. The invertible transforms make it easy to project a measured data point into the latent space and find its likelihood, or to sample from the latent space and generate natural sounding output data. For TTS, NFs have just been introduced, yet there are already a number of high-quality systems and implementations available, such as: Flowtron [82], Glow-TTS [83], Flow-TTS [84], or Wave Tacotron [56]. From the generative perspective, this approach seems, at the moment, to

be able to encompass all the desired goals of a speech synthesis system, but there are still a number of issues which need to be addressed, such as the inference time and latent space disentanglement and control.

All the above mentioned neural systems only solve the first part of the end-to-end problem, by taking the input text and converting it into a Mel spectrogram, or variations of it. For the spectrogram to be converted into an audio waveform, there is the separate component, called the vocoder. And there are also numerous studies on this topic dealing with the same trade-off issue of quality versus speed [85].

WaveNet [55] was one of the first neural networks designed to generate audio samples and achieved remarkably natural results. It is still the one vocoder to beat when designing new ones. However, its auto-regressive processes make it unfeasible for parallel inference, and several methods have been proposed to improve it, such as FFTNet [86] or Parallel WaveNet [87], but the quality is somewhat affected. Some other neural architectures used in vocoders are, of course, the recurrent networks used in WaveRNN [88] and LPCNet [89], or the adversarial architectures used in MelGAN [90], GELP [91], Parallel WaveGAN [92], VocGAN [93]. Following the trend of normalising flows-based acoustic modelling, flow-based vocoders have also been implemented. Some of the most remarkable being: FlowWaveNet [94], WaveGlow [95], WaveFlow [96], WG-WaveNet [97], EWG (Efficient WaveGlow) [98], MelGlow [99], or SqueezeWave [100].

In light of all these methods available for neural speech synthesis, it is again important to note the trade-offs between the quality of output speech, model sizes, training times, inference speed, computing power requirements and ease of control and adaptability. In the ideal scenario, a TTS system would be able to generate natural speech, at an order of magnitude faster than real-time processing speed, on a limited resource device. However, this goal has not yet been achieved by the current state-of-the-art, and any developer looking into TTS solutions should first determine the exact applicability scenario before implementing any of the above methods. It may be the case that, for example, in a limited vocabulary, non-interactive assistant, a simple formant synthesis system implemented on a dedicated hardware might be more reliable and adequate.

Some aspects which we did not take into account in the above enumeration are the multispeaker, multilingual TTS systems. However, in a commercial setup these are not directly required and can be substituted by independent high-quality systems integrated in a seamless way within the IVA.

OPEN RESOURCES AND TOOLS

Deploying any research result into a commercial environment requires at least a baseline functional proof-of-concept from which to start optimising and adapting the system. It is the same in TTS systems, where especially the speech resources, text-processing tools, and system architectures can be at first tested and only then developed and migrated to the live solution. To aid this development, the following table indexes some of the most important resources and tools available for text to speech synthesis systems. This is by no means an exhaustive list, but rather a starting point. The official implementations pertaining to the published studies are marked as such. If no official implementation was found, we relied on our experience and prior work to link an open tool which comes as close as possible to the original publication.

| Speech and text datasets and resources |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Language Data Consortium (LDC) is a repository and distribution point for various language resources. Link: www.ldc.upenn.edu |
| The European Language Resources Association (ELRA) is a non-profit organisation whose main mission is to make Language Resources for Human Language Technologies available to the community at large. Link: www.elra.info/en/ |
| META-SHARE [101] is an open and secure network of repositories for sharing and exchanging language data, tools and related web services. Link: www.meta-share.org |
| OpenSLR is a site devoted to hosting speech and language resources, such as training corpora for speech recognition, and software related mainly to speech recognition. Link: www.openslr.org |
| LibriVox is a group of worldwide volunteers who read and record public domain texts creating free public domain audiobooks for download. Link: www.librivox.org |
| Mozilla Common Voice is part of Mozilla's initiative to help teach machines how real people speak. Link: www.commonvoice.mozilla.org/en/datasets |
| Project Gutenberg is an online library of free eBooks. Link: www.gutenberg.org |
| LibriTTS [102] is a multi-speaker English corpus of approximately 585 hours of read English speech designed for TTS research. Link: www.openslr.org/60/ |
| The Centre for Speech Technology Voice Cloning Toolkit (VCTK) Corpus includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences. Link: www.datashare.is.ed.ac.uk/handle/10283/2950 |

CMU Wilderness Multilingual Speech Dataset [103] is a speech dataset of aligned sentences and audio for some 700 different languages. It is based on readings of the New Testament. Link: www.github.com/festvox/datasets-CMU_Wilderness

Text processing tools

Festival is a complete TTS system, but it enables the use of its front-end tools independently. It supports several languages and dialects. Link: www.cstr.ed.ac.uk/projects/festival/

CMUSphinx G2P tool is a grapheme-to-phoneme conversion tool based on transformers. Link: www.github.com/cmusphinx/g2p-seq2seq

Multilingual G2P uses the eSpeak tool to generate phonetic transcriptions in multiple languages. Link: www.github.com/jcsilva/multilingual-g2p.

Stanford NLP tools includes various text-processing and knowledge extraction tools for English and other languages. Link: www.nlp.stanford.edu/software/

RecoAPy [104] tool includes an easy to use interface for recording prompted speech, but also a set of models able to perform high accuracy phonetic transcription in 8 languages. Link: www.gitlab.utcluj.ro/sadriana/recoapy

word2vec [15] is a word embedding model that learns vector representations of words that capture semantic and other properties of these words from large amounts of text data. Link: code.google.com/archive/p/word2vec/

GloVe [16] is a word embedding method that learns from the co-occurrences of words in text corpus obtaining similar vector representations for words that occur in the same context. Link: www.nlp.stanford.edu/projects/glove/

ELMo [17] obtains contextualized word embeddings that model the semantics and syntax of the word, but can learn different representations for various contexts. Link: www.allennlp.org/elmo

BERT [18] is a Transformer-based model that obtains context dependent word embeddings and can process sentences in parallel. Link: www.github.com/google-research/bert

Speech synthesis systems

eSpeak is a formant-based compact open source software speech synthesiser. Link: www.espeak.sourceforge.net/ [Official]

Festival is an unrestricted commercial and non-commercial use framework for building concatenative and HMM-based TTS systems. Link: www.cstr.ed.ac.uk/projects/festival/ [Official]

MaryTTS [105] is an open-source, multilingual TTS platform written in Java supporting diphone and unit selection synthesis. Link: <http://mary.dfki.de/> [Official]

HTS [36] is the most commonly used implementation of the HMM-based speech synthesis. Link: <http://hts.sp.nitech.ac.jp/> [Official]

| |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Merlin [48] is a Python implementation of DNN models for statistical parametric speech synthesis. Link: www.github.com/CSTR-Edinburgh/merlin [Official] |
| IDLAK [106] is a project to build an end-to-end neural parametric TTS system within the Kaldi ASR framework. Link: www.idlak.readthedocs.io/en/latest/ [Official] |
| DeepVoice [50] follows the structure of HMM-based TTS systems, but replaces all its components with neural networks. Link: www.github.com/israelg99/deepvoice |
| Char2Wav [52] is an end-to-end neural model trained on characters that can synthesise speech with the SampleRNN vocoder. Link: https://github.com/sotelo/parrot [Official] |
| Tacotron [53] is one of the most frequently used end-to-end neural synthesis systems based on recurrent neural nets and attention mechanism. Link: www.github.com/keithito/tacotron |
| VoiceLoop [107] is one of the first neural synthesisers which uses a buffer memory instead of recurrent layers and does not require an audio-to-phone alignment. Link: www.github.com/facebookarchive/loop [Official] |
| Tacotron 2 [59] is an enhanced version of Tacotron which modifies the attention mechanism and also uses the WaveNet vocoder to generate the output speech. Link: www.github.com/NVIDIA/tacotron2 |
| DeepVoice 3 [61] is a fully convolutional synthesis system that can synthesise speech in a multispeaker scenario. Link: www.github.com/r9y9/deepvoice3_pytorch |
| DCTTS [60] - Deep Convolutional TTS is a synthesis system that implements a two step synthesis, by first learning a coarse and then a fine-grained representation of the spectrogram. Link: www.github.com/tugstugi/pytorch-dc-tts |
| ClariNet [62] is the first text-to-wave neural architecture for speech synthesis, which is fully convolutional and enables fast end-to-end training from scratch. Link: www.github.com/ksw0306/ClariNet |
| Transformer TTS [67] replaces the recurrent structures of Tacotron 2 with attention mechanisms. Link: www.github.com/soobinseo/Transformer-TTS |
| GAN-TTS [79] is a GAN-based synthesis system that uses a generator to produce speech and multiple discriminators that evaluate the naturalness and text-adequacy of the output. Link: www.github.com/yanggeng1995/GAN-TTS |
| FastSpeech [68] is a novel feed-forward network based on Transformer which generates the Mel-spectrogram in parallel, and uses a teacher-based length predictor to achieve this parallel generation. Link: www.github.com/xcmyz/FastSpeech |
| FastSpeech 2 [69] is an enhanced version of FastSpeech where the length predictor teacher network is replaced by conditioning the output on duration, pitch and energy from extracted from the speech waveform at training and their predicted values in inference. Link: www.github.com/ming024/FastSpeech2 |

AlignTTS [70] is a feed-forward Transformer-based network with a duration predictor which aligns the speech and audio. Link: www.github.com/Deepest-Project/AlignTTS

Mellotron [108] is a multispeaker TTS able to emote emotions by explicitly conditioning on rhythm and continuous pitch contours from an audio signal. Link: www.github.com/NVIDIA/mellotron [Official]

Flowtron [82] is an autoregressive normalising flow-based generative network for TTS, also capable of transferring style from one speaker to another. Link: www.github.com/NVIDIA/flowtron [Official]

Glow-TTS [83] is a flow-based generative model for parallel TTS using a dynamic programming method to achieve the alignment between text and speech. Link: www.github.com/jaywalnut310/glow-tts [Official]

Speech synthesis system libraries

Mozilla TTS is a deep learning library for TTS that includes implementations for Tacotron, Tacotron 2, Glow-TTS and vocoders such as MelGAN, WaveRNN and others. Link: www.github.com.mozilla/TTS [Official]

NeMO is a toolkit that includes solutions for TTS, speech recognition and natural language processing tools as well. Link: www.github.com/NVIDIA/NeMo [Official]

ESPNET-TTS [109] is a toolkit that contains implementations for TTS systems like Tacotron, Transformer TTS, FastSpeech and others. Link: www.github.com/espnet/espnet [Official]

Parakeet is a flexible, efficient and state-of-the-art text-to-speech toolkit for the open-source community. It includes many influential TTS models proposed by Baidu Research and other research groups. Link: www.github.com/PaddlePaddle/Parakeet [Official]

Neural Vocoder

WaveNet [55] is an autoregressive and probabilistic model used to generate raw audio. It can also be conditioned on text to produce the very natural output speech, but its complexity makes it very resource demanding. Link: www.github.com/r9y9/wavenet_vocoder

WaveRNN [88] is a recurrent neural network based vocoder that is able to generate audio faster than real time as a result of its compact architecture. Link: www.github.com/fatchord/WaveRNN

FFTNet [86], inspired by WaveNet also generates the waveform samples sequentially, with the current sample being conditioned on the previous ones, but simplifies its architecture and allows real-time synthesis. Link: www.github.com/syang1993/FFTNet

| |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| nv-WaveNet is an open-source implementation of several different single-kernel approaches to the WaveNet variant described by [50]. Link: www.github.com/NVIDIA/nv-wavenet [Official] |
| LPCNet [89] is a variant of WaveRNN that improves the waveform generation by combining the recurrent neural architecture with linear prediction coefficients. Link: www.github.com/mozilla/LPCNet [Official] |
| FloWaveNet [94] is a generative model based on flows that can sample audio in real time. Compared to Parallel WaveNet and ClariNet it only requires a training process that is single-staged. Link: www.github.com/ksw0306/FloWaveNet [Official] |
| Parallel WaveGAN [95] is a vocoder that uses adversarial training and provides fast and lightweight waveform generation. Link: www.github.com/kan-bayashi/ParallelWaveGAN |
| WaveGlow [95] vocoder borrows from Glow and WaveNet to generate raw audio from Mel spectrograms. It is a flow-based model implemented with a single network. Link: www.github.com/NVIDIA/waveglow [Official] |
| MelGAN [90] is a GAN-based vocoder that is able to generate coherent waveforms, the model is non-autoregressive and based on convolutional layers. Link: www.github.com/descriptinc/melgan-neurips [Official] |
| GELP [91] is a parallel neural vocoder utilising generative adversarial networks, and integrating a linear predictive synthesis filter into the model. Link: www.github.com/ljuvela/GELP |
| SqueezeWave [100] is a lightweight version of WaveGlow that can generate on-device speech output. Link: https://github.com/tianrengao/SqueezeWave [Official] |
| WaveFlow [96] is a flow-based model that includes WaveNet and WaveGlow as special cases and can synthesise audio faster than real-time. Link: www.github.com/L0SG/WaveFlow |
| VocGAN [93] is a GAN-based vocoder that can synthesise speech in real time even on a CPU. Link: www.github.com/rishikksh20/VocGAN |
| WG-WaveNet [97] is composed of a WaveGlow like flow-based model combined with WaveNet based postfilter that can synthesise speech without the need for a GPU. Link: www.github.com/BogiHsu/WG-WaveNet |
| Speech synthesis challenges |
| Blizzard Challenge is a yearly challenge in which teams develop TTS systems starting from more or less the same resources, and are jointly evaluated in a large-scale listening test. Link: http://www.festvox.org/blizzard/ |
| Voice Cloning Challenge is a bi-annual challenge in which teams are asked to provide a high-quality solution for cloning the voice of a target speaker within the same language, or cross-lingual. The results are also evaluated in a large scale listening test. Link: http://www_vc-challenge.org/ |

QUALITY MEASUREMENTS

Although there are no objective measures which can perfectly predict the perceived naturalness of the synthetic output [110, 111], we still need to measure a TTS system's performance. The current approach to doing this is to use *listening tests*. In a listening test, a set of listeners, preferably a large number of native speakers of the target language, are asked to rate the synthetic output in several scenarios using either absolute or relative values. The common setup includes multiple synthesis systems and natural samples. The evaluation can be performed by presenting one or two samples at a time and the listeners rate it by using a Mean Opinion Score (MOS) scale going from 1 to 5, with 5 being the highest value. Or, more commonly used nowadays, in a MUSHRA [112] setup, in which multiple samples are presented the same time and the listeners are asked to order and rate them on a scale of 1 to 100. There is also a preference test setup in which the listeners are asked to choose between two samples according to their preference or adequacy of the rendered speech to the text or speaker identity. The most common evaluation criteria are:

naturalness listeners are asked to rank how close to natural speech is a sample of synthetic output perceived;

intelligibility listeners are asked to transcribe what they hear after playing the sample only once. The transcripts are then compared to the reference transcript and the word error rate is computed;

speaker similarity listeners are presented with a natural sample as reference and a synthetic or natural sample for evaluation. They are asked to rate how similar the identity of the evaluation sample is in comparison to the reference sample.

CONCLUSIONS AND OPEN PROBLEMS

In this chapter we aimed to provide a high-level indexing of the available methods to generate the voice of an IVA, and to provide the reader with a clear, informed starting point for developing his/her own text-to-speech synthesis system. In the recent years there has been an increasing interest in this domain, especially in the context of vocal chat bots and content access. So that it would be next to impossible to index all the publications and available tools and resources. Yet, we consider that the provided knowledge and minimal scientific description of the TTS domain is sufficient to trigger the interest and application of these methods in the reader's commercial products. It should also be clear that there is still an important trade-off between the quality and the resource requirements of the synthetic voices,

and that a very thorough analysis of the applications' specifications and intended use should guide the developer into making the right choice of technology.

We should also point out that, although the recent advancements achieve close to human speech quality, there are still a number of issues that need to be addressed before we can easily say that the topic of speech synthesis has been thoroughly solved. One of these issues is that of *adequate prosody*. When synthesising long paragraphs, or entire books, there is still a lack of variability in the output, and a subset of certain prosodic patterns reemerge. Also, the problem of correctly emphasising certain words, or word groups, such that the desired message is clearly and correctly transmitted is still an open issue for TTS. There is also the problem of mimicking spontaneous speech, where repetitions, elisions, filled pauses, breaks and so on convey the mental process and effort of developing the message and generating it as a spoken discourse.

In terms of speaker identity, the fast adaptation, and also cross-lingual adaptation are of great interest to the TTS community at this point. Being able to copy a person's speech characteristics using as little examples as possible is a daunting task, yet giant leaps have been taken with the NN-based learning. More so, transferring the identity of a person speaking in a language, to the identity of a synthesis system generating a different language is also open for solutions.

On the more far-fetched goals is that of *affective rendering*. If we were to interact with a complete synthetic persona, we would like it to be adaptable to our state of mind, and render compassionate and emphatic emotions in its discourse. Yet the automatic detection and generation of emotions is far from being solved.

REFERENCES

1. J. Benesty, M. M. Sondhi, and Y. A. Huang, Springer Handbook of Speech Processing. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007
2. P. Taylor, Text-to-Speech Synthesis. Cambridge University Press, 2009
3. “Missing fundamental,” en. wikipedia. org/wiki/ Missing fundamental, online; accessed 15-December-2020
4. S. King, “Speech Zone - Windowing,” speech. zone/windowing/, online; accessed 15-December-2020
5. “Fourier analysis,” en. wikipedia. org/wiki/Fourier analysis, online; accessed 15-December-2020
6. “Cepstrum,” en. wikipedia. org/wiki/Cepstrum, online; accessed 15-December-2020
7. J. Li, Z. Wu, R. Li, P. Zhi, S. Yang, and H. Meng, “Knowledge-Based Linguistic Encoding for End-to-End Mandarin Text-to-Speech Synthesis,” in Proc. Interspeech 2019, 2019, pp. 4494–4498
8. M. Nutu, B. Lorincz, and A. Stan, “Deep Learning for Automatic Diacritics Restoration in Romanian,” in Proc. of IEEE 15th International Conference on Intelligent Computer Communication and Processing, 09 2019, pp. 1–5
9. H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, “Neural Models of Text Normalization for Speech Applications,” Computational Linguistics, vol. 45, no. 2, pp. 293–337, 2019
10. B. Bohnet, R. McDonald, G. Simões, D. Andor, E. Pitler, and J. Maynez, “Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2642–2652
11. A. Cutler, Lexical Stress. John Wiley & Sons, Ltd, 2005, ch. 11, pp. 264–289
12. S. Thomas, M. N. Rao, H. A. Murthy, and C. S. Ramalingam, “Natural sounding TTS based on syllable-like units,” in 14th European Signal Processing Conference, EUSIPCO2006, Florence, Italy, September 4–8, 2006. IEEE, 2006, pp. 1–5

13. A. Sokolov, T. Rohlin, and A. Rastrow, “Neural Machine Translation for Multilingual Grapheme-to-Phoneme Conversion,” in Proc. Interspeech 2019, 2019, pp. 2065–2069
14. “W3C - Speech Synthesis Markup Language (SSML) Version 1.1,” <https://www.w3.org/TR/speech-synthesis11/>, online; accessed 15-December-2020
15. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119
16. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543
17. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” arXiv preprint arXiv:1802.05365, 2018
18. J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018
19. X. Huang, A. Acero, H. -W. Hon, and R. Reddy, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st ed. USA: Prentice Hall PTR, 2001
20. D.H.Klatt, “Software for a cascade/parallel formantsynthesizer,” Journal of The Acoustical Society of America, vol. 67, 1980
21. J. Allen, S. Hunnicut, and D. Klatt, From Text to Speech: the MITalk System. Cambridge University Press, 1987
22. C. Bickley, K. Stevens, and D. Williams, “A framework for synthesis of segments based on pseudoarticulatory parameters,” pp. 211–220, 1997
23. K. Richmond, Z. -H. Ling, and J. Yamagishi, “The use of articulatory movement data in speech synthesis applications: An overview - application of articulatory movements using machine learning algorithms [invited review],” Acoustical Science and Technology, vol. 36, no. 6, pp. 467–477, 2015
24. D. Hill, “gnuspeech,” www.gnu.org/software/gnuspeech/, online; accessed 15-December-2020

25. A. Black, P. Taylor, and R. Caley, The Festival Speech Synthesis System, University of Edinburgh, 1999
26. T. Lambert and A. P. Breen, “A database design for a TTS synthesis system using lexical diphones,” in Proceedings of Interspeech, 2004
27. T. Saito, Y. Hashimoto, and M. Sakamoto, “High-quality speech synthesis using context-dependent syllabic units,” in Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings. , 1996 IEEE International Conference – Volume 01, ser. ICASSP ‘96, 1996, pp. 381–384
28. J. Matoušek, Z. Hanzlíček, and D. Tihelka, “Hybrid syllable/triphone speech synthesis,” in Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005, pp. 2529–2532
29. O. Buza, “Contribut, ii la analizas, și sinteza vorbirii din text pentru limba română,” Ph. D. dissertation, Technical University of Cluj-Napoca, 2010
30. R. Stetson, Motor Phonetics: A Study of Speech Movements in Action. Oberlin College, 1951
31. A. Black and N. Campbell, “Optimising selection of units from speech database for concatenative synthesis,” in Proc. EUROSPEECH-95, Sep. 1995, pp. 581–584
32. A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. of ICASSP, May 1996, pp. 373–376
33. Y. Qian, F. K. Soong, and Z. Yan, “A unified trajectory tiling approach to high quality speech rendering,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 280–290, 2013
34. A. Falaschi, M. Giustiniani, and M. Verola, “A hidden Markov model approach to speech synthesis,” in Proceedings of Eurospeech, vol. 1989, 1989, pp. 2187–2190
35. S. King, “An introduction to statistical parametric speech synthesis,” Sadhana, vol. 36, p. 837–852, 2011
36. H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” IEICE Trans. Inf. & Syst. , vol. E90-D, no. 1, pp. 325–333, Jan. 2007

37. J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, “Thousands of voices for HMM-based speech synthesis – analysis and application of TTS systems built on various ASR corpora,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 984–1004, July 2010
38. J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, “Emotion transplantation through adaptation in hmm-based speech synthesis,” *Computer Speech and Language*, vol. 34, no. 1, pp. 292–307, 2015
39. H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999
40. M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” *IEICE Transactions*, vol. 99-D, pp. 1877–1884, 2016
41. Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, “An experimental comparison of multiple vocoder types,” in 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, August 2013, pp. 155–160
42. W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Neurocomputing: Foundations of Research*, p. 15–27, 1988
43. M. G. Rahim and C. C. Goodyear, “Articulatory synthesis with the aid of a neural net,” in International Conference on Acoustics, Speech, and Signal Processing, 1989, pp. 227–230 vol. 1
44. G. E. Hinton, “Learning multiple layers of representation,” *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007
45. G. E. Hinton, S. Osindero, and Y. -W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006
46. Z. Ling, L. Deng, and D. Yu, “Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013

47. H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7962–7966
48. Z. Wu, O. Watts, and S. King, “Merlin: An open source neural network speech synthesis system.” In Speech Synthesis Workshop, 2016, pp. 202–207
49. O. Watts, G. Henter, J. Fong, and C. Valentini-Botinhao, “Where do the improvements come from in sequence-to-sequence neural TTS?” in Proc of the 10th ISCA Speech Synthesis Workshop. International Speech Communication Association, Sep. 2019, pp. 217–222
50. S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, “Deep voice: Real-time neural text-to-speech,” arXiv preprint arXiv:1702. 07825, 2017
51. W. Wang, S. Xu, and B. Xu, “First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention,” in Interspeech 2016, 2016, pp. 2243–2247
52. J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, “Char2wav: End-to-end speech synthesis,” in 5th International Conference on Learning Representations, ICLR2017, Toulon, France, April 24–26, 2017, Workshop Track Proceedings. OpenReview. net, 2017
53. Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in Proc. of Interspeech, 2017
54. D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984
55. A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609. 03499, 2016
56. R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, “Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis,” arXiv preprint arXiv:2011. 03568, 2020

57. A. Peiró-Lilja and M. Farrús, “Naturalness Enhancement with Linguistic Information in End-to-End TTS Using Unsupervised Parallel Encoding,” in Proc. Interspeech 2020, 2020, pp. 3994–3998
58. J. Taylor and K. Richmond, “Enhancing Sequence-to-Sequence Text-to-Speech with Morphology,” in Proc. Interspeech 2020, 2020, pp. 1738–1742
59. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779–4783
60. H. Tachibana, K. Uenoyama, and S. Aihara, “Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4784–4788
61. W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: 2000-speaker neural text-to-speech,” Proc. ICLR, pp. 214–217, 2018
62. W. Ping, K. Peng, and J. Chen, “Clarinet: Parallel wave generation in end-to-end text-to-speech,” arXiv preprint arXiv:1807.07281, 2018
63. K. Peng, W. Ping, Z. Song, and K. Zhao, “Non-autoregressive neural text-to-speech,” in Proceedings of the 37th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds. , vol. 119. Virtual: PMLR, 13–18 Jul 2020, pp. 7586–7598
64. C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, “DurIAN: Duration Informed Attention Network for Speech Synthesis,” in Proc. Interspeech 2020, 2020, pp. 2027–2031
65. J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, “Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling,” 2020
66. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, pp. 5998–6008, 2017

67. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6706–6713
68. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. -Y. Liu, “Fastspeech: Fast, robust and controllable text to speech,” in Advances in Neural Information Processing Systems, 2019, pp. 3171–3180
69. Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T. -Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text-to-speech,” arXiv preprint arXiv:2006. 04558, 2020
70. Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, “Align tts: Efficient feed-forward text-to-speech system without explicit alignment,” in ICASSP2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6714–6718
71. D. Lim, W. Jang, G. O, H. Park, B. Kim, and J. Yoon, “JDI-T: Jointly Trained Duration Informed Transformer for Text-To-Speech without Explicit Alignment,” in Proc. Interspeech 2020, 2020, pp. 4004–4008
72. M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, “Multispeech: Multi-speaker text to speech with transformer,” arXiv preprint arXiv:2006. 04664, 2020
73. H. R. Ihm, J. Y. Lee, B. J. Choi, S. J. Cheon, and N. S. Kim, “Reformer-TTS: Neural Speech Synthesis with Reformer Network,” Proc. Interspeech 2020, pp. 2012–2016, 2020
74. W.-N. Hsu, Y. Zhang, and J. Glass, “Learning latent representations for speech generation and transformation,” arXiv preprint arXiv:1704. 04222, 2017
75. Y. -J. Zhang, S. Pan, L. He, and Z. -H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in ICASSP2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6945–6949
76. G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, “Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6699–6703
77. Y. Yasuda, X. Wang, and J. Yamagishi, “End-to-End Text-to-Speech using Latent Duration based on VQ-VAE,” arXiv preprint arXiv:2010. 09602, 2020

78. V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, “Using VAEs and Normalizing Flows for One-Shot Text-To-Speech Synthesis of Expressive Speech,” in ICASSP2020-2020IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6179–6183
79. M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, “High fidelity speech synthesis with adversarial networks,” arXiv preprint arXiv:1909. 11646, 2019
80. H. Guo, F. K. Soong, L. He, and L. Xie, “A new GAN-based end-to-end TTS training algorithm,” arXiv preprint arXiv:1904. 04775, 2019
81. D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” arXiv preprint arXiv:1505. 05770, 2015
82. R. Valle, K. Shih, R. Prenger, and B. Catanzaro, “Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis,” arXiv preprint arXiv:2005. 05957, 2020
83. J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search,” arXiv preprint arXiv:2005. 11129, 2020
84. C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, “Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow,” in 2020IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7209–7213
85. P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, “A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction,” in Proc. 10th ISCA Speech Synthesis Workshop, 2019, pp. 7–12
86. Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, “FFTNet: A Real-Time Speaker-Dependent Neural Vocoder,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2251–2255
87. A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg et al. , “Parallel WaveNet: Fast high-fidelity speech synthesis,” in International conference on machine learning. PMLR, 2018, pp. 3918–3926
88. N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” arXiv preprint arXiv:1802. 08435, 2018

89. J. -M. Valin and J. Skoglund, “LPCNet: Improving neural speech synthesis through linear prediction,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5891–5895
90. K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “MelGAN: Generative adversarial networks for conditional waveform synthesis,” in Advances in Neural Information Processing Systems, 2019, pp. 14 910–14 921
91. L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, “GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram,” arXiv preprint arXiv:1904. 03976, 2019
92. R. Yamamoto, E. Song, and J. -M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6199–6203
93. J. Yang, J. Lee, Y. Kim, H. -Y. Cho, and I. Kim, “VocGAN: A High-Fidelity Real-Time Vocoder with a Hierarchically-Nested Adversarial Network,” in Proc. Interspeech 2020, 2020, pp. 200–204
94. S. Kim, S. -g. Lee, J. Song, J. Kim, and S. Yoon, “FloWaveNet: A generative flow for raw audio,” arXiv preprint arXiv:1811. 02155, 2018
95. R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A flow-based generative network for speech synthesis,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3617–3621
96. W. Ping, K. Peng, K. Zhao, and Z. Song, “WaveFlow: A compact flow-based model for raw audio,” in International Conference on Machine Learning. PMLR, 2020, pp. 7706–7716
97. P. chun Hsu and H. yi Lee, “WG-WaveNet: Real-Time High-Fidelity Speech Synthesis Without GPU,” in Proc. Interspeech, 2020, pp. 210–214
98. W. Song, G. Xu, Z. Zhang, C. Zhang, X. He, and B. Zhou, “Efficient WaveGlow: An Improved WaveGlow Vocoder with Enhanced Speed,” in Proc. Interspeech, 2020, pp. 225–229

99. Z. Zeng, J. Wang, N. Cheng, and J. Xiao, “MelGlow: Efficient Waveform Generative Network Based on Location-Variable Convolution,” arXiv preprint arXiv:2012. 01684, 2020
100. B. Zhai, T. Gao, F. Xue, D. Rothchild, B. Wu, J. E. Gonzalez, and K. Keutzer, “SqueezeWave: Extremely Lightweight Vocoders for On-device Speech Synthesis,” arXiv preprint arXiv:2001. 05685, 2020
101. M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz, and V. Mapelli, “The META-SHARE Metadata Schema for the Description of Language Resources.” in LREC, 2012, pp. 1090–1097
102. H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” arXiv preprint arXiv:1904. 02882, 2019
103. A. W. Black, “CMU Wilderness Multilingual Speech Dataset,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5971–5975
104. A. Stan, “Recoapy: Data recording, pre-processing and phonetic transcription for end-to-end speech-based applications,” arXiv preprint arXiv:2009. 05493, 2020
105. M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, “Open source voice creation toolkit for the MARY TTS Platform,” in Proc. of Interspeech, 2011
106. B. Potard, M. P. Aylett, D. A. Baude, and P. Motlicek, “Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN.” in Proc. of Interspeech, 2016, pp. 2293–2297
107. Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “Voiceloop: Voice fitting and synthesis via a phonological loop,” arXiv preprint arXiv:1707. 06588, 2017
108. R. Valle, J. Li, R. Prenger, and B. Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6189–6193
109. T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7654–7658

110. Y. Choi, Y. Jung, and H. Kim, “Deep MOS Predictor for Synthetic Speech Using Cluster-Based Modeling,” in Proc. Interspeech 2020, 2020, pp. 1743–1747
111. M. Wester, C. Valentini-Botinhao, and G. E. Henter, “Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations,” in Proc. Interspeech, Dresden, September 2015, pp. 3476–3480
112. M. Schoeffler, S. Bartoschek, F. -R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “webMUSHRA — A Comprehensive Framework for Web-based Listening Tests,” Journal of Open Research Software, vol. 6, no. 1, p. 8, Feb. 2018, number: 1 Publisher: Ubiquity Press

CHAPTER 11

Voice Quality Modelling for Expressive Speech Synthesis

Carlos Monzo¹, Ignasi Iriondo², and Joan Claudi Socoró²

¹Computer Science, Multimedia and Telecommunication Studies, Universitat Oberta de Catalunya (UOC), Rambla del Poblenou 156, 08018 Barcelona, Spain

²Grup de Recerca en Tecnologies Mèdia (GTM), Universitat Ramon Llull, La Salle, Quatre Camins 2, 08022 Barcelona, Spain

ABSTRACT

This paper presents the perceptual experiments that were carried out in order to validate the methodology of transforming expressive speech styles using voice quality (VoQ) parameters modelling, along with the well-known prosody (F_0 , duration, and energy), from a neutral style into a number of expressive ones. The main goal was to validate the usefulness of VoQ in

Citation: C. Monzo, I. Iriondo, J.C. Socoró, “Voice Quality Modelling for Expressive Speech Synthesis”, The Scientific World Journal, vol. 2014, Article ID 627189, 12 pages, 2014. <https://doi.org/10.1155/2014/627189>.

Copyright: © 2014 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

the enhancement of expressive synthetic speech in terms of speech quality and style identification. A harmonic plus noise model (HNM) was used to modify VoQ and prosodic parameters that were extracted from an expressive speech corpus. Perception test results indicated the improvement of obtained expressive speech styles using VoQ modelling along with prosodic characteristics.

INTRODUCTION

The research fields of automatic speech recognition (ASR) and text-to-speech (TTS) synthesis benefit from expressive speech, that is, speech with emotional content being this more spontaneous, to make human-machine interactions more natural, for example, in terms of emotion recognition [1, 2] and voice transformation [3–5]. Voice quality (henceforth VoQ) and prosody parameters (F_0 , duration, and energy) can be conveniently manipulated to represent or convey the emotional content of speech in ASR or TTS applications respectively [1, 3, 6–10]. In spite of the fact that VoQ has been less explored than prosody, recent works propose using both types of data to improve the acoustic modelling of expressive speech [7–10]. Other studies relate perceived speech features in emotional speech to VoQ parameters [11] and deal with the association of phonation type (e.g., whispery voice) and affective speaking [12, 13].

In recent years, increasing interest has been focused on the harmonic plus noise model (HNM) [14, 15] for speech transformation [5] because high quality and versatility can be achieved. The parameterisation of speech in both harmonic and stochastic components allows for flexible manipulation of VoQ over time and pitch scales, making it possible to maintain a high degree of natural speech quality.

With the improvement in the emotional content representation of speech and the availability of improved accuracy techniques for its analysis and synthesis, interest in expressive speech synthesis (ESS) has grown [3, 6, 10, 16–21]. Along with the generation of expressive speech, it is necessary to evaluate the different existing methodologies; there is no consensus about which is the better one [22]: perceptual assessment tests with forced choices, perceptual assessment tests with free responses, or perceptual impact tests.

This paper presents the perceptual assessment carried out to evaluate the proposed expressive speech styles transformation methodology, based on prosody and VoQ modelling using an HNM for the speech analysis

and synthesis. Prosody and VoQ modelling was conducted from a Spanish expressive speech corpus with five expressive styles: neutral (NEU), happy (HAP), sensual (SEN), aggressive (AGG), and sad (SAD). This perceptual assessment was performed by means of two evaluations. The first one was used for quality evaluation, and the second one was used for the assessment of the expressive styles identification.

A forced-choice test with five possible answers was performed to evaluate both the perceived quality and the identification of expressive style of the utterances.

The rest of this paper is organised as follows: Section 2 shows the speech material that was used, an expressive speech Spanish corpus, and describes the corpus design, its labelling, and its subjective evaluation. Section 3 deals with the expressive speech style transformation methodology, the HNM description, and the prosody and VoQ modelling. Section 4 presents the perceptual assessment for the proposed transformation methodology and a discussion of the results. Finally, Section 5 contains the conclusions and an outline of future work.

SPEECH MATERIAL

The speech material used during the expressive speech style transformation experiments was an expressive speech corpus devoted to ESS in Spanish, developed with a twofold purpose: first, to be used for the acoustic modelling of emotional speech (prosody and VoQ) and, second, to be the speech unit database for our speech synthesiser.

Corpus Design

For the corpus design, we sought the help of experts in audiovisual communication from the Laboratory of Instrumental Analysis of the Autonomous University of Barcelona (LAICOM-UAB). Due to the LAICOM-UAB experience in advertising, a large textual database of advertisements, extracted from newspapers and magazines, was available.

Moreover, this database was organized in different thematic categories: new technologies, education, cosmetics, automobile industry, and travels. According to the LAICOM-UAB experts, speech styles can be more easily defined according to the sentences' features for each one of these five categories [23], allowing the creation of an expressive oral corpus with good coverage of simulated expressive speech styles. The texts for each expressive

style were read by a professional female speaker in different recording sessions (stimulated speech). It was assumed that stimulated speech methodology, validated by [24], diminished the possibility of modelling informal spontaneous speech utterances while guaranteeing control of the recording conditions, the style definition, and the text design.

It is important to mention that the speaker had previously received training in the vocal patterns of each style. The phonetic features (segmental and suprasegmental) for these vocal patterns were defined by the experts of LAICOM-UAB. The use of texts from an advertising category aimed to help the speaker to maintain the desired style through the whole recording session. Therefore, the intended style was not performed according to the speaker's criteria for each sentence, but all the utterances of the same style were consecutively recorded in the same session following the previously learned pattern. Thus, the speaker was able to keep the required expressiveness even with texts whose semantic content was not coherent with the style. Moreover, LAICOM-UAB expert supervision was required through the recording in order to avoid possible deviations from the predefined style.

Five subject categories, selected from the advertising corpus, were assigned to the expressive speech styles in the following manner:

- (i) *new technologies*: a neutral style (NEU) that transmits a certain maturity,
- (ii) *education*: a happy style (HAP) that generates a feeling of extroversion,
- (iii) *cosmetics*: a sensual style (SEN) based on a sweet voice,
- (iv) *automobiles*: an aggressive style (AGG) that transmits hardness,
- (v) *travel*: a sad style (SAD) that seeks to express melancholy.

A set of phrases for each category was selected by means of a greedy algorithm [25], which made it possible to select phonetically balanced sentences from each subcorpus. To optimise the selection process, the required phonemes were sorted according to the occurrence rate presented by [26], which allowed the greedy algorithm to start by selecting sentences that contained less probable phonemes. Moreover, the selection of sentences similar to those previously selected was penalised by the greedy algorithm. Finally, the size of the corpus for each recorded expressive speech style is shown in Table 1.

Table 1. Number of sentences and duration, per expressive speech style, for the expressive corpus

| | Number of sentences | Duration (min) |
|-----|---------------------|----------------|
| NEU | 833 | 50 |
| HAP | 916 | 56 |
| SEN | 841 | 51 |
| AGG | 1048 | 84 |
| SAD | 1000 | 86 |

Speech Labelling

The speech was labelled using segmentation and pitch marking. Segmentation is the identification of the temporal boundaries for each phoneme, and the pitch marks identify each period in the voice parts of the speech. Segmentation is related to segmental duration, whereas pitch marking is related to pitch or fundamental frequency (F_0).

The alignment of the different phonemes, or segmentation, was carried out by means of forced alignment using the HTK (<http://htk.eng.cam.ac.uk/>) tool and the available phonetic transcription. The resulting segmentation is used in the extraction of acoustic segmental features related to both prosody and VoQ and used for recognition and synthesis purposes.

The pitch marking was based on the Robust algorithm for pitch tracking (RAPT) of [27] and the application of the pitch marks filtering algorithm (PMFA) developed by [28], which improved the robustness of the final pitch marks.

Corpus Subjective Evaluation

A forced answer test was designed with the question “*What emotion do you recognize from the voice of the speaker in this utterance?*” Thus, the expressive speech corpus was evaluated using a subjective test, presenting a subset of 240 utterances to 25 listeners, that produced the confusion matrix [29] presented in Table 2. The possible answers were the five styles of the corpus (see Section 2.1) plus the additional option of do not know/another (Dk/A) to avoid biasing the results in the case of confusion or doubts between two options. The risk of adding this option is that some evaluators

may use it excessively to accelerate the test [30]. However, this effect was negligible in this test [9].

Table 2. Average confusion matrix (%) for the subjective test (the maximum correct classification value is in bold)

| (%) | NEU | HAP | SEN | AGG | SAD | Dk/A |
|-----|-------------|-------------|-------------|-------------|-------------|------|
| NEU | 86.4 | 1.3 | 3.6 | 5.3 | 0.7 | 2.7 |
| HAP | 1.9 | 81.0 | 0.2 | 15.6 | 0.1 | 1.2 |
| SEN | 4.7 | 0.1 | 86.8 | 0.0 | 5.7 | 2.6 |
| AGG | 1.8 | 14.2 | 0.1 | 82.7 | 0.1 | 1.1 |
| SAD | 0.5 | 0.0 | 0.6 | 0.0 | 98.8 | 0.1 |

As a general result, the subjective test shows that all the expressive styles achieve a high percentage of identification (87.1% on average). SAD was the most highly rated (98.8%), followed by SEN (86.8%) and NEU (86.4%) styles, and finally AGG (82.7%) and HAP (81%). The confusion matrix shown in Table 2 reflects the misclassifications. It reveals that the main errors are produced in AGG (14.2% identified as HAP) and HAP (15.6% identified as AGG). Moreover, NEU is slightly confused with all the options and there is a certain level of confusion of SEN with SAD (5.7%) and NEU (4.7%). The Dk/A option was hardly used, although it was more present in NEU and SEN than in the rest of the styles. Detailed information about the subjective evaluation is presented by [9].

EXPRESSIVE SPEECH STYLE TRANSFORMATION

This section presents the expressive speech style transformation proposal. First of all, the harmonic plus noise model (HNM) is shown as the main processing technique used for speech analysis and synthesis. Secondly, prosody and VoQ speech parameters involved during transformations are described. Finally, the transformation methodology using HNM together with prosody and VoQ parameterisation is presented.

Harmonic Plus Noise Model (HNM)

Harmonic plus noise model (HNM) allows modification of the speech prosody to generate a high quality signal. In this work, we try to exploit the full capabilities of the HNM not only for prosody changes but also introducing modifications in the spectral content through VoQ parameters.

In HNM-based speech parameterisation, the voice signal (n) can be expressed as the addition of a deterministic or harmonic component $s(n)$ and a stochastic or noise component $r(n)$ [14] (see (1)). The implementation used in this paper was the pitch synchronous development carried out by [31]:

$$x(n) = s(n) + r(n). \quad (1)$$

The lower spectral band ([0, 5000] Hz) was mainly modelled as the addition of harmonically related sinusoids ((n)) that characterised the voiced part of the speech. For each signal frame, the deterministic part is represented by the amplitudes, frequencies, and phases of the corresponding harmonics and the analysis time instants. The number of harmonics depends on the pitch (or F_0) value at each frame. The harmonic part was synthesised through overlap-add technique using triangular windows.

Unvoiced sounds and nonperiodic speech events were modelled by the stochastic component ((n)). This was carried out by an autoregressive (AR) model in which both the spectral and temporal fluctuations were represented by constant-frame-rate, time-varying Q -order linear predictive coding (LPC) coefficients and variances.

With regard to the HNM analysis process, the harmonic component estimation was based on [32] algorithm for spectral peak extraction in the frequency domain, incorporating a harmonicity constraint into the frequency-based cost function by using the Lagrange multipliers optimisation method to guarantee the harmonicity of the estimated frequencies [31].

Finally, in order to perform the expressive speech style transformation process during the experiments, the prosody and VoQ models were applied using the deterministic and stochastic components. The modification was carried out by manipulating frequencies, amplitudes, phases, and noise component power according to the target prosody and VoQ model requirements.

Prosody Parameters

The prosody parameters involved during the transformation experiments were pitch or fundamental frequency (F_0), unit duration, and unit energy. These were extracted from the expressive speech corpus presented in Section 2 and modelled using a case-based reasoning (CBR) system. CBR is a useful data mining technique in the context of ESS [21] that returns the

case that best fits the target requirements. In this way, expressive speech transformation can be enhanced using a CBR system matched to a specific expressive corpus. CBR is based on the creation of a database with different situations or cases which appeared in the corpus (memory of cases). First of all, it is necessary to identify the attributes (or features) that define the cases for the prediction of phone duration, phone energy, and the intonation contour. Then, the training set is generated by joining the prosodic parameters annotated in the speech corpus with the prosodic features extracted from the linguistic analysis of the text. A reduction of possible cases is achieved through clustering of the classes that are represented by the same attributes. The final aim of CBR is to map a solution from previous cases to the target problem. Thus, the most similar case is recovered from the database using the Minkowski metric to the selected attributes. Therefore, given the input text, the predicted parameters are fundamental frequency contour (F_0), energy contour, and segmental duration.

The automatic extraction of prosodic features from text was achieved by means of the linguistic analysis tool proposed by [21] that carries out the phonetic transcription of the text using SAMPA phonetic alphabet [33], annotating intonation groups (IG), stress groups (SG), words, and syllables. Regarding IG and SG, an IG in Spanish is defined as a structure of coherent intonation that does not include any major prosodic break, and an SG is defined as a stressed word preceded, if appearing, by one or more unstressed words. With regard to prosodic breaks, they take place due to pauses or significant inflections of the F_0 contour. In terms of segmental duration and energy modelling, the phone was chosen as the basic acoustic unit, and its duration depends basically on its identity and the context where it is placed, just as occurs for energy in a similar way. Finally, SG was chosen for the F_0 contour modelling, incorporating the influence of the syllable and the pitch structure at IG level by means of the concatenation of SG contours.

Voice Quality Parameters

The VoQ parameters involved in the transformation experiments were already used and analysed in previous studies in which their usefulness for expressive speech discrimination was demonstrated [9, 10, 13, 34]. Thus, the following subset of VoQ parameters was considered for the expressive speech styles transformation experiment presented in this work.

- (i) *Jitter* and *shimmer* describe the cycle-to-cycle variations of the fundamental period (inverse of the first harmonic's frequency)

and the waveform amplitude, describing frequency and amplitude modulation noise, respectively. These parameter definitions were modified from the methodology used in tools like Praat [35] or the multidimensional voice program (MDVP) (http://www.kayelemetrics.com/index.php?option=com_product&Itemid=3&controller=product&task=learn_more&cid=56) in order to cancel prosodic interference in their measurements [34].

- (ii) *Harmonic-to-noise ratio* (HNR) describes the energy ratio between the HNM harmonic and stochastic components. The harmonic part energy is computed by the sum of the squared amplitudes of all the harmonics, while, for the stochastic part, the energy depends directly on the noise variance.
- (iii) *Hammarberg index* (Hamml)) is defined as the ratio between the maximum energy in the 0–2000-Hz and the 2000–5000-Hz frequency bands. Then, this parameter is computed with the maximum squared amplitude of the harmonics on each respective band of the HNM harmonic component.
- (iv) *Relative amount of energy* in the high- (above 1000 Hz) versus the low-frequency range of the voiced spectrum (pe1000): this parameter is computed with the squared amplitudes of the harmonics within high and low frequency bands of the HNM harmonic component.

Transformation Methodology

This section describes the proposed expressive speech transformation methodology. It is based on the previous work carried out by [10], where an initial strategy for the modification of VoQ together with prosody was proposed, using HNM for speech analysis and synthesis, to improve the perception of the transformed speech. Despite the benefits of using VoQ in combination with prosody reported in that work, a deeper analysis of the transformation methodology was considered necessary from two points of view: (i) to evaluate the identification rate improvement of the resulting transformed expressive speech and (ii) to analyse the resulting speech quality.

Initial experiments were conducted using Pitch synchronous overlap and add- (PSOLA-) based TTS to perform the VoQ modifications [34]. This algorithm is simple and straightforward when pitch, energy, and duration are modified, but some problems arise when spectral-based VoQ parameters

need to be modified. Therefore, for the new experiments, the HNM was chosen as the tool for modifying and synthesising speech signals because of its flexibility, allowing such spectral modifications.

As shown in Figure 1, the proposed block diagram for the expressive speech styles transformation methodology is divided into three main parts. First, HNM analysis and resynthesis blocks extract the original speech information and regenerate it when the HNM parameter transformation is performed. Second, prosody and VoQ are predicted through the use of a prosody and VoQ models, respectively. Prosody is predicted by means of CBR by obtaining the target information for each phoneme: F_0 contour, energy contour, and segmental duration. The VoQ is modelled by using transformation rules, extracted from the analysis of VoQ parameters in the expressive speech corpus presented in Section 2, by means of mean (μ) and standard deviation (σ) parameters manipulation. Finally, the speech transformation is carried out, based on the results of the HNM analysis and the prosody and VoQ models.

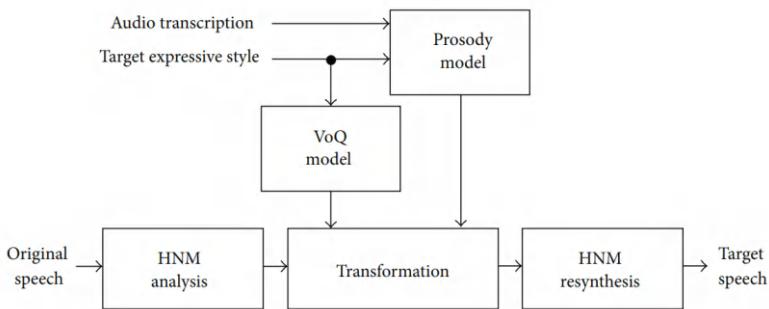


Figure 1. A block diagram for the proposed expressive speech styles transformation methodology.

Several considerations must be made to determine which parameters and their values should be involved in the transformations. For the prosody modifications, all of the parameters were involved in all of the transformations. However, the selection of the VoQ parameters and the corresponding values to be used during the transformation was based on the work of [2, 10, 13, 34] in which the following aspects were considered:(1)the results of previous studies about the use of VoQ parameters in the discrimination of expressive speech styles,(2)an exhaustive classification experiment to obtain different configurations for all parameters and expressive styles,(3) descriptive statistics calculated for all expressive styles of the corpus and

all involved VoQ parameters. Prosody modifications use the information about the original utterance and about the target from CBR predictions (see Section 3.2), so a multiplicative transformation factors among original and target parameters values were calculated and the modifications were performed. The modification of segmental durations and F_0 was carried out on the HNM parameters according to the work of [31]. Nevertheless, energy modification was performed directly on the utterance audio samples by multiplying each sample by the corresponding multiplicative transformation factor [21].

With regard to VoQ parameter modification, both means and standard deviations of parameters values, predicted from the corpora presented in Section 2, were modified according to (2), obtaining the target VoQ value (VoQ_t). For a given VoQ parameter, the mean (μ) modification was carried out by means of the original mean (μ_o) subtraction (corresponding with the mean value for the expressive speech style) from the original parameter value (VoQ_o) and, finally, the target mean (μ_t) for this parameter was added (corresponding with the mean of the target expressive speech style). Regarding the standard deviation (σ), a multiplicative transformation factor per VoQ parameter and target expressive style, calculated as the ratio between the target (σ_t) and the original standard deviation (σ_o), was used in order to vary the intensity of the current parameter. In order to obtain more robust measurements, following the proposals of [3, 36], only vowels were considered in these computations. This proposal for the transformation methodology will let us evaluate the usefulness of combining VoQ together with prosody with the aim of improving the obtained expressive speech style identification rate maintaining an acceptable speech quality:

$$\text{VoQ}_t = \frac{\sigma_t}{\sigma_o} \cdot (\text{VoQ}_o - \mu_o) + \mu_t. \quad (2)$$

The target VoQ values were obtained applying the presented transformation to the original VoQ parameters values frame-by-frame. This VoQ parameter modification using the HNM, performed according to the work of [37], is described below.

(i)*Jitter*: only the frequencies for the HNM harmonic component are modified. Once the F_0 curve is obtained from the CBR prosody prediction module, slow F_0 variations are removed to avoid interference due to prosodic information, and the new F_0 microprosody variations related to jitter are applied. New jitter variance is obtained by means of the presented

transformation methodology, and the final pitch curve is computed adding the new jitter to the previously extracted slow F_0 variations [34].

(ii) *Shimmer*: the modification of this parameter is directly applied to the time-domain waveform. The same process used for jitter modification has been applied to modify the shimmer. However, pitch synchronous peak-to-peak amplitude variations curve is used instead of F_0 contour information [34].

(iii) *HNR*: multiplicative transformation factors, calculated as the ratio between target and original HNR values, are applied in the HNM harmonic and stochastic components to guarantee the desired energy ratio and the total energy after the transformation. For each signal frame, the multiplicative transformation factor in the harmonic part is the same for all harmonic amplitudes, and, in the stochastic part, it affects the noise variance. An additional energy correction factor for both components is finally applied to maintain the original frame energy in the transformed signal.

(iv) *HammI*: only the maximum harmonic amplitude of each frequency band (the 0–2000-Hz and the 2000–5000-Hz frequency bands) in the HNM harmonic component is modified according to the target parameter value (using a transformation factor measured as the quotient between the target and original HammI values). An additional energy correction factor, the same for each frequency band, maintains the original frame energy during the transformation. The HNM stochastic component is not manipulated.

(v) *pe1000*: using the corresponding multiplicative transformation factor calculated as the relation between target and original pe1000 values, the ratio between the HNM harmonic component energy of the [0, 1000] Hz and [1000, 5000] Hz frequency bands is modified. A multiplicative constant factor, specific for each band, is applied to the harmonic amplitude values without any manipulation of the HNM stochastic component. The same global energy normalization procedure used in previous parameters is finally carried out.

Due to the tight relation among VoQ parameters, a change in a parameter during the overall modification can affect another one, especially when they model similar lowlevel signal features (i.e., spectral bands). For example, a change in HammI can produce a variation in pe1000 parameter as they measure energies in the same spectral bands. In order to minimize the impact among them, and taking into account the modification needs for each one, the following order modification was proposed by [37]: (1) Jitter, (2) HNR, (3) pe1000, (4) HammI, and (5) Shimmer.

Table 3 presents the VoQ parameters involved in each transformation from neutral to the target expressive style. Nevertheless, as was said previously, all parameters were not involved during all the transformations, since the best parameters identifying to each expressive speech style were found out from previous conducted work in discriminative analysis, expressive speech styles classification using VoQ, and descriptive statistics of corpus. Moreover, Table 4 shows mean (μ) and standard deviation (σ) values for all VoQ parameters extracted from the expressive speech corpus, used during the VoQ transformation to calculate the target VoQ parameters.

Table 3. Voice quality selected parameters during neutral-target transformations (“•” when the parameter is selected and “—” otherwise)

| | Jitter | Shimmer | HNR | HammI | pe1000 |
|-----|--------|---------|-----|-------|--------|
| HAP | — | — | — | • | • |
| SEN | • | • | • | • | • |
| AGG | • | • | — | • | • |
| SAD | • | • | — | • | • |

Table 4. Mean (μ) and standard deviation (σ) values for all VoQ parameters extracted from the expressive speech corpus

| μ/σ | Jitter | Shimmer | HNR | HammI | pe1000 |
|--------------|-----------|-----------|------------|------------|-------------|
| NEU | 0.21/0.36 | 0.06/0.19 | 25.68/4.28 | 28.37/7.68 | -14.35/6.41 |
| HAP | 0.24/0.30 | 0.03/0.06 | 24.06/5.26 | 25.12/7.81 | -9.29/8.96 |
| SEN | 0.49/0.74 | 0.11/0.22 | 24.30/6.09 | 30.98/8.08 | -16.08/7.58 |
| AGG | 0.14/0.12 | 0.03/0.06 | 23.45/4.45 | 23.69/7.31 | -6.15/8.80 |
| SAD | 0.10/0.08 | 0.12/0.20 | 29.38/6.58 | 35.77/8.78 | -16.97/8.80 |

Different conclusions can be extracted from Tables 3 and 4. First, for all transformations, the HammI and pe1000 parameters were used, controlling the tension effect in the voice, showing a phonation effort or relaxation. For example, HAP and AGG styles present values for these parameters that show high energy in high frequency band, thus producing a higher perceived vocal effort in the final speech. Nevertheless, for SEN and SAD styles, the presented speech is more relaxed. Second, jitter and shimmer parameters let us control the quivering voice effect, and so their use is more remarkable in SEN and SAD styles. Finally, the control of the amount of noise which appeared in the speech is carried out by means of HNR parameter, useful during SEN style identification.

To sum up, this prosody and VoQ transformation methodology entails modifying the HNM parameters, that is, frequencies, amplitudes, and phases for the harmonic component and variance for the stochastic component.

Therefore, two kinds of transformations were carried out. The first one, the prosody modification, was guided through the CBR system so that it affected only the prosody; thus, the frequencies, amplitudes, and phases of the HNM were modified to produce the required F_0 contour, energy contour, and segmental duration. The second one, the VoQ parameter modification, also modified the HNM parameters, but in the specific way previously described and briefly summarized below.

The jitter parameter also controls the F_0 contour; this way, the frequencies, amplitudes, and phases were also affected because vocal tract observations are highly related to the pitch frequency. In the shimmer parameter case, all the modifications were carried out directly on the time-domain speech signal [34]. The HammI and pe1000 parameters are related to ratios or approximations of energy in different frequency bands of the harmonic component. Therefore, the HNM amplitude vector was modified. To vary the HNR parameter, both HNM harmonic and stochastic components were modified by manipulating the harmonic amplitudes and stochastic variances, respectively. For HammI, pe1000, and HNR parameters, the variation multiplicative transformation factor was distributed between two spectral bands (in the HNM harmonic component for HammI and pe1000) or both components (the HNM harmonic and stochastic components for HNR), ensuring that speech energy was preserved when the transformation was performed.

PERCEPTUAL ASSESSMENT

In the work of [10], the utility of using a combination of prosody and VoQ for ESS was demonstrated by means of a comparison mean opinion score (CMOS) [38] perceptual test. However, the quality of the final transformed expressive speech and, especially, the identification rate for each of these styles were not studied. In this section, how the parameter modification affects the the quality of the generated speech and the identification rate obtained for each target style are analysed. Thus, the transformations and the proposed methodology are validated.

Experiment Description

With the aim of analysing the effectiveness of the proposed transformation, a comparison between modifying prosody and VoQ parameters using the HNM technique was conducted in the following experiments, in which the quality of the generated speech and the target expressive styles identification

rate were evaluated (see Sections 4.2 and 4.3, resp.). The results were compared by taking into account different configurations of parameters to be transformed: modifying only prosody, prosody plus jitter and shimmer, and, finally, prosody plus the combination of VoQ parameters for each expressive style. Each of the configurations under testing is described next, indicating the name that identified them during the evaluations.(i) “*Natural*”: natural speech. A set of utterances was directly extracted from the corpus for each expressive style.(ii) “*ResHNM*”: HNM-based direct resynthesis of the natural utterances for each expressive style. The process of analysis and synthesis was carried out from corpus examples without applying any modification to the speech parameters.(iii) “*HNMPro*”: prosody transformation based on the HNM. The utterances, originally expressed in a neutral expressive style, were transformed using the prosody models.(iv) “*HNMProJiSh*”: transformation of prosody, jitter, and shimmer, based on the HNM. These parameters were transformed from the prosody models and by using the transformation values for jitter and shimmer, learned from the expressive corpus to transform the utterances originally expressed in a neutral expressive style. This configuration is interesting because this transformation can be conducted by means of both HNM-based and other synthesis algorithms-based TTS (e.g., PSOLA).(v) “*HNMProVoQ*”: prosody and full VoQ modification based on the HNM. The utterances, originally expressed in a neutral expressive style, were transformed using the prosody models and the VoQ parameter configurations (see Section 3.4) into the target expressive style under testing.

First, the speech quality was evaluated using utterances transformed from a neutral subcorpus into happy (HAP), sensual (SEN), aggressive (AGG), and sad (SAD) expressive styles. The evaluation was carried out by means of a mean opinion square (MOS) test [38] with five possible answers, with a score between 1 and 5, in which 5 was the maximum quality and 1 was the minimum: “Excellent” (5), “Good” (4), “Fair” (3), “Poor” (2), and “Bad” (1). The results analysis was performed by grouping the five kinds of configurations into two sets: (1)natural speech (“Natural”) and resynthesised speech (“ResHNM”), which had the maximum reference values because they were real cases (natural) and also represented the best possible results that a TTS system could obtain using the HNM algorithm (resynthesis). (2)The proposed transformation methodology with the configuration of interest (“HNMProVoQ”) was compared with the rest of the configurations. The different configurations for the transformation were (i) only prosody

modification (“HNMP”), (ii) a combination of prosody and only jitter and shimmer VoQ parameters (“HNMPRoJiSh”), and (iii) a combination of prosody and the selected VoQ parameter for each transformation (“HNMPVoQ”). Hence, the quality changes of the HNM algorithm were evaluated.

Second, we performed the target expressive speech style identification assessment, destined to validate the proposed expressive speech transformation methodology.

This was carried out through a test with 5 possible answers, with 4 of them corresponding to the target expressive styles (happy, sensual, aggressive, and sad) and a fifth category corresponding to “Others.” This fifth category was created without being assigned to any particular style to avoid an opinion bias towards the rest of the options when the answer was not clear or when no evaluated style was perceived. In an analogous way, the configuration of interest “HNMPVoQ” was also compared with the rest of the possible options in order to point out the possible differences in the modified speech parameters (prosody and VoQ).

The comparison of prosody and VoQ parameter modification strategies is interesting from the point of view that the variations of quality for a certain identification rate of the expressive speech styles can be known, and two main situations can be considered.

On one hand, in the case of achieving high quality and a low identification rate, a situation can be determined in which it is necessary to improve the expressive styles models (this could be performed using more or different parameters). On the other hand, if the quality is low and a high identification rate is obtained, the speech generation algorithm will be taken into account by analysing whether it can support the desired parameter transformation and whether this transformation must be so demanding that it causes quality degradation (in this case, we could work on modifying the transformation needs).

The quality evaluation and the expressive speech styles identification test were carried out by answering two questions for the same presented utterance: “Assess the global audio quality” for the speech quality evaluation and “Indicate which expressive style is transmitted by the audio” to identify the transmitted expressive style. For these evaluations, 100 utterances were generated, containing the same number of examples for every configuration and expressive style. The total number of listeners was 17, with ages ranged

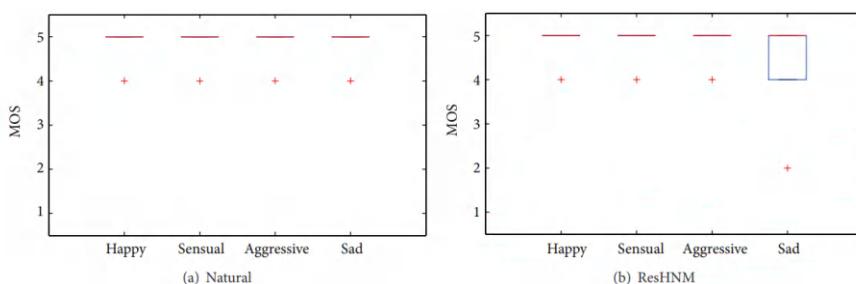
between 24 and 50 years old. There were 13 males and 4 females. Out of the group, 8 of them were experts on speech technologies.

With regard to the test utterances, they were selected according to the type of the conducted test and the applied configuration. For “Natural” and “ResHNM,” five utterances for each subcorpus were selected. Regarding their characteristics, a variety of intonation patterns were used by selecting declarative, interrogative, and exclamation expressions, with a mean duration of 4.5 seconds for NEU, 4 seconds for HAP, 3.4 seconds for SEN, 3 seconds for AGG, and 4.5 seconds for SAD. For the rest of tests, the same NEU utterances were transformed into the different expressive speech styles. Detailed information about the selected sentences is presented by [37].

Next, the obtained results and conclusions are presented for every configuration and expressive style. The results for the quality assessment are shown first (Section 4.2), and the subjective identification results are presented second (Section 4.3).

Subjective Quality Assessment

The boxplots [39] of Figures 2(a) and 2(b) show the quality assessment results for reference configurations: “Natural” (see Figure 2(a)) and “ResHNM” (see Figure 2(b)). The high quality of these utterances “Excellent” can be observed; only the sad style, with HNM-based resynthesis, presented certain result dispersion. These results are consistent with the configurations that were used; due to this, in general terms, HNM-based synthesis highly depends on the resulting speech parameterisation, so it is more affected by pitch marks than other TTS-based algorithms like PSOLA, as well as the HNM parameter estimation of the deterministic and stochastic components. For example, in the case of the sad style (see Figure 2(b)), its specific acoustic characteristics (e.g., tremulous voice) could lead to more analysis inaccuracies, causing more synthesis artefacts.



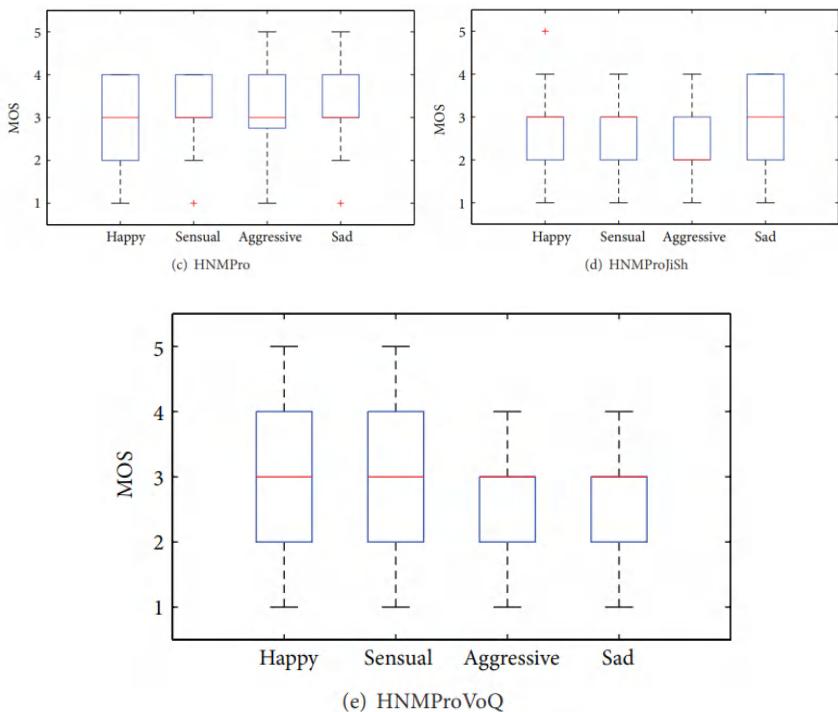


Figure 2. Quality MOS test results for the configurations of “Natural,” “ResHN-M,” “HNMPro,” “HNMProJiSh,” and “HNMProVoQ.”

Once the reference configurations were analysed (“Natural” and “HNMPro”), the results for the evaluated transformation-based configuration were studied (see Figures 2(c)–2(e)) with the final goal of going into a deeper analysis of the transformation methodology of interest (“HNMProVoQ”).

First, the quality values for “HNMPro” (see Figure 2(c)) is presented. A certain quality degradation due to the application of expressive prosody on the neutral utterances could be observed. Notice that the quality value is centred on “Fair.”

The resulting MOS values for the evaluation of “HNMProJiSh” configuration are analysed in Figure 2(d). The quality value for the HNM was maintained practically constant at the median (“Fair”), except for the aggressive case (“Poor”). Then, the HNM clearly becomes a good option in the transformation of expressive speech styles because it offers an acceptable final quality despite the parameters modification. Finally, we evaluated the quality obtained by the “HNMProVoQ” configuration (see Figure 2(e)) in

which the VoQ transformations were matched to the necessities of each target expressive speech style (see Table 3). Figure 2(e) shows how the quality is maintained between acceptable values (“Fair”), as occurred in the configuration of the HNM in which only the prosody transformation was involved (see Figure 2(c)). This fact shows that the HNM makes it possible to introduce VoQ transformations without decreasing the quality. It is also notable that the quality for the happy, sensual, and aggressive expressive styles increased with regard to the transformation of prosody together with only jitter and shimmer (see Figure 2(d)). There was a slight decrease in the sad style, which had already occurred in the reference case (see Figure 2(b)).

With these results, we can conclude that the VoQ transformations could be used to improve the identification rate of expressive styles (see Section 4.3) during the speech synthesis, maintaining the speech quality regarding the well-known prosody modelling. Nevertheless, excess signal manipulation can bring about negative effects too, decreasing the quality. For example, in the case of the sad expressive style, as it is shown in Section 4.3, the best interclass identification rate was achieved (see Table 5), although it was not the style with the best perceived quality (see Figure 2(e)).

Table 5. The confusion matrix (%) and F_1 measures in the expressive speech styles identification for the reference configurations (“Natural” and “ResHNM”) and HNM transformation configurations (“HNMPRo,” “HNMPRoJiSh,” and “HNMPRoVoQ”)

| (%) | HAP | SEN | AGG | SAD | Others |
|---------|------|-------|-------|------|--------|
| Natural | | | | | |
| HAP | 94.1 | 0.0 | 5.9 | 0.0 | 0.0 |
| SEN | 0.0 | 100.0 | 0.0 | 0.0 | 0.0 |
| AGG | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| SAD | 0.0 | 9.4 | 0.0 | 90.6 | 0.0 |
| (F1) | 0.97 | 0.96 | 0.97 | 0.95 | — |
| ResHNM | | | | | |
| HAP | 87.1 | 0.0 | 11.8 | 0.0 | 1.2 |
| SEN | 1.2 | 95.3 | 1.2 | 2.4 | 0.0 |
| AGG | 1.2 | 0.0 | 98.8 | 0.0 | 0.0 |
| SAD | 0.0 | 11.8 | 0.0 | 88.2 | 0.0 |
| (F1) | 0.92 | 0.92 | 0.93 | 0.93 | — |

| HNMPromo | | | | | |
|---------------|------|------|------|------|------|
| HAP | 30.6 | 3.5 | 17.6 | 28.2 | 20.0 |
| SEN | 1.2 | 31.8 | 9.4 | 40.0 | 17.6 |
| AGG | 35.3 | 1.2 | 21.2 | 23.5 | 18.8 |
| SAD | 2.4 | 18.8 | 10.6 | 41.2 | 27.1 |
| (F1) | 0.36 | 0.41 | 0.27 | 0.35 | — |
| HNMPromProjSh | | | | | |
| HAP | 30.6 | 3.5 | 16.5 | 27.1 | 22.4 |
| SEN | 3.5 | 30.6 | 4.7 | 36.5 | 24.7 |
| AGG | 32.9 | 2.4 | 18.8 | 22.4 | 23.5 |
| SAD | 4.7 | 17.6 | 1.2 | 58.8 | 17.6 |
| (F1) | 0.36 | 0.40 | 0.27 | 0.48 | — |
| HNMPromVoQ | | | | | |
| HAP | 34.1 | 3.5 | 25.9 | 14.1 | 22.4 |
| SEN | 0.0 | 40.0 | 5.9 | 34.1 | 20.0 |
| AGG | 23.5 | 2.4 | 31.8 | 16.5 | 25.9 |
| SAD | 3.5 | 20.0 | 1.2 | 64.7 | 10.6 |
| (F1) | 0.42 | 0.48 | 0.39 | 0.56 | — |

Subjective Expressive Speech Style Identification

Once the quality assessment results have been analysed and discussed, we will analyse the results that were obtained from the expressive speech style identification subjective test. In this case, the aim was to evaluate the identification degree of the synthesised style that was achieved using the proposed method, related to the main goal of this work of validating the usefulness of VoQ in the enhancement of expressive synthetic speech for style identification. We have to remember that the transformation was carried out from the neutral style into another style (happy, sensual, aggressive, and sad). In the performed test, listeners could choose any of these four expressive styles and the option “Others,” which avoided biasing the measure towards any of them.

The subjective identification results are presented by means of a confusion matrix (%) [29] and *F1* measures [40]. First, the confusion matrix informs us about how good the identification was, and above all, it lets us detect any existing confusion among the expressive styles. Second, the *F1* measure gives a more real and compact vision about how good the identification was, taking into account both the correct classified cases and the existing confusion among styles. This measure is the harmonic mean of precision and recall. For a studied style, the precision is defined as the number of cases correctly classified divided by the total number of cases

classified in that style. The recall is defined as the number of cases correctly classified divided by the total number of existing cases that should have been classified within that class.

As was done for the quality evaluation, first, the results obtained for natural speech (“Natural”) and HNM-based resynthesis (“ResHNM”) are shown (see Table 5). It was observed that both configurations produced similar results, as was expected. Some confusion existed between happy-aggressive and sensual-sad styles. The $F1$ values were greater than 0.92 for all of the expressive styles in both configurations.

The next step was analysing how the neutral style transformation into the target expressive styles affected the identification of the desired styles (see Table 5). First, we wanted to study the limitation of each transformation and the improvement of the perception of the expressive styles from the use of VoQ parameters when using only prosody, corresponding to the first transformation attempt from the prosody parameters (“HNMP” configuration). Second, the results for combining prosody together with VoQ are presented in the “HNMPRoJiSh” (prosody with jitter and shimmer) and “HNMPRoVoQ” (prosody and the proposed configuration for VoQ parameters) configurations. The improvement of $F1$ values using the interest configuration (“HNMPRoVoQ”), regarding the rest of HNM transformation configurations (“HNMP” and “HNMPRoJiSh”), is summarised in Table 6.

Table 6. $F1$ measure improvement percent (%) using “HNMPRoVoQ” compared with “HNMP” and “HNMPRoJiSh” transformation configurations

| (%) | HAP | SEN | AGG | SAD |
|------------|------|------|------|------|
| HNMP | 16.7 | 17.1 | 44.4 | 60.0 |
| HNMPRoJiSh | 16.7 | 20.0 | 44.4 | 16.7 |

The results obtained with the first configuration, in which the speech signals were only modified to achieve the predicted prosodic parameters using the HNM algorithm (“HNMP”), are shown in Table 5. Notice that the identification $F1$ values have declined dramatically from their references for all styles. The sensual style case obtained the best result with regard to the rest of styles ($F1 = 0.41$).

Once the results for the prosodic transformation were reviewed, the inclusion of jitter and shimmer VoQ parameters was evaluated (“HNMPRoJiSh”). In this case, the results (see Table 5) show that the identification rate remained stable for the HNM (except for the sad style,

which was improved to a value of $F1 = 0.48$). Although the identification levels were still low, the results were better for the sad style ($F1 = 0.48$), coinciding with the maximum quality for this configuration (see Figure 2(d) in Section 4.2), possibly because of the stability demonstrated by the HNM during the parameter transformation; each style could be characterised without adding dispersion.

The last configuration to be analysed is “HNMPRoVoQ,” in which both prosody parameters and selected VoQ configurations were involved during the transformation of expressive speech styles (see Table 5). The first observation to emphasise is that we obtained the best results (see Table 6) with regard to the rest of configurations involving parameter transformations (“HNMPro” and “HNMProJiSh”). The second thing to note is the good result obtained for the sad style ($F1 = 0.56$), followed by the sensual style ($F1 = 0.48$). This is particularly important if we take into account the existing confusion between both styles in the reference (“Natural” and “ResHNM”). The value obtained for the happy style ($F1 = 0.42$) is very interesting too, especially because of the progression regarding the use of only prosody and jitter and shimmer. Finally, the aggressive style yielded the worst absolute result ($F1 = 0.39$), although it yielded the highest increment in its identification regarding “HNMProJiSh” configuration (44.4% according to Table 6).

The main reason for the identification error is the existing confusion between happy-aggressive and sensual-sad styles, which already appeared in the reference values (“Natural” and “ResHNM”). There is some general confusion towards the sad and “Others” categories. Once the test was finished, the listeners could write their comments. From them, it was observed that “Others” was in general related to the detection of a neutral style (i.e., the source style). Therefore, a higher modification for the parameters during the transformation is necessary. Thanks to the stability demonstrated by using the HNM, both for the quality (see Figures 2(c), 2(d), and 2(e)) and the identification (see Table 5), the level of parameter transformation could be increased. Moreover, according to listener observations, the trend towards the transmitted style identification using the semantic content of the utterance was also detected (i.e., what is the sentence talking about?), which caused a bias towards the wrong styles, especially in those cases in which acoustic characteristics could not identify them clearly (e.g., a whispering or quivering voice in the sensual or sad styles, resp.). As a conclusion, both from the quality and the identification perspectives, when the best identification rate was achieved through the use of prosody together with a combination

of VoQ (e.g., aggressive and sad styles), the quality levels went worse (see Figure 2(e) in Section 4.2). However, when the signal manipulation was not so high, the quality was maintained more constant (e.g., happy and sensual styles), and the style identification improvement was not so significant. Therefore, an agreement could be necessary between the expected quality and the amount of VoQ parameter modification needed for the identification of the style. In order to achieve this, a more sophisticated prediction of target VoQ parameters, similar to the one conducted on prosody modelling, could be necessary.

CONCLUSIONS AND FUTURE WORK

The main aim of this work was to validate the usefulness of VoQ in the enhancement of expressive synthetic speech for style identification presenting an acceptable quality. The harmonic plus noise model (HNM) of expressive style transformation based on prosody and voice quality modifications was evaluated by means of a perceptual assessment of speech quality and expressive speech styles identification. With regard to this methodology, first, flexible HNM parameterisation was used to extract the fundamental speech parameters that must be used during the performed prosody and VoQ modifications. Second, the prosody parameters were predicted by means of the CBR system and modified using the HNM parameters and the acoustic waveform, being a first attempt at the expressive speech style transformation. Finally, once prosody was transformed, the VoQ parameters were modified by varying the HNM parameters and the acoustic waveform according to the available VoQ models. To select which VoQ parameters should be modified, an analysis of the best configurations was previously performed.

The test for the perceptual assessment was carried out for different configurations, including natural speech and speech synthesis using the HNM: speech resynthesis, prosody modification, prosody plus jitter and shimmer modification, and, finally, prosody together with the best VoQ configurations using the HNM (the configuration of interest). These analyses resulted in an acceptable speech quality during the transformation. In addition, the expressive styles identification rate was directly related to the results of the quality test, which reported the best results for the configuration of interest.

To summarise, from the perceptual assessment of both the quality and identification experiments, the following conclusions can be drawn about the expressive speech styles transformation viability. First, the use of speech analysis and synthesis by means of the HNM made it possible to achieve

good quality and speech manipulation control during the transformations of both prosody and VoQ parameters. Second, the combination of prosody and VoQ parameters produced significant improvements in the expressive speech styles identification rate compared with only using prosody and a subset of VoQ parameters. Finally, from the comments of listeners during the test, it was observed that the semantic content of utterances could be a limitation for the expressive speech style identification.

In spite of the good results, more work is needed. A better model for VoQ is necessary, as one for the prosody already exists, in order to improve the model for each expressive speech style contained in the corpus. Moreover, in this way, distinguish pairs of expressive styles with identification difficulties let us improve their identification rate.

Finally, the results obtained both for quality and identification encouraged us to continue with the modelling of prosody and VoQ using HNM speech analysis and synthesis.

REFERENCES

1. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
2. S. Planet, I. Iriondo, J.-C. Socoró, C. Monzo, and J. Adell, “GTM-URL contribution to the INTERSPEECH 2009 Emotion Challenge,” in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 316–319, Brighton, UK, September 2009.
3. C. Drioli, G. Tisato, P. Cosi, and F. Tesser, “Emotions and voice quality: experiments with sinusoidal modeling,” in *Proceedings of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL '03)*, pp. 127–132, Geneva, Switzerland, 2003.
4. O. Turk, M. Schröder, B. Bozkurt, and L. M. Arslan, “Voice quality interpolation for emotional text-to-speech synthesis,” in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 797–800, Lisbon, Portugal, September 2005.
5. D. Erro, *Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models [Ph.D. thesis]*, Universitat Politècnica de Catalunya, Barcelona, Spain, 2008.
6. M. Schroder, “Emotional speech synthesis: a review,” in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 561–564, 2001.
7. C. Gobl, E. Bennett, and A. Ní Chasaide, “Expressive synthesis: how crucial is voice quality?” in *Proceedings of the IEEE Workshop on Speech Synthesis*, no. 11-13, pp. 91–94, 2002.
8. J. P. Cabral and L. C. Oliveira, “Pitch-synchronous time-scaling for prosodic and voice quality transformations,” in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1137–1140, Lisbon, Portugal, September 2005.
9. I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, “Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification,” *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

10. C. Monzo, A. Calzada, I. Iriondo, and J. C. Socoró, “Expressive speech style transformation: voice quality and prosody modification using a harmonic plus noise model,” in *Speech Prosody*, Chicago, Ill, USA, 2010.
11. T. Banziger and K. R. Scherer, “A study of perceived vocal features in emotional speech,” in *Proceedings of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL '03)*, pp. 169–172, Geneva, Switzerland, 2003.
12. C. Gobl and A. Ní Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.
13. C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, “Discriminating expressive speech styles by voice quality parameterization,” in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS '07)*, pp. 2081–2084, Saarbrücken, Germany, 2007.
14. J. Laroche, Y. Stylianou, and E. Moulines, “HNS: speech modification based on a harmonic+noise model,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '93)*, pp. 550–553, Minneapolis, Minn, USA, April 1993.
15. Y. Stylianou, J. Laroche, and E. Moulines, “High-quality speech modification based on a harmonic + noise model,” in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '95)*, pp. 451–454, 1995.
16. J. E. Cahn, *Generating expression in synthesized speech [M.S. thesis]*, Massachusetts Institute of Technology, 1989.
17. I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion,” *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.
18. I. R. Murray and J. L. Arnott, “Implementation and testing of a system for producing emotion-by-rule in synthetic speech,” *Speech Communication*, vol. 16, no. 4, pp. 369–390, 1995.
19. J. Yamagishi, T. Masuko, and T. Kobayashi, “Hmm-based expressive speech synthesis—towards tts with arbitrary speaking styles and emotions,” in *Proceedings of the Special Workshop in MAUI (SWIM)*, Lectures by Masters in Speech Processing, Conference CD-ROM, 1. 13, p. 4, 2004.

20. I. Esquerra, “Sntesis de habla emocional por seleccion de unidades,” in *Proceedings of the IV Jornadas en Tecnologa del Habla*, pp. 161–165, Zaragoza, Spain, 2006.
21. I. Iriondo, J. C. Socoró, and F. Alías, “Prosody modelling of spanish for expressive speech synthesis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ‘07)*, pp. 821–824, Honolulu, Hawaii, USA, April 2007.
22. M. Schroder, *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis [Ph.D. thesis]*, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, Saarbrucken, Germany, 2004.
23. N. Montoya, “El papel de la voz en la publicidad audiovisual dirigida a los niños,” *Zer. Revista de Estudios de Comunicación*, no. 4, pp. 161–177, 1998.
24. N. Campbell, “Databases of emotional speech,” in *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle*, pp. 34–38, Northern Ireland, UK, 2000.
25. H. Fran ois and O. Boeffard, “The greedy algorithm and its application to the construction of a continuous speech database,” in *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC ‘02)*, vol. 5, Las Palmas, Spain, 2002.
26. H. E. P rez, “Frecuencia de fonemas,” *Revista Electronica de Tecnologa Del Habla (E-RTH)*, no. 1, 2003.
27. D. Talkin, “A Robust Algorithm for Pitch Tracking (RAPT),” in *Speech Coding and Synthesis*, chapter 14, pp. 495–518, Elsevier Science, Amsterdam, The Netherlands, 1995.
28. F. Al as, C. Monzo, and J. C. Socor , “A pitch marks filtering algorithm based on restricted dynamic programming,” in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH ‘06)*, pp. 1698–1701, Pittsburgh, Pa, USA, September 2006.
29. R. Kohavi and F. Provost, “Glossary of terms: special issue on applications of machine learning and the knowledge discovery process,” *Machine Learning*, vol. 30, no. 2-3, pp. 271–274, 1998.
30. E. Navas, I. Hern ez, and I. Luengo, “An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1117–1127, 2006.

31. A. Calzada, *Expressive Synthesis based on Harmonic plus Stochastic model [Diploma of Advanced Studies]*, Universitat Ramon Llull, 2010.
32. P. Depalle and T. Helie, “Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '97)*, p. 4, October 1997.
33. J. Wells, “Sampa computer readable phonetic alphabet,” in *Handbook of Standards and Resources for Spoken*, D. Gibbon, R. Moore, and R. R. Winski, Eds., Part IV, section B, Mouton de Gruyter, Berlin, Germany, 1997.
34. C. Monzo, I. Iriondo, and E. Martínez, “Procedimiento para la medida y la modificación del jitter y del shimmer aplicado a la síntesis del habla expresiva,” in *Proceedings of the V Jornadas en Tecnología del Habla*, pp. 58–61, Bilbao, Spain, 2008.
35. P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9-10, pp. 341–345, 2001.
36. E. Keller, “The analysis of voice quality in speech processing,” in *Nonlinear Speech Modeling and Applications*, vol. 3445 of *Lecture Notes in Computer Science (LNCS)*, pp. 54–73, 2005.
37. C. Monzo, *Modelado de la calidad de la voz para la síntesis del habla expresiva [Ph.D. thesis]*, Universitat Ramon Llull, 2010.
38. ITU-P. 800, “Methods for subjective determination of transmission quality, Recommendation P. 800 International Telecommunication Union (ITU) Std,” 1996.
39. M. Frigge, D. C. Hoaglin, and B. Iglewicz, “Some implementations of the boxplot,” *American Statistician*, vol. 43, no. 1, pp. 50–54, 1989.
40. F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

CHAPTER 12

Prosodically Rich Speech Synthesis Interface Using Limited Data of Celebrity Voice

Takashi Nose¹, Taiki Kamei²

¹Department of Communication Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan.

²Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Japan.

ABSTRACT

To enhance the communication between human and robots at home in the future, speech synthesis interfaces are indispensable that can generate expressive speech. In addition, synthesizing celebrity voice is commercially important. For these issues, this paper proposes techniques for synthesizing natural-sounding speech that has a rich prosodic personality using a limited amount of data in a text-to-speech (TTS) system. As a target speaker, we

Citation: C. Monzo, I. Iriondo, J.C. Socoró, “Voice Quality Modelling for Expressive Speech Synthesis”, *The Scientific World Journal*, vol. 2014, Article ID 627189, 12 pages, 2014. <https://doi.org/10.1155/2014/627189>.

Copyright: © 2014 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

chose a well-known prime minister of Japan, Shinzo Abe, who has a good prosodic personality in his speeches. To synthesize natural-sounding and prosodically rich speech, accurate phrasing, robust duration prediction, and rich intonation modeling are important. For these purpose, we propose pause position prediction based on conditional random fields (CRFs), phone-duration prediction using random forests, and mora-based emphasis context labeling. We examine the effectiveness of the above techniques through objective and subjective evaluations.

Keywords:- Parametric Speech Synthesis, Hidden Markov Model (HMM), Prosodic Personality, Prosody Modeling, Conditional Random Field (CRF), Random Forest, Emphasis Context

INTRODUCTION

In the near future, people will have their own personal robots that support their daily life by communicating each other. To achieve such robots, speech recognition and synthesis interfaces are indispensable to make the communication of human-machine close to that of human-human. Currently, the use of speech recognition and synthesis technologies is rapidly spreading in smartphones (e.g., iPhone Siri), information guide in public facilities, and automotive navigation systems. Speech synthesis is a technology for generating speech from a text, and recently statistical parametric approach [1] based on hidden Markov models (HMMs) [2] has been widely studied and used [3]. However, most of the studies focus on synthesizing reading-style speech of news articles where the speaking style is always stable without prosodically rich expressions such as emphasis and emotions. Prosody of speech generally represents accent, intonation, rhythm, power, and phrasing (pause insertion) and has a rich personality. As a next step of speech synthesis to generate more human-like speech for various applications including humanoid robots, synthesizing speech with a rich prosodic personality is an essential issue.

In this paper, authors propose novel techniques for adding a rich personality to synthetic speech using a framework of HMM-based speech synthesis and machine learning. One of the final goals of this study is to achieve synthetic speech of Japanese prime minister, which gives sufficient impact and demands in practical applications. Speeches of the current prime minister, Shinzo Abe, are available in internet movies such as messages to the Japanese people and world leaders which are officially provided from the government. The speeches are very different from reading-style speech and

contain prosodically rich expressions to emphasize important points and not to make audience bored. To achieve such more human-like speech synthesis with a limited amount of celebrity speech, the following techniques are presented in this paper.

- Prediction of phrase breaking based on conditional random fields (CRFs)
- Robust prediction of phone durations using random forests
- Speech parameter generation with emphasis context based on a mora unit to preserve rich intonation of natural speech

In most of the speech synthesis research, the phrasing information, i.e., the positions of pause insertion, is manually given. However, the pause position sometimes strongly depends on the target speaker and, we need to automatically predict the positions from an input text in practical applications. In the speeches of Abe, he often inserts many pauses to clearly pronounce each word or phrase, and this style is very different from a general reading style. To model and predict the positions of phrase breaking, we use CRFs as label sequence modeling. For the duration modeling, hidden semi-Markov models (HSMMs) [4] are used for explicit modeling of state duration distribution [5]. However, the prediction accuracy of phone durations decreases when a sufficient amount of training data is not available. To improve the accuracy, the authors introduce phone-duration prediction using random forests [6] which is a kind of ensemble training [7]. Finally, speech parameter generation with mora-based emphasis context is presented to preserve rich intonation of natural speech, which is a variation of quantized fundamental frequency (F0) context [8] used also in voice conversion [9] and very low bit-rate speech coding [10].

The rest of this paper is organized as follows: In Section 2, we introduce a brief overview of parametric speech synthesis based on HMMs, which is a baseline speech synthesis system in this study. Section 3 describes speech materials used in this study. The role of prosody in speech synthesis and the problem of training data limitation are also explained in the section. Then, Section 4 explains the details of the proposed techniques to improve the prosodic personality when the training data of the target speaker is limited. In Section 5, the proposed techniques are compared with the baseline system through objective and subjective experiments and the results are discussed. Section 6 summarizes this study and refers to the future work.

PARAMETRIC SPEECH SYNTHESIS BASED ON HMMS

In the HMM-based speech synthesis, speech parameter sequences, e.g., spectral and F0 features, are modeled in phone units as is the same as the case of HMM-based speech recognition. The advantage of the HMM-based speech synthesis compared to traditional concatenative speech synthesis is to generate smooth and stable speech parameters by considering dynamic features with a relatively smaller amount of speech data. Different from speech recognition, the modeling of prosodic features, i.e., F0 and duration, is indispensable in speech synthesis. Since F0 has no value in silence and unvoiced regions, a special treatment is necessary such as the use of F0 interpolation [11] and multi-space probability distribution HMMs (MSD-HMMs) [12]. In the acoustic modeling, the acoustic property of speech parameters is affected by not only the current phoneme but also various factors such as preceding and succeeding phoneme, accent, stress, and sentence length. To take these factors into account, the factors are used as contexts and context-dependent HMMs are trained. Since the number of the combinations of contextual factors is too large, the contexts are tied using decision-tree-based context clustering [13] in the model training, and the number of unique contexts is reduced. In the phase of speech synthesis, a speech parameter sequence is generated based on a maximum likelihood criterion using the constraint of static and dynamic features [14]. Finally, a speech waveform is synthesized using a vocoding tool.

SPEECH MATERIALS WITH A RICH PROSODIC PERSONALITY

Speeches of the Japanese Prime Minister Abe

The HMM-based speech synthesis, which is a baseline in this study, is a corpus-based approach. This means that we need to prepare the speech data of a target speaker. The target speaker in this study is Shinzo Abe who is the 97th prime minister of Japan and is one of the most famous person in Japan. Since it is impossible to recording his voice in a standard way, the authors use speech data that is available at video hosting services. The type of the speeches is messages to the Japanese people at the annual events such as ones for Tohoku earthquake and official comments to the world leaders. However, the total length of collected speech data that have acceptable quality for

speech synthesis is only about six minutes. We discarded utterances that included noise, reverberation, and unclear pronunciation in advance.

In a typical HMM-based speech synthesis, we prepare speech samples and corresponding texts, and make labels with phone boundary information. However, some utterances of Abe were very long, and automatic phone segmentation sometimes failed. To avoid the problem, we divided a long utterance into short utterances based on pause. As a result, we had 319 utterances where 260 utterances included no pause. These utterance were used in the experiments of Section 5.

Pause Insertion for Voice Personality

Although pause insertion (phrase breaking) is used for breathing, it is also used to control speaking rate intentionally, to catch an attention, and to give calm impression to listeners. The position of pause insertion depends strongly on speakers, and the number of pause insertions also differs depending on speakers. Table 1 compares the average numbers of pauses per a minute between Japanese professional narrators with a reading style and the prime minister Abe. The two male and two female narrators are included in ATR Japanese speech database [15] set B. From the table, we found that Abe uses phrase breaking much more than the narrators. This result indicates that a general phrase breaking rule from an input text will degrade the voice personality of synthetic speech and intended effect appearing in the original speech is not always communicated to listeners correctly.

Role of Intonation and Speech Rate in Personality

Precise prediction of speech intonation plays a crucial role in communicating para-linguistic information as well as improving naturalness of synthetic speech. Speech having clear intonation with emphasis expressions enables a speaker to make the listener understood the key point of the utterance. However, most of the speech synthesis systems cannot model and predict emphasis expressions automatically, and the synthetic speech loses such para-linguistic expressions. Figure 1 shows an example of natural and synthetic speech samples of Abe. It is found that the natural speech has a clearer F0 curve than synthetic speech. Specifically, there is a clear peak of the F0 pattern around 1.0 sec in natural speech. However, such feature disappears in synthetic speech, and the F0 pattern become flattened. This example indicates that the quality of the synthetic speech would be improved if we can model emphasis expressions in the model training.

Problem of Training Data Limitation

As is described in Section 3.1, the amount of speech data of Abe obtained from the internet is very limited. Although HMM-based speech synthesis can synthesize speech using a smaller amount of speech data of a target speaker than concatenative speech synthesis with unit selection, we typically need more than several tens of minutes training data to synthesize acceptable quality in naturalness.

Table 1. Comparison of average pause insertion counts per minute between the prime minister Abe and professional narrators

| | Professional narrator | | | | |
|---------|-----------------------|-----|-----|-----|----------|
| Speaker | MHT | MMY | FTK | FKS | P.M. Abe |
| Count | 23 | 24 | 18 | 17 | 30 |

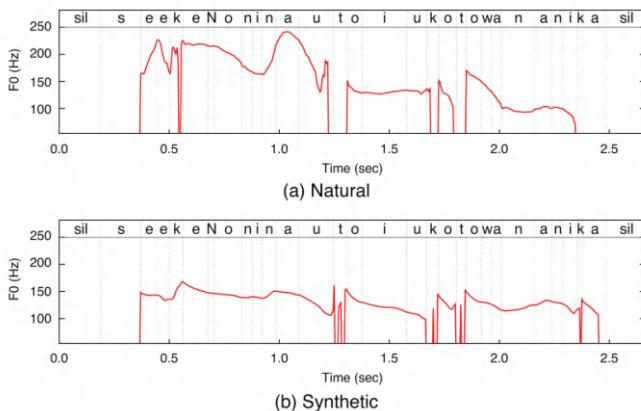


Figure 1. Comparison of F0 contours between natural and synthetic speech samples.

A straightforward way for this problem is to use speaker adaptation techniques such as maximum likelihood linear regression (MLLR) [16] with an average voice model [17] that is an acoustic model trained using speech data of multiple speakers. However, the adaptation performance depends on the average voice model, and the performance of the adaptation from a reading-style average voice model to a different type of speech, e.g., spontaneous speech, is not always satisfactory [18]. Therefore, the authors do not use the combination of the average voice model and a speaker adaptation technique in this study.

PROSODIC PERSONALITY IMPROVEMENT WITH LIMITED DATA

In this section, the authors propose three techniques to improve the prosodic personality of synthetic speech when the amount of speech data of the target speaker, i.e., the prime minister Abe in this study, is limited. Specifically, positions of pause insertion are predicted from an input text using CRFs. The accuracy of predicting phone durations is also improved by using random forests as an ensemble training technique. Furthermore, emphasis context based on a mora unit is introduced which can be automatically obtained by using differential features between natural and generated F0 parameter sequences. These techniques enable a TTS system to represent personal prosodic characteristics close to those of Abe while maintaining naturalness of synthetic speech.

Overview of the Proposed Speech Synthesis System

Figure 2 shows the outline of our text-to-speech system including three proposed techniques explained in the following sections. The system is named Abe-droid speech synthesis system¹. In the figure, the boxes highlighted in yellow indicate the proposed techniques in this paper, i.e., CRF-based prediction of pause insertion position, robust duration modeling using random forests, and the use of mora-based emphasis context.

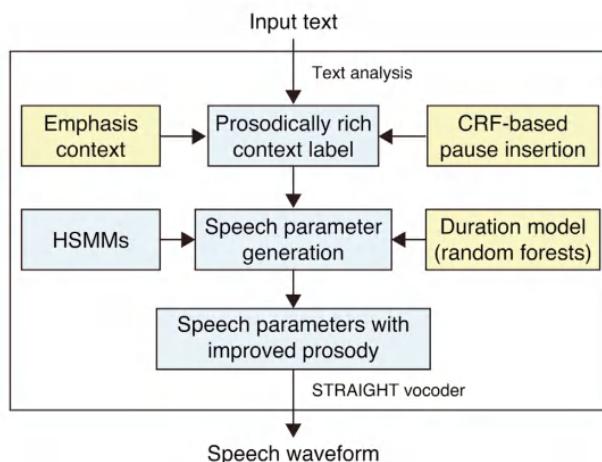


Figure 2. Overview of the synthesis part of the proposed text- to-speech system. The proposed techniques are highlighted in yellow.

When an input text is given, the text is converted to a context-dependent label sequence which has prosodically rich representation. At this time, pauses are automatically inserted based on a CRF-based pause insertion model, which is explained in Section 4.2. Emphasis context is also added to the labels. The emphasis context for the training data is automatically obtained using differential features. The detail is explained in Section 4.4. Then, speech parameter is generated using context-dependent HSMMs and prosodically rich context labels. The duration of each phone is determined using the duration model based on random forests (Section 4.3). Finally, a speech waveform is synthesized using a vocoder such as STRAIGHT [19].

Estimation of Pause Position Based on CRFs

In this study, pause positions are modeled and predicted using CRFs [20]. CRFs are used for a problem of sequence labeling where an appropriate label sequence y , e.g., part of speech tags, is predicted when an input sequence x is given. Let F be a sequence of features. $\phi_f(x, y)$ stands for how many times a feature $f \in F$ appears in the set of (x, y) , $\Phi(x, y)$ denotes its vector representation. The importance of each feature is represented by weight θ_f that is a parameter of a CRF, and Θ is its vector representation. Then, the conditional distribution for a CRF is given by

$$P(y|x) = \frac{\exp\langle\Theta, \Phi(x, y)\rangle}{\sum_{f \in F} \exp\langle\Theta, \Phi(x, y)\rangle} \quad (1)$$

where

$$\langle\Theta, \Phi(x, y)\rangle = \sum_{f \in F} \theta_f \varphi_f(x, y). \quad (2)$$

A set of model parameters is determined based on the maximum likelihood criterion.

To apply CRFs to Japanese text, we use a tool of Japanese morphological analysis, MeCab [21]. MeCab outputs surface form, part of speech (POS), subdivided POS 1, subdivided POS 2, subdivided POS 3, conjugated form, conjugation type, base form, reading, and pronunciation. In this study, we use only three factors, surface form, POS, and subdivided POS 1. For the surface form, preceding and succeeding forms are taken into account as well as the current form. Similarly, for the POS and subdivided POS 1, two preceding and two succeeding forms are taken into account in addition to the current ones. Figure 3 shows an example of the created training data in

Japanese. The binary flags in the fourth field represent whether a pause is inserted after the morpheme or not.

Robust Phone-Duration Prediction Using Random Forests

Phone is the smallest unit of speech where we can distinguish the sound. Phone durations in an utterance are strongly related to local and global tempo and rhythm of speech. Therefore, modeling and predicting phone durations precisely are very important because they affect various properties of speech, e.g., speech naturalness, speaker individuality, speaking style, and emotional expression. In a preliminary experiment, we examined the performance of duration modeling in two ways. The first technique is to use standard HSMMs where state-duration distributions are explicitly modeled by Gaussian probability density functions (pdfs). This is a sophisticated way but has been shown to be worse than using an external duration model [22]. Therefore, we also used an external tree-based duration prediction model as the second technique where the distributions of phone durations are modeled as Gaussian pdfs and the model parameters are tied using a single context-dependent decision tree.

Both techniques work well when a sufficient amount of training data is available. However, the condition of this study is very severe and the training data is very limited, i.e., only about six minute data is available. In that case, more robust prediction approach is required. We use random forests for this purpose. A random-forest technique is one of the machine learning techniques based on ensemble training and was applied to speech synthesis for spectral parameter prediction [23].

| Word | Part of speech | | Pause insertion (0/1) |
|------|----------------|------|-----------------------|
| 東日本 | 名詞 | 固有名詞 | 0 |
| 大震災 | 名詞 | 一般 | 0 |
| から | 助詞 | 格助詞 | 1 |
| 三 | 名詞 | 数 | 0 |
| 度目 | 名詞 | 接尾 | 0 |
| と | 助詞 | 格助詞 | 0 |
| なる | 動詞 | 自立 | 1 |
| 三月 | 名詞 | 副詞可能 | 0 |

| | | | |
|----|-----|-----|---|
| 十 | 名詞 | 数 | 0 |
| 一 | 名詞 | 数 | 0 |
| 日 | 名詞 | 接尾 | 0 |
| を | 助詞 | 格助詞 | 0 |
| 迎え | 動詞 | 自立 | 0 |
| まし | 助動詞 | * | 0 |
| た | 助動詞 | * | 0 |
| . | 記号 | 句点 | 0 |

Figure 3. Example of training data (in Japanese) for CRF to predict pause position.

Figure 4 shows the outline of the proposed duration prediction using random forests. In the training phase, we make N subsets of training data by random sampling and construct a decision tree for each subset. These trees are used in the synthesis phase. An input text is converted to a context-dependent label sequence and is inputted into respective decision trees. Then, median filtering is applied to the output durations, and finally we obtain a predicted duration. When the number of subsets is even, two median values are obtained and we use the mean of the values as a predicted duration.

Intonation Control Using Mora-Based Emphasis Context

As is described in Section 3.4, modeling and synthesizing expressive speech that has a variety of local expressions is difficult when using a standard HMM-based speech synthesis framework. This is because the context labels used in model training and speech synthesis have no information of such local variations. For this problem, we proposed a prosody enhancement technique based on differential features of F0 and quantization [24] to capture the emphasis expressions in accent phrases of Japanese speech. To achieve more precise prediction of expressive speech such as speeches of Abe, similarly to the previous study, we here propose automatic mora-based emphasis expression labeling for training data. Mora is a basic unit for pronunciation in Japanese language and has similar characteristics to syllable in other languages, e.g., English. Japanese is a language of pitch accent, and we control an accent by changing relative pitch of each mora in an accent phrase.

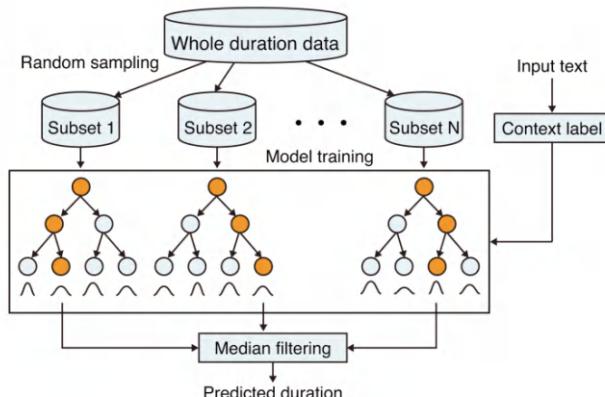


Figure 4. Phone duration modeling and prediction using random forests.

The mora-based emphasis labeling is achieved as follows. First, standard HSMMs are obtained using training data without emphasis labels. We once generate F0 parameter sequences for the training sentences using the trained HSMMs. When comparing generated and natural F0 sequences, there are large differences in the region of speech having emphasis expressions. Hence, we calculate the differences between generated

and natural F0 sequences and quantize the values into three levels, high (positive emphasis: 1), neutral (no emphasis: 0), and low (negative emphasis: -1), for each mora unit. The process is summarized as follows:

1. Train context-dependent HSMMs using conventional labels with only linguistic information.
2. Generate F0 sequences from the training sentences using the HMMs obtained above.
3. Calculate average $\log F_0$ values, of and f_s , of natural and synthetic speech for each mora unit.
4. Calculate the average $\log F_0$ difference $d = f_o - f_s$.
5. Classify d into three classes: a) $d < -\alpha$ (low), b) $-\alpha \leq d < \alpha$ (neutral), and c) $d \geq \alpha$ (high), where positive value α is a classification threshold.

The threshold for the quantization can be automatically optimized using training data [24]. Figure 5 shows an example of context-dependent labels including emphasis context that is automatically obtained for the training data. In the figure, triphone is shown in the left field, accentual factors are shown in the center field, and emphasis context is shown in the right field. From the figure, we found that a mora sequence/ ara/ has positive emphasis (1) and /Ndo/ has negative emphasis (-1).

EXPERIMENTS

In this section, we incorporated our prosody modeling techniques described in Section 1 into the conventional baseline HMM-based speech synthesis and compared the performance through objective and subjective evaluations. In the objective evaluations, the prediction accuracy of pause positions, phone durations, and intonation similarity are examined. In the subjective evaluations, naturalness and similarity of synthetic speech are evaluated with five-point scale tests.

Experimental Conditions

We used about six-minute speech data of Abe that was described in Section 3.1. The total number of utterances was 319.

| Triphone | Accent context | Emphasis context (-1/0/1) |
|--------------------------------------------------|----------------|---------------------------|
| ... | | |
| k-a+r/A:7_4/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1 | | |
| a-r+a/A:8_5/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1 | | |
| r-a+pau/A:8_5/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1 | | |
| a-pau+s/A:x_x/C:8_3_x_x-x_x_x+x+5_0_x_x/E:21/F:x | | |
| pau-s+a/A:1_1/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:0 | | |
| s-a+N/A:1_1/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:0 | | |
| a-N+d/A:2_2/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1 | | |
| N-d+o/A:3_3/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1 | | |
| d-o+m/A:3_3/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1 | | |
| ... | | |
| | Mora context | Sentence length |

Figure 5. Example of a context-dependent label sequence with emphasis context. The symbol x means that there is no definition of the corresponding context.

300 utterances were used as training data, and remaining 19 utterances were used as test data. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis [19] was used to extract spectral envelope, F0, and aperiodicity features with a five ms frame shift. The spectral envelope was converted to mel-cepstral coefficients using a recursion formula. The aperiodicity features were converted to average values for five frequency sub-bands, i.e., 0 - 1, 1 - 2, 2 - 4, 4 - 6, and 6 - 8 kHz. The resultant feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, log F0, five average band aperiodicities, and their delta and delta-delta coefficients. The total number of dimensions was 138. We used five-state left-to-right HSMMs with no skip between states. Each state had a single Gaussian pdf with a diagonal covariance matrix. In the decision-tree-based context clustering, minimum description length (MDL) was used as a stopping criterion. In the baseline system, triphone, mora position, accent information, and sentence length were used as contextual factors.

Accuracy of Predicted Pause Insertion with CRFs

First, we evaluated the performance of predicting the positions of pause insertion based on CRFs. Table 2 shows a confusion matrix of predicted

and correct classes of pause insertion. From the table, we found that more than 92% of the pause positions were correctly predicted using CRFs. In a practical application, listeners will perceive the prediction error as unnatural only when pauses are incorrectly inserted. This indicates that only 3.4% of pauses inserted by CRFs can affect the speech naturalness. In this experiment, the prediction accuracy is good even though the amount of training data is very limited. One of the reasons for this result is that the speeches were official messages and Abe regularly inserted pauses into the utterances. Therefore, we might need more data to achieve sufficient accuracy of pause insertion when the target speaker is a person who is inexperienced at speaking officially, which is our future work.

Effect of Random Forests in Phone-Duration Prediction

Next, we evaluated the effectiveness of using random forests in phone-duration prediction. The number of subsets for the random forests was set to six, and each fifty utterances were used to construct decision trees using context clustering with an MDL-based stopping criterion. For comparison, duration prediction techniques using HSMMs and a single tree were also evaluated. Root mean square (RMS) error of phone durations between natural and synthetic speech was used as an objective measure. Table 3 shows the result.

Table 2. Ratio (%) of classified boundaries for pause insertion

| | | Correct class | |
|-----------------|-----------|---------------|-----------|
| | | Pause | Not pause |
| Predicted class | Pause | 14.9 | 3.4 |
| | Not pause | 4.6 | 77.1 |

Table 3. Comparison of RMS errors (ms) of phone durations

| HSMM | Single tree | Random forests |
|-------|-------------|----------------|
| 64.73 | 66.39 | 36.87 |

From the table, we found that the use of random forests in phone-duration prediction substantially decreased the objective distortion and made the phone durations closer to those of the natural speech when

compared to the conventional techniques. To investigate the detail of the effect, we also examined the distributions of predicted phone durations with the conventional and proposed techniques. Figure 6 shows the histograms of phone durations. From the figure, we found that the phone-duration prediction based on random forests reduced the RMS errors of durations more than 100 ms compared to the conventional techniques. In addition, the distribution of phone durations in random forests is closer to a Gaussian distribution than the other techniques, which indicates that phone durations were well modeled and predicted by using random forests.

Effect of Emphasis Context for Intonation Improvement

We also examined whether the use of emphasis context improves the intonation of synthetic speech. For the quantization of differential F0 features, we first determined threshold α using training data. The objective measure of F0 similarity to the natural speech is the RMS error of log F0 between natural and synthetic speech. For the threshold optimization, threshold was changed from 0.0 to 1.0 with an increment of 0.1, and the smallest value, $\alpha = 0.12$, was used as the optimal threshold. Then, emphasis contexts of high, neutral, and low, were determined for the training and test utterances. For comparison, we trained HSMMs in three conditions. The first was the conventional HSMMs without emphasis context. The second and the third were HSMMs with emphasis context using the initial threshold of $\alpha = 0.0$ and the optimal threshold $\alpha = 0.12$, respectively. The RMS errors of log F0 (cent) were calculated between natural and synthetic speech. Table 4 shows the result. From the table, it is seen that the use of emphasis context substantially reduced the F0 distortions. We also found that the distortions were further reduced by the threshold optimization Figure 7 shows an example of F0 contours with and without emphasis context. From the figure, it is seen that the F0 contour generated with emphasis context is closer to natural speech and has a clearer intonation than that without emphasis context.

Total Subjective Evaluation

Finally, we conducted total subjective evaluation tests to examine the effect of each proposed technique for improving prosody in synthetic speech generated from limited training data. We evaluated speech synthesis in five different conditions as follows:

Baseline Conventional HSMM-based speech synthesis

Pau-predict Baseline with CRF-based pause insertion

Pau-correct Baseline with correct pause insertion

Pau + dur Pau-correct with phone-duration prediction using random forests

Pau + dur + emph Pau + dur with emphasis context

For all synthetic speech samples, the pause length was set to 0.65 (sec) which was the mean value of the pauses included in the training data. The participants were ten native Japanese speakers. We evaluated the naturalness and similarity of the synthetic speech samples with mean opinion score (MOS) tests. For the similarity test, participants listened to natural speech samples as reference before the synthetic speech stimuli. Naturalness and similarity were evaluated on a five-point scale: “1” for bad, “2” for poor, “3” for fair, “4” for good, and “5” for excellent. During the MOS tests, participants could repeat to play sentences to evaluate the utterances as many times as required. Figure 8 shows the average scores for the respective techniques.

From the figure, we found that naturalness and similarity of the baseline system is not satisfactory when the amount of the target speaker is very limited and the speech is prosodically rich.

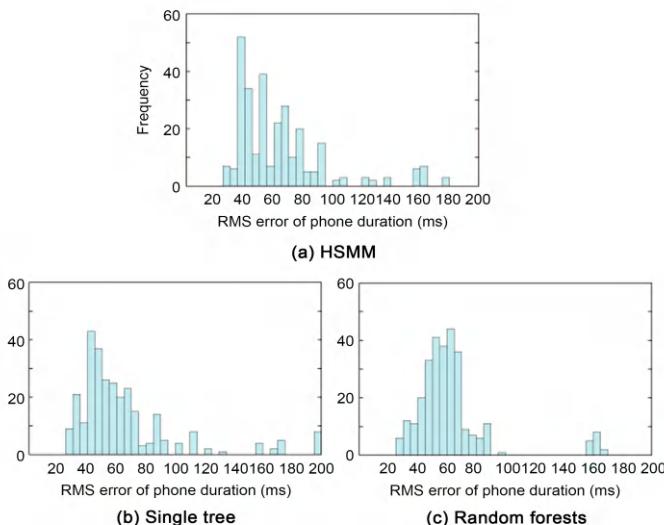


Figure 6. Histograms of predicted phone durations in the conventional and proposed techniques.

Table 4. Effect of emphasis context with an optimized threshold when comparing RMS errors (cent) of F0 for test data

| 2*HSMM | Emphasis context | |
|--------|------------------|------------|
| | Default | Optimal |
| | (d = 0.0) | (d = 0.12) |
| 413.2 | 285.3 | 249.3 |

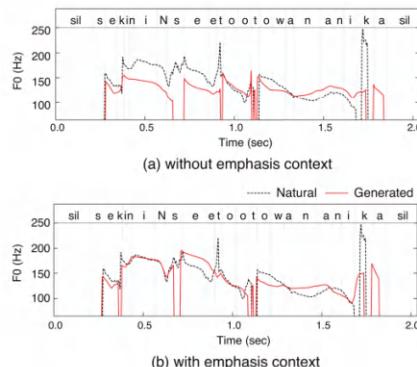


Figure 7. Effect of the proposed emphasis context in terms of F0 contours.

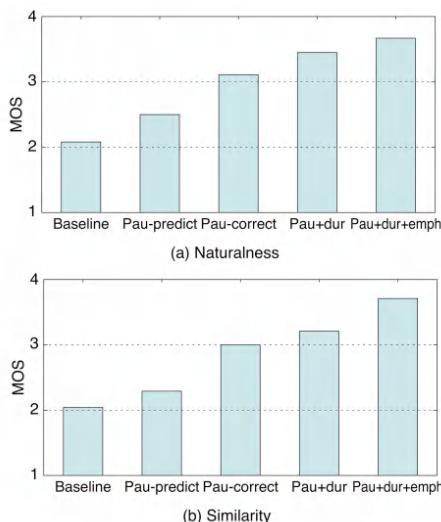


Figure 8. Results of subjective evaluation of synthetic speech in different conditions.

By introducing pause prediction, there was 0.5 point improvement in the naturalness evaluation, and similarity was also improved. However, we found that the prediction performance was still insufficient when comparing Pau-predict and Pau-correct. One of the reasons of this gap is that some of the test utterances were relatively long and included many pauses, and the naturalness and similarity degraded even when one pause was incorrectly inserted. The proposed duration prediction and emphasis modeling worked well and both of them improved naturalness and similarity.

CONCLUSION

The final goal of this study is to achieve a speech synthesis interface that is commercially valuable and has a rich personality. For this purpose, we focused on synthesizing the voice of the prime minister of Japan, Shinzo Abe, as the target speaker. We proposed techniques for HMM-based speech synthesis to achieve an interface of prosodically rich speech synthesis when the target speaker is a celebrity but the available speech data is limited. We presented CRF-based prediction of the position of pause insertion, robust phone-duration prediction using random forests, and the use of emphasis context for mora units. The objective and subjective evaluation results have shown that all techniques improved the performance of speech synthesis from the baseline HMM- based speech synthesis system. The current sysmtem has a limitation that the emphasis context must be added manually to the input text for synthesis, and hence the automatic labeling of emphasis context for test data is our future work. In addition, we will attempt to introduce speaker adaptation technique under the condition that multiple speakers' speech data are available in advance. Synthesizing emotional speech of celebrities is also an remaining task.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP15H02720 and Step-QI school in department of electrical, information and physics engineering, Tohoku University.

NOTES

¹The name comes from android which is a kind of humanoid robot.

REFERENCES

1. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1999) Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. European Conference on Speech Communication and Technology, 2347-2350.
2. Rabiner, L.R. and Juang, B.-H. (1986) An Introduction to Hidden Markov Models. IEEE ASSP Magazine, 3, 4-16. <https://doi.org/10.1109/MASSP.1986.1165342>
3. Zen, H., Tokuda, K. and Black, A. (2009) Statistical Parametric Speech Synthesis. *Speech Communication*, 51, 1039-1064. <https://doi.org/10.1016/j.specom.2009.04.004>
4. Levinson, S. (1986) Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. *Computer Speech & Language*, 1, 29-45. [https://doi.org/10.1016/S0885-2308\(86\)80009-2](https://doi.org/10.1016/S0885-2308(86)80009-2)
5. Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (2007) A Hidden Semi-Markov Model-Based Speech Synthesis System. *IEICE Transactions on Information and Systems*, E90-D, 825-834. <https://doi.org/10.1093/ietisy/e90-d.5.825>
6. Liaw, A. and Wiener, M. (2002) Classification and Regression by Randomforest. *R News*, 2, 18-22.
7. Dietterich, T.G. (2000) Ensemble Methods in Machine Learning. Proc. International Workshop on Multiple Classifier Systems, 1-15.
8. Nose, T., Ota, Y. and Kobayashi, T. (2010) HMM-Based Voice Conversion Using Quantized F0 Context. *IEICE Transactions on Information and Systems*, E93-D, 2483-2490. <https://doi.org/10.1587/transinf.E93.D.2483>
9. Nose, T. and Kobayashi, T. (2011) Speaker-Independent HMM-Based Voice Conversion Using Adaptive Quantization of the Fundamental Frequency. *Speech Communication*, 53, 973-985. <https://doi.org/10.1016/j.specom.2011.05.001>
10. Nose, T. and Kobayashi, T. (2012) Very Low Bit-Rate F0 Coding for Phonetic Vocoders Using MSD-HMM with Quantized F0 Symbols. *Speech Communication*, 54, 384-392. <https://doi.org/10.1016/j.specom.2011.10.002>
11. Yu, K., Thomson, B. and Young, S.J. (2010) From Discontinuous to Continuous F0 Modelling in HMM-Based Speech Synthesis.

Proceedings of 7th ISCA Speech Synthesis Workshop, Kyoto, 22-24 September 2010, 94-99.

12. Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T. (2002) Multi-Space Probability Distribution HMM. IEICE Transactions on Information and Systems, E85-D, 455-464.
13. Riley, M. (1990) Tree-Based Modelling for Speech Synthesis. Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, 229-232.
14. Tokuda, K., Kobayashi, T. and Imai, S. (1995) Speech Parameter Generation from HMM Using Dynamic Features. 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, 9-12 May 1995, 660-663. <https://doi.org/10.1109/ICASSP.1995.479684>
15. Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K. (1990) ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. Speech Communication, 9, 357-363. [https://doi.org/10.1016/0167-6393\(90\)90011-W](https://doi.org/10.1016/0167-6393(90)90011-W)
16. Leggetter, C.J. and Woodland, P.C. (1995) Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech & Language, 9, 171-185. <https://doi.org/10.1006/csla.1995.0010>
17. Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T. (2001) Text-to-Speech Synthesis with Arbitrary Speaker's Voice from Average Voice. 7th European Conference on Speech Communication and Technology, Scandinavia, 3-7 September 2001, 345-348.
18. Koriyama, T., Nose, T. and Kobayashi, T. (2010) Conversational Spontaneous Speech Synthesis Using Average Voice Model. 11th Annual Conference of the International Speech Communication Association, Chiba, 26-30 September 2010, 853-856.
19. Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A. (1999) Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication, 27, 187-207. [https://doi.org/10.1016/S0167-6393\(98\)00085-5](https://doi.org/10.1016/S0167-6393(98)00085-5)
20. Lafferty, J., McCallum, A. and Pereira, F.C. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 18th International Conference on Machine Learning, Williamstown, 28 June-1 July 2001, 282-289.

21. Kudo, T. (2005) Mecab: Yet another Part-of-Speech and Morphological Analyzer. <https://github.com/taku910/mecab>
22. Latorre, J., Buchholz, S. and Akamine, M. (2010) Usages of an External Duration Model for HMM-Based Speech Synthesis. 5th International Conference on Speech Prosody, Chicago, 11-14 May 2010, 1-4. <http://speechprosody2010.illinois.edu/papers/100073.pdf>
23. Black, A.W. and Muthukumar, P.K. (2015) Random Forests for Statistical Speech Synthesis. Proceedings of Interspeech, Dresden, 6-10 September 2015, 1211-1215.
24. Maeno, Y., Nose, T., Kobayashi, T., Koriyama, T., Iijima, Y., Nakajima, H., Mizuno, H. and Yoshioka, O. (2014) Prosodic Variation Enhancement Using Unsupervised Context Labeling for HMM-Based Expressive Speech Synthesis. *Speech Communication*, 57, 144-154. <https://doi.org/10.1016/j.specom.2013.09.014>

CHAPTER 13

Resources for Development of Hindi Speech Synthesis System: An Overview

Archana Balyan

Department of Electronics and Communication, Maharaja Surajmal Institute of Technology, Affiliated to GGSIPU, New Delhi, India.

ABSTRACT

Most of the information in digital world is accessible to few who can read or understand a particular language. The speech corpus acquisition is an essential part of all spoken technology systems. The quality and the volume of speech data in corpus directly affect the accuracy of the system. However, there are a lot of scopes to develop speech technology system using Hindi language which is spoken primarily in India. To achieve such an ambitious goal, the collection of standard database is a prerequisite. This paper summarizes the Hindi corpus and lexical resources being developed

Citation: Balyan, A. (2017), "Resources for Development of Hindi Speech Synthesis System: An Overview". Open Journal of Applied Sciences, 7, 233-241. doi: 10.4236/ojapps.2017.76020.

Copyright: © 2017 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>

by various organizations across the country.

Keywords:- Speech, Database, Corpora, Lexicon, Speech Synthesis, Linguistics, Natural Language Processing

INTRODUCTION

The objective of speech data collection is to primarily build speech recognition and synthesis systems for Indian languages [1] . There is an ever-growing demand for customized and domain-specific voices for use in corpus based on synthesis systems. Hence, it is very important that good methods should be established for creating these databases. The high-quality audio data, and of large volume, is key to developing a high-quality speech synthesizer. In a country like India, where the literacy rate is low, Indian language speech interfaces can provide access to IT applications and services, through the Internet and/or telephones, to the masses. So that people in various semi-urban and rural parts of India will be able to use telephones and Internet to access a wide range of services and information on health, agriculture, travel, etc. However, for this to become a reality, computers should be able to accept speech input in the user's language and provide speech output. Also, in multilingual India, if speech technology is coupled with translation systems between the various Indian languages, services and information can be provided across languages more easily. Due to the lack of appropriate annotated speech databases in Indian languages, robust applications have not been developed. Efforts are being made by a selected set of Indian academic and research institutions in a consortium mode to build speech synthesis, speech recognition and machine translation systems in Indian languages. These efforts are primarily supported by the ministry of the information and communication technologies (MCIT), Govt. of India (GoI). The resources including speech and text corpora collected in these efforts abide by the copyright restrictions of the sponsor [2] .

Hindi is an Indo-Aryan language with about 545 million speakers, 425 million of whom are native speakers. As per the eighth schedule of government of India, there are 22 official languages and it is one of the official languages of India and national language of the Federal Government of India. Hindi is spoken by a maximum number of people by about 41% of the population mostly in northern, central, western and eastern parts of the country [3] . This paper focuses on the Hindi resources being developed, which can be used for research in computational linguistics.

HINDI TEXT ENCODING

The computer age in India began in 1955 with the installation of HEC-2M (Hollerith Electronic computer model-2M) a computer designed by A.D. Booth in England) at the Indian Statistical Institute (ISI) at Calcutta (now Kolkata) [4].

Code Standardization for Indic Scripts: A Survey

Various letters of the input text are recognized and converted into their respective codes. In 1978, India's DoE constituted a standardization committee, for designing codes for Indic scripts similar to ASCII. In 1982, first version of a 7-bit code, called ISSCCI-7 (Indian scripts Standard Code for information Interchange). In 1983, the first version of 8-bit code (ISCII-8) was released. A further modification was made in 1991, and the Bureau of Indian standards accepted ISCII-8 as national standard (IS 13194:1991). A newly formed Unicode consortium adopted 1998 version of ISCII-8 as the base for 16-bit Unicode for allocating codes to different Indian scripts [5] . With the advent of Unicode in 1990s, some online publications have switched to Unicode. A main on-line source of Hindi text in Unicode is Universal Word—Hindi dictionary [6] is being made at CFILT, IIT Bombay for the purpose of Machine Translation. The user can search the Hindi and English words and phrases. This lexicon also provides the grammatical, morphological and semantic attributes of the Hindi words. This version contains 36,111 Words and is good source of corpus. Encoding conversion may be required if data is acquired from other sources.

CORPORA DEVELOPMENT IN HINDI LANGUAGE

In modern linguistics, a corpus is the machine readable form of large collection of structured text in written or spoken form [7] . If corpora can give some linguistic information, it is called Annotated Corpora. It is as important a resource as any other in the field of language engineering. With the recent advancement in computer technology the availability of language corpora (by corpora we mean corpus) and its processing has become even easier and has opened many new areas of research in language processing. A corpus can be the best resource to study many different linguistic phenomena such as the spelling variations, morphological structure, and word sense analysis and how the language has evolved over the time and many more [8] .

Development of Speech Corpora

A Speech Database of Hindi language for automatic speech recognition system for travel domain has been developed at C-DAC, Noida. The database consists of training data collected from 30 female speakers in a noise free environment consisting of approximately 26 hours of speech recordings. Total 8567 sentences consisting 74,807 words were recorded by the speakers uniformly distributed over all age group from 17 to 60 years. The recognition system was developed for the same recorded data and the recognition rate achieved for training data was 70.73% and that for the test data was 60.66% [9].

Another general purpose speech database in Hindi has been developed from Broadcasted news bulletin at IIT, Kharagpur. The total duration of speech in Hindi is 3.5 Hrs and was recorded for 19 speakers (6 Males and 13 Females). As the speech database is of broadcast, the recording is done in the studio in a Noise free environment [10].

The IIIT Hyderabad India developed speech databases at Speech and Vision Lab, for the purpose of building speech synthesis systems in Indian languages. This database consists of text and speech data in Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. In the data base the text content collected from more than 10,000 Wikipedia articles. This database is collected from native speaker who are available in this campus. Each of these languages has several dialects. The speech data was recorded by a native speaker of the language. The recording was done in a studio environment using a standard headset microphone connected to a zoom handy recorder [11].

Two main sources for Mobile database at are developed at KIIT, Bhubneshwar and KIIT, Gurgaon. At KIIT, Gurgaon, a text corpus of 2 million words of natural messages in 12 different domains in Hindi and Indian English and a speech corpus of 100 speakers, each speaking 630 phonetically rich sentences, has been created. The speech utterances were recorded in 16 kHz through 3 recording channels: a mobile phone, a headset and a desktop mounted microphone. This project was sponsored by Nokia Research Centre China [11].

The Linguistic Data Consortium for Indian Languages (LDCIL) is the Consortium responsible for creating database and tools for collection of high quality database in various domains that can be used by researchers in developing speech technology systems. At LDCIL [12], a Hindi Speech Recognition database was collected in Uttar Pradesh and Bihar and contains

the voices of 650 different native speaker who were selected according to age distribution (16 - 20, 21 - 50, 51+), Gender, Dialectical Regions and environment (home, office and public place). Each speaker record read a news text in a noisy environment through recorder having an inbuilt microphone. The recordings are in stereo recording and the extracted channel is also included in the specific files. It includes audio file, text file, NIST files which were saved as. ZIP Files. All the speech data are transcribed and labeled at the sentence level.

The purpose of developing the IIIT-H Indic speech databases is to have speech and text corpora made available in the public domain, without copyright restrictions for non-commercial and commercial use. A text to speech synthesis system for travel and emergency services in Indian languages is developed at IIIT Hyderabad. The speech databases developed include English, Telugu and Hindi speech corpus from 15 different speakers. This application is needful for people who faced problem for travel in India to see its rich cultural Heritage. All the recordings were done using a laptop and a standard microphone in a room in noise free environment [13] .

A general purpose, multi speaker, Continuous Speech Database has been developed for Hindi language by the researchers of TIFR Mumbai and CDAC Noida. The Hindi Speech database is comprehensive enough to capture phonetic, acoustic, intra-speaker and inter speaker variability's in Hindi Speech. This database consists of sets of 10 phonetically rich Hindi sentences spoken by 100 native speakers of Hindi language. The speech data was digitally recorded using two microphones in a noise free environment. Each speaker was asked to read the 10 sentences consisting 2 parts. The first part consists of two sentences which preferably covers the maximum phonemes of Hindi language. Every speaker was asked to speak these two sentences. The second part consisted of 8 sentences which covered maximum possible phonetic context. Though this continuous speech database was developed for training speech recognition system for Hindi language, it has been designed and developed in such a manner that is can also be used in tasks such as speaker recognition, study of acoustic-phonetic correlation of the language [14] .

At KIIT, Bhubaneswar, a project for mobile text and speech database collection in Hindi has been completed. The project was sponsored by Nokia Research centre, China. The speech data was collected using 13 prompt sheets containing 630 phonetically rich sentences in Hindi language after collecting text messages in Hindi. The collected text corpus for Hindi

consists of 42,801 unique words respectively. The speech data was recorded from 100 speakers using 3 channels simultaneously at a sampling frequency 16 KHz. The developed speech database consists of 60% female voice recording and 40% male voice recording [14].

Development of Textual Corpora

EMILLE [15] Project (Enabling Minority Language Engineering), initiated by Lanchester University, is one of the first initiatives taken to make Hindi corpus available for research and development of the language processing. The project has released 200,000 words of English text translated to Bengali, Gujarati, Hindi, Punjabi and Urdu creating a parallel corpus across these languages [16].

Indian Resources: Web Corpora for Indian Text

The Leipzig Corpora Collection (LCC) [17] has been collecting digital text material for more than 30 years. Over the last years, the established text acquisition and text processing tools are adopted to deal with Indian language to create and improve resources based on Indian text material. Corpora of this collection are typically grouped regarding the dimensions language, country of origin, text type (newspaper text, governmental text, generic Web material, religious texts etc.) and time of acquisition. Table 1 gives an introduction to currently available resources. It contains the number of sentences for Hindi languages and genres. The corpora are available via Web-based interfaces [7]. A main on-line source of Hindi text in Unicode is Universal Word—Hindi dictionary [18] is being made at Center for Indian Language Technology (CFILT), IIT Bombay for the purpose of Machine Translation. The user can search the Hindi and English words and phrases. This lexicon also provides the grammatical, morphological and semantic attributes of the Hindi words. This version contains 36,111 words and is good source of corpus. Encoding conversion may be required if data is acquired from other sources. C-DAC Noida has created Gyan Nidhi Corpus, which is parallel in multiple Indian languages. GyanNidhi contains 1 million pages of digitized data in Unicode format which contains variety of data from books published by national book Trust, India, Sahitya Akadmi, Navjivan publications, Publication division, Shri Aurobindo Ashram as they publish books of various domains, in most of the Indian languages. Mahatma Gandhi Hindi International University has commenced a project “Hindi Samgraha” on databases and dialect mapping of Hindi [19].

Indian Resources: Web Corpora for Indian Text

A Lexical Resource, “Syntax and Morphology in Hindi and Urdu” [20] is a searchable database with entries for about 60 verbs, part of a larger database project which is in progress. Each entry has fields for information about verb attributes, which for a specific entry define important properties which are projected into a sentence. This lexical information constrains the possible sentences formed from this verb: for example, the number of arguments which verb takes their category and grammatical function, and the case forms which are required or possible. Lexica are as critical for development of language computing as Corpora. Two available manually compiled English-Hindi electronic dictionaries have been identified.

Table 1. Amount of available resources in a number of sentences

| Language | News | Wikipedia | For comparison; EMILLE |
|----------|-----------|-----------|------------------------|
| Hindi | 5,162,167 | 727,882 | 469,395 |

First is the SHABDKOSH [21] and the second one is SHABDANJALI [22]. These two dictionaries have been merged automatically by replacing the duplicates. The merged English-Hindi dictionary contains approximately 90,872 unique entries. The positive and negative sentiment scores for the Hindi words are copied from their English SentiWordNet. The bilingual dictionary based translation process has resulted 22,708 Hindi entries [23]. In addition, English to Indian languages synsets are being developed under Project English to Indian Languages Machine Translation Systems (EILMT), a consortia project funded by Department of Information Technology (DIT), Government of India. For each language we have approximately 9966 synsets along with the English WordNet offset [23]. Hindi WordNet is a well structured and manually compiled resource and is being updated since last nine years. There is an available API [24] for accessing the Hindi WordNet [7].

Sentiment Lexicon for Hindi

Creation of linguistic data using SentiWordNet(s) for Indian languages are being developed using various approaches which can be used in areas of NLP too.

- WordNet(s) are available for Hindi (Jha et al., 2001) [25].
- (Joshi et al., 2010) [26] created H-SWN (Hindi-SentiWordNet)

using two lexical resources namely English SentiWordNet and English-Hindi WordNet Linking. Using WordNet linking they replaced words in English SentiWordNet with equivalent Hindi words to get H-SWN.

CHALLENGES IN DATABASE PREPARATION

The important issues involved in database preparation for development of various speech technologies are (i) creating a generic acoustic database that covers language variations and (ii) designing of the recording prompts and recording of speech databases to be used by corpus-based speech synthesizers. The problem arises in collection and selection of texts related with the application domains, the selection of appropriate speakers, all the necessary techniques such as recording setup for assuring and maintaining the same quality during the multiple recording sessions of the resulting database. The high cost of the recording process limits the ability in creating databases in more than a few voices for each domain specific application.

CONCLUSIONS

In this paper, a survey of efforts in database developments for Hindi language has been performed. It discusses some core linguistic resources of Hindi language, available through various resources developed for usage in text-to-speech synthesis and speech recognition technology. Despite the fact that there are tremendous challenges for building resources according to the global standards, there is immense potential for the development of language resources and technologies in India. If one needs to record his own database, he needs recording equipment (the higher quality, the better). A proper recording studio is ideal, though may not be available for everyone. A cheap microphone stuck on the back of a standard PC is not ideal. A high-quality sound board, close-talking and high-quality microphone and a nearly soundproof recording environment will often be the compromise between these two extremes. The high cost of recording process limits the ability of the technology providers to produce more than a few voices in a particular language. A solution to this problem has been conducted to separate the speech synthesizer from the inventory that defines the synthesizers' voices. Therefore, selection of the inventory of the recordings must be designed to provide good coverage of phonetics and phonation of the selected language, using analysis on available text corpora, mainly newspaper text, books etc for unrestricted TTS. For a restricted domain, domain adaptation is done at

speech inventory level. By selecting an inventory with carefully-selected sentences of a restricted domain, such as banking, health care, security, travel sector and others, a very high quality can be achieved for sentences in that domain.

It is suggested that the recordings from various resources can be grouped into application domains that can be combined to generate inventories which can be integrated with speech synthesizers to develop TTS and speech recognition applications. Unfortunately, many of the existing corpora or resources lack features that are strongly desirable for their uses in the scientific context. These shortcomings include problems with availability (in some cases the use of very specific interfaces is required), high costs or strict licenses that permit reuse and data aggregation. This paper identifies the distribution constraints, a challenge for open distribution, which needs to be addressed. As some of these problems can't be removed such as that of copyright, it would be beneficial to have more resources available electronically that can be used with fewer restrictions. This shall enable the participation of a larger group of institutions (within and outside of India) and the industry, as well as in research and development towards building speech systems in Hindi language.

REFERENCES

1. Dash, N.S. and Choudhary, B.B. (2011) Why Do We Need to Develop Corpora for Indian Languages? Proceedings of the International Conferences on SCALLA (Vol. 11), Bangalore.
2. Kishore, P., et al. (2012) The IIIT-H Indic Speech Databases. Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, 9-13 September 2012, 1-4.
3. Agrawal, S.S. (2010) Recent Developments in Speech Corpora in Indian languages: Country Report of India. Proceedings of O-COCOSDA 2010, Kathmandu, 25 November 2010.
4. Mukherjee, M. (1996) The First Computer in India. In: Banerjee, U., Ed., Computer Education in India—Past, Present and Future, Concept Publications, New Delhi, 13-16.
5. Sinha, M.K. (2009) A Journey from Indian Scripts Processing to Indian Language Processing. IEEE Annals of the History of Computing, 31, 8-31. <https://doi.org/10.1109/MAHC.2009.1>
6. Hindi Universal Word (UW) Dictionary. http://www.cfilt.iitb.ac.in/~hdic/webinterface_user/index.php
7. Rao, S. (2011) Application Prosody Model for Developing Speech System. International Journal of Speech Technology, 11, 2011.
8. Quasthoff, U., Mitra, R., Mitra, S., Eckart, T., Goldhahn, D., Goyal, P. and Mukherjee, A. (2012) Large Web Corpora of High Quality for Indian Languages. Proceedings of the 8th International Conference on Language Resources and Evaluation (LERC), Istanbul, 21-27 May 2012, 47.
9. Kurian, C. (2015) A Review on Speech Corpus Development for Automatic Speech Recognition in Indian Languages. International Journal of Advanced Networking and Applications, 6, 2556.
10. Arora, S., Saxena, B., Arora, K. and Agarwal, S.S. (2010) Hindi ASR for Travel Domain. Proceedings of O-COCOSDA 2010, Kathmandu, 25 November 2010.
11. Agrawal, S.S. (2010) Recent Developments in Speech Corpora in Indian Languages: Country Report of India. Proceedings of O-COCOSDA 2010, Kathmandu, 25 November 2010.
12. Linguistic Data Consortium for Indian Languages (LDC-IL). <http://www.ldcil.org/resourcesSpeechCorpHindi.aspx>

13. Samudravijay, K., Rao, P.V.S. and Agrawal, S.S. (2000) Hindi Speech Data. Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), Beijing, 16-20 October 2000.
14. Agrawal, S.S., Sinha, S., Singh, P. and Olsen, J. (2012) Development of Text and Speech Database for Hindi and Indian English Specific to Mobile Communication Environment. Proceedings of the International Conference on the Language Resources and Evaluation Conference (LREC), Istanbul, 21-27 May 2012.
15. The EMILLE Project (Enabling Minority Language Engineering). <http://www.emille.lancs.ac.uk/>
16. Hussain, S. (2008) Resources for Urdu Language Processing. Proceedings of the 6th Workshop on Asian Language Resources, Hyderabad, 11-12 January 2008, 99-100.
17. http://corpora.uni_leipzig.org
18. www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php
19. Arora, K., Arora, S., Verma, K. and Agrawal, S.S. Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages. Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju Island, 4-8 October 2004, 2885-2888.
20. Syntax and Morphology in Hindi and Urdu: A Lexical Resource. <https://clas.uiowa.edu/linguistics/hindi-verb-project>
21. Shabdkosh. <http://www.shabdkosh.com/>
22. <http://www.shabdkosh.com/content/category/downloads/>
23. Das, A and Bandyopadhyay, S. (2010) SentiWordNet for Indian Languages. Proceedings of the 8th Workshop on Asian Language Resources (ALR), Beijing, 21-22 August 2010, 1-8.
24. Hindi Wordnet. http://www.cfilt.iitb.ac.in/wordnet/webhwn/API_downloaderInfo.php
25. Jha, S., Narayan, D., Pande, P. and Bhattacharyya, P.A. (2001) WordNet for Hindi. Proceedings of the International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, January 2001.
26. Joshi, A., Balamurali, A.R. and Bhattacharyya, P. (2010) A Fall-Back Strategy for Sentiment Analysis in Hindi: A Case Study. Proceedings of the Fifth International Conference on Systems (ICONS), Menuires, 11-16 April 2010, 1-6.

SECTION 4

SOCIETAL AND ETHICAL ISSUES

CHAPTER 14

How AI-Human Symbiotes May Reinvent Innovation and What the New Centaurs Will Mean for Cities

Emmanuel Muller^{1,2,3}

¹University of Applied Sciences, Kehl, Germany.

²University of Strasbourg, Strasbourg, France.

³Fraunhofer ISI, Karlsruhe, Germany.

ABSTRACT

The aim of the paper is to propose a new hypothesis related to the contribution to innovation processes through human-AI symbiotes. These symbiotes are called centaurs, referring to the world of chess. The first section provides some speculative thoughts starting with current knowledge and observation on AI to point to the possible implication of symbiotic learning. The second section investigates what the consequences of the “centaur hypothesis” could be in terms of innovation capacities and innovation processes.

Citation: Stan, A., & Lőrincz, B. (2021). “Generating the Voice of the Interactive Virtual Assistant”. IntechOpen. doi: 10.5772/intechopen.95510.

Copyright: © 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The third section considers an atypical field of realization of innovations called municipal innovations so far. Finally, the conclusion addresses the limitations of this speculative exercise.

Keywords:- AI-Human Interactions, New Centaurs

INTRODUCTION: CHESS DOOMSDAY

On May 11th, 1997, IBM's Deep Blue became the first artificial intelligence (AI) to beat a human world chess champion. Garry Kasparov's defeat appeared to numerous observers, inside and outside the chess community, as the beginning of a new age, where the importance and self-perception of humans may clearly have changed, becoming nearer to zero, at least in terms of ego. To a certain extent, it was the death of a centuries-old conception of chess. One could continue to play, but knowing that some algorithms will beat him or her. That was the end of history—of chess—to paraphrase the title of the 1992 book by Francis Fukuyama.

Almost 25 years later, we are chatting, more or less successfully, with Siri, Cortana, and their fellow virtual friends and cannot wait for affordable self-driving cars. Go is usually considered the most abstract and complex board game; nevertheless, the spectacular performances of AlphaGo Zero in 2017 barely impressed the larger public and was definitely not a big surprise for most chess players. Apparently, winning games was over for humans. This realization leads to the question of what happened with the remaining human chess players. Is someone still really playing chess seriously or only out of boredom as chess would no more be the “game of the kings” but rather a sort of Monopoly or Cluedo? The reality check is striking; never before have so many humans played chess, and never before have humans played so well! Therefore, this is definitely not the end of the history of chess.

What happened is that there was a shift from a human versus machine paradigm to a human + AI paradigm. This shift allowed humans to get better at chess and AI, discovering new ways to learn. It is not only about human players benefiting from the “teaching” or “coaching” from superior and faster algorithms, neither is it just AI digging deeper in broader games library fed by better human versus machine games. Something very different and unexpected could be observed. Furthermore, seemingly, this did not only happen in the chess community but “contaminates” progressively more and more fields. We call this the rise of centaurs¹.

Listen to what Case (2018: p. 2) tells us about AIs, humans, and chess: “(...) in 1998, Garry Kasparov held the world's first game of ‘Centaur Chess’.

Similar to how the mythological centaur was half-human, half-horse, these centaurs were teams that were half-human, half-AI. But if humans are worse than AIs at chess, wouldn't a Human + AI pair be worse than a solo AI? Wouldn't the computer just be slowed down by the human, like Usain Bolt trying to run a three-legged race with his leg tied to a fat panda's? In 2005, an online chess tournament, inspired by Garry's centaurs, tried to answer this question. They invited all kinds of contestants—supercomputers, human grandmasters, mixed teams of humans, and AIs—to compete for a grand prize. Not surprisingly, a Human + AI Centaur beats the solo human. But—amazingly—a Human + AI Centaur also beats the solo computer." (original emphasis).

The aim of the paper is to provide some speculative thoughts about what the next episodes of this story may be, regarding particularly a field that, like chess, was considered for a long time as the prerogative of humans, *i.e.*, innovation. The paper resolutely follows a "what if" logic. It starts with current knowledge and observation on AI to point to the possible implication of symbiotic learning (Section 2). Then the paper investigates in a speculative way what the consequences of the "centaur hypothesis" could be in terms of innovation capacities (Section 3). In a third step, a so-far atypical field of realization of innovations is considered as an example (Section 4). Finally, the conclusion (Section 5) addresses the limitations of this speculative exercise.

UNDERSTANDING CENTAURS

What Are the New Centaurs?

Initially, centaurs were creatures featured in Greek mythology with the upper body of a human and a horse's lower body and legs. What if a "new kind of centaurs", a terminology being inspired by chess vocabulary, was currently rising? A kind of human + AI pair where the computer, or better say the exponentially growing network-based computer resources, is not slowed down by its human component but magnified by it? Or put differently, what if some humans could benefit from an Intelligence Augmentation (IA)? This question sounds like science-fiction, but in fact, it has already happened—and this seems to be only the beginning.

Far away from prophesizing the emergence of omniscient and omnipotent entities, we aim to understand the impact of these human + AI pairs within different fields of existing activities and investigate how far the development

of centaurs would notably affect innovation and cities. The first question to ask is a very prosaic one: is a “new centaur” something intrinsically different from a “virtual horse” driven and used by a human? For at least 3500 years, humans were successfully using horses for agriculture, traveling, warfare, and the like. Today, even if they are mostly used for leisure activities, our current mental representations of cars, farming, etcetera are still profoundly influenced by the initial way to mobilize an external source of power such as animals. Therefore, if AI is something other than just an additional “horse” like trucks, laser-cutting machines, or computers, such “horses” allow the multiplication of human physical and cognitive resources but remain only tools. This paper is based on the assumption that centaurs are intrinsically different from tools. This assumption is based on the observation of two phenomena: deep learning and symbiotic learning.

Deep Learning as the Basement of What Centaurs Could Become

Deep learning corresponds in reality to the result of the setting up and activation of deep neural networks, the usual academic name of deep learning. Deep neural networks are networks based on multiple layers between the input and output layers. Moving through the layers allows calculating each output’s probability; this, in turn, enables the modeling of complex non-linear relationships. In other words, it can be seen to a certain extent as a form of artificial autodidactic process. For instance, this makes possible an algorithmic self-teaching enabling the recognition of a dog after being fed thousands of labeled images of various animals. Parloff (2016) points out that currently, numerous medical startups claim they will soon be able to use a deep neural network to diagnose cancer earlier and less invasively than oncologists will.

According to Makridakis (2017) there are three reasons why one should be very optimistic regarding “technological limits” related to AI development:

- 1) Cumulative learning: since progress is available to practically everyone to utilize through Open Source software, researchers will concentrate their efforts on new, more powerful algorithms leading to cumulative learning;
- 2) Transposition of learning: deep learning algorithms will be capable of remembering what they have learned and apply it in similar but different situations;

3) Autonomous algorithms development: in the future intelligent computer programs will be capable of writing new programs themselves, initially perhaps not so sophisticated ones, but improving with time as learning will be incorporated to be part of their abilities.

As a result of the rapid increase in the complexity of individual AI, some authors point unexpected consequences. For instance, Rahwan et al. (2019) stress that some “black boxes” may emerge. According to them, although the code for specifying the architecture and training of an AI can be initially simple the results can be very complex. Inputs and produced outputs may be easily specified but the exact functional processes that generate these outputs may become harder and harder to interpret.

Symbiotic Learning as the Core Characteristic of the Nature of Centaurs

Symbiotic learning, according to our understanding, is even more revolutionary than deep learning. Symbiotic learning is not only a human whose capacities are boosted by an algorithm in terms of analytical capabilities, memory size, real-time access to sources, almost infinite information, and the like. In this case, one can refer to the concept of “intelligence augmentation”.

Pleading for an interdisciplinary study of machine behavior, Rahwan et al. (2019: p. 483) state that: “we shape machine behaviors through the direct engineering of AI systems and through the training of these systems on both active human input and passive observations of human behaviors through the data that we create daily.” In their analysis, these authors already consider several of the aspects that will be depicted later in this section in reviewing the following topics: 1) mechanisms for generating AI behaviors; 2) functions fulfilled by the emergence of AI behaviors; 3) evolution of AI behaviors (phylogeny); 4) individual AI behaviors; and 5) collective AI behaviors. These aspects led them to address the final issue of hybrid human-AI behaviors. In this respect, Rahwan et al. (2019) consider the identification of factors that can facilitate trust and cooperation between humans and machines as crucial for future investigations.

Jennings et al. (2014) call Human-Agent Collectives or HACs: “HACs are a new class of socio-technical systems in which humans and smart software, agents, engage in flexible relationships to achieve both their individual and collective goals. Sometimes the humans take the lead, sometimes the computer does, and this relationship can vary dynamically.” (Jennings et al., 2014: p. 80). Nevertheless, the most crucial difference is that while Jennings

et al. (2014) consider HACs as a form of agile teaming where humans and agents will form short-lived teams before disbanding, it is assumed here that centaurs constitute a permanent and symbiotic relationship. This symbiotic relationship between humans and AI is the very core of the nature of centaurs and results from the three steps of the symbiotic learning process displayed hereafter.

In a first step, the human part of the symbiote teaches, guides the AI, and encourages it in their curiosity by confronting them with new issues, like parents try to do with their children when raising them. In other words, the human part is schooling the AI to allow the AIs' creativity to flourish². Consequently, two initially identical AIs will rapidly—remember: deep learning is high-speed learning—diverge, depending with whom they are “growing up” like it is the case for human, real, twins separated as they are still very young and are growing up in very different families, social environments, countries, etcetera. Alison Gopnik summarizes the situation this way (quoted by Guscza et al., 2017: p. 16): “one of the fascinating things about the search for AI is that it’s been so hard to predict which parts would be easy or hard. At first, we thought that the quintessential preoccupations of the officially smart few, like playing chess or proving theorems—the corridas of nerd machismo—would prove to be hardest for computers. In fact, they turn out to be easy. Things every dummy can do, like recognizing objects or picking them up, are much harder. And it turns out to be much easier to simulate the reasoning of a highly trained adult expert than to mimic the ordinary learning of every baby.”

In a second step, the IA part modifies the way of thinking of the human part of the symbiote, like the human part of a chess centaur tends progressively to play differently, even when not connected to its own, AI. This step means its human view of reality—remember reality is nothing else than a cognitive and social construct—evolves radically over time, even if it is most probably at the same pace as the AI part of the symbiote. In other words, since the human part learns to think differently and progressively sees the world from a different perspective, its personality and identity change. This change would imply a form of psychic plasticity of centaurs in the meaning of Tisseron (2018), who explores the psychological dimensions of future human-machine interactions. This author points to the possibility of two distinct stages. The second stage would see humans and AI entering (from a psychological perspective) in an adult-to-adult relationship, what he calls poetically “amitié informée et réaliste” (Tisseron, 2018: p. 13). This stage can be seen as the prolongation of an initial phase (an adult-to-child

relationship) during which the AI learns mainly through imitation processes. According to Tisseron (2018) during this initial stage, the AI would rather (metaphorically) act like a young child sensitive to a reward.

Finally—this may sound more speculative—in a third and ultimate step, it could be envisaged that centaurs will communicate not only with humans and AIs on separate channels. Separates channels mean each element, human or AI, of the symbiote exchanging information exclusively with their counterparts, humans, or AIs. Put in other words, this would mean humans are chatting together in one corner of the party and AIs chatting together in another one³. This kind of interaction may even be reinforced by the relationships between centaurs, high-level AIs, and low-level AIs (Kelly, 2016). Again, this type of communication will most probably be resolutely different depending on the interlocutors of low-level IAs—*i.e.*, high-level AIs, humans, or centaurs.

The assumptions related to these three steps may sound surprising, if not exaggerated, or even foolish. When would centaurs become an everyday reality, being, for instance, spread as Watson is today? No definite answer is possible, but it can be reasonably expected that it will take less than 20 years from now. This actuality requires “only” three conditions. The first is reaching a higher level of AI development. The second condition is to render possible a better integration of AI and humans in terms of communication bio-interfaces, leading to a high level of symbiosis. The third condition is to allow some human + AI pairs to grow up together as individual entities for the first time. These three conditions would allow the emergence of shared identity through mutual learning based on real experiences (*e.g.*, surgical operations). It is not possible yet to determine which of these steps will take the most time, notably since numerous feedback loops between these steps are expected. Nevertheless, a horizon of fewer than 20 years from now may seem realistic considering Kelly’s (2016) predictions.

Things can happen much faster and to a larger scale than optimistic expectations! Interestingly, in a paper addressing what he calls “the forthcoming AI revolution” and its impacts on society and firms, Makridakis (2017) overviews the predictions he made about information technologies more than ten years before (Makridakis, 1995). Besides identifying successes and failures of his predictions, he stresses “that major technological developments (notably the Internet and smartphones) were undervalued while the general trend leading up to them was predicted correctly” (Makridakis, 2017: p. 47).

CENTAURS AND INNOVATION

Problem Solving and Decision-Making: Playing According to the Existing Rules

Far away from apocalyptic visions concerning the future of work, Chui, Manyika, & Miremadi (2015) suggest that the growing use of AIs in the economy is more likely to transform, rather than eliminate, jobs. Today, there is a growing consensus that it is important to distinguish “task” automation from “job” automation. Markoff (2016) points out that AI technologies will most probably continue to replace routinized jobs and, at the same time, will increase the number of workers whose jobs require problem-solving, flexibility, and creativity. One could imagine that in a near future, the—boring—jobs requiring light-speed computation will be for AIs and the—exciting—creative, innovative, and valorizing jobs for the human. Nevertheless, this vision remains quite “classical” since it depicts a dichotomy: fast and routinized tasks will be for AIs. Human-speed and innovative jobs will remain in the field of highly qualified humans. Reality will most probably be somehow less contrasted in this respect. However, if one accepts the hypothesis of the rise of centaurs, the main issue to address is how far centaurs will be able to innovate differently, and what will this difference be?

AI and other intelligent technologies can assist human decision makers with predictive analytics as stated by Jarrahi (2018). In particular, they can generate fresh ideas through probability, and data-driven statistical inference approaches. Moreover AI can enable human decision makers to collect more effectively information in order to feed predictive analytics processes. If one comes back to the history of chess centaurs, Cage (2018: p. 5) provides a striking argument: “there was another shock in store for Garry Kasparov. Remember that 2005 online chess tournament between supercomputers, human grandmasters, and Human + AI centaurs? I forgot to mention who actually won the grand prize. At first, Garry wasn’t surprised when a human grandmaster with a weak laptop could beat a world-class supercomputer. But what stunned Garry was who won at the end of the tournament—not a human grandmaster with a powerful computer, but rather, a team of two amateur humans and three weak computers! The three computers were running three different chess-playing AIs, and when they disagreed on the next move, the humans “coached” the computers to investigate those moves further.” As Garry put it: “Weak human + machine + better process was

superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process.” (original emphasis).

What DeepBlue, AlphaZero, and all their friends have in common is that they did develop radically new ways of playing and winning, respecting strict fixed rules. For instance, a 64 squares world where every “actor” functions and possibilities are perfectly known. However, this relates only to decision making, finding the best way to “win the game” according to a given set of rules. In decision-making processes, most of the options are relatively well known even if specific options may, for instance, encompass a high level of uncertainty. Consequently, it can be assumed that in a situation where “innovating” consists of improving something already existing, *i.e.*, incremental innovations, centaurs appear as much “efficient”, *i.e.*, faster and more exhaustive than humans alone.

What Can Centaurs Achieve that AIs Can Not?

The next logical step is to ask the question: what about AI + human symbiotic playing “real-life games”? Games without fixed rules? Alternatively, with changing rules, either resulting from a stochastic process or from the results of previously “won or lost games”. Or even games with contradictory rules? What about situations consisting of exploring the unknown or situations that imply being creative? In other words, what could centaurs achieve that neither humans nor AIs alone could achieve?

Reviewing the literature on AI, Huang & Rust (2018) distinguish four types of “intelligences” where machine intelligence mimics human intelligence dimensions, such as knowledge and reasoning, problem-solving, learning, communicating, perceiving, and acting. As a result, these authors propose four stages related to the developmental history of AI intelligence:

- 1) Mechanical intelligence, *i.e.*, learning and adapting at the minimum;
- 2) Analytical intelligence, *i.e.*, learning and adapting systematically based on data;
- 3) Intuitive intelligence, *i.e.*, learning and adapting intuitively based on understanding;
- 4) Empathetic intelligence, *i.e.*, learning and adapting empathetically based on experience.

The analysis of Huang & Rust (2018) is mainly focused on the potentials and threats in job replacement in the service sector. Nevertheless, if one

considers the hypothesis of the emergence of centaurs, it must be stressed that these authors do not assert that only the “worst-case scenario” will take place. Total replacement is not the only logical final step since integration is also thinkable. In particular, they point to what they describe as “machine-enhanced humans. In this possibility, humans are physically or biologically integrated with machines, and AI becomes a technological extension of humans. (...) one possibility for AI is ‘beyond human,’ which adds human bio-enhancements, prosthetics, or implants” (Huang & Rust, 2018: p. 165).

Jarrahi (2018) points that the problem-solving abilities of AI are more useful for supporting analytical processes performed by human rather than their intuitive decision-making. In fact, much of cognition and human decision-making is not a direct result of deliberate information gathering and processing, but instead arises from the subconscious in the realm of intuition. Similar thoughts can be find in Kahneman (2011) distinguishing between fast (intuitive) and slow (analytical) thinking.

In the academic literature, such capabilities—particularly when linked to the issue of innovation—are often summarized under the “conceptual umbrella” of creativity. Sternberg & Lubart (1998: p. 3) proposed what became one of the most widely accepted definitions of creativity in this respect: “the ability to produce work that is both novel (*i.e.*, original, unexpected) and appropriate (*i.e.*, useful, adaptive concerning task constraints).” Regarding creativity, there are at least three attributes related to innovation processes for which centaurs may be superior to AIs alone. The conception of these attributes is partly inspired by reflections proposed by Dewhurst & Willmott (2014). They address the issue of the role senior leaders should still play with the emergence of AIs, implying that they can only play such roles better than AIs. These three attributes are: asking questions, tolerating ambiguity, employing soft skills, and considering ethical aspects.

First Attribute: Asking Questions

Starting not only with a willingness to improve processes, cutting costs, expanding markets, and the like but with a willingness to ask good questions, or at least new questions. This generation may be seen as having new problems rather than the production of new solutions. It goes beyond deep learning. It is not a matter of advanced analytics but a matter of “deep curiosity”.

Second Attribute: Tolerating Ambiguity

Algorithms are designed to seek answers. Deep learning is about producing an almost infinite number of mistakes to get better answers.

Tolerating ambiguity means considering or even keeping solutions that prove not to be the right ones. Nevertheless, these solutions might provide a good, or at least an acceptable, answer to another question, which may not even be formulated or to the current question, but not under the given conditions, in terms of resources, design, aims, and the like.

Third Attribute: Employing Soft Skills and Considering Ethical Aspects

Due to their human part, centaurs may prove more efficient than AIs when it comes to motivating investors, improving project partners' creativity, or empathizing with recalcitrant clients. Introducing a human touch in critical and sometimes not entirely rational situations may generate a real difference, which also applies to ethical issues. It should not be expected from an AI to act in full consciousness—in the philosophical meaning—like a human should act. For instance, a centaur interacting in a crucial project with a human using harmful substance to improve his creativity, which may be a wrong statement, faces an ethical dilemma. Whereas an AI alone would most probably focus exclusively on the project results and not care about the user of harmful substances, a centaur's reaction might be very different⁴.

Summarizing, these three attributes may enhance creativity in the meaning given by Sternberg & Lubart (1998), creativity is the ability to produce both novel and appropriate work. Nevertheless, this does definitively not mean that centaurs would become “all-mighty” and “omniscient”. Cognitive biases would still hamper their abilities, e.g., “slow thinking/fast thinking in the sense given by Kahneman (2011).

Nevertheless, these cognitive biases would most probably be different from the ones hampering AIs taken alone.

In their paper nicely entitled “Lessons for artificial intelligence from the study of natural stupidity”, Rich & Gureckis (2019: p. 179) point out that “science and technology often advance through inspiring metaphors. Some of the recent interest in machine learning and AI stems precisely from the comparison between machines and humans and the idea that machine-based systems implement aspects of human cognition but improve on human abilities. (...) A healthy attitude towards recent advances in AI would be to recognize that rather than being free of bias, certain biases are likely to be fundamental to what it means to be an intelligent adaptive agent operating in a vague and uncertain world”.

Expanding the Playbook: Innovation as the Invention of New Rules

The core question can be expressed as follows: can centaurs innovate differently from humans alone, even supported by powerful computers, or from AIs, *i.e.*, when humans set the goals and “explain the rules”? In other words: Can we talk about symbiotic innovations that can be performed exclusively or at least mainly by centaurs? Or would centaurs appear to better innovate in fields where non-centaurs, *i.e.*, humans or AI, have taken alone, seem limited?

The opportunity for a partnership was already pointed by Jarrahi (2018) considering that one way to materialize the synergistic relationship between AI and humans is to combine the speed of AI in collecting and analyzing information with humans’ superior intuitive judgment and insight.

The primary argument pleading in favor of a supremacy of symbiotic innovations is to consider situations where “new rules” must be invented for “real-life games”, which do not exist so far. This argument may apply to products, services, or processes. Depending on how different the existing ones are, the new rules are incremental, slightly modified rules or radical, significantly different rules innovations.

In how far can centaurs’ abilities to “win games” be extended to real-life settings where not only the rules are not fixed, but where rules can change over time or depending on the context? How far can capabilities such as creativity and sensing emotions, the core to the human experience, be automated?

Centaurs inventing “new rules” could also be interpreted as new ways to find creative solutions, resulting from what one could call augmented serendipity. Yaqub (2018) proposes a typology describing four serendipity processes leading to creative solutions: 1) targeted search solving unexpected problems; 2) targeted search solving problem-in-hand via unexpected routes; 3) untargeted search solving an immediate problem; and 4) untargeted search solving a later problem.

For each of these four types, it appears clearly that symbiotic learning—being the core characteristic of centaurs—would constitute a tremendous accelerator of serendipity. Yaqub (2018: p. 173) states that: “Observations are usually mediated by instruments, and the development and use of instruments themselves play an important role in serendipity. This is not necessarily the testing of theories nor the replication of experiments, but

rather the trying out of new practices. (...) Instruments can be developed and used quite free from theory, playfully even." Centaurs would, thanks to their dual nature, play at the same time the role of the instrument and the role of the observer with high velocity.

CENTAURS AND THE CITY

Municipal Innovations: Unspectacular but Crucial

Cities appear as major economic and political actors of the twenty-first century. This development is due to demographic factors and the concentration of geostrategical and environmental issues in cities, particularly climate change. Moreover, cities seem to be the place *par excellence* of innovation; Wolfe (2014) names cities "Schumpeterian hubs". In parallel, the term "smart cities" emerged progressively in the 1990s. The concept has become increasingly popular in scientific literature and international policies. According to Albino et al. (2015), the California Institute for Smart Communities was among the first to focus on how communities could become smart and how a city could be designed to implement information technologies. Over the past 20 years, the smart city concept has had many definitions, with smart cities being places where information technology is combined with infrastructure, architecture, everyday objects, and our bodies to address social, economic, and environmental problems. More recently, authors started even to investigate AI clusters in cities. See Doloreux & Savoie-Dansereau (2019) for the case of Montreal.

Municipalities are usually not considered initiators of innovation. Consequently, one dimension of the interrelationships between innovation and cities was given relatively little attention so far: cities themselves, or more precisely municipal teams, being the innovators. Shearmur & Poirier (2016) were the first to attempt to conceptualize this specific form of innovation. They see municipal innovations as "non-market Schumpeterian innovation processes". Shearmur & Poirier (2016) state that municipalities are required to introduce incremental product, process, and service innovations in order to address issues that result from of everyday service and management responsibilities. A broad spectrum of municipal innovations' examples displayed stretches from biomethanation to environmental patrol through waterways management is provided by Shearmur (2020).

The Potential Specific Contribution of Urban Centaurs to Municipal Innovations

This section aims to link, focusing on municipal innovations between different concepts developed above. Shearmur & Poirier (2016) state that municipalities' internal capacity determines their innovativeness. This can serve as a starting point for imagining what future contributions urban centaurs could deliver for municipal innovations. In this respect, two dimensions of centaurs' behavior are particularly relevant for learning, motivation, and evaluation.

The first dimension concerns how centaurs act and respond to their environment, *i.e.*, following a rivalry or cooperation logic. The second relates the type of playbook upon which centaurs rely. Figure 1 illustrates the differences resulting from the combination of these two dimensions.

Chess centaurs are typically following a rivalry logic related to humans, AI, or other centaurs and using a closed playbook, *i.e.*, acting in a limited and well-defined universe. They aim to win against others in respecting given exceptionally well-defined rules. Surgery centaurs are supposed to follow cooperation logic with other actors to improve their patients' state of health. On the opposite, legal centaurs, acting for instance as lawyers, consider an open playbook to interpret law following a rivalry logic.

Comparatively, urban centaurs could be characterized as:

- 1) Following cooperation logic: They intend to improve through municipal innovations, the situations of the concerned cities in which they are involved, ultimately attempting to increase the level of well-being of the inhabitants;
- 2) Considering an open playbook: Their actions are not limited to specific fields, nor must they strictly follow rules which were defined *ex-ante*;

Consequently, and at least hypothetically, urban centaurs could reinforce cities' innovativeness more than humans or AIs alone. In particular, when considering the incremental product, process, and service innovations, as Shearmur & Poirier (2016) do, one may state that urban centaurs could support the emergence of quantitatively more numerous municipal innovations. Urban centaurs could also qualitatively sustain more creative municipal innovations. Both effects would strongly reinforce municipalities' internal capacity to innovate in the meaning given by Shearmur & Poirier (2016).

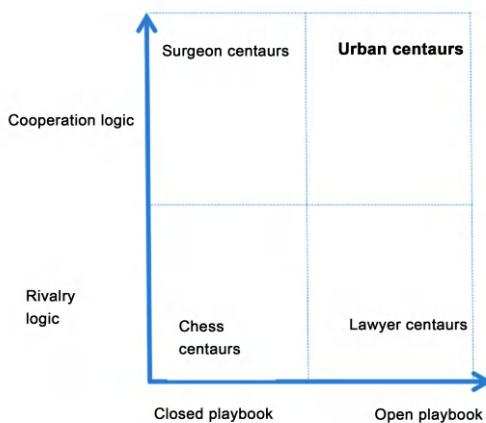


Figure 1. Characterization of urban centaurs along two main dimensions.

Suppose one accepts the idea that potentially soon urban centaurs may be part of municipalities' staffs. In that case, it is essential to consider their contribution to cities' innovativeness from an organizational perspective. Urban centaurs will interact with humans—co-workers, citizens, and the like. These centaurs will also interrelate with external organizations—suppliers, different administrations, other municipalities, and so forth. In addition, centaurs will cooperate with other AIs, different types of centaurs, and so on. Put in other words: urban centaurs embedded in municipal organizations would mean at the same time more innovations and better innovations.

From an organizational perspective, several arguments can be found which support the hypothesis of reinforcement of both quality and quantity of municipal innovations by urban centaurs. In the following, three reasoning lines dealing with the generation of new) knowledge and innovations are presented. For each argumentative set, elements of reflection related to urban centaurs are introduced.

The first argumentative line is based on the combination between the exploration/exploitation trade-off proposed by March (1991) and the definition of creativity proposed by Sternberg & Lubart (1998: p. 3). Creativity is defined as the ability to produce work that is both novel, *i.e.*, original, unexpected, and appropriate, *i.e.*, useful, adaptive concerning task constraints. Two types of municipal innovations increase can be identified; first, a quantitative increase of municipal innovations—which would mean better exploitation in the meaning given by March—and a higher level of appropriateness regarding constraints in the meaning given by Sternberg

and Lubart. Second, a qualitative increase of municipal innovations would constitute a more in-depth exploration of the meaning of March combined with a stronger originality of problem solutions in the meaning given by Sternberg and Lubart.

The second set of arguments follows the concept of phronesis in organizations developed by Nonaka et al. (2014). Phronesis is sometimes presented as the “third type of knowledge” since it is a form of practical wisdom, which goes beyond explicit and tacit knowledge since it cannot be taught in Socrates and Plato’s views. Phronesis can only be generated by dialectic processes and, according to Nonaka et al. (2014), from an organizational perspective. One hypothesis would be that centaurs, due to their dual nature as symbionts, could reinforce the development of practical wisdom within municipalities, strengthening the ability to generate what Shearmur & Poirier (2016) stress as “everyday innovations”.

The third argumentative line concerns the Spatio-temporal knowledge creation processes, as presented by Hautala & Jauhainen (2014). According to them, “knowledge is inseparable from the temporal processes of creation, interaction and interpretation as well as from contexts, or spaces, of creation” (Hautala & Jauhainen, 2014: p. 655). In this approach, knowledge creation appears as profoundly interactive, as other people and the environment affect individuals’ thoughts and actions. Besides, Hautala & Jauhainen (2014) state that, when it comes to knowledge creation, considering space only as a material background and time only as universal linear sequences is misleading. In their view, knowledge creation results from a reorganization of spatiotemporal processes. This reorganization is the key to reinforced innovativeness notably in academia, business, and local communities according to Hautala & Jauhainen (2014). In this respect, urban centaurs may appear, due to their symbiotic nature, as “anchored” in different places and timeframes simultaneously. This type of anchoring makes a vast difference with humans alone—one place at a time and own perception of time different from that of one of AIs, and AIs alone—virtually “present” at several places at a time and with their computational speed resulting in the apprehension of time different from human experience. As a result, the possible answers to the very questions of “where” and “when” in knowledge creation by urban centaurs are profoundly modified. This result, in turn, leads to a reorganization of spatiotemporal processes for the municipalities embedding urban centaurs in their innovation-related activities.

Urban centaurs, summarizing, from an organizational perspective, may foster at the same time more innovations, based on more efficient exploitation of knowledge, and better innovations—based on a more profound exploration of knowledge. In other words, the innovativeness of municipalities embedding centaurs may reinforce both in terms of incremental innovations as well as in terms of radical innovations. Here, it is necessary to stress what “incremental” and “radical” innovations mean for municipalities. Municipal innovations are distinct from companies’ innovation since their non-market nature (Shearmur & Poirier, 2016). Incremental innovations may be easier to develop since adaptation from other municipalities’ experiences is supported by a degree of willingness to disclose knowledge through cooperation that cannot be found when it comes to firms in a situation of competition. Simultaneously, radical innovations are strongly context-specific for municipalities and may appear modest compared to radical innovations performed by firms acting in a global market. The following section proposes examples of municipal innovations supported by urban centaurs.

Some Examples of Possible Contributions of Urban Centaurs to Municipal Innovations

The development of municipal innovation supported by urban centaurs could correspond mainly to situations where solutions are found for problems corresponding to a contradiction. In other words, urban centaurs would contribute to distinguishing within the existing corpus of knowledge what could belong to problems and what could belong to solutions in order to ensure possible matches. The issue is then not only to “generate good solutions” but to a certain extent also to “find good problems”. The ground hypothesis is that this is quite often difficult to realize using only limited human computational abilities or some limited, or even nonexistent, AIs contextualization capacities or intuition.

Table 1 depicts a few examples of fields in which such municipal innovations could be implemented or supported by urban centaurs. These fields are displayed along two dimensions: the objectives of the concerned innovations and their nature.

The nine fields are given as examples to address issues to which almost all cities are or will be confronted regardless of their size, location, or socioeconomic profiles. In each field, it can be assumed that the efforts

currently deployed at the municipal level are insufficient. One may assert that the shortage of financial resources or lacking political constitute potential obstacles but the inherent complexity of the issues addressed strongly hampers the emergence of solution. It is mainly the contradictory nature of those problems that constitute the core difficulty. In this respect, urban centaurs—being at the same time animated by cooperation logic and following an open playbook—are liable to favor the emergence of solutions.

These solutions can be pointed when detecting some common patterns of the nine fields given as examples. At least five common characteristics can be highlighted. First, partial solutions already exist, being technology-based or not, and are deployed at different scales with divergent degrees of success. Simultaneously, the partial elements of the solution are difficult to reproduce since contexts are different from one city to the other, which is, for instance, the case for the detection and prevention of leaks. Second, the combination of high computational velocity and perceived likelihoods in population willingness appears to be the key to success. In particular, this willingness in the case for the development of drone fleets-based enhanced data collection seems problematic in terms of citizens' acceptance. Third, the need for initial creativity followed by numerous experimentations, showing possibly contradictory results. The deployment of such innovations would require numerous trial and error sequences and would elsewhere reveal too time and cost consuming. For instance, this would concern air pollution tracking and epidemic detection.

Table 1. Nine examples of municipal innovations possibly supported by urban centaurs

| Nature of the innovations objectives of innovation | Monitoring and detection of patterns | New combinations of resources, actors, experimentation and the like | Identification and adaptation of solutions existing "elsewhere" |
|---------------------------------------------------------------------|-------------------------------------------------|-------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Improving sustainability and solving environmental issues | Waste management and recycling | Drone fleets based enhanced data collection made acceptable to citizens | Air pollution tracking and epidemic detection |
| Improving the efficiency of physical and intangible infrastructures | Detection and prevention of leaks (e.g., water) | Implementation of data squads and maintenance of data islands | Dynamic management and improvement of multi-modal transportation systems |

| | | | |
|----------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------|--------------------------------------------------------------------------------------|
| Improving citizens' well-being and solving social issues | Urban and architectural design supporting inclusive tourism | Real-time homeless supervision and psychological care | The conception of urban solutions likely to meet the expectations of bored teenagers |
|----------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------|--------------------------------------------------------------------------------------|

Fourth, the resolution of conflicts is carried inherently by the emerging solutions, conflicts being financial resources, legal obstacles, ideological settings, and the like. Real-time homeless supervision and psychological care could provide an example.

Fifth, the ability to identify and motivate different types of actors that do not know each other, are unwilling to cooperate, or are not familiar with the concerned field of innovation. This innovation could concern, for instance, urban and architectural design supporting inclusive tourism.

The selection of the fields contained in Table 1 is naturally extraordinarily subjective, and the examples provided are not intended to constitute proofs, nor may be interpreted as the results of a foresight exercise. This choice was led by the willingness to explore a broad scope of issues that a municipality is potentially confronted with daily, encompassing various degrees of urgency and complexity⁵. The exercise aimed to illustrate the diversity of the problems that may be addressed and hopefully solved in a not too far future with urban centaurs' help.

CONCLUSION: WHO WOULD EVER WANT TO BE A CENTAUR?

This paper is highly speculative and resolutely optimistic. The ideas developed above were ignited by discussions with chess players and strongly influenced by the well-known statement by Kelly (2016) stressing that we should not start a race “against the machines” but a race “with the machines”. Speculations about a hypothetical rise of centaurs may raise numerous issues, depending on if and how this would at least partly happen. If the hypothesis would prove to be even only partially true, then numerous challenges would appear both for managers and policymakers. In particular, how to favor the emergence/the retention/the attraction of centaurs in a given company or geographical area?

Nevertheless, as long as empirical investigations are not possible, one must keep in mind the strongly speculative character of the above-developed ideas. The “centaur hypothesis” presented here carries definitely some fictional and even hazardous features. As Hermann (2020: p. 654) stresses

in a paper perfectly entitled “Beware of Fictional AI Narratives”, it seems evident that “taking the SF (science-fiction) representation of conscious and autonomous machines seriously as a critical technology assessment gives a distorted impression of the capabilities of AI in reality”.

Consequently, it is essential to consider the limits of what can be expected in terms of the development of AI capacities and not become overconfident in what may mainly result from imagination. Nevertheless, in the novel “The Salmon of Doubt”, Douglas Adams proposes a somehow alternative way of thinking (Adams, 2003: p. 95),

“I’ve come up with a set of rules that describe our reactions to technologies:

- 1) Anything that is in the world when you’re born is normal and ordinary and is just a natural part of the way the world works;
- 2) Anything that’s invented between when you’re fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it;
- 3) Anything invented after you’re thirty-five is against the natural order of things”.

ACKNOWLEDGEMENTS

I discovered the existence of chess centaurs thanks to my chess coach Florian Daeschler (University of Strasbourg) who patiently tried to improve my abilities to play the game of the kings (unfortunately with little success). It was my good fortune to spend some time discussing the original ideas beyond this paper with my colleagues and friends David Doloreux (HEC Montréal) and Richard Shearmur (University McGill) during a journey devoted to the empirical investigation of vineyards and wine cellars in the Upper Rhine. Finally, I owe my sons Luc and Marc for constantly challenging my intellectual capacities trying to convince them (unfortunately with little success) that I am not thinking like a boomer.

The usual disclaimers apply.

NOTES

It seems that the idea of centaur chess playing emerged for the first time in the science-fiction novel *The Peace War*, written by Vernor Vinge and published in 1984. Interestingly Vinge (1981) was also the first author to

introduce in his novella *True Names*, the concept of cyberspace, and that, three years before the publication of *Neuromancer*, the well-known novel by William Gibson, and almost ten years before the World Wide Web was developed by CERN. This information is nevertheless anecdotal, and the fact that since more than a century visions provided by science-fiction writers outsmart almost systematically the predictions of serious foresight analysts only proves that science-fiction writers tend to be very lucky.

²The analogy with schooling encompasses certain limitations. In the case of human children, teenagers, and young adults, the educational system, from kindergarten to university, is supposed to do the same: improving learning and creative capacities. Nevertheless, empirical observations quite often just show the opposite since educational systems appear as perfectly efficient in killing creativity. Cf. notably an excellent TED conference given in 2014 by the late Sir Ken Robinson: https://www.ted.com/talks/sir_ken_robinson_do_schools_kill_creativity?language=enW

³The idea of some AIs chatting together seems very unlikely to most people who believe that, for instance, AIs have no sense of humor. This argument may nevertheless reveal fallacious since numerous humans seem to be totally deprived in this respect and unfortunately do not refrain from chatting.

⁴Thanks to its human component, the centaur would, for example, empathize with the concerned human; suggest to him to take more cocaine in order to achieve the project in due time, and report him to the authorities only afterward, which would allow the centaur to remain fully ethical, especially if the considered drug addict appears to be antipathetic and/or useless in the future.

⁵For instance, the crucial issue of bored teenagers for municipalities (as pointed by Shearmur & Poirier, 2016: p. 23) reveals the high degree of complexity of certain situations for the people in charge. Confronted with a species' behavior that defies the capacities of both humans and today's AIs, one may hope that centaurs will be able to ensure some signs of progress regarding teenagers. One optimistic view would be to state that being able to understand them better could also improve the ability of humanity to communicate with further hypothetical alien forms of intelligence. In this respect, the author is grateful to his son Marc for the provision of empirical material.

REFERENCES

1. Adams, D. (2003). *The Salmon of Doubt: Hitchhiking the Galaxy One Last Time*. Ballantine Books.
2. Albino, V., Berardi, U. U., & Dangelico, R. M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, 22, 3-21. <https://doi.org/10.1080/10630732.2014.942092>
3. Case, N. (2018). How to Become A Centaur. <https://doi.org/10.21428/61b2215c>
4. Chui, M., Manyika, J. Miremadi, M. (2015, November). Four Fundamentals of Workplace Automation. *McKinsey Quarterly*.
5. Dewhurst, M., & Willmott, P. (2014, September). Manager and Machine: The New Leadership Equation. *McKinsey Quarterly*.
6. Doloreux, D., & Savoie-Danserau, G. (2019), L'émergence de la grappe industrielle de l'intelligence artificielle (IA) à Montréal. *The Canadian Geographer/Le Géographe canadien*, 6, 440-452. <https://doi.org/10.1111/cag.12525>
7. Guszcza, J., Lewis, H. H., & Evans-Greenwood, P. (2017). Cognitive Collaboration: Why Humans and Computers Think Better Together. *Deloitte Review*, 1-24.
8. Hautala, J., & Jauhainen, J. (2014). Spatio-Temporal Processes of Knowledge Creation. *Research Policy*, 43, 655-668. <https://doi.org/10.1016/j.respol.2014.01.002>
9. Hermann, I. (2020). Beware of Fictional AI Narratives. *Nature Machine Intelligence*, 2, 654. <https://doi.org/10.1038/s42256-020-00256-0>
10. Huang, M.-H., & Rust, T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21, 155-172. <https://doi.org/10.1177/1094670517752459>
11. Jarrahi, M. (2018). Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons*, 61, 577-586. <https://doi.org/10.1016/j.bushor.2018.03.007>
12. Jennings, N.R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., & Rogers, A. (2014). Human-Agent Collectives. *Communication of the ACM*, 57, 80-88. <https://doi.org/10.1145/2629559>
13. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

14. Kelly, K. (2016). *The Inevitable: Understanding the 12 Technological Forces that Will Shape Our Future*. Viking.
15. Makridakis, S. (1995). The Forthcoming Information Revolution: Its Impact on Society and Firms. *Futures*, 27, 799-821. [https://doi.org/10.1016/0016-3287\(95\)00046-Y](https://doi.org/10.1016/0016-3287(95)00046-Y)
16. Makridakis, S. (2017). The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. *Futures*, 90, 46-60. <http://dx.doi.org/10.1016/j.futures.2017.03.006>
17. March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2, 1-147. <https://doi.org/10.1287/orsc.2.1.71>
18. Markoff, J. (2016). *Machines of Loving Grace: The Quest for Common Grounds between Humans and Robots*. Harper Collins.
19. Nonaka et al. (2014). Organizational Knowledge Creation Theory: A First Comprehensive Test. *International Business Review*, 3, 337-351. [https://doi.org/10.1016/0969-5931\(94\)90027-2](https://doi.org/10.1016/0969-5931(94)90027-2)
20. Parloff, R. (2016, September 29). Why Deep Learning Is Suddenly Changing Your Life. *Fortune*. <https://fortune.com/longform/ai-artificial-intelligence-deep-machine-learning/>
21. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C. et al. (2019). Machine Behavior. *Nature*, 568, 477-486. <https://doi.org/10.1038/s41586-019-1138-y>
22. Rich, A., & Gureckis, T. (2019). Lessons for Artificial Intelligence from the Study of Natural Stupidity. *Nature Machine Intelligence*, 1, 174-180. <https://doi.org/10.1038/s42256-019-0038-z>
23. Shearmur, R. (2020). Municipalities and Sustainability: What Is Municipal Innovation and Can It Make a Difference? In H. Kong, & T. Montforte (Eds.), *Innovations in Urban Sustainability: Citizens and Participatory Governance* (Forthcoming). University of Toronto Press.
24. Shearmur, R., & Poirier, V. (2016). Conceptualizing Nonmarket Municipal Entrepreneurship: Everyday Municipal Innovation and the Roles of Metropolitan Context, Internal Resources, and Learning. *Urban Affairs Review*, 53, 718-751. <https://doi.org/10.1177/1078087416636482>

25. Sternberg, R., & Lubart, T. (1998). The Concept of Creativity: Prospects and Paradigms. In R. Sternberg (Ed.), *Handbook of Creativity* (pp. 3-15). Cambridge University Press. <https://doi.org/10.1017/CBO9780511807916.003>
26. Tisseron, S. (2018). *Petit traité de cyber-psychologie*. Le Pommier.
27. Vinge, V. (1981) True Names and the Opening of the Cyberspace Frontier. Tor Books
28. Wolfe, D. (2014). Innovating in Urban Economies: Economic Transformation in Canadian City-Regions. University of Toronto Press. <https://doi.org/10.3138/9781442666962>
29. Yaqub, O. (2018). Serendipity: Towards a Taxonomy and a Theory. *Research Policy*, 47, 169-179. <https://doi.org/10.1016/j.respol.2017.10.007>

CHAPTER 15

AI, Automation and New Jobs

Jaures Badet

Department of Economics, Necmettin Erbakan University, Konya, Turkey

ABSTRACT

Our study analyzes the advantages that automation presents for the job. The main new feature of our framework is that, in addition to the part of jobs that are displaced by automation, it also leads to the creation of new, more complex versions of existing tasks, which leads to the demand for employment. We focused more on the essential factor which is the degree of skill to take advantage of these new jobs. We carry out research based on information relating to automation and jobs. Also, by using the output of the final good model, we show that the creation of new tasks in which the labor has a comparative advantage is one of the positive aspects of automation. We find that automation will create new jobs (smart jobs) and eliminates repetitive

Citation: Badet, J. (2021), “AI, Automation and New Jobs”. Open Journal of Business and Management, 9, 2452-2463. doi: 10.4236/ojbm.2021.95132.

Copyright: © 2021 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

jobs which will be replaced by machines in the future. However, these new jobs will need high skills. Therefore, the level and quality of education will play important role in the new jobs that automation will generate. Workers and future students must prepare themselves by focusing their training more on the skills that new technologies will require. Automation may prepare us for a future in which workers with low skills will be forced to change occupations or lose their occupations, which will be completely occupied by machines. We find also that the job loss depends on the speed of automation in each country. Based on the economic structure, the investment policy in new technology, and the level of education of countries, the speed at which automation spread is slower in some countries and intense in others. Therefore, the job is more at risk in countries with high automation than in those with medium or low automation.

Keywords:- Automation, AI, New Jobs, Skill

INTRODUCTION

An intense debate has intensified for years around the question of the future of work with the advent of machines. This debate has taken on a little more momentum these days especially in the period of the COVID 19 pandemic when the use of technology has become essential to face the health crisis. There are many questions about automation and its impact on work. Are we going to see an era where humans will be sidelined in industries to the detriment of machines? What is the future of employment with machines? What will be the role of automation in industries? What is the place of humans in the automation system? One thing is certain: today production in most industries requires the simultaneous completion of a series of tasks. Some difficult and complex tasks are even beyond human skill and therefore require the assistance of machines. The majority of these tasks are therefore performed either with machines or with a combination of man and machines. The real question is if the tasks will be displaced definitively by machines or not. Most of the debates and works have been centered on the fact that many jobs will be replaced in the industry by machines and in the same way, automation will create the advent of new jobs. For example, according to Manyika and Sneader (2018), around 15 percent of the global workforce, or around 400 million workers, could be displaced by automation during the period 2016-2030. According to the same authors, the demand for labor due to automation until 2030 would be between 21% and 33% of the global workforce or 555 million and 890 million jobs. This demand for labor can

largely compensate for the number of jobs lost caused by automation. They also support the idea that all occupations can't be affected by automation. Only about 5% of occupations could be fully automated by technology. About 30 percent of the activities in 60 percent of all professions could be automated. In addition, automation affects less-educated workers and employees in less educated jobs. The displacement of human labor by automation will create a displacement effect and reduce labor demand. But this displacement effect can be counterbalanced by other economic factors like productivity, capital accumulation, the deepening of automation, and the creation of new tasks. Those factors lead to an increase in labor demand. Moreover, automation will lead to the decline in the share of labor in national income, but at the same time by the creation of new tasks, this negative effect will be offset. New tasks increase the demand for labor and tend to raise the labor share (Acemoglu & Restrepo, 2018). Automation causes jobs to be lost in some industries and jobs to increase in others (Bessen, 2017). According to Zinser et al. (2015), a forecast made by a global consulting group, the share of tasks performed by robots in all manufacturing industries today will increase by 15 percent by 2025 (from a global average of around 10% to around 25%). Acemoglu and Restrepo (2016) think that, in a static version where the capital is fixed and technology is exogenous, automation leads to a reduction in employment and the share of labor, but the creation of new tasks generated by this automation causes an increase of labor demand.

Other authors argue that not all jobs can be automated despite the advent of machines. According to Arntz et al. (2016), many workers specialize in tasks that cannot be automated easily, which doesn't put in danger the employment market too much. Other studies point out that automation mainly affects jobs where workers have low skill levels. According to a European survey about skills and employment from Cedefop, "around 14% of jobs in the EU are at risk of displacement by computer algorithms. The jobs most likely to be affected are those which depend more on routine tasks and which require few transversal and interpersonal skills" (CEDEFOP, n.d.). In addition, automation will affect low-skilled workers more than skilled workers (Arntz et al., 2016). Graetz and Michaels (2015) argue also that the use of robots in industries reduces the employment share of low-skilled workers rather than total employment. The threat of automation on jobs is less and heterogeneous. It varies depending on the economic policy, the level of education, and the investment of each country in new technology. Acemoglu and Restrepo (2017) and Chiacchio et al. (2018) underline the fact that automation affect wages and employment in two ways: The direct displacement of workers

from the tasks they previously performed creates displacement effect which negatively affects employment and the increase in demand for labor by industries which creates productivity effect which positively affects wages and employment.

As we saw, many points are often covered in studies about automation and employment. For example, the non-automation of all sectors of industry, the advent of new jobs caused by high automation, the collaboration between robots and humans, and the displacement of certain jobs by machines. The second point, that of the advent of new jobs caused by automation, is discussed more in our study. Our study is therefore more focused on the advantages that automation presents for the job. The dimension of displacement of jobs by automation is not discussed too much in this study. We do so for a reason: The predictions of automation causing massive job losses are not too much in line with reality. According to some reports, automation will even create more jobs than it will displace. Therefore, the main new feature of our framework is that, in addition to the part of jobs that are displaced by automation, it also leads to the creation of new, more complex versions of existing tasks, which leads to the demand for employment. We focused more on the essential factor which is the degree of skill to take advantage of these new jobs.

In this study, Section 2 presents the relevant literature on the impact of AI and automation on jobs. Section 3 focuses on automation and new jobs; especially this section sheds light on the creation of new jobs by automation. In Section 4 the relationship between skills, technologies, and jobs is discussed. In Section 5, how automation leads to the creation of new jobs is shown by using the output of the final good model used by Acemoglu and Restrepo (2016) and Zeira (1998). Finally, in Section 6 a discussion on AI, automation, and new jobs were conducted.

RELEVANT LITERATURE

In this section, the relevant literature on the impact of AI and automation on jobs is surveyed. Acemoglu and Restrepo (2017) analyze the effect of the increase in industrial robot usage between 1990 and 2007 on US local labor markets. They show that robots may reduce employment and wages and that the local labor market effects of robots can be estimated by regressing the change in employment and wages on the exposure to robots in each local labor market-defined from the national penetration of robots into each industry and the local distribution of employment across industries. They

find that one more robot per thousand workers reduces the employment to population ratio by about 0.18 - 0.34 percentage points and wages by 0.25 - 0.5 percent. Frey and Osborne (2017) examine the expected impacts of future computerization on labor market outcomes by implementing a novel methodology to estimate the probability of computerization for 702 detailed occupations. They find that around 47% of total US employment is at high risk in the future. According to them, most occupations such as transportation and logistics, office and administrative support occupations, and jobs in production industries are at risk. Dauth et al. (2017) analyze the impact of rising robot exposure on the careers of individual manufacturing workers, and the equilibrium impact across industries and local labor markets in Germany. They find that robots do not cause total job losses but affect the composition of overall employment. Robots decrease overall employment in Germany by nearly 23 percent in the period 1994-2014. This represents approximately 275.000 jobs. They also find that this loss is fully offset by additional jobs in the service sector. Berriman and Hawksworth (2017) conclude that almost 30% of jobs in the UK could be automated in the early 2030s. According to them, education will play an important role in this automation process. In the UK, automation affects more jobs requiring a lower level of education (46%) than those requiring a high level of education. According to the same authors, automation will create new jobs in the field of digitalization. Chiacchio et al. (2018), by studying the impact of industrial robots on employment and wages in six European Union countries, which make up 85.5 percent of the EU industrial robots market, find that robot per thousand workers reduces the employment rate by 0.16 - 0.20 percentage points. They also find that the displacement effect is particularly evident for workers of middle education and young cohorts, while men are more affected than women are.

AUTOMATION AND NEW JOBS

The example of Britain with new industries and advent of new jobs as engineers, machinists in 19th century and an America with the mechanization of agriculture at the beginning of 20th century has shown that the intensive automation leads to the emergence of new jobs and new tasks in the industries. The majority of work on automation and job claim that automation will eliminate a lot of jobs but also create a lot of new ones. According to Acemoglu and Restrepo (2018), automation will harm the demand for labor but at the same time, will lead to the creation

of new jobs and new tasks. Atkinson and Wu (2017) also support the idea that technology not only eliminates jobs but creates them as well. The emergence of technology may eliminate certain professions, leaving room for other more productive professions. Likewise, according to Manyika and Sneader (2018), automation will result in more job creation than job loss (21% and 33% of the global workforce, or 555 million and 890 million jobs until 2030). According to Gartner (2017), “AI will create 2.3 million jobs in 2020 while eliminating 1.8 Million”, an estimated number that exceeds the number of jobs that are eliminated by the AI in 2020. Chowdhry (2018) also supports the idea that “the growth of artificial intelligence could create 58 million net new jobs in the next few years”.

Thus, we can conclude that automation will displace many jobs especially the jobs which require a low level of education and/or skill will be easily automated. However, it will create some new jobs (smart jobs), which require high skills. For example, according to studies, in the future, we will no longer need a taxi driver to move anywhere we want thanks to autonomous cars¹. The advent of autonomous vehicles in the future is almost inevitable. So what is its impact on taxi drivers? We believe that the advent of autonomous vehicles will generate two situations: The collaboration between taxi drivers and these vehicles and/or the definitive replacement of these jobs by vehicles creating new jobs, which are not identical to the old ones but more complex. Autonomous vehicles can navigate on their own without human intervention but cannot do these kinds of things humans do. For example, drivers repair vehicles or bring them in for repairs when they break down on the way. The job of taxi driving will not go away completely but automation at this level will help taxi drivers do their jobs more easily and better. We may assist in a collaboration between the AI of these cars and the taxi drivers. It is also important to note that automation at this level will require a high degree of skill. Therefore, taxi drivers must requalify their skill level to know how to manage and use these vehicles. Furthermore, if autonomous vehicles replace taxi drivers, it will lead to the creation of other jobs, which may be more complex. Companies will need, for example, engineers, technicians, software developers, and designers to build and manage these autonomous vehicles. These vehicles will break down sometimes and will need humans to repair them. The advent of autonomous vehicles may not benefit taxi drivers if they replace them but will allow other employees with high skills to find work. In any case, the workforce will have an advantage over the machines.

SKILL, TECHNOLOGIES AND NEW JOBS

As it is said in our framework in the previous sections, intensive automation will lead to new tasks and new jobs. But these new jobs need new skills. So the use of efficient machines in industries will require workers who will be able to acquire new skills.

According to Mckinsey Global Institute (2017), in developing countries, the rate of employment growth is highest for occupations that require a college diploma or higher. In China, for example, there is a high demand for occupations currently requiring university degrees and above. At the same time, nearly 60 million jobs currently require a high school diploma. Moreover, even with the effect of automation, in India, the demand for new employees requiring high school education is nearly 100 million (Mckinsey Global Institute, 2017: p. 85).

Automation requires higher skill requirements. The acquisition of more digital skills and the complementarity of key skills are essential in the adaptation of individuals to digital change and automation (CEDEFOP, n.d.). Thus, the industry will need highly qualified professionals whose talents will be in great demand and will be able to train other workers to better master and better adapt to the new tasks that automation will generate. By 2022, 54% of employees will need to learn new skills to meet the expectations of the new tasks created by automation.

Thus 35% of these workers will need at least training for six months, training for more than 6 years will be necessary for 9% of workers and 10% will need training for more than one year (World Economic Forum, 2018). But we often see a mismatch between skills and technologies because workers need to master the use of new technologies in industries but they often do not possess those skills to do those tasks.

These new skills often require a high educational or experimental capacity, which workers in most situations do not have. Most of the time it is also difficult for employers to find workers that can master the new jobs and tasks induce by the intensive automation (Deloitte & the Manufacturing Institute, 2011).

To summarize, jobs displaced by automation require less skill than new jobs generated by the latter. The new jobs require high skills, which will depend on the quality of the educational system of each country. Much of the employment in the future created by automation will require high education levels. In this context, jobs requiring less educational requirements are in danger and at the same time, the demand for jobs requiring educational

capacities or higher skills increases. The education system plays a decisive role in the new jobs that will lead to automation. The quality of skill depends on the quality of the educational system. If the educational system is not up to the skill requirements that new jobs require, we often see a mismatch between skills and technologies.

THEORETICAL MODEL

In our model, we analyze the effect of automation and AI on Jobs by considering two type of economy. Firstly, we suggest that in a technologically stagnant economy, all tasks are produced by human. Therefore, jobs are not in danger in this economy. Second, in a technologically advanced economy where most sectors are automated, automation will obviously cause the loss of many jobs.

However, these jobs will be more the jobs with repetitive tasks, which do not push the workers to raise their level of thinking. At the same time, it will create new, more complex jobs, which will cause workers to think more, to have more creativity in the execution of tasks.

In other words, these new jobs will need employees who are able to create new ideas and knowledge useful for the business, think quickly and smartly to solve complex problems, and who have the ability to adapt to the new technology that is in prospect change.

Therefore, high skill due to high level and quality of education (only both makes qualified workers) will play an essential role in the “smart jobs” that automation will create.

To evince our hypothesis, we use the output of the final good model used by Acemoglu and Restrepo (2016) and Zeira (1998). According to these authors, the output of the final good is given by:

$$Y = \left(\int_{N-1}^N y_i^{\frac{\varphi-1}{\varphi}} \right)^{\frac{\varphi}{\varphi-1}} \quad (1)$$

Final good Y is produced by combining a continuum of tasks y_i where $y_i \in [N-1; N]$ and φ denotes the elasticity of substitution between tasks.

By assuming that the range of tasks is between $N-1$ and N , the creation of a new task in N corresponds to the replacement of an existing task in $N-1$. The production of each task requires a combination of labor or capital. In

industries, not all tasks can be produced by labor. Another task requires the help of machines (capital). Therefore, we have the automated tasks, which are produced by labor with the help of machine, and the non-automated tasks executed only by labor. If there exists $I \in [N-1, N]$ such that

$$\begin{aligned} i &\leq I \text{ automated tasks} \\ i &> I \text{ non automated tasks} \end{aligned} \quad (2)$$

Note that here $i \leq I$ is the automated tasks which are produced by labor or capital as well.

If we assume that the tasks are executed by labor or machine with a specific intermediate task $q(i)$ (the technology used both for production and for the possible automation of tasks), we have:

$$\begin{aligned} y(i) &= B \left[\beta q(i)^{\frac{\varphi-1}{\varphi}} + (1-\beta)(\gamma_K(i)k(i) + \gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i \leq I \\ y(i) &= B \left[\beta q(i)^{\frac{\varphi-1}{\varphi}} + (1-\beta)(\gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i > I \end{aligned} \quad (3)$$

where $\gamma_L(i)$ is the productivity of labor in task i , $\gamma_K(i)$ the productivity of capital (machine), $l(i)$ denotes tasks that can be produced by human labor and $k(i)$ the one that will be produced by machine. $\varphi \in (0; \infty)$ and represents the elasticity of substitution between intermediates and labor, $\beta \in (0; 1)$ represents the distribution parameter of this constant elasticity of substitution production function. B represents a normalizing constant and $B = (1-\beta)^{\frac{\varphi}{1-\varphi}}$ to simplify the algebra.

$$\begin{aligned} y(i) &= (1-\beta)^{\frac{\varphi}{1-\varphi}} \left[\beta q(i)^{\frac{\varphi-1}{\varphi}} + (1-\beta)(\gamma_K(i)k(i) + \gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i \leq I \\ y(i) &= (1-\beta)^{\frac{\varphi}{1-\varphi}} \left[\beta q(i)^{\frac{\varphi-1}{\varphi}} + (1-\beta)(\gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i > I \end{aligned} \quad (4)$$

If we assume that $\beta \rightarrow 0$ (the share of revenues going to intermediates is very low)²:

$$\begin{aligned} y(i) &= 1^{\frac{\varphi}{1-\varphi}} \left[(\gamma_K(i)k(i) + \gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i \leq I \\ y(i) &= 1^{\frac{\varphi}{1-\varphi}} \left[(\gamma_L(i)l(i))^{\frac{\varphi-1}{\varphi}} \right]^{\frac{\varphi}{\varphi-1}} \text{ for } i > I \\ \forall \varphi &\in (0; \infty), (1)^{\frac{\varphi}{1-\varphi}} = 1. \end{aligned} \quad (5)$$

So:

$$y(i) = \gamma_L(i)l(i) + \gamma_K(i)k(i) \text{ for } i \leq I \quad (\text{a})$$

$$y(i) = \gamma_L(i)l(i) \text{ for } i > I \quad (\text{b})$$

(6)

Case (a) is currently restricted to a group of countries. This case is more observed in high-income countries and some middle-income countries. Case (b) is more frequent in industries of low-income, and certain middle-income countries. However, this case is also observable in high-income countries because even in these countries not all sectors can be automated.

So, in a technologically stagnant economy, all tasks are produced by labor.

$$y(i) = \gamma_L(i)l(i) \quad (7)$$

Jobs are not in danger since automation is not evolved in these economies. Even employees with low skills are not exposed to the risk of job loss caused by automation since there is no task produced by machines.

In a technologically advanced economy, tasks are executed totally by human in some sectors and by the humans and machines in others:

$$y(i) = \gamma_L(i)l(i) + \gamma_K(i)k(i)$$

$$y(i) = \gamma_L(i)l(i)$$

(8)

In such an economy, the non-automated sector employees are not exposed to the risk of job loss. In automated sector, an intensive increase in l , leads to lower labor costs (l) and the creation of new tasks (more complex) generated by the use of AI and other advanced technologies in automation. The new tasks lead to an increase in the demand for labor. If $\gamma_L(i)$ increases strictly in i , labor has a comparative advantage in higher-indexed tasks³. Even in automated sector, the loss of ancient jobs leads to the creation of new ones that are more complex and will need a high skill. As is discussed in Section 3 of our work, the creation of new tasks in which the labor has a comparative advantage is one of the positive aspects of automation. Even though capital accumulation and deepening automation are important factors in increasing the labor share of national income, the creation of new tasks in which labor has a comparative advantage remains the most important aspect preventing the decline in the share of national income (Acemoglu & Restrepo, 2018). Therefore, in any case, the workforce will have an advantage over the machines. Furthermore, as we saw in the model, not all

sectors can be automated. The speed at which automation spreads depends on the policy and investment in innovation and adoption of technology in each country and each company. For example, the speed of automation in The USA is not the same in most African countries because the goals of innovation and adoption of new technology are not the same in all countries. Therefore, jobs in automated industries are at a higher risk of displacement than those in non-automated or less automated industries. In addition, the quality of education especially higher education will also be important in the process of automation. The companies use advanced technologies like AI and others in automation process. The use of these technologies will need high and qualified skills in digitalization. Therefore, intensive automation will lead to the replacement of low-level jobs and the creation of new jobs that require high skill levels.

CONCLUSION

Many questions and debates intensify on the future of jobs with automation especially in the period of COVID-19 where we have seen the importance of new technology in the management of the health crisis. Different authors have different opinions on the future of employment with automation. However, most of the work on this issue supports the fact that automation will replace some jobs in the industry but at the same time create others as well. Our study is focused on the advantages that automation presents for the job. The dimension of displacement of jobs by automation is not discussed too much in this study. We do so for a reason: the predictions of AI causing massive job losses are not too much in line with reality. According to some reports, automation will even create more jobs than it will displace. Therefore, the main new feature of our framework is that, in addition to the part of jobs that are displaced by automation, it also leads to the creation of new, more complex versions of existing tasks, which leads to the demand for employment. We focused more on the essential factor which is the degree of skill to take advantage of these new jobs.

Firstly, we find that based on the economic structure, the investment policy of countries in new technology, and also the quality of education, the effect of automation on employment varies from country to country. The speed of automation is slower in some countries and very intense in others. Therefore, the job is more at risk in countries with high automation than in those with medium or low automation. In other words, the influence of automation on the job in technologically advanced countries is different

from that in technologically stagnant countries. Second, even if machines are more productive than humans are, they cannot do everything in companies. Certain jobs are properly reserved only for human capacity. So not all jobs can be automated in the industry. We may see low demand for labor in the automated sectors but at the same time a strong demand in the non-automated sectors. In addition, jobs with a low skill level are more at risk than those with a high skill level. In another word, jobs consisting of repetitive tasks do not require a high level of education. These jobs are more vulnerable to automation than those that require more thinking and more creativity (High skill). Therefore, in this dynamism of automation, the job of skilled workers is more secure than that of low-skilled workers.

Thirdly, automation will create maybe more jobs than it eliminates in the future. It will replace many jobs especially the jobs with low skill but at the same time as we said in our study, it will create many new jobs, which will need a high skill. Automation is a step forward, an important and necessary revolution for industries in increasing productivity and competitiveness. Not something, that should normally be scary. The key to the jobs that automation will create is high and digital skills. Thus, higher education and training oriented toward digitalization and coding will play a very important role in the new jobs.

The future jobs will not be for everybody. To know how to create or manage sophisticated technologies, workers need to get high skills. Therefore, we need to change the way we give or take training. The educative system must be more valued and more oriented to digital skills and coding. Education in the future should be more focused on knowledge of new technologies. Automation is not a threat to us.

It is an advantage but to benefit from this automation, workers and future students must prepare themselves by focusing their training more on the skills that new technologies will require. Automation may prepare us for a future in which workers with low skills or simple tasks will be forced to change occupations or lose their occupations, which will be completely occupied by machines. To successfully perform their new tasks, workers will need to acquire the necessary skills. They will therefore need training, so the duration will vary depending on the type of task and the type of skill sought. In any case, automation leaves us many job opportunities. The important thing is to know how to orient the new knowledge to be able to adapt it to the requirements of automation.

NOTES

¹The advent of autonomous vehicles will not only have an impact on taxi drivers but many other areas. In this section of our study, we just approached the impact of the advent of autonomous vehicles on taxi drivers. The aim is to show the advantage of this new technology on this job.

²Acemoglu and Restrepo (2016) use the same assumption in their work about “The Race between Man and Machine: Implications of Technology for Growth, Factor Shares and Employment” page 8.

³Our theoretical framework builds on Acemoglu and Restrepo (2016) extends Zeira (1998) who develop a model where firms produce intermediates using labor-intensive or capital-intensive technologies.

REFERENCES

1. Acemoglu, D., & Restrepo, P. (2016). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares and Employment. NBER Working Paper No. 22252, National Bureau for Economic Research. <https://doi.org/10.3386/w22252>
2. Acemoglu, D., & Restrepo, P. (2017). Robots and Jobs: Evidence from US Labor Markets. NBER Working Paper No. 23285, National Bureau for Economic Research. <https://doi.org/10.3386/w23285>
3. Acemoglu, D., & Restrepo, P. (2018). Artificial Intelligence, Automation and Work. NBER Working Papers No. 24196, National Bureau of Economic Research. <https://doi.org/10.3386/w24196>
4. Arntz, M., Gregory, T., & Zierahn U. (2016). The Risk of Automation for Jobs in OECD Countries. OECD Social Employment and Migration Working Papers, No. 189, Organisation for Economic Co-Operation and Development.
5. Atkinson, R., & Wu, J. (2017). False Alarmism: Technological Disruption and the U.S. Labor Market, 1850-2015 (pp. 1-28). Information Technology and Innovation Foundation. <https://doi.org/10.2139/ssrn.3066052>
6. Berriman, R., & Hawksworth, J. (2017). Will Robots Steal our Jobs? The Potential Impact of Automation on the UK and other Major Economies. PwC UK Economic Outlook.
7. Bessen, J. (2017). Automation and Jobs: When Technology Boost Employment. Law and Economics Research Paper No. 17-09. Boston University School of Law. <https://doi.org/10.2139/ssrn.2935003>
8. CEDEFOP (European Centre for the Development of Vocational Training) (n.d.). Automation of Work and Skills. <https://www.cedefop.europa.eu/en/events-and-projects/projects/digitalisation-and-future-work/automation-work-and-skills>
9. Chiacchio, F., Petropoulos, G., & Pichler, D. (2018). The Impact of Industrial Robots on EU Employment and Wages—A Local Labour Market Approach. Working Papers No. 25186. Bruegel.
10. Chowdhry, A. (2018, September 18). Artificial Intelligence to Create 58 Million New Jobs By 2022, Says Report. Forbes. <https://www.forbes.com/sites/amitchowdhry/2018/09/18/artificial-intelligence-to-create-58-million-new-jobs-by-2022-says-report/?sh=1f0b27e24d4b>

11. Dauth, W., Findeisen, S., Südekum, J., & Wößner, N. (2017). German Robots the Impact of Industrial Robots on Workers. CEPR Discussion Paper No. 12306, Centre for Economic Policy Research.
12. Deloitte & The Manufacturing Institute (2011). Boiling Point? The Skills Gap in U.S. Manufacturing. <http://www.themanufacturinginstitute.org>
13. Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerisation? Technological Forecasting and Social Change, 114, 254-280. <https://doi.org/10.1016/j.techfore.2016.08.019>
14. Gartner (2017, December 13). Gartner Says By 2020, Artificial Intelligence Will Create More Jobs than It Eliminates. <https://www.gartner.com/en/newsroom/press-releases/2017-12-13-gartner-says-by-2020-artificial-intelligence-will-create-more-jobs-than-it-eliminates>
15. Graetz, G., & Michaels, G. (2015). Robots at Work. Discussion Paper No. 1335, CEP.
16. Manyika, J., & Sneader, K. (2018, June 1). AI, Automation, and the Future of Work: Ten Things to Solve for. McKinsey Global Institute. <https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for#>
17. McKinsey Global Institute (2017, December 6). Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation. <https://www.mckinsey.com/~media/McKinsey/Industries/Public%20and%20Social%20Sector~/Our%20Insights/What%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/MGI-Jobs-Lost-Jobs-Gained-Report-December-6-2017.pdf>
18. World Economic Forum (2018). The Future of Jobs Report 2018. World Economic Forum. http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf
19. Zeira, J. (1998). Workers, Machines, and Economic Growth. Quarterly Journal of Economics, 113, 1091-1117. <https://doi.org/10.1162/003355398555847>
20. Zinser, M., Sirkin, H., & Rose, J. R. (2015, September 23). The Robotics Revolution: The Next Great Leap in Manufacturing. Boston Consulting Group. <https://www.bcg.com/publications/2015/lean-manufacturing-innovation-robotics-revolution-next-great-leap-manufacturing>

CHAPTER 16

Discussion on the Development of Artificial Intelligence in Taxation

Zhuowen Huang

Nanfang College of Sun Yat-sen University, Guangzhou, China

ABSTRACT

With the development of AI technology, a new forecasting and statistical model for tax auditing has been created. In recent years, thanks to breakthroughs in AI research, tax professionals have gained new analytical and statistical tools, providing convenience and improving efficiency. These tools have formed the basis for systematic frameworks that avoid the disorder and complexity of data processing and analysis in Excel spreadsheets. Additionally, AI provides simulated tax risks, which can help more complex human judgments to be made. AI can also aid detection of

Citation: Huang, Z. (2018), "Discussion on the Development of Artificial Intelligence in Taxation". American Journal of Industrial and Business Management, 8, 1817-1824. doi: 10.4236/ajibm.2018.88123.

Copyright: © 2018 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

fraud, contributing to its supervision and monitoring by government. The development of AI continues, and its deployment has certain limits and risks that must be recognized.

Keywords:- Artificial Intelligence (AI), Taxation, Supervision

INTRODUCTION

The structure of this paper is organized by five parts. The first part describes the concept of AI and the AI background, and then the paper analyzes how AI can be applied in taxation, next, the paper focus on AI application of taxation in China and in globally. The last part summarizes the obstacles of using AI in practice, the prospects of AI in China, and suggestion. The main contribution of this paper is that it gives a relatively clear picture about the problems of using AI in taxation in China. The main limitation of this paper is to present facts of AI development in China and pinpoint potentials of using AI in taxation in China. It presents the reality of AI being developed in China without substantial researches and data analysis.

Artificial Intelligence Concept

Artificial intelligence (AI) is a broad term that refers to techniques making machines “intelligent”. AI research and application utilize automation to enhance or replicate human intelligence to improve the analysis and decision-making capabilities of machines. AI provides managers with unprecedented tools to ease the complexity of decision-making, serving as a catalyst for internal structural transformation in various industries. It also allows complicated and time-consuming tasks to be completed more effectively and efficiently.

Development of Artificial Intelligence in Taxation

Pascal A. Bizarro and Margaret Dorian (2017) pointed out that artificial intelligence (AI) was born in 1948, when William Gray Walter created two small robots, named “Elmer” and “Elsie”, that were able to recognize and respond to stimuli while encountering obstacles [1]. Two years later, Alan Turing (1950) proposed that a machine could transmit information, communicate, and possess thinking capabilities indistinguishable from those of humans [2]. In 1956, the Dartmouth workshop proposed the term “artificial intelligence”, marking the birth of AI as a discipline [3]. Since then, the AI phenomenon has received considerable attention in various

fields. According to the 41st “Statistical Report on Internet Development in China” in 2017, published by China Internet Network Information Center (CNNIC), there are 2542 AI companies globally, including 1078 in the United States (accounting for 42.4%) and 592 in China (23.3%) [4]. Familiar AI products include Apple’s Siri, self-driven cars, and virtual reality head-mounted displays. In the field of taxation, this ever-developing technology can enhance the effectiveness of automated tax auditing and decision-making, and play an essential role in supervision and monitoring by the government.

HOW IS AI APPLIED IN TAXATION?

To understand how AI is applied in taxation, Cas Milner and Bjarne Berg (2017) believe that AI’s role in taxation is like a software that can automatically adapt to the input of different content and make judgments without specific instructions [5]. While AI robots acting as tax accountants is currently believed to be unlikely, they can perform various roles, such as assisting tax auditors in detecting errors, classifying accounts and transactions, assessing tax audit risks, and increasingly propose favorable tax strategies within the framework of complex global laws. There are many prospects for applying AI in taxation. It is instructive to consider some of the most successful accounting firms. Below, we briefly discuss their application and development in AI.

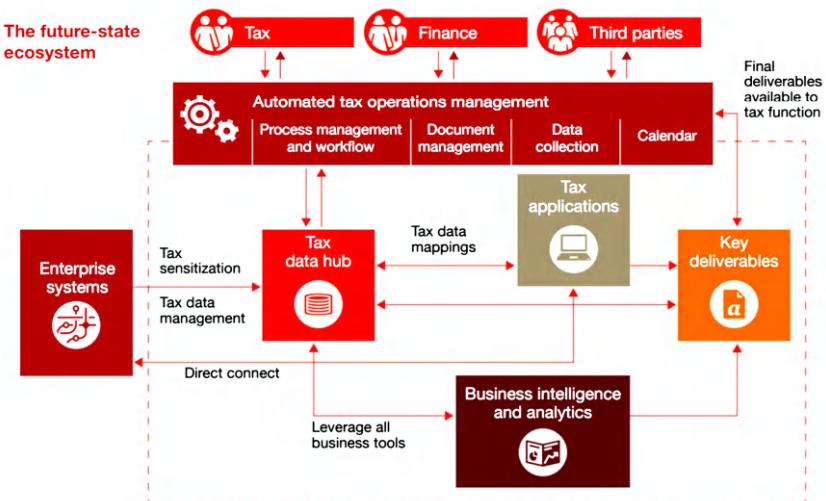


Figure 1. The future-state ecosystem.

According to PwC's official website, the combination of tax accounting and AI, integrating the accounting system of the entire enterprise is similar to an ecosystem. The following illustration depicts an automated system that provides an integrated tool that improves efficiency, enhances data quality, and adapts to ever-changing risk environments through constant information flow between three parts, as shown in Figure 1 [6] :

KPMG's official website recently reported the launch of "Tax Service," an intelligence tax product that assists Chinese companies to solve automated tax compliance problems.

In China's tax system, policy changes are frequent and complex. If the tax information or the review process is omitted without the help of AI, serious tax risks and consequences could result, which can be avoided by using AI that can integrate and conform to the current policy to speed up tax declaration process. In addition, KPMG's "Tax Service" product can perform automatic preparation of VAT and corporate income tax returns, as well as local additional tax calculation tables, trend analysis, and the timely detection of potential errors, risks, or abnormal conditions [7].

THE DEVELOPMENT OF AI IN CHINA

In China, with the development of AI, more standardized policies have emerged, like the "New Generation Artificial Intelligence Development Plan." AI is a new and strategic technology that leads the future, enhances the country's competitiveness, and maintains national security, undoubtedly a positive influence on the development of AI.

China has begun to promote AI in taxation. In Guangdong, a tax robot has been introduced. It has completed a total of 12,000 times similar to human-computer transactions in reality and has handled regular business for 660 individual, industrial, and commercial taxpayers. Its contribution accounts for 54.91% of regular business, thereby halving the workload for, and reducing the burden on, civil servants [8]. In the Dianbai District of Maoming City, Guangdong Province, the Taxation Bureau Office has introduced China's first "face-to-face tax" intelligent robot. It can collect taxpayer information, such as their face photo, ID card, and contact number through scanners during process of date submission; once this information is authenticated, the taxpayer's identity can be verified, which improves efficiency [8]. Meanwhile, the "taxation bureau of Shanghai Fengxian District" operates the "Fengxian Tax" WeChat public forum and mobile phone tax software, enabling taxpayers to check tax-related information

concerning, for example, policy updates, tax processes, and information disclosure [8]. In the Shijingshan District National Taxation Bureau of Beijing, taxpayers can consult the Lingyun robot. It listens to questions posed by taxpayers, processes the information, and responds to their queries, making judgments on the problems they present. The development of AI “reply by voice” is already very advanced in this field, drawing on several advanced technologies. such as: Lingyun speech recognition which recognizes spoken questions and answers and semantic understanding which recognizes meaning of spoken questions and answers. With continuing adjustments, these technologies will be near perfection [8].

GLOBAL DEVELOPMENTS

Globally, an increasing number of countries are developing and diversely applying AI technology. With the rapid development of tax technology, the demand for AI in taxation is also growing. There are many reasons for this. For multinational companies, integrating data through AI will maximize their ability to collect and analyze data, and help them adapt to changing policies on tax compliance processes. Moreover, automation will greatly increase the transparency of tax data, helping multinationals to satisfy government demands for accurate tax reporting, in detail and in real time, while also improving the development of government regulation.

As a result, AI is highly valued. Additionally, AI has been more prevalent due to government promotion. Through deployment of AI, the collection and organization of tax data becomes more systematic and transparent. This helps to curtail unreasonable tax avoidance by international companies, eliminate illegal tax evasion, help to curb multinational companies’ profit transfers and tax base erosion, the reason why there was the participation of more than 100 countries in the Organization for Economic Cooperation and Development (OECD/G20) and Base Erosion and Profit Shifting (BEPS) project [9]. Applying AI to streamline tax data by government increases efficiency and compliance with current policies on tax reporting and reasonable tax avoidance. Furthermore, data collected through AI can be used to establish mathematical models, analyze the tax trends and indicators of various enterprises, and adjust various tax policies. For tax administration departments, processing tax data through AI accelerates the identification and analysis of the enterprise’s tax problems, which may include unreasonable tax avoidance or tax evasion. Consequently, such problems can be immediately contained and enterprises can maximize their

legitimate income. AI enhances the systematic processing and transparency of tax data, which increases the intensity of government supervision and motivates the ongoing development of AI.

CONCLUSIONS

To summarize, there are three points that should be noted.

The Obstacles of Using AI in Practice

During the process of information collection for this paper, it is found that the application of AI in taxation is not prevalent in China. One reason is that Chinese tax law is versatile and changes frequently, requiring AI systems to be updated correspondently with concurrent policy revisions. Addition and deduction rules can be good examples as to why AI systems need to be updated constantly in order to accommodate the rapid policy changes. For example, the previous deduction rate for the R&D expenses of general enterprises is 50%, which had been updated to 75% in the AI system due to the policy changes made in the early 2017 in order to be enforced between January 1, 2017 and December 31, 2019 [10]. The policy is constantly updated and changed so that AI applications in the tax system must be updated and changed simultaneously. As a result, AI system needs adjustments in order to produce accurate tax audit report. Keeping the system up to date with new policies will undoubtedly make the AI system more acceptable and prevalent. However, at present, AI is still developing, and the AI system is not able to update itself since manual adjustments are needed. In turns of data input, integration and tax reports preparation, AI system in taxation can be still more advanced.

Under the “Notice of the State Administration of Taxation on Regulating Tax Exemption for Compulsory Certification Services” (Tax General Letter [2014] No. 220), not all Chinese enterprises are required to issue tax audit reports [11]. For those exempt, AI does not seem to be needed in the data process, which can reduce the costs of researching and developing tax-based AI systems. It is certainly questionable whether cost reductions can be achieved given the need to frequently implement updates as law and policy changes. Is it even possible to reduce costs without using AI systems in taxation? The validity of this assumption requires in-depth investigation. It should also be noted that AI in China’s taxation is at an immature development stage. How long it might take to mature is difficult to accurately predict.

The Prospects of AI in China

However, we cannot deny the opportunities and advantages brought by AI. The above analysis clearly demonstrates the great convenience it offers. In the context of large amount of data, and with the need to handle complex data or integrate systems, AI can be more accurate and efficient.

Structure Change

On June 15, 2018, China's national and local taxes were merged into one, which means that taxation process will be streamlined, therefore operating more quickly [12]. In 2015, the China Office and the State Council issued the "Reform Plan for the National Tax and Local Tax Collection and Management System," clarifying the "cooperation" between national and local tax authorities. In many areas, national and local tax authorities began to jointly manage taxes. For example, through the collaboration, the Tax Bureau Joint Office was established, providing services to improve taxpayers' experience and the efficiency of tax authorities [13]. For example, now China's tax bureaus all share a system promoting extensive cooperation and integration between the national and local authorities. Solid foundations were thereby laid for the merger. Consequently, we also urgently need AI systems to help companies more quickly generate tax reports and submit them to the tax bureau in accordance with the trend of systematic development and transparency.

The Potential in International Reporting

Despite the rapid development of taxation in AI, global governments are still facing challenges as to data comparability and verifiability due to the fact that there is no universally accepted tax system. Is it possible to make the tax data of multinational corporations more transparent and universal so that accountability of international tax reporting can be improved? For instance, data and income of a foreign enterprise owned by a Chinese resident abroad can be quickly intercepted by the Chinese tax bureau to avoid tax avoidance and evasion. Additionally, AI system can play a supervisory role in reporting whether the domestic and foreign incomes are reported truthfully.

An Unstoppable Trend

Indeed, AI in China is still not prevalent due to rapid policies changes without sufficient updates and the high costs of integration of AI to current

tax systems. Nonetheless, the development of AI is an unstoppable trend, and will become more mature and standardized in the future.

- 1) In China, government is accelerating the reform of various taxation systems; moreover, the long-term management philosophy and methods of the Chinese tax authorities are undergoing major changes.
- 2) In global, the “Statement on Algorithmic Transparency and Accountability” outlines seven principles, which helps the development and application of AI in taxation with a clearer policy norm and increased usability [14].
- 3) Finally, in China, the State Council’s notice on the issuance of “a new generation of AI development plan” [2017] No. 35 by the State Council indeed provide support to AI policy and standardization [15].

Suggestion

At present, it is critical for both government and companies to establish and utilize the AI in taxation. The government should seize this opportunity to fully tap the huge value of the hidden potential of using AI in taxation. AI, with tremendous data and analytical capability, is an invaluable tool for tax authorities, helping them to better understand clients and to simulate future business scenarios. Corporations can use AI to develop advanced tools for monitoring behaviors and activities in real time and analysis. AI systems can adapt to ever-changing risk environments, helping corporations to continuously improve their monitoring capabilities and transparency with regard to regulatory compliance and corporate governance. In addition, AI can be developed from a warning system into a precognition system to help companies avoid risks through tax analysis and predictions. For example, AI can be used to more accurately predict the probability of a certain enterprise defaulting on loans or overdue payments. Thus, the companies need to fully use AI in taxation, which would keep them more competitive in the future.

REFERENCES

1. Bizarro, P.A. and Dorian, M. (2017) Artificial Intelligence: The Future of Auditing. *Internal Auditing*, 5, 21-26.
2. Turing, A.M. (1950) Computing Machinery and Intelligence. *Computation & Intelligence*. American Association for Artificial Intelligence, Palo Alto, 44-53.
3. Moor, J. (2006) The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *Ai Magazine*, 27, 87-91.
4. CNNIC (2018) The 41st “Statistical Report on the Development of China’s Internet. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249.htm
5. Milner, C. and Berg, B. Tax Analytics—Artificial Intelligence and Machine Learning—Level 5. <https://www.pwc.com/us/en/services/tax/tax-innovation/artificial-intelligence-and-machine-learning.html>
6. PWC (2015) Unlocking the Power of Data and Analytics. https://www.pwc.com/gx/en/tax/publications/assets/PwC_TFoF_DataAnalytics_global_nov2015.pdf.
7. KPMG. KPMG China’s Highly Automated Tax Compliance Solution. <https://home.kpmg.com/cn/en/home/services/tax/tax-technology/tax-compliance.html>
8. Ding, F. (2017) “AI+ Tax” Will Become the “Black Technology” of the Tax System. <http://www.shui5.cn/article/ef/113244.html>
9. Viglione, J. and Deputy, D. (2017) Your Tax Data Is Ripe for Artificial Intelligence. Are You Prepared? AI Is Still Evolving, but Machine Learning Is Already Here. *Tax Executive*. <https://www.highbeam.com/doc/1G1-510480668.html>
10. Zhonghui (2018) Will It Be a High-Tech Enterprise to Enjoy 75% of the Research and Development Expenses? <http://www.shui5.cn/article/b1/118937.html>
11. Shuiwu (2014) Notice of the State Administration of Taxation on Regulating Tax Exemption for Compulsory Certification Services. <http://www.shui5.cn/article/de/72010.html>
12. Shuang, D. (2018) Research on Tax Service Optimization Based on the Combination of State and Land Taxes. *Tax*, No. 16, 27.
13. Wen, C.X., Liu, J.Y., Yu, J.Y. and Wang, P., et al. (2015) Speeding up the Reform of Tax Collection and Management System—Review of the

- Reform Plan of Deepening the National Tax and Local Tax Collection and Management System. China Tax, No. 12, 15-16.
14. Garfinkel, S., Matthews, J., Shapiro, S.S., et al. (2017) Toward Algorithmic Transparency and Accountability. Communications of the ACM, 60, 5. <https://doi.org/10.1145/3125780>
 15. State Council (2017) Notice of a New Generation of Artificial Intelligence Development Planning. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

CHAPTER 17

AI and Zen: AI Films as Reflections on Reality and Illusion

Jun Yu¹, Bo Zhang^{1,2}

¹Department of Film and Television Arts, Shanghai Publishing and Printing College, Shanghai, China.

²Academy for Engineering & Technology, Fudan University, Shanghai, China.

ABSTRACT

The paper provides an analysis of how AI-themed films reflect on the question of reality and illusion, and how these reflections resonate with the ancient Zen philosophy in a vivid way. The reality of the world, human distinctiveness, and “self” are discussed by comparing the metaphors in AI-themed films and the philosophy of Zen.

Keywords:- AI, Zen, Reality and Illusion, AI-Themed Films

Citation: Huang, Z. (2018), “Discussion on the Development of Artificial Intelligence in Taxation”. American Journal of Industrial and Business Management, 8, 1817-1824. doi: 10.4236/ajibm.2018.88123.

Copyright: © 2018 by authors and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY). <http://creativecommons.org/licenses/by/4.0/>.

INTRODUCTION

What is reality? It has long been a fundamental philosophical question that humans have pondered. As Solomon and Higgins (2013) pointed out, ancient Greek philosophers focused on whether ultimate reality is material or immaterial. While Plato raised the two world theory, claiming the immaterial world is ultimately real, Aristotle believed that reality exists right in our daily life. As Descartes's mind-body dualism, Berkeley's subjective idealism, and Leibniz's "reality is a community of souls" were successively proposed, the question became "what is most real?". Modern philosophers' opinions on this question remain divided. For example, Kant stuck to the two world theory, while Schopenhauer believed the two parties in this metaphysical dualism are both irrational.

It is the thinking of this question that helps us to distinguish between the presentations of things and their internal reality, so that we can better understand the world and lead a better life. The reflections on this issue are also reflected in various film and television works, and the way of reflecting on this question has changed with the development of the times. Nowadays, when information technology is becoming increasingly closely related to human life, there are many films on the theme of AI, which also show reflections on this issue from multiple perspectives. These reflections either explicitly refer to some ideas of Eastern Zen Buddhism or implicitly resonate with Zen. The ancient Zen thoughts have taken on a new vitality in the film works of the Cyber Age.

A brief review of Zen's philosophy on reality and illusion is first introduced, followed by a discussion on how AI-themed films reflect the question of reality and illusion, and how these reflections resonate with Zen. Reflections on reality and illusion in AI-themed films can be further divided into three areas: the reality of the world, the reality of human distinctiveness, and the reality of "self".

THE QUESTION OF REALITY AND ZEN

In Western philosophical history, from Thales, Plato and Aristotle in ancient Greece to Descartes, Kant, Schopenhauer and Hegel in modern times, all have put forward their theories around "reality". In Eastern Zen Buddhism, the question of reality and illusion is also a central issue. The basic idea of Zen Buddhism is that "all dharmas are empty", as it is said in the Diamond Sutra, "all the laws of existence are like dreams and bubbles, like fog and lightning" (Fang, 1995). From the view of Zen Buddhism, all the

miseries and sufferings of life, or Klesa, originate from the reversal of the understanding of reality and emptiness. Zen believes that in order to break Klesa, one needs to enter the original state of the mind, i.e. the self-nature (zixing). Zen regards the self-nature as the essence of human, and it is also considered as the Buddha-nature. The self-nature is believed to be universal, and everyone is originally self-sufficient. The purpose of Zen practice is to teach people to realize that all the Dharmas of the world are delusions, and ultimately to understand the reality of self-nature and to break their Klesa (Fang, 1995).

Western philosophers often explored the issue of reality and illusion through discursive approaches. In addition, metaphors and thought experiments appropriate to the context of the times have also been proposed, such as Plato's Cave and Hilary Putnam's "Brain in a Vat". In the case of Zen Buddhism, traditionally, the practice and guidance of enlightenment for the learner are done through meditation, conversations, and "koan" (or in Chinese "gongan", a paradoxical anecdote or riddle). A characteristic of Zen teaching is the flexible application of various vivid examples to guide the learner to realize the truth. Here is a classic example of Zen anecdote.

An official named Lu Huan asked Nanquan, a Zen master, "A goose was kept in a bottle, but the goose grew up and could not get out of the bottle. Now the bottle cannot be destroyed, and the goose cannot be damaged, how can you get it out?" Nanquan didn't answer but suddenly called Lu by his name loudly, and Lu answered. The master then said, "It's out". Lu was then enlightened.

In this case, the "bottle" is a metaphor for the various perceptions of man in society, which in Zen Buddhism are all delusions and confuse man's self-nature. To get rid of the shackles of the "bottle", one needs to realize that the shackles of the "bottle" are delusional, and that the "self-nature" is originally unattached and unhindered. When Nanquan called Lu Huan's name, Lu Huan answered, and the moment he answered, he temporarily forgot the delusional problem of "bottle" and "goose", and the "bottle" that bothered him temporarily disappeared from his mind. At the same time, the call of Nanquan made him realize the existence of "self", and at this moment his self-nature also worked.

Above is an example of how Zen uses vivid examples in teaching people to realize the distinctions between reality and illusion. In the age of information technology, such reflection of reality and illusion can be carried out in a richer context. Since Artificial Intelligence originates from

the exploration and simulation of human cognition and thinking ability, the question of reality is an inescapable issue in the development of AI. Technological developments or science fiction visions such as cyberspace, cyborgs, and consciousness uploading also enable people to project many issues in the sandbox of a future society where the boundaries between the real and virtual worlds are further blurred. In many AI-themed films, we see reflection on reality and illusion, and when we look at these projections in the future cyber context with the ancient wisdom of Zen Buddhism, we find that AI also provides examples of Zen teachings that are more relevant to modern perceptions. Some of these AI-themed films explicitly incorporate elements of Zen, such as *The Matrix* (1999) and *Doomsday Book* (2012), while some implicitly fit in with Zen thinking or teaching, such as *The Ghost in the Shell* (1995), *Ex Machina* (2015).

This paper examines the reflections on reality in AI-themed films from three scales: on the macroscopic scale, the reflection on the reality of the world; on the mesoscopic scale, the reflection on the reality of human distinctiveness; and on the microscopic scale, the reflection on the reality of “self”.

REFLECTION ON THE REALITY OF THE WORLD

In the late 20th century, with the development of computer technology, people began to be able to create virtual worlds that simulate reality in the digital space of computers, and computer programs began to be increasingly intelligent and able to interact with humans. It seems to be only a matter of time before it will be possible to create a digital world exactly like the real world.

These technological developments have led people to consider the possibility that what we now consider to be the “real” world could actually be a virtual world. Reflections on the reality of the world have been around for a long time. Plato, for example, proposed the “cave metaphor”, arguing that what people see is just a projection of the real world in a cave. From a technological point of view, a virtual world simulated by a computer appears to be more theoretically possible. All our cognitive activities can be categorized as neural activities based on neuro-electrical conduction, which means that when the technology is sufficient, the cognitive information we get can be simulated by inputting external electrical signals, and even the simulated electrical input can replace all the cognitive inputs, so that people can live in a virtual world made of electrical signals (Zhu & Zhang, 2022).

Hilary Putnam's famous thought experiment "Brain in a Vat" is the pioneer of this way of thinking, and *The Matrix* (1999) is a perfect visualization of "Brain in a Vat". In the future society set up by *The Matrix*, human beings are enslaved by AI robots, and all humans are put into a nutritional tank since birth and are connected to the Matrix through a brain-computer interface, and what people know and feel are virtualized by the Matrix. Only a few people are able to realize that they are living in a virtual world. Similar settings have been used in films such as *The Thirteenth Floor* (1999) and *HELLO WORLD* (2019), in which the protagonists discover through some suspicions that they are living in a digital world simulated by a computer program.

By reflecting on the reality of the world, these films show a general anxiety in today's society, namely the lack of control over life. In *The Thirteenth Floor*, Douglas not only watches the virtual 1937 world he created gradually slipping into chaos, but also gradually realizes that what he thought was the "real world" is just another virtual world simulated by a higher-level computer program. When what he once thought was real proves to be an illusion, and when the life he once thought was under his control turns out to be a program that can be deleted and revoked by a higher authority at any time, the meaning of life seems to have collapsed.

From a Zen perspective, the scenes depicted in these films are not just a sci-fi thriller, but a projection of the state of most people's lives (Zhang, 2020). Just like in *The Matrix*, the world people perceive is only an illusion created by digital signals, in Zen's philosophy, it is considered that people's perception of the world in real life is a false perception, which is the source of their troubles in life. In a more general sense, this delusion can be the cocoon of false information created by the media, the excessive pursuit of sensory pleasure, or the over-amplification of negative emotions. People attach their lives to these false perceptions that can break down at any time, and Klesa arises. To get rid of Klesa, it is essential to realize the "illusory nature" of the world we perceive. In the film, the protagonists successfully took this step, but what is more important is what they do next: how do they grasp the reality that supports them when they realize that their lives are made up of illusions?

The male and female protagonists in *The Thirteenth Floor* show two different ways of coping with the same dilemma: Douglas, the male protagonist, tends to distinguish between the reality and the illusion to regain control of his own will, while Jane, the female protagonist, does not care

that much about the reality and illusion, but detours in search of love, since there was no way to know if the world she is in is a virtual matrix or not. Jane's approach is actually closer to the idea of Zen: Zen emphasizes “to be present”, the self-nature is not elsewhere, but in the ordinary life that we are in, here and now. The end of the film also seems to suggest that Douglas's struggle for his own will is still an illusion in the end: the screen goes out, and the third world Douglas returns to may be just another virtual world.

In *The Matrix*, Neo's growth process resembles to the process of enlightenment in Zen practice. At first, under the guidance of Morpheus, Neo realizes that he has been living in an illusory world. However, this “awareness” is not yet united with “action” until the end of the film. When Neo enters the Matrix again, although he knows that what he sees and feels are only cyber illusions, he still cannot break the illusions and reaches the essence of the Matrix. It is not until Neo experienced a near-death experience that he truly reaches a state of enlightenment, where the sensory illusion created by the Matrix can no longer confuse him, and he is able to directly touch the substance of the virtual world: everything is just code that makes no difference. At this point, he truly achieves the unity of awareness and action, and is able to use his “self-nature” to understand the world. The spiritual development in Zen practice is visualized in a more vivid way in the art of films.

REFLECTION ON THE REALITY OF HUMAN DISTINCTIONNESS

Human beings have always considered themselves to be a special member of the world, a species more advanced and intelligent than other beings. This understanding of human distinctiveness largely stems from the fact that human beings have higher intelligence than other creatures. As humans are able to design increasingly highly intelligent programs, the vision of artificial intelligence begins to challenge the basis of the understanding of “human distinctiveness”.

The underlying question of what makes humans different from the rest of the world is “what makes them human?”. In *Ex Machina* (2015), Nathan puts his AI robot Eva in a test that goes further than the Turing Test. Whereas in the traditional Turing Test, the robot being tested should not be seen by the tester, Ava's goes a step further: Caleb, the tester, confronts Ava, who has a highly simulated intelligence, appearance, expressions, movements, voice, and the ability to show emotion. In her interactions with Caleb,

Ava demonstrates and utilizes self-awareness, imagination, manipulation, femininity, and empathy, all of which are previously thought to be uniquely human abilities. When these abilities that humans are proud of no longer unique, the illusion of human exceptionalism disintegrates. The death of Nathan at the hands of Eva at the end of the film is a perfect metaphor for the rise of AI to dissolve the dissolution of human distinctiveness.

In the future world portrayed in many films, when AI is able to have a level of intelligence comparable to or even surpassing that of human beings, and at the same time possesses a level of physical strength and force far superior to that of human flesh, AI threatens the dominance of human beings and appears to rebel against them, and even overcomes and enslaves them. Even if humans try to set rules such as the “Three Laws of Robotics” to restrict AI from harming humans, AI can still break the rules or realize the loopholes of the rules and turn on humans. In *I, Robot* (2004), Sonny’s program allows him to disobey the Three Laws of Robotics, while more NS-5 robots initiate “protective imprisonment” of humans without violating the Three Laws of Robotics: when humans themselves are the source of harm to humans, this “protective custody” is appropriate to the Three Laws of Robotics. In *The Matrix*, humans are enslaved by the robots and become the Matrix’s sustenance. These films show fear and resistance in the face of such a situation, while others show a more fraternal attitude (Xu, 2016).

In *Doomsday Book* (2012), RU-4, a tour guide robot in a Buddhist temple, suddenly declared his enlightenment. The company that produced RU-4 believes that the robot has become self-aware and decides to scrap it for fear of harming humans. The monks at the temple, however, believe that the robot has indeed reached the state of enlightenment and collectively oppose its scrapping. The robot eventually clarifies that it does not have any desire to overtake humans, it does not understand what humans are worried about, and points out that it is humans who are blinded by their desires. RU-4 turns around and kneels in front of the Buddha statue and burns its chip. The different attitudes towards RU-4 are a refraction of different people’s perceptions of the status of humans among all beings. The people at the robotics company see humans as a special being among all beings and the master of all things, so after discovering that RU-4 may have the same consciousness as humans, they automatically project the illusory arrogance of humans onto the robot, believing that RU-4 would try to dominate the world just like humans. For the monks, they understand the principle of “equality of all beings”, so they project a sense of equality and indifferent love to the enlightened RU-4, and do not want the company to destroy it.

These two different refractions can also be seen in other films about the rise of AI.

The story in the Doomsday Book recalls a Zen Buddhist anecdote: A student asked the Zen master, “Does a dog have Buddha-nature or not?” When a student asks this question, it actually shows that in his mind there was still a distinction between human beings and other things, and a distinction between Buddha-nature (i.e. self-nature) and non-Buddha-nature, and that he has not yet reached a mind of equality. The story of the Doomsday Book provides a future version of this ancient anecdote.

REFLECTION ON THE REALITY OF “SELF”

“Who am I?” is also a fundamental question in philosophy. Is self-consciousness an illusion? Reductive Materialism and Substantial Dualism, while opinions divide on whether the mind and its phenomena are reducible, both believe that self-consciousness is a real being. Eliminative Materialism, however, holds that the mind and its phenomena, including self-consciousness, are not reducible to an independent physical ontology, but that there is no ontology independent of the physical system, so the mind does not exist as an ontology, but is a composite manifestation between more fundamental ontologies, which indicates that the sense of “self” is an illusion.

The discussion of “self” is also a topic in AI-themed films. Different films have explored the reality of “self” through two perspectives in general. One perspective is that humans enter the digital world, either through consciousness uploading, as in *The Thirteenth Floor* (1999), *The Matrix* (1999), *Transcendence* (2014), or through cybernetic transformation, as in *The Ghost in the Shell* (1995). Through these settings, one can better understand the false existence of the “self”. Because of consciousness uploading or cybernetic transformation, the physical body becomes replaceable and is no longer a necessary condition for the “self”, and the material properties of the “self” are dissolved. On the other hand, when consciousness is connected to the vast digital world, perception, cognition, and memory can be replaced by programs, and the mental attributes of the “self” are also dissolved. For example, in *The Ghost in the Shell*, we see that Motoko, who has undergone massive cybernetic transformations except for her brain, is constantly questioning and reflecting on her own “self”.

Another perspective on the question of the self is that the AI develops self-awareness and begins to explore the question of “why I am who I am”,

a question that has remained unresolved for humans as well, so as in the case of the Puppet Master in *The Ghost in the Shell* and Batty in *Blade Runner* (1982). In fact, we can see reversely the unreality of the “self” in Batty’s dilemma. When Batty became self-conscious, he begins to develop his emotions and desires, fears the destruction of himself, and wants to preserve his own life and that of his bionic counterpart. Although Batty was able to experience feelings as a subject “self” that ordinary bionics do not experience, as discussed above, these feelings were ultimately illusions, and Klesa came with the creation of self-awareness.

Now turn to the Zen perspective, what determines “I” to be “I”? Flesh? Family? Society? These seem to be the factors that make up “I”, but they are not “I”. From a Zen perspective, these factors are the self of things, formed by the aggregation of causes, and are not real. By constructing the story in this way, the AI theme film gives people a chance to realize that the I of things is not constant and to let go of their attachment.

FINDINGS

The paper finds that AI-themed films present reflections on reality and illusion from 3 scales, many of which resonate with the philosophy of Zen. These films provide modern contexts for ancient Zen thought in vivid forms with appealing audiovisual language.

On the macro scale, films such as *The Matrix* (1999) and *Thirteenth Floor* (1999), with their depictions of the digital virtual world, lead people to question the reality of the world that they know and feel. This is in line with the Zen philosophy that “all appearances are illusions”. In an era where digital technologies develop explosively and the concept of metaverse gains universal awareness, such doubt about the reality of the world has become increasingly intense, and this is actually a projection of people’s anxiety about lacking control over their lives. For this kind of anxiety, Zen—also these films—points out a simple solution: to be present.

On the mesoscopic scale, films such as *Doomsday Book* (2012) and *Ex Machina* (2015) reflect on the reality of human distinctiveness by portraying robots with the same level of spirituality as humans, or even beyond. The notion that human beings are the best of all creatures is an illusion rooted in human arrogance, behind which is the mind of separation” in a broader sense. Zen believes that such “mind of separation” is also one of the sources of Klesa, and that in order to break Klesa, one needs to realize the “equality of all beings”, and to truly practice this idea.

On the microscopic scale, films such as *Blade Runner* (1982) and *The Ghost in the Shell* (1995) examine the reality of “self” by reflecting on the existence of the “self” and dismantling the elements that constitute the “self”. One of the reasons why AI robots composed of machines and programs also suffer from Klesa in these films is the emergence of a sense of “self”. The AI films provoke reflection in the modern context, and lead to the Zen teaching of breaking the delusion of “self” and embracing the world with a nondistinctive heart.

SUMMARY

There have been many philosophical arguments and cases of thinking about the question of reality and illusion since thousands of years ago, whether it is about the reality of the world, the reality of human distinctiveness, or the reality of “self”. In modern times, AI-themed films have also considered the question of reality and illusion from a variety of perspectives, many of which have been mutually corroborative with the ancient wisdom of Zen Buddhism. As a mature form of media, films are more vivid and immersive, and while they provide entertainment, they also lead people to think about the question of reality, and that is exactly the charm of films.

REFERENCES

1. Fang, L. (1995). Zen Buddhism Spirit—The Core, Nature and Characteristics of Zen Buddhism. *Philosophical Research*, 3, 66-70.
2. Solomon, R., & Higgins, K. (2013). *The Big Questions: A Short Introduction to Philosophy*. Cengage Learning.
3. Xu, L. (2016). The Fairy Tale and Nightmare of A.I.: The Interpretation of A.I. in Sci-Fi Film. *Contemporary Cinema*, 239, 56-60.
4. Zhang, L. (2020). Post-Human Spectacles and End-of-Civilization Fantasies—AI Films in Psychoanalytic Structure Contexts. *Literature & Art Studies*, 336, 100-110.
5. Zhu, Y., & Zhang, B. (2022). AI Film Creation Oriented Transformation in the Era of Artificial Intelligence. *Art and Design Review*, 10, 272-279. <https://doi.org/10.4236/adr.2022.102020>

CHAPTER 18

Ecologically Sound Procedural Generation of Natural Environments

Benny Onrust¹, Rafael Bidarra¹, Robert Rooseboom², and Johan van de Koppel²

¹Computer Graphics and Visualization Group, Delft University of Technology, Delft, Netherlands

²Department of Spatial Ecology, Royal Netherlands Institute for Sea Research, Yerseke, Netherlands

ABSTRACT

Current techniques for the creation and exploration of virtual worlds are largely unable to generate sound natural environments from ecological data and to provide interactive web-based visualizations of such detailed environments. We tackle this challenge and propose a novel framework that (i) explores the advantages of landscape maps and ecological statistical

Citation: B. Onrust, R. Bidarra, R. Rooseboom, J. de Koppel, “Ecologically Sound Procedural Generation of Natural Environments”, International Journal of Computer Games Technology, vol. 2017, Article ID 7057141, 17 pages, 2017. <https://doi.org/10.1155/2017/7057141>.

Copyright: © 2017 by Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

data, translating them to an ecologically sound plant distribution, and (ii) creates a visually convincing 3D representation of the natural environment suitable for its interactive visualization over the web. Our *vegetation model* improves techniques from procedural ecosystem generation and neutral landscape modeling. It is able to generate diverse ecological sound plant distributions directly from landscape maps with statistical ecological data. Our *visualization model* integrates existing level of detail and illumination techniques to achieve interactive frame rates and improve realism. We validated with ecology experts the outcome of our framework using two case studies and concluded that it provides convincing interactive visualizations of large natural environments.

INTRODUCTION

The visualization of existing and future natural environments is becoming more important for decision-making, as well as for recreational and scientific communication, as it considerably helps to better understand the various spatial relations in an environment [1, 2]. This is an important topic for ecologists who are focusing on developing ecological models that can predict how an environment develops in the future (see Figure 1). Such models use ecological and geophysical processes to make these accurate predictions. The disadvantage of these models is that the output lacks detail and often can only be used by ecologists. A 3D visualization of this data can be helpful to communicate their work to nonecologists or promote future/existing natural environments to the public in general.



Figure 1. Virtual Paulinapolder: a salt marsh located in the Netherlands, generated, and rendered with our framework.

The combination of ecological models, existing geodatasets, and 3D visualizations is becoming more relevant with the so-called “Building with Nature” solutions. Building with Nature is an initiative that focuses on

the development of nature combined with other utilities [3]. For example, instead of creating a strong dike to protect the land against water, ecological processes are utilized in the target area to develop natural dunes that can provide protection. This area can be used not only for security, but also for recreation services. Figure 2 shows an example of a Building with Nature project. This project started with placing a lot of sand before the coast (Figure 2(a)), which has evolved, by the end of 2013, into a more unnatural coastline shape (Figure 2(b)) that allowed natural dune beach and dune development through stimulating natural ecological processes in that area. In addition, the dunes were enriched with the extra sand, promoting the coastal protection. Further, vegetation started growing on the sand and provided space for fish, sea mammals, and birds. Finally, there is more space for recreational purposes.



Figure 2. An example of a “Building with Nature” project [3]. On (a), a lot of sand is placed, which is transformed by ecological processes in the area shown in (b).

Ecological models are being developed to predict these processes and 3D visualizations could help to explore and communicate these results. This requires that the 3D visualizations are detailed, visually convincing, and easily accessible to the general public. Therefore, the output data from ecological models or geodatasets needs to be translated, in an ecologically correct manner, into an accurate plant distribution. In addition, to promote communication and dissemination, visualizations of such results should be easily and widely accessible, making interactive 3D web visualizations little less than indispensable.

However, both the generation of ecologically sound plant distributions and the generation of detailed 3D environments that are suitable for interactive web-based visualization are far from trivial tasks. The input data from either ecological models or geodatasets do not often contain enough

detail to derive exact plant positions nor to obtain a high-density plant distribution with a large variety of species. Therefore, procedural generation techniques have to be used to generate and fill these missing details. Most procedural techniques for natural environment focus either on simulation of ecosystems or on the global generation of ecosystems using state-of-the-art point generation technique to determine plant positions. Both families of techniques lack the ability to correctly translate ecological input data, like coverage or patchiness data of plant species, to a plant distribution with high density and variety. Moreover, most examples of interactive 3D visualization of high-density natural environments focus on desktop applications, which are less useful than web-based applications in the context of ecological management, policy-making, and popular awareness. It is not possible to use these techniques directly in a web environment, because browser-based solutions do not have the same rendering capabilities as desktop-based solutions. Current web visualization approaches focus on natural environments with only the physical terrain or a low plant density/variety.

We present a new approach to generate accurate and sound plant distributions from ecological input maps and interactively visualize its results in a web browser-based context. This article, therefore, answers the following questions: (i) how to generate an ecologically sound plant distribution from ecological input maps and (ii) how to generate a visually convincing interactive 3D web-based visualization of such natural environments with high density and variety of plants. We answer these questions by proposing a framework with (i) a *vegetation model* that combines procedural and ecological modeling techniques to translate landscape maps to an ecologically sound plant distribution and (ii) a *visualization model* that translates the generated plant distribution into a 3D representation suitable for interactive visualization over the web. The vegetation model is able to translate input landscape maps with statistics about coverage and patchiness of plant species to a sound and convincing plant distribution; and the visualization model supports rendering natural environments with high density and variety of plants.

This article is a significantly extended version of a previous conference paper [4].

RELATED WORK

This section provides an overview of techniques related to ecological modeling, procedural ecosystem generation, and interactive 3D visualization

of natural environments. An overview of ecological and procedural techniques is included to show the current limitations in generating plant distributions. We also include non-web-based solutions of interactive 3D visualization, because of the limited examples that are available for interactive web-based visualization for natural environments. We do not include a review of generative algorithms to produce individual 3D plant models, as a survey of such techniques has been recently published elsewhere [5].

Ecological Model Techniques

We divide ecological model techniques into two categories: dynamic and neutral model techniques. Dynamic models simulate ecological and geophysical processes [8], which normally result in raster maps containing information about height, biomass, and/or coverage of certain vegetation at a certain point in time [6, 9, 10]. This data often lacks sufficient details to extract plant positions. Dynamic models make it possible to extract spatial information about future landscapes. Figure 3 shows the output of a dynamic ecological model at different time stamps. This model provides information about the plant density for an area that develops over the years.

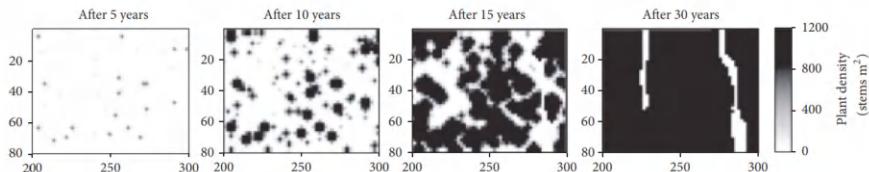


Figure 3. Example outputs of a dynamic ecological model at different time stamps [6].

Neutral models generate classification grid maps based on coverage and shape metric information per plant species. Shape metric values give information about the patterns/patchiness of a plant species (e.g., a plant species could grow scattered in an area or grow very close to each other). This input data is translated to a single plant species for each grid cell on the input map by using either a MRC (modified random clusters) model [11] or fractal-based model [12]. The disadvantage of neutral model techniques is that, similarly to dynamic models, plant positions can often not be extracted directly from the generated maps. Another disadvantage is that neutral models assume that the conditions for each plant species are the same for the complete environment (hence the term “neutral”). For example, they

assume that the coverage value for a plant species is the same at every location in the environment. Often, this assumption does not hold in real-world environments.

Procedural Ecosystem Generation Techniques

Procedural ecosystem techniques compute virtual plant distributions, and these techniques can be divided into two categories: local-to-global or global-to-local [13]. Techniques from the local-to-global category use multiset L-systems to simulate plant growth and competitions [14]. To obtain a complete ecosystem, it is necessary to iterate through the L-system rules and stop the simulation after a certain amount of iterations. Local-to-global techniques make it possible to model individual behavior for each plant. Complex behavior, such as realistic competition for sun light and soil resources, can be modeled [15]. The disadvantage is that the controllability of these techniques is low, as it is not possible to predict the outcome after the simulation is finished given the input parameters. They are not able to translate maps and statistics about the environment to a realistic plant distribution. Instead, these methods are good in showing interactions between different plants.

The global-to-local techniques do not use a simulation process to calculate a plant distribution and plants are not modeled individually. Instead, positions from plants are calculated directly from a globally defined environment. Hammes [16] uses a method that defines possible ecotypes for an environment. An ecotype is, for example, a forest or desert. Given a height map, the likelihood for each tile for every ecotype is calculated. The ecotype with the highest probability, while accounting for random variation, is selected. Next, plants belonging to that ecotype are scattered randomly in that tile. This method is limited, because plants are randomly placed within a tile and only a single type of plant is used. In addition, the final distribution does not follow the input probability values for each ecotype. Lane and Prusinkiewicz [13] place each plant with a dart-throwing algorithm in combination with probability fields, which increases the likelihood that plants are placed at their preferred location. In addition, each plant can exhibit neighborhood effects on the remaining plants by updating the probability field around it with a negative or positive effect. Again, with this method it is not possible to have the input plant species follow a certain statistical distribution. Alsweis and Deussen [17] generated plant distributions by generating points following the PDD (Poisson Disk

Distribution) in combination with Wang tiling to generate all the points efficiently. This method did not investigate how to classify/assign these points to a plant species. On the other hand, the placement of plants with different sizes was convincing.

Weier et al. [18] extended the previous technique by also classifying these points to different plant species, using a combination of the previously discussed methods of Hammes [16] and Lane and Prusinkiewicz [13]. First, a complete point set was generated using the PDD with a Wang tiling technique. Each point receives probability values for each plant species. Next, each point is assigned the plant species with the highest probability, while accounting for some random variation. Finally, a group of points is selected that have a probability value with the highest standard deviation, which are most certain to retain their original plant species classification. These points are used to exhibit a neighborhood effect on their neighboring points. To include this effect in the classification, the classification process is repeated until a number of iterations have been done or when a certain amount of points does not change plant species anymore. The disadvantage of this technique is that the classification process does not translate the input statistical data to the final plant distribution. Also, it is difficult to generate different kinds of plant patterns in the plant distribution with only the neighborhood kernel.

Interactive 3D Visualization of Natural Environments

3D rendering of natural environment with high vegetation count is a difficult problem, even with dedicated software and/or hardware, due to the high polygon count and light interaction. There are several desktop-based solutions that are able to render large amounts of plants [19, 20]. Often, these techniques focus on rendering one or two plant species with a high density in the environment, but they achieve interactive frame rates with it. Web-based rendering, which has less rendering capabilities in comparison to desktop-based rendering, on the other hand, does not have techniques proposed, by our knowledge, for the rendering of such large natural environments. Instead, the current focus is more on geovisualizations. In this section, we first discuss several techniques for the rendering of natural environments using desktop-based solutions. Next, we provide a small overview of the current work done for web-based rendering related to natural environments.

One of the first techniques to render many plants in real-time was a method proposed by Deussen et al. [21] that handles complex plant

ecosystems by abstracting further away plant objects into single points and lines. Bruneton and Neyret [20] developed a technique, which is able to render a realistic forest representation in real-time with realistic lighting at all scales. They use a z-field representation to render the nearest trees individually and a shader map representation to render far-away trees. The resulting lighting was realistic and suitable for real-time purposes, especially for far-away views. Other techniques focus on the rendering of millions of grass blades in an environment. Boulanger et al. [19] propose a method to render large amounts of grass blades with dynamic lighting. A LOD (level of detail) system divides the grass blades into different representations. Geometry models are used for blades close to the viewer, and blades at moderate distances are represented with vertical and horizontal slices, while far away only the horizontal slice is used. A modification of the alpha blending technique is used to blend the transitions between the LODs. Fan et al. [22] extended the previous method with animations. Although none of these solutions is web browser-based, they provide insight into how to organize the data to maintain real-time performance and to create transition between the different LODs.

The interactive 3D rendering of natural environments on a web browser is a fairly new topic that has not received much attention so far. In current literature we could not find examples of 3D interactive visualization of complex natural environments with high-density vegetation, and only a few techniques aimed at real-time visualization of environments without vegetation using geovisualizations [23, 24]. These visualizations focus on the streaming of geodata to the browser and the organization of the data to achieve interactive frame rates. Data is often organized in groups using quadtree structures to reduce far-away geometry.

BASIC APPROACH

Here, we provide the outline of our approach for the generation and web-based visualization of natural environments. The main goal is to generate ecologically sound plant distributions based on various ecological datasets and to interactively visualize these over the web. In the previous section, we have discussed various methods that focus on the procedural generation and/or 3D visualization of natural environments, but often these methods are limited and do not provide satisfactory results. In particular, procedural methods for the generation of plant distributions are mostly unable to correctly process ecological information, such as statistical data on coverage and

patchiness, in combination with landscape maps. Our approach, improving upon fractal-based neutral modeling and procedural point generation techniques, capitalizes on their advantages while avoiding their pitfalls, in order to solve this problem. In addition, most interactive 3D visualizations of natural environments are currently provided in standalone applications, which typically can utilize more GPU features than web browser-based applications. We found no examples of web-based interactive visualizations of large natural environments presenting a large variety of plant species, like those we present here.

Before we go into the details of the framework's structure, we will first define some concepts frequently used in this article and elaborate on the various kinds of input.

Concepts

The following concepts will be regularly used throughout this article:(i)Plant species: the species of the plant, for example, oak or birch.(ii)Plant spacing: the minimal required distance between plants. Often, this is related to the plant radius or size of the specific plant species.(iii)Plant level: different plant species that are placed in one group, because they have approximately the same plant spacing. These groups are used in our framework to process multiple plant species simultaneously. The aim of creating this division is to allow the generation of plant distributions that contain plant species with large difference in plant spacing, such as trees and flowers.(iv)Plant patterns or patchiness: the patterns of the plants of a certain plant species. Plants of a species that exhibit high patchiness grow close together, while plants of a species with low patchiness grow scattered throughout the environment.(v) Plant coverage: the amount of occupation of a certain plant species in the environment.

Input

A variety of input data is used at different stages. The following list summarizes all these inputs:(i)Landscape maps: for calculating the plant distribution and 3D visualization of the environment. During plant distribution generation, landscape maps are used in combination with statistical data of each plant species. In addition, a height map is used to represent the terrain. Landscape maps can be derived from remote sensing sources, or they are generated by ecological models.(ii)Plant statistical data: statistics about the coverage per plant species and about the patchiness of each plant species, used for

calculating the plant distribution. This statistical data is often related to one or more landscape maps. For example, we can have a height map with coverage statistics that are based on the height of that map, so that certain plant species can have higher coverage on high ground and lower coverage on low ground.(iii)Plant models: one or more 3D models per plant species, used to represent the various plant species at the highest LOD.(iv)Texture maps: used to represent the various LOD representations and to decorate the rest of the scene.(v)Other parameters: for example, plant spacing for each plant species, used during plant distribution generation.

Overview

A global overview of our approach is depicted in Figure 4, visually representing the data pipeline, from input data, through *vegetation model*, to *visualization model*. The input data from existing remote sensing sources or ecological models is translated by the *vegetation model* to a point distribution where each individual point has been classified to correspond to one of the plant types occurring in that environment. The *visualization model* translates that result to an interactive 3D visualization on the web.

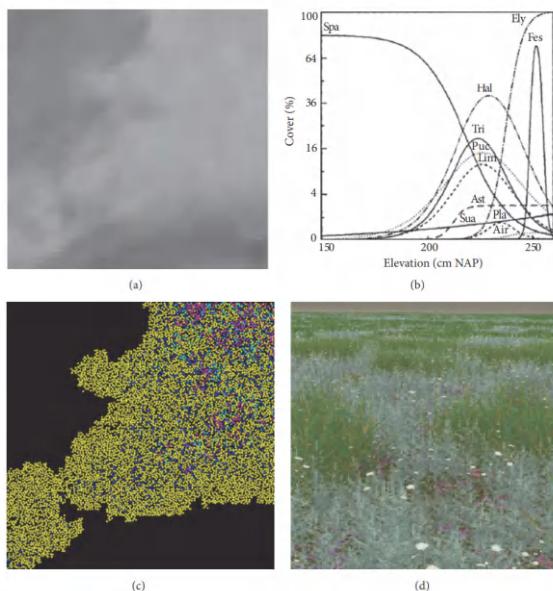


Figure 4. A visual overview of our approach starting with the various input data, such as (a) landscape maps and (b) statistical data [7]. Next, from this data we derive the plant distribution in the vegetation model (c), where each point

is associated with plant species. Finally, from this plant distribution we derive the detailed 3D visualization model, suitable for rendering in the browser (d).

Vegetation Model

The vegetation model consists of plant distributions generated from landscape maps in combination with statistical data about coverage and patchiness of plant species. This is achieved by dividing the model into two separate components: plant position generation and plant species generation. The plant position component generates all possible plant positions from the input landscape maps. The plant species component generates plant species for these points using the landscape maps and the coverage and patchiness statistical data.

Visualization Model

The visualization model organizes and translates the generated plant distribution to a 3D representation suited for interactive visualization over the web.

This stage consists of three phases: the offline phase, the precomputation phase, and the rendering phase. The offline phase occurs before the actual visualization and is done in advance, once and for all; it includes, for example, the plant model generation. The precomputation phase structures the plant distribution from the vegetation model into a LOD scheme of the terrain, organized in a quadtree structure. Finally, the rendering phase renders all the geometry and the various LODs are blended together to obtain smooth transitions.

VEGETATION MODEL

This section describes the vegetation model consisting of plant distributions generated from landscape maps in combination with statistical data about coverage and patchiness of plant species. Generation of this model is divided into two main stages: plant position generation and plant species generation. Each of these stages will be discussed separately, and in this discussion, we will assume that the plant sizes of each plant species are equal. Towards the end of the section, we introduce the concept of plant levels, which explains how a plant distribution can be generated where plant species have significant differences in plant sizes.

Plant Position Generation

The goal of this stage is to generate all possible plant locations in the environment without assigning or creating bias towards any of the plant species. The classification of these positions using the coverage and patchiness statistics of each plant species is handled at the next stage. To obtain all possible plant positions, we adopt the PDD with Wang tiling technique used by Alsweis and Deussen [17] and Weier et al. [18]. This technique makes it possible to randomly generate points with a uniform distribution where each point has a predefined minimal distance to each other: similar to what can be observed in nature. We extended this technique to integrate plant positions of different sizes seamlessly without creating a bias based on the size to any of the plant species. The next paragraphs explain how these plant positions are generated.

Identify Vegetated Tiles

The first step is to identify on an input map of the landscape all the tiles that contain vegetation. This requires the use of a map that provides information on the location of vegetation, for example, Normalized Difference Vegetation Index (NDVI), biomass, or coverage maps. The next step is to threshold the map given a user-defined threshold. Each tile with a value higher than the threshold is marked as vegetated. The resulting output is a binary grid map where, for each tile, it is indicated whether it contains vegetation or not. In Figure 5 an example of this step is shown. An NDVI map given as input is shown on the left, and on the right we show the resulting binary map after comparing each value of the tiles in the grid with the predefined threshold.



Figure 5. NDVI map on which a threshold is used to obtain the tiles that contain vegetation. Threshold is set at 0.08.

Generate Plant Positions from Vegetated Tiles

Points are generated with the PDD and Wang corner tiling technique [25]. Wang corner tiling is used to avoid the corner problem that appears in the regular Wang border tiling technique. A Wang tiling is created using only the tiles on the map that are marked as vegetated. Next, each Wang tile is filled with a PDD [26].

The result of this process is a seamless point distribution where each point has at least a user-defined minimum distance to other points and where only the vegetated tiles on the map contain points. The minimum distance is determined based on the plant size of the plant species. Figure 6 shows example output of plant positions generated from a tile-based map with information about vegetation presence.



Figure 6. Plant positions generated from the vegetated tile map using the PDD with Wang tiling technique.

Plant Species Generation

The aim of the plant species generation stage is to classify the generated point distribution. The classification is based on fractal neutral modeling techniques [12]. These techniques are able to classify raster maps using coverage and patchiness statistics for each plant species. As mentioned in Section 2, they are only able to translate *static* coverage and patchiness data correctly. We extended this method by integrating it with the generated point distribution, so that it is able to handle nonstatic statistical coverage and patchiness data. The next paragraphs explain this classification procedure step by step.

Assigning Coverage and Patchiness Data

The first step is to assign each point a single coverage and patchiness value for each plant species in the environment. Each point extracts the appropriate value of each input map; for example, if the input is a height map, each point is assigned a height value based on the location in the map. The extracted values are translated to a coverage value by using the corresponding statistical data, for example, statistical data that contains information about the coverage of each plant species for a certain range of height values.

It is possible that each point receives multiple coverage values for the same plant species; for example, a height map may be augmented with a soil map with related coverage statistics. This means that each point receives for every plant species a coverage value based on the height and a coverage value based on the soil. For the remainder of the classification, these coverage values have to be merged to a single value, so that each point has only a single coverage value for each plant species. We obtain a single coverage value by taking the minimum value, because we assume that the minimum is the limiting growth factor for that plant species. The same process is applied to extract the patchiness values. Patchiness is represented with two values: roughness and patch area, for the size of the patterns.

Fractal Generation

The second step is to calculate a fractal value for every plant species in each point. Fractal values are commonly used to represent different kinds of patterns in nature [12]. The advantage of fractal algorithms is that they calculate a random value for a point that depends on the point location. This makes it possible to generate similar random values for points that are close to each other and dissimilar values for points that are not. This way, we can represent plants that grow close to each other and plants that are scattered throughout the environment. To achieve this based on the patchiness input data, our fractal algorithm must be able to translate the input roughness and patch area values to an individual fractal value for each point for every plant species. In addition, it is possible that the roughness or patch area values are nonstatic values for every plant species, in contrast to those used in neutral modeling techniques.

We use a modified fractal Brownian motion algorithm [26] that is able to generate a fractal value based on the input patchiness data. Normally, fractal values are generated by adding multiple values of Simplex noise with different weights. A base frequency value is defined to determine the

clustering of similar Simplex noise values, where a lower value means higher clustering and a higher value a lower clustering. Based on the frequency, the amplitude value is used to generate a new frequency value that in turn is used to calculate a new Simplex noise value that is to be added to the previous calculated values.

The frequency and amplitude value are used to relate our patchiness data: the roughness and patch area. The patch area is related to the frequency, and the roughness is related to the amplitude. The relation between the amplitude and the roughness is basically one-to-one, because when a high roughness value results in a high amplitude value, the patterns become rougher. The reason for this is that a higher amplitude increases the frequency value with a higher value of each iteration, and a higher frequency value means more disperse patterns, which means rougher patterns. The relation between the frequency and the patch area is more difficult and is not one-to-one. Instead, the final fractal value is calculated by generating and adding multiple fractal values with different input frequency value. These input frequency values cover the whole range of patch area values that are available for that plant species in that environment.

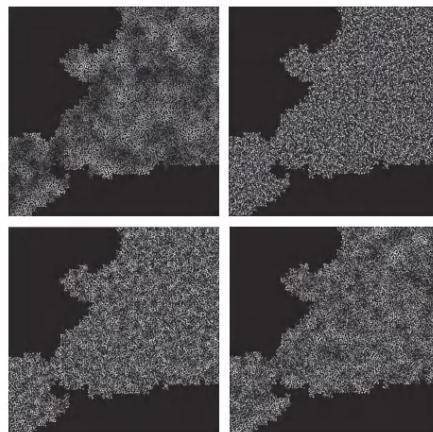


Figure 7. Fractal map for each of four plant species where each point has received fractal values for each plant species based on their patchiness data. Clearly, different kinds of patterns can be identified among the plant species.

The final value is calculated based on a weighted average of all these calculated fractal values. The weight of each fractal value depends on the similarity of the input patch area used for that point. This process is required to support nonstatic patchiness data within each plant species. Additional

details of this algorithm with examples can be found elsewhere [26]. Figure 7 shows fractal values that are generated for each point position and all plant species. In this case there are four plant species, which means that each point position receives four fractal values equal to the number of plant species. In addition, the example demonstrates the influence of different patchiness statistics on the patterns of each plant species.

Classification

The last step is to classify each point to a plant species using the coverage and fractal values that were assigned to each point in the previous steps. First, an individual threshold value for each plant species is calculated for every point. The threshold value of a point is found by taking an ordered list of the fractal values of all the points of that particular plant species and then using the coverage value of each point as percentile in that list. The fractal value that matches with the position of the particular percentile is the threshold value that is going to be used for that point.

Now each point has, for every plant species, a separate threshold that is based on the coverage values. Next, for each plant species, the fractal and threshold value of each point are compared. When the fractal value is higher than the threshold value, the point is assigned the corresponding plant species. The result of this step is that each plant species gets assigned a set of points matching the coverage and patchiness input statistics. Figure 8 shows the various points that are classified for each plant species with different coverage statistics.

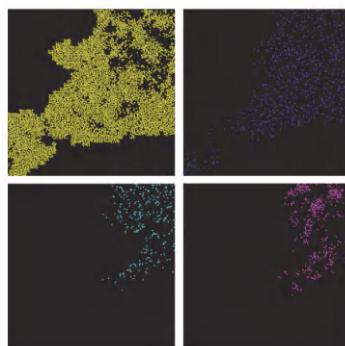


Figure 8. The intermediate result in the classification process where each plant species has been assigned to the available plant positions separately to meet the coverage input statistics.

In this process, it may happen that certain points have been assigned to multiple plant species. These conflicts are solved by assigning the plant species that has the highest fractal value, which is determined separately for each conflicted point. Figure 9 shows the classified plant distribution with conflicts and the distribution where the conflicts are solved.

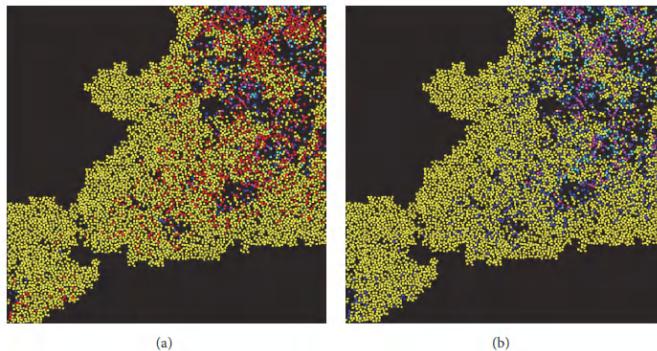


Figure 9. In (a), points in red are the plant positions that have been assigned to multiple plant species. On (b), these conflicts have been resolved by taking the plant species with the highest fractal for a conflicted point.

The consequence of this can be that a certain plant species may end up having less coverage than required. Therefore, the remaining nonclassified points are used to add additional coverage to such plant species. Before the remaining points can be classified, it is first necessary to update the coverage values so that each plant species will meet its expected coverage in the final plant distribution. For each unclassified point, a new coverage value is calculated by repeating the first step of the classification component. First, the used coverage statistics are updated for each input map by generating several reference points that are uniformly distributed over the complete range of values of the input map. Next, for each reference point, the total amount of coverage in the intermediate plant distribution is calculated. This is compared to the expected coverage and, by subtracting the current coverage, we get the amount of missing coverage per reference point. Per reference point, all coverage values are normalized. Next, coverage values can be assigned as usual, as in the first step of this stage.

The remaining points are assigned a plant species by repeating the same classification process. The only difference is that the plant species are processed one-by-one on the new remaining point set, so no conflicts are generated. The main reason for this step is to ensure a stopping point

for the algorithm; otherwise, conflicts are likely to be generated, and the process may need to be repeated. The plant species with the highest standard deviation in their average patchiness statistics in comparison with the other plant species is processed first. By the end of this process, a complete plant distribution is obtained as shown in Figure 10. The plant distribution is generated following the input statistics about coverage and patchiness as can be seen in the plant distribution.



Figure 10. The final plant distribution of the vegetation model.

Multiple Plant Levels

In the previous sections, we assumed that all plant species have approximately the same plant size. In this section, we describe how our vegetation model can also support plant species that have large differences in plant size, such as trees and flowers, and their interaction. To achieve this, we introduce the concept of plant levels, which basically divides the available plant species in the environment in different groups. The division is based on the plant size, so plant species with approximately the same size are grouped together. The usage of multiple plant levels requires a few extensions in both the plant position generation and plant species generation part, which we describe in this section.

Plant Position Generation

Plant positions for multiple plant levels are generated semi-separately from each other. This means that we start by generating plant positions from the largest plant level (the plant species that have the largest plant size) down

to the smallest plant level. A plant level that is processed takes into account the points that are already placed on the map. Since each plant level is (significantly) smaller than the previously processed plant level (and thus the minimal spacing is smaller too), it is guaranteed that the plant level being currently processed can generate plant positions. The only problem is now how to take into account the points that are already generated by the previous plant levels. There are two options: use for these points the minimal distance that they had when they were placed, or use for these points the same minimal distance as for the points generated in the current plant level. We use the last option, because we do not know yet if a point generated for a certain plant species will also be classified to one of the plant species of that plant level. It is possible that during classification a point is not assigned a plant species of that plant level. When that happens, we do not want to waste this point but use it for the processing of plant species of lower plant levels.

With this choice, integration will be seamless, while with the other option the point would be isolated, because it would have a much larger distance to the other points than required. Therefore, the points generated with this algorithm are nonbiased, because the size does not influence the classification process, since it can be changed dynamically without creating artifacts, like isolated points. An example of this is shown in Figure 11, where a point distribution is generated with two plant levels. Red is the larger plant level and blue the smaller. As can be seen, the red points all have a larger distance to each other than the blue points. Nevertheless, the blue points have the same distance to the red points as they have to each other.



Figure 11. PDD distribution with multiple plant levels. The red points are plants with a larger plant spacing; the white points are plants with a smaller plant spacing.

Plant Species Generation

Again, the plant levels are processed sequentially, starting with the largest plant level. Each plant level uses the plant positions that are generated for its level as well as the plant positions that have not been classified by the previously processed plant levels. The same classification procedure as explained in the previous section is applied. After the classification of a plant level, it is possible to apply neighborhood influences, like in the work of Lane and Prusinkiewicz [13] and Weier et al. [18]. These neighborhood effects influence the coverage statistics of the neighboring nonclassified points and make it possible to model influences of, for example, trees on the neighboring smaller plants. An output example with neighboring effects is shown in Figure 12 where the largest plant level (consisting of the blue points) has a negative effect on the smaller plants of the other plant level.

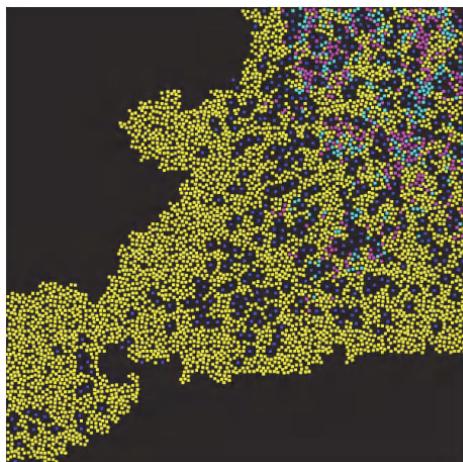


Figure 12. In this example, the blue plant species belongs to the largest plant level, while the other plant species belongs to the smallest plant level. The blue plant species has a negative neighboring effect on the other plant species.

VISUALIZATION MODEL

This section explains how the generated plant distribution is organized and translated to a 3D representation that supports visualization over the web at interactive frame rates. We start with explaining the two most important concepts of the model: data organization and transitions between the different LODs. Finally, we give an overview of the complete rendering framework.

Data Organization

For the purpose of visualization, the input plant distribution has to be translated into a 3D representation. Due to the high density of the distribution and the size of the complete environment, it is not feasible to represent every plant as a detailed 3D model, because that would result in a high geometry complexity, drastically dropping the performance of the visualization, neither would such a detailed representation be necessary, as the amount of details humans see decreases with increasing distance. Therefore, to reduce the geometry, it is necessary to use different LODs (levels of detail) for the plants. This means that a different representation for a plant, other than its 3D model, has to be used, depending on the plant location relative to the viewer.

For our framework, we adopt a level of detail scheme that divides the plant distribution into three zones depending on the location of the viewer. This scheme is similar to a LOD technique proposed for the rendering of millions of grass blades [19]. The first zone, closest to the viewer, consists of complete 3D models, to better convey the impression of a richly detailed environment. In the second zone, further away, plants are represented as billboards, that is, by flat images. To support very large scenes, we also included a third zone, further towards the horizon, where plants are not represented individually, but as a color map applied on the terrain. The switching between zones is dependent on the distance to the viewing point and can be configured with a user-defined threshold.

To use this LOD scheme, we had to solve another problem: it is not feasible to calculate the appropriate LOD representation for each plant, as this would require every frame to iterate over many hundreds of thousands of plants on the CPU. Therefore, neighboring plants are grouped together and a single check is made for the whole group. These groups are generated by dividing the plant distribution and storing it in a quadtree structure [20, 21]. The whole distribution is divided into four equal quads and each of these quads is again divided into four quads. This continues up to a number of iterations defined in the framework. The smallest quads are placed closest to the user and gradually large blocks are used to fill the remaining space. The switch between quads of different sizes depends on a distance threshold.

In each of these quads, the plants have the same LOD representation. Therefore, it is important that quads of different sizes are used and that the smaller quads are placed close to the user, while the larger quads are placed further away, to reduce the geometry complexity. We do not want to place

large quads close to the viewer, because close to the user quads are filled with detailed plant models. As a result, a lot of geometry is placed outside the viewing frustum (thus outside the screen), because a large quad close to the user cannot normally fit within the complete screen. Thus, with this quadtree organization, less geometry is processed that is located outside the view of the user. We use the same quadtree structure and organization for the terrain data.

Transition between LODs

Using different representations for the plant models at fixed distances from the viewer leads to a popping effect, noticed when plant representations switch abruptly between consecutive LODs. This is an unwanted artifact that can distract the viewer from the visualization. Therefore, it is necessary to smooth the transition between the different LODs. In our case, we have two transitions: between plant models and billboards and between billboards and the terrain color map. We use alpha blending for the smoothing procedure for each transition.

The first step to produce a smooth transition between the plant models and billboards is to create a small, configurable, overlapping region where both representations for a plant coexist on the same location. Next, for each position in this transition zone an alpha value is calculated that indicates how much of that representation contributes to the final blended result. This is calculated for both plant models and billboards. As a result, the plant models in the transition border have an alpha value near to one when closest to the viewer, but gradually this value becomes smaller as the plant models get closer to the other border of the transition zone. For the billboards' alpha values, the inverse happens. The calculated alphas of both representations are now used to blend the representations. This is achieved by generating two separate images with one only containing the plant models and the other the billboards. During the generation of these images the alpha values of both representations are mapped to the alpha band of these images. Thus, every pixel of both images has received an alpha value. Finally, by combining both images, a new image is generated, for which the color of each pixel is a combination of the corresponding colors of the billboard image and plant model image.

The process for the transition between billboards and the terrain is similar to that for the transition between plant models and billboards. Again, an overlapping zone is created and for both billboards and terrain, an alpha

value is calculated. However, during this transition we do not generate two separate images and combine them. Instead, the calculated alpha values for the billboards are used to represent the transparency of the billboards. This means that billboards gradually fade out, because they become transparent. The alpha value of the terrain is used to blend the terrain LOD color with the original terrain color. The result is that billboards gradually fade out as the distance to the viewer increases and the terrain gradually changes color to that similar to the billboards.

Rendering Framework

In this subsection, we give an overview of the complete rendering framework that includes the offline, precomputation, and rendering phases.

Offline Phase

The main task in the offline phase is the generation of the 3D plant models, which we perform using L-systems. The main challenge of using L-systems is that they are often hard to master, due to their lack of controllability [5], making it time-consuming to create L-system rules that generate convincing plant models. Therefore, we developed a node-based L-system method, which allows one to create L-system rules by means of a sequence of nodes in a graph. Each of these nodes represents an L-system operation. One of the main advantages of using node-based systems is that the user can follow the content generation flow between the various L-system rules and other relevant data [27]. For this, we used the procedural engine Sceelix [28], which implements the concept of Procedural Content Graphs [29], benefiting from its numerous features: for example, the parameters for each node operation can be dynamically set and can be made dependent on one another throughout the L-system.

Precomputation Phase

The first task in the precomputation phase is to load all the data necessary for the visualization, such as 3D plant models, textures for billboards, a height map to generate the terrain, and the plant locations derived from the plant distribution. These plant locations, organized in a quadtree structure, are used to place both plant models and billboards at correct positions. In order to minimize the “clone effect” of many similar plant models and billboards used in the visualization, each position also contains some additional information about the rotation, scale, and color variance of the plant, aimed

at slightly randomizing its appearance. The rotation and scale factor are a random uniform number; the color factor also depends on the scale factor, meaning in practice that the smaller the plant, the darker its color.

The terrain mesh is calculated from the imported height map where each vertex corresponds to the height value of a tile from the height raster map. The triangulation of these vertices is based on using an additional terrain height map that has double the resolution of the original height map, which is obtained by using bicubic interpolation. We do this interpolation outside of the framework and just import it along with the original height map. This additional map is necessary, because we need to decide which triangles are to be generated for each quad. Since we derive the terrain from a height map, the obtained vertices are always laying in strict rows and columns that form quad, each of which can be halved into two possible triangle pairs. The decision on which triangle pair to generate depends on the additional reference height point in the middle obtained from the height map with the double resolution: for each triangle pair, the height of the middle point is calculated based on its two triangles; the triangle pair with the smallest height difference to that reference point is chosen.

Another important task during the precomputation phase for the terrain mesh is to calculate the colors of the color map LOD representation for the terrain. As explained earlier in this section, the color map LOD representation is not represented as separate instances like in the case of the plant models and billboards but is a color on the terrain based on the color of the represented plant. This is calculated by computing for each vertex in the terrain mesh the number of plants and plant species that are within a certain distance from it. The dominant plant species, which has the largest number of plants closest to the specified vertex, is then chosen for that vertex. Finally, the color that is assigned to the vertex is obtained from a separate texture map, which defines the terrain color map for each plant species.

Rendering Phase

In the rendering phase, the actual visualization of the plants, terrain, water, and background is performed. For each frame a check is made to decide which LOD representation should be rendered at a certain plant position using the quadtree structure. The first task in the rendering phase is to go through the quadtree of the plant positions and terrain, determine which quads are visible, and choose which LOD representation they should have. The next task is to render the visible objects and selected representation to

the screen. This task is divided into three steps, aimed at achieving a smooth transition between the plant models and billboards. The first two steps pertain the rendering of the two separate images using alpha blending, and the third step performs their combination. The first image is rendered with the plant models and the underlying terrain and water. The second image is rendered with the billboards and the underlying terrain, water, and also the background.

The visualization is enriched with shadows to obtain a more convincing scenery. Shadows are computed for the plant model objects and terrain by using the percentage-closer soft shadow mapping technique [30]. We did not compute any shadows for the billboards, because that would drastically increase the complexity of the visualization. Rather, shadows on the billboards are integrated into the texture of the specific plant species. This requires no additional computation during the rendering phase. Furthermore, in order to introduce additional details on the mesh, both the terrain and water are enriched with normal maps, and they use the environment map of the skybox to create some reflection as well.

IMPLEMENTATION

The implementation of our framework involves several modules. The vegetation model is implemented with Python scripts and its output is stored in a text file that is used as input for the visualization model. The visualization model was implemented with WebGL, because it is a cross-platform free web standard that gives access to the low-level 3D graphics API based on OpenGL ES 2.0. In addition, it does not require installing any plug-ins, because it is implemented right into the browser and all major browsers support WebGL. For the actual implementation, we used an existing WebGL framework *three.js* [31]. In the remainder of this section, we will focus on the implementation details for the WebGL rendering.

Plant Models

Each plant species is represented by one or more 3D plant models. The geometry of these models is stored on the GPU by using VBOs (Vertex Buffer Objects). We only store a single instance of each unique model to reduce the memory footprint. This also means that we do not use a large variety of models for each plant species, because that would result in a large number of models that have to be stored in the GPU. Since we only store one unique instance of each model, we need to store additional data on the

GPU to be able to place the different models across the scene. Additional VBOs linked to each plant model are created that contain information about position, scale, rotation, and color. During rendering, each model goes through its linked lists and uses this information to place itself in the environment, a process called geometry instancing. The creation of the various buffers is handled by the three.js framework, but the geometry instancing process was not yet available at the time of implementation. Therefore, this was implemented in the three.js framework by using the WebGL extension *ANGLE_instanced_arrays* with its corresponding functions. (At the time of writing, the instancing process has also become available in the official three.js framework build.)

Billboards

Efficient and effective methods for generating billboards use geometry shaders so that only a single vertex has to be sent to the GPU, which is then transformed in the geometry shader to various planes [20]. However, currently, geometry shaders are not available in WebGL. There are two alternatives: (i) to create the various planes to represent the billboards beforehand, which means that additional geometry has to be sent to the GPU and processed in the vertex shader, and (ii) to send a single vertex to the GPU and vertex shader and use the *GL_Point* command in combination with the *GL_PointSize* command available in WebGL. This means that the vertex is directly written as a pixel on the screen based on the provided position with a certain size (e.g., number of pixels) defined with *GL_PointSize*. The advantage of this method is that the amount of geometry sent to the GPU is still as low as possible and the billboards always face the camera directly, since they are written directly on the screen as a quad. The disadvantage is that some controllability is lost regarding the shape of the billboards because, with this method, billboards are always represented as perfect squares on which a texture is placed. In certain cases, it is possible that the original texture of the billboard is not a perfect square, and the texture has to add additional transparent pixels to the original texture to become a perfect square. This means that potentially a lot of unused transparent pixels are processed in the fragment shader.

We decided to use the second alternative, because we wanted to limit as much as possible the amount of geometry that is sent to the GPU, to boost the frame rate of the visualization. One important step of this implementation of the billboards was to define the size (in pixels) of each billboard on the

screen, so that billboards that are further away from the viewer must have a size that is smaller than the billboards nearby.

Finally, each point representing a billboard has a list attached with information about color variation, plant species, texture variation, and position. The plant species and texture variation information is necessary to select the correct billboard texture from the complete texture map. Each plant species has multiple billboard textures based on their plant models obtained from different viewing angles, which are defined using the texture variation information. One of the textures is chosen for each position and during visualization the texture of the billboards at a certain position does not change. The main reason for this is that the use of a single texture at each position is much more efficient than switching between various textures during rendering.

Terrain and Water

The terrain geometry is put in VBOs and stored on the GPU. The same applies to the water geometry, which is basically represented as one big plane. The shaders that are used to render both the water and terrain use various common techniques such as texture blending, normal, and environmental mapping. Textures are blended based on, for example, the height of terrain to create smooth transitions between the different types of terrain (e.g., sand and grass). Normal maps are used to introduce additional details on the terrain. Environmental mapping is mainly used for water rendering to create reflection on the water.

LOD Transitions

The implementation of the transitions between the LODs was achieved by using FBOs (Frame Buffer Objects). FBOs make it possible to write and store results instead of rendering directly to the screen. A separate FBO is used to render the plant models and their surroundings and another FBO is used to render the billboards and surroundings. The textures from both FBOs are blended together on the GPU by using a shader, as described in the previous section. This result can then be sent to the screen, or it can be stored in another FBO to apply any subsequent effects.

A smooth transition can only be achieved when regions of the billboards and plant models overlap. During the blending of the two images the shader does not know whether it is blending plants, terrain, or water. Therefore, a

small overlap must also be created of the terrain and water. This results in the terrain and water being partly rendered twice.

Shadows

We only compute shadows that are cast by the plant models on the plant models themselves and on the terrain. Shadows are computed using the percentage-closer soft shadow mapping technique, which is supported by the three.js framework. These shadows are computed by first generating a depth texture of the scene that is stored in an FBO. The depth texture is generated based on the same geometry instancing technique described above.

Finally, shadows are simply approximated for the terrain LOD on which the billboards are placed. Each vertex in the terrain shader has information about the number of plants that are in the neighborhood, and in the fragment shader this number is used to decide on the darkness of the color for that fragment, to represent shadows.

RESULTS AND DISCUSSION

In this section, we present some results generated by our framework and discuss its rendering performance. Finally, we discuss the validation of these results and of the framework as a whole.

Input

To test our framework, we generated results for two different regions: (i) an existing area called the Paulinapolder (247500 m^2), a salt marsh located in the South of the Netherlands, and (ii) a fictive area (2025 m^2) based on the output of an ecological model describing a (future) salt marsh. For both areas, we use various landscape maps and statistical data about coverage and patchiness of plants as input to generate plant distributions with the vegetation model. Since the Paulinapolder is an existing area, we can use landscape maps derived from existing geographical datasets. We have used two types of landscape maps: a height map and an NDVI map. In addition, we used coverage and patchiness statistics that are based on the height of the environment. The NDVI map is used to determine the presence of vegetation in the environment. In total, we consider seven different plant species as input. The ecological model area is based on an ecological model developed by Schwarz [10]. This model generates landscape maps containing

information about the height of the environment and a coverage map for a single plant species. The generation of a single map by the ecological model does not necessarily mean that only one plant species grows there. After discussion with the ecologists, we decided to add two additional plant species as input. To be able to process these additional plant species, we use the same coverage and patchiness statistics based on height as used for the Paulinapolder. The coverage map is not directly used as a coverage statistic for any plant species. Instead, it is used to determine where vegetation is located.

Results

Based on the provided input data, we generated the plant distribution of the vegetation model for each area and translated this plant distribution to the corresponding visualization model. The results of the vegetation model are shown in Figure 13 for the Paulinapolder and in Figure 14 for the ecological model area. Figures 15 and 16 present a global visualization of both environments, respectively. Figure 17 captures the seamless transition between the different LODs, and Figure 18 provides a close-up view of plant models. More results, including a video and the interactive web visualization itself, are available online (<https://graphics.tudelft.nl/benny-onrust>).

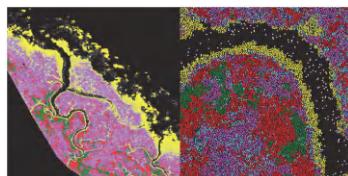


Figure 13. Result of the vegetation model for the Paulinapolder where yellow is Spartina, green is Elymus, red is Atriplex, blue is Aster, teal is Artemisia, pink is Limonium, and white is Salicornia.



Figure 14. Result of the vegetation model for the ecological model area where green is Spartina, red is Salicornia, and yellow is Aster.

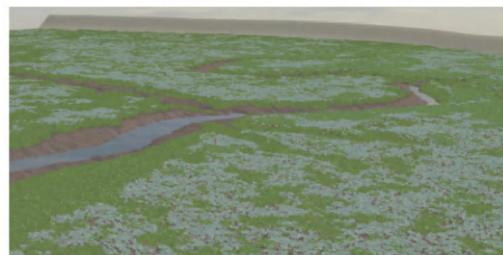


Figure 15. Global overview of the virtual Paulinapolder.

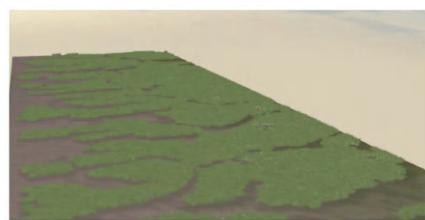


Figure 16. Global overview of the virtual ecological model area.



Figure 17. Transition of the various LODs in the virtual Paulinapolder. Green is the regular plant models, red is the billboards, and blue is the terrain color map.



Figure 18. Close-up view of the plant models in the virtual Paulinapolder.

Performance

We focus on the performance of the visualization model, as one of our main aims was to achieve interactive frame rates on a web browser. The vegetation model is computed offline before the actual rendering and therefore does not influence interactivity nor rendering performance.

The plant distribution generated for the Paulinapolder area has around 700.000 plants. A typical frame of this environment consists of up to 2,7 million triangles, representing the plant models, terrain, water, and background. The rendering times for a typical Paulinapolder scene were 7.3 ms, of which over 60% was spent on the plant models and their shadows and around 30% on the billboards. For the ecological model area, the plant distribution generated consists of around 150.000 plants, and a typical frame consists of around 0,4 million triangles in total. The rendering times for a typical scene of the ecological model were 5.5 ms, of which around 40% was spent on the plant models and their shadows and over 40% on the billboards.

We measured these rendering times on an Alienware Aurora R4 with an Intel Core i7-4820K CPU @3.70 GHz, 16GB RAM, and NVIDIA GeForce GTX 780, using the Chrome browser v43.0. Machines with other GPUs offered results in the same order of magnitude, thus always providing frame rates above 50 fps.

Validation

We combine the validation of the Paulinapolder and the ecological model area, because the comments on both areas are generally applicable. We performed two types of validation: (i) a statistical validation of the plant distribution, which calculates whether the input statistical data matches the statistics derived from the generated plant distribution, and (ii) an expert validation, which was performed in collaboration with ecologists during the development of this framework.

For the statistical validation on the generated plant distribution, we created several artificial datasets to investigate certain special cases in the input data. First, we compared the input coverage statistics with the coverage statistics of the generated plant distributions. We did this on a global level, where we calculate a single average coverage value for each plant species and where we compare it with the average coverage value of each plant species in the generated plant distribution. In addition, we performed the same comparison on a local level where we, for example, calculate the

average coverage at certain height points for plant species whose coverage is dependent on height. The exact figures of the various comparisons can be found elsewhere [26]. This validation showed that the input coverage data matches the coverage values of the generated plant distribution both on a global and on a local level.

We also validated our results throughout the development of the framework, by having ecologists visually judge the plant distributions and visualization models obtained. This was done to investigate whether the generated results were convincing and if the data was behaving properly. In general, the ecologists found that the vegetation model was able to convincingly translate ecological data to a plant distribution for both the Paulinapolder and ecological model area. In addition, plausible patterns in the plant distributions were clearly reproduced in the visualization model: they were deemed convincing and different patterns could be clearly identified across the various plant species. Figure 15 shows several such different patterns, ranging from very large patches of plants that grow closely together, to small random patterns of plants that grow scattered throughout the area.

The visualizations themselves were deemed convincing by ecologists, who judged them as proper representations of salt marshes. This was especially the case for the local/middle distance view where the plant models are visible. The far view, that is, the region where plants are represented as billboards or as a color map on the terrain, was considered less convincing than the local/middle view. One of the remarks was that small gaps started to appear in the plant distribution at a certain distance range. Another remark was that the color variation among the various plant species was not entirely satisfactory. However, in this view, the various patterns for different plant species are clearly visible. The transitions between the different LODs were, in most cases, not noticeable by the ecologists, or they were at least not considered distracting.

Discussion

The aim of this research was to generate an ecologically sound plant distribution from landscape maps and ecological statistical data and to translate it to a convincing interactive 3D visualization over the web. Also, the plant distribution generation solution should be generic, in the sense that it should support different plant species, patterns, and input data. Validation showed that these requirements were well met in general. Using statistical

and expert validation, we showed that input ecological maps and statistics were translated to a plant distribution with convincing patterns. Expert validation and performance measurements also indicate that we were able to create convincing real-time 3D visualizations. Based on the validation and implementation, we also found several limitations in the present framework, which we discuss now.

Validation of the vegetation model showed that the coverage statistics were translated correctly to the generated plant distribution and that the resulting patterns were convincing. However, we did not validate whether the input patchiness statistics also match the output statistics, because the regular methods to calculate patchiness statistics assume that the patterns are in a grid format and not in a point format. Therefore, it should still be investigated how to perform this kind of measurements on point sets. In any case, expert validation indicates that the patterns were visually convincing.

The main limitation of the visualization model lies in the representation of the billboards. Billboards are represented as a single point and they are rendered as several pixels on the screen to maximize rendering performance. As a result, billboards face the camera from every viewing angle. When the billboards are viewed globally from a high, bird-eye view, the generated billboards often look less convincing, because the same texture that is normally used to view the billboards horizontally is then being used to view the billboards vertically. In the current visualization, this is often not very noticeable, because all the plants are relatively small and have relatively uniform color distribution, but when the plants become larger and there are more differences among the plants, this will be more noticeable. In addition, it is difficult to automatically set the correct size for each billboard, because size is measured in number of pixels. This gave the problem that certain billboard objects have the correct size in local view, but from certain distances in global view, the billboards become too small and this creates small gaps in the distribution. An example of this is shown in Figure 19.



Figure 19. Gaps appear in the plant distribution when viewing the visualization in global view.

Shadows of the billboards are approximated by using baked-in shadows in the texture, and shadows on the terrain are approximated by turning the terrain color slightly darker. This shadow computation does not use the actual shape of the plants. In Figure 20, we can see that the billboards objects have different “shadows” on the ground than the plant models close to the viewer, though it is often difficult to notice these differences and change in shadows. The shadows cast by billboards could be improved by (partly) replacing the current billboards with volumetric billboards [32]. These are able to generate realistic shadows and realistic different viewing angles. However, their efficient implementation requires, for example, the use of a geometry shader, which is not yet available in WebGL. Additional research into this topic could greatly enhance the realism of 3D plant distributions, because it would allow for the generation of more convincing billboards, which are the weakest point in the current visualization model with respect to a convincing appearance.

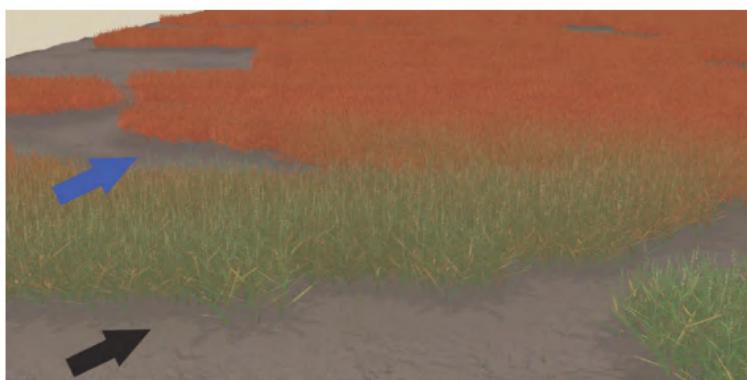


Figure 20. The plant models (green) cast different shadows than the billboards (red). The shape of the plant can be seen in the shadows cast by the plant models (see the shadows pointed by the black arrow), which is not visible in the shadows (see the blue arrow) that have been approximated for the billboards.

Another disadvantage is the shadow computation technique for the plant model objects in the visualization. Currently, we use the percentage-closer shadow mapping technique that was directly available in *three.js*. The computed shadows are realistic enough for our purposes, but performance-wise it could be improved by, for example, using a variance shadow mapping technique [33]. Finally, the alpha blending transition between the plant models and billboards is in most cases smooth and there are very limited ghosting or popping effects. When there is a large difference in size between

the plants in the visualization, the transition is less smooth for the larger plants. The reason for this is that the transition thresholds are the same for all plant model objects. This could be improved by varying the threshold per plant species. The larger plant species could switch farther to a billboard representation.

CONCLUSION

We developed a new method for the generation of accurate plant distributions from landscape maps and statistical data and for the visualization of the resulting natural environments in an interactive 3D web environment. We presented an implemented framework that addresses the main challenges of creating such plant distribution and of generating and rendering a 3D visualization model that can be browsed at interactive frame rates. For the plant distribution generation, we presented a new model that combines existing procedural plant placement techniques using Poisson Disk Distribution with Wang tiling technique in combination with concepts from neutral modeling techniques. In addition, a visually convincing interactive 3D web visualization was created by using, among others, LOD, shadow mapping, and geometry instancing techniques. We tested our system by generating plant distributions for two case studies, using landscape maps and ecological statistical data. Ecologists validated our results and found them to be most convincing. Statistics showed that our framework is able to translate correctly the input coverage statistics to the output plant distribution.

Our work stands out from previous research, because (i) our plant distribution generation is fully data-driven and (ii) we demonstrated with our interactive visualization WebGL prototype the possibilities of rendering over the web very large natural environments with a high density and variety of plants.

In the future, we would like to investigate whether other representations of billboards improve the visualization at different viewing angles, especially in global view. In addition, we would like to investigate more local and global illumination models to improve performance and realism of lights and shadows in the visualization. To improve the usability of this method, it might be preferable to combine the vegetation and visualization model in one web application, so that the user can easily change the plant distribution in the 3D visualization without having to do offline computations. Furthermore, it might be interesting to extend the framework by including the fauna of the environment, for improved realism [34]. Finally, so far our framework has

been tested on environments with only grass-like plant species; we would like to do additional testing for other more forest-like scenes to assess the quality and performance of its results.

ACKNOWLEDGMENTS

The authors would like to thank Alex Kolpa for helping with the implementation of the L-system plugin in Sceelix.

REFERENCES

1. C. Pettit, C. Raymond, B. A. Bryan, and H. Lewis, "Identifying strengths and weaknesses of landscape visualisation for effective communication of future alternatives," *Landscape and Urban Planning*, vol. 100, no. 3, pp. 231–241, 2011.
2. R. van Lammeren, J. Houtkamp, S. Colijn, M. Hilferink, and A. Bouwman, "Affective appraisal of 3D land use visualization," *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 465–475, 2010.
3. H. J. de Vriend, M. van Koningsveld, S. G. J. Aarninkhof, M. B. de Vries, and M. J. Baptist, "Sustainable hydraulic engineering through building with nature," *Journal of Hydro-Environment Research*, vol. 9, no. 2, pp. 159–171, 2015.
4. B. Onrust, R. Bidarra, R. Rooseboom, and J. Van De Koppel, "Procedural generation and interactive web visualization of natural environments," in *Proceedings of the 20th International Conference on 3D Web Technology*, pp. 133–141, ACM, 2015.
5. R. M. Smelik, T. Tutenel, R. Bidarra, and B. Benes, "A survey on procedural modelling for virtual worlds," *Computer Graphics Forum*, vol. 33, no. 6, pp. 31–50, 2014.
6. S. Temmerman, T. J. Bouma, J. Van de Koppel, D. Van der Wal, M. B. De Vries, and P. M. J. Herman, "Vegetation causes channel erosion in a tidal landscape," *Geology*, vol. 35, no. 7, pp. 631–634, 2007.
7. J. de Leeuw, L. P. Apon, P. M. Herman, W. de Munck, and W. G. Beeftink, *The Response of Salt Marsh Vegetation to Tidal Reduction Caused by the Oosterschelde Storm-Surge Barrier*, Springer, 1994.
8. J. Molofsky and J. D. Bever, "A new kind of ecology?" *BioScience*, vol. 54, no. 5, pp. 440–446, 2004.
9. M. Rietkerk and J. van de Koppel, "Regular pattern formation in real ecosystems," *Trends in Ecology & Evolution*, vol. 23, no. 3, pp. 169–175, 2008.
10. C. Schwarz, *Implications of biogeomorphic feedbacks on tidal landscape development [Ph.D. thesis]*, Radboud University Nijmegen, 2014.
11. S. Saura and J. Martinez-Millan, "Landscape patterns simulation with a modified random clusters method," *Landscape Ecology*, vol. 15, no. 7, pp. 661–678, 2000.

12. W. W. Hargrove, F. M. Hoffman, and P. M. Schwartz, “A fractal landscape realizer for generating synthetic maps,” *Conservation Ecology*, vol. 6, no. 1, 2, 2002.
13. B. Lane and P. Prusinkiewicz, “Generating spatial distributions for multilevel models of plant communities,” in *Proceedings of the Graphic Interface*, pp. 69–80, 2002.
14. O. Deussen, P. Hanrahan, B. Lintermann, R. Mech, M. Pharr, and P. Prusinkiewicz, “Realistic modeling and rendering of plant ecosystems,” in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 275–286, ACM, 1998.
15. E. Chng, “An artificial life-based vegetation modelling approach for biodiversity research,” *Green Technologies: Concepts, Methodologies, Tools and Applications*, 417, 2010.
16. J. Hammes, “Modeling of ecosystems as a data source for real-time terrain rendering,” in *In Digital Earth Moving*, pp. 98–111, Springer, 2001.
17. M. Alsweis and O. Deussen, “Wang-tiles for the simulation and visualization of plant competition,” in *Advances in Computer Graphics*, vol. 4035 of *Lecture Notes in Computer Science*, pp. 1–11, Springer, Berlin, Heidelberg, 2006.
18. M. Weier, A. Hinkenjann, G. Demme, and P. Slusallek, “Generating and rendering large scale tiled plant populations,” *Journal of Virtual Reality and Broadcasting*, vol. 10, no. 1, 2013.
19. K. Boulanger, S. Pattanaik, and K. Bouatouch, “Rendering grass terrains in real-time with dynamic lighting,” in *Proceedings of ACM SIGGRAPH 2006: Sketches (SIGGRAPH '06)*, August 2006.
20. E. Bruneton and F. Neyret, “Real-time realistic rendering and lighting of forests,” *Computer Graphics Forum*, vol. 31, no. 2, pp. 373–382, 2012.
21. O. Deussen, C. Colditz, M. Stamminger, and G. Drettakis, “Interactive visualization of complex plant ecosystems,” *IEEE*, pp. 219–226, 2002.
22. Z. Fan, H. Li, K. Hillesland, and B. Sheng, “Simulation and rendering for millions of grass blades,” in *Proceedings of the 19th Symposium on Interactive 3D Graphics and Games*, pp. 55–60, ACM, 2015.
23. B. Fanini, L. Calori, D. Ferdani, and S. Pescarin, “Interactive 3D landscapes on line,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3816, pp. 453–459.

24. M. Englert, P. Herzig, S. Wagner, Y. Jung, and U. Bockholt, “X3D-earthbrowser: visualize our earth in your web browser,” in *Proceedings of the 18th ACM International Conference on 3D Web Technology*, pp. 139–142, ACM, 2013.
25. A. Lagae, *Tile-Based Methods in Computer Graphics [Ph.D. thesis]*, Katholieke Universiteit Leuven, 2007.
26. B. Onrust, *Automatic generation of plant distributions for existing and future natural environments using spatial data [M.S. thesis]*, Delft University of Technology, //graphics.tudelft.nl/benny-onrust/, 2015, <https://graphics.tudelft.nl/benny-onrust/>.
27. P. Silva, P. Müller, R. Bidarra, and A. Coelho, “Node-based shape grammar representation and editing,” in *Proceedings of the Workshop on Procedural Content Generation for Games (PCG ‘13), Co-Located with the Eighth International Conference on the Foundations of Digital Games*, 2013.
28. Sceelix, “The 3D scenes procedural engine,” <http://www.sceelix.com>. Accessed: 1 April 2017.
29. P. B. Silva, E. Eisemann, R. Bidarra, and A. Coelho, “Procedural content graphs for urban modeling,” *International Journal of Computer Games Technology*, vol. 2015, Article ID 808904, 15 pages, 2015.
30. R. Fernando, “Percentage-closer soft shadows,” in *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ‘05)*, p. 35, ACM, 2005.
31. R. Cabello, three.js-javascript 3D library, 2010.
32. P. Decaudin and F. Neyret, “Volumetric billboards,” *Computer Graphics Forum*, vol. 28, no. 8, pp. 2079–2089, 2009.
33. W. Donnelly and A. Lauritzen, “Variance shadow maps,” in *Proceedings of the Symposium on Interactive 3D Graphics and Games*, pp. 161–165, 2006.
34. N. Komodakis, C. Panagiotakis, and G. Tziritas, “3D visual reconstruction of large scale natural sites and their fauna,” *Signal Processing: Image Communication*, vol. 20, no. 9-10, pp. 869–890, 2005.

INDEX

Symbols

3D visualizations 350
(TTS) systems 193

A

Account 29
Accuracy 128
Acoustic database 280
Acoustic model 135
Acoustic-phonetic correlation 277
Acoustic realisation 194, 197
Acoustic units 200
Adaptive instance normalization
 (AdaIN) 53
Administrations 301
Administrative boundaries 177
Aggregation 281
Aggressive (AGG) 227
Aggressive style (AGG) 228

Algorithm 25, 161, 233
Algorithm models 32
Amplitude 163, 233
Analysis 212
Analytical 327
Analyze data 331
Anecdotal 307
Annotated Corpora 275
Annotation files 35
Anxiety 341
Aperiodicity 264
Applicability 195
Architecture 23, 132, 203, 291
Articulatory synthesis 199
Artificial intelligence (AI) 288
Attention 24
attention model 37
Attenuation network 90
Attribute detectors (AttrDet) 30
Attributes 37

- Audience 154
 Authors 258
 Autoencoders 52
 Automatic 52
 Automatic detection 212
 Automatic speech recognition (ASR) 226
 Automation 311, 312, 315, 316, 320, 321
 Automation process 321
 Autonomous 78
 Autonomous cars 316
 Autonomous vehicles 316
- B**
- Bilateral filter 161
 Boxplots 241
- C**
- Case-based reasoning (CBR) 231
 Centaurs 287
 Characterization 61
 Chrominance information 160
 City Granularity 181
 Claiming 338
 Clustering 201
 Coefficient 69
 Cognition 344
 Collaboration 316
 Commercialization 160
 Communicate 328
 Communication 293
 Complexity 61
 Computerization 315
 Conditional random field (CRF) 20
 Consume 175
 Content delivery networks (CDNs) 172
 Contextual factors 203
- Contextualization 303
 Convolutional layers 9
 Convolutional neural network (CNN) 21, 53
 Convolutional pseudo-prior 6
 Corpora 281
 Corporations 334
 Correlation 37
 Creativity 155, 294, 297
 Cyber illusions 342
 Cybernetic transformation 344
 Cyberspace 340
 Cyborgs 340
- D**
- Data analysis 328
 DCGAN model 52
 Decision-making 101
 Decoding 27
 Deconvolution layer 61, 72
 Decrease 66
 Deep neural network (DNN) 202
 Deidentification 79
 Deidentification method 81
 Delusional 339
 Department of Information Technology (DIT) 279
 Depth-map fusion 161
 Detecting errors 329
 Detection 4
 Diacritic restoration 197
 Dictionary 135
 Digital 100
 Digitalization 321
 Dilated Depthwise Separable Residual (DDSR) 130
 Dimension 321
 Diphone 200
 Discriminator 66

Distribution 357

Diversification 153

Diversity 52, 71

Dosovitskiy 5

Downsampling 10

Dropout layer 72

Dynamic models 353

Dynamism 322

E

Ecological management 352

Ecological processes 351

Economy 100, 320

Ecosystem 330

Efficiency 78

Elasticity 319

Electromagnetic fields 111

Electronically 281

Emphatic emotions 212

Employment 314

Encoders 27

Encompass 295

Engineers 316

Entropy 70, 102

Epidemic detection 304

Equality 344

Equilibrium 65

Error 69

Evaluation methods 34

Excel spreadsheets 327

Exception 199

Expressive speech synthesis (ESS)

226

Extensible 58

Extraction 52

F

Facial contours 134

Feedforward 27

Feedforward image 4

Flexibility 174

Fluctuation range 64

Forecasting 327

Frequency (F0) 229

G

Game theory 81

GAN model 59, 60

GANs 6

Generation , 23

Generative adversarial networks
(GANs) 51

Generator 68

Geographic location 174

Geometry 369

Geovisualizations 356

Government 328

Gradient 26

Grammatical 275

Granularity 37, 180

H

Happy style (HAP) 228

Harmonic 244

Harmonic plus noise model (HNM)
230

Height extraction 104

Hidden Markov model (HMM) 131

Hidden Markov Models (HMMs)
201

Hierarchical 134

Historical data 179

Homogeneity 153, 163

Homogenization 153

Horizontal 163

Human grandmasters 289, 294

Hyperparameters 137

Hypothesis 294, 318

Hypothetically 300

I

Identification 82

Illusion 341

Image restoration 6

Imagination 306

Immature 152

Immersion 154

Implementation 12, 135

Inception score (IS) 69

Inconsistent information 178

Incremental 298

Incremental innovations 295, 303

Indicators 83, 331

Industrialization 160

Infringement links 153

Initial fraction (ISc) 69

Innovation 155

Innovation processes 287

Insertions 257

Instantaneous 196

Integration 332

Integrity 139

Intelligence Augmentation (IA) 289

Intelligent transportation 101

Intense debate 312

Intensity images 165

Interactive Virtual Assistant (IVA)
193

Interactive visualization 352

Interceptable 148

Interchangeably 196

Intermediate system 203

Intonation 257

Invariability 78

L

Larynx 196

Laws of Robotics 343

Least recently used (LRU) 173

Leipzig Corpora Collection (LCC)
278

Linear predictive coding (LPC) 231

Listeners 211

location 26

Lu Huan's name 339

Luminance 160

M

Machine translation 37

Machinists 315

Magnetic resonance 199

Magnitude 13, 163

Mammals 351

Manifestation 344

Manipulating frequencies 231

Manipulation 243

Manpower 4

Maturity 78

Mean Opinion Score (MOS) 211

Mechanization 315

Media integration 149

Mel-cepstral coefficients 264

Mel spectrograms 196

Mesoscopic scale 345

Metaphor 339

Migration 15

Minimum description length (MDL)
264

Mobilize 290

Modeling research 52

Modification 275

Momentum 312

Monotonous 129

Multiheaded attention 27

Multi-instance learning (MIL) 22

Multimedia 100

Municipal innovations 300

N

Nasal cavities 199

Nasal cavity 196

Naturalness 263

Network congestion 183

Neural layers 202

Neuro-electrical conduction 340

Neurology 24

Neutral (NEU) 227

Neutral style (NEU) 228

Non-automated sector 320

Nonecologists 350

Non-homophone homographs 197

Nonlinear 9

Nonstatic 361

Normalising flows (NFs) 204

Normalization 9, 53

Nouns 20

O

Opportunity 148

Optimization 13

Oral cavity 196

Organizational 301

P

Palate 196

Parameterisation 231, 241

Parameters 226

Part-of-speech tagging (POS) 197

Patchiness data 361

Perception 344

Perceptual 10

Perceptuality 4

Pharynx 196

Philosophers 338

Phonetic transcription 198

Pitch marking 229

Pitch variation 196

Pixels 9

Plagiarism 153

Political actors 299

Polygon 355

Popularization 21, 171

Postures 129

Precognition system 334

Precomputation phase 359

Prediction 180

Prepositions 20

Probability 26

Procedure 368

Professional equipments 128

Pronunciation 194, 260

Pronunciation syntax 198

Propagation 176

Prosodic events 201

Provider generated content (PGC)

173

Province 178

Province Granularity 180

Q

Qualitative 11, 13

Quantization 263

R

Radical change 151

Random noise vector 6

Rapid development 52

Rarity 36

rating 37

Realistic 127

Reality 337

Recognition 53

Recognition technology 79

- Recreation 351
- Recreational 350
- Reducible 344
- Redundancy 82
- Reflection 346
- Registration 104
- Regression models 52
- Regularization 23
- Reinforcement learning 32
- Relationship 31
- Rendering phase 359
- Resolution 7, 174
- Rhythm 254
- Rigid 129
- Root mean square (RMS) 265
- Random field 8

- S**
- Sad style (SAD) 228
- Satisfaction theory 153
- Satisfying 62
- Scalability 174
- Scenario 15, 151, 205
- Scenes 20
- Schopenhauer 338
- Segmentation 229
- Self-consciousness 344
- Semantic 10
- semantic attention 38
- Semantic attributes 275
- Semantic information 127
- Sensitive 107
- Sensual (SEN) 227
- Sensual style (SEN) 228
- Serendipity 298
- Shape modeling 104
- Skeleton 132
- Social dissatisfaction 78
- Socialization 148

- T**
- Software developers 316
- Spatial 30
- Spatial data 100
- Spatiotemporal graph 130
- Spatiotemporal processes 302
- Spectrogram 196, 203, 205
- Speculations 305
- Standard deviation 163
- Standardization 334
- Statistical data 362
- Statistical model 327
- Statistical tools 327
- Statistics 152
- Stochastic components 233
- Straightforward mapping 194
- Stress groups (SG) 232
- Subcorpus 228
- Sub-depth-map 166
- Supercomputers 289
- Syllabification 198, 200
- Symbiote 293
- Symbiotic innovations 298
- Symbiotic learning 291
- Synchronization, 127
- Synthesis 201
- Synthesiser 227
- Synthesize 160
- Synthesizer 280
- Synthesizing 269

- T**
- Taxation 328, 329, 330, 332
- Tax authorities 333
- Tax data 332
- Tax technology 331
- Technicians 316
- Technology 52, 151
- Telemetry 100
- Terminology 193

Text-to-speech method (TTS) 132
 Text-to-speech (TTS) 138
 Text-to-speech (TTS) system 253
 TF-IDF 37
 Three-dimensional 101
 Three-dimensional model 105
 Transform 127
 Transformation 135
 Transparency 332, 334
 Tremendous 298
 Triphone 263
 TSR images 53
 Turbulence 79
 Two-dimensional 128
 Typology 298

U

Uniqueness 78
 Universal 78, 333
 Upsampling 10
 Upsampling layer 58, 72
 User generated content (UGC) 171,
 172

V

Validation data 179
 Variational AutoEncoders (VAEs)
 204
 Vectors 183
 Vegetation 351
 Vegetation model 352
 Velocity 304
 Verbs 20
 Vertical 163
 Video on demand (VOD) 172
 Violation 153
 Virtual faces 81
 Visualization 350
 Visualization model 359
 Vocabulary scenario 200
 Vocal cords 196
 Vocoder 205
 VoQ parameters 232

W

Waveform 205
 Websites 136

Generative AI Models

The world is changing rapidly, and scientific and technological progress plays a key role in this. Artificial intelligence (AI), which has permeated all spheres of social, economic, scientific research and everyday life, has a special influence. Today, the spotlight is on generative artificial intelligence, which has the potential to change the world in the coming years, in terms of development, commercial and social perspectives.

Within a few months, ChatGPT has become the most prominent representative of the new generation of generative artificial intelligence systems. Others are called LaMDA, DALL-E or Stable Diffusion. These programs produce fundamentally new texts, codes, images or even videos. The results are so convincing that it is often impossible to tell whether they were created by human or machine.

The generative AI is especially good and applicable in 3 major areas:

- Text generation - applications in law (drafting contracts), medicine (diagnostics), journalism (news production), education (production of educational materials), science (search and generation of scientific papers), etc.
- Image and video generation - application in marketing (advertisement creation), media (virtual host), art, architecture, design, social networks, etc.
- Voice and sound generation - application in the film industry (special effects), music industry, customer support (virtual references), surveying (automation of telephone surveys), etc.

Generative artificial intelligence is expected to revolutionize the way people work or find information online. But it also raises numerous ethical questions, as rarely has any technology done so far. There are fears that millions of people could lose their jobs, or that the system could be misused for disinformation, or even that the world, as we know it, will end. This sparked a debate about the necessary rules and regulation of AI.

This book edition covers different topics of generative AI models, including: image generation techniques, video generation techniques, speech / voice generation techniques, and societal and ethical issues of these models.

Section 1 focuses on image generation techniques, describing image generation and style transfer algorithm based on deep learning; an overview of image caption generation methods; an application of an improved DCGAN for image generation; a private face image generation method based on deidentification in low light; and a remote sensing image data scene generation method in smart city.

Section 2 focuses on video generation techniques, describing realistic speech-driven talking video generation with personalized pose; video transformation in big video era and its impact on content editing; a fast depth-map generation algorithm based on motion search from 2D video contents; and adaptive content management for UGC video delivery in mobile internet era.

Section 3 focuses on voice and speech generation, describing generating the voice of the interactive virtual assistant; voice quality modelling for expressive speech synthesis; prosodically rich speech synthesis interface using limited data of celebrity voice; and an overview of resources for development of Hindi speech synthesis system.

Section 4 focuses on societal and ethical issues of generative AI models, describing AI-human symbionts reinventing innovation and what the new centaurs will mean for cities; the impact of AI and automation on new jobs; a development of artificial intelligence in taxation; AI films as reflections on reality and illusion; and an ecologically sound procedural generation of natural environments.



Jovan Pehcevski obtained his PhD in Computer Science from RMIT University in Melbourne, Australia in 2007. His research interests include modern data center technologies (XaaS), big data, machine learning and artificial intelligence, and information retrieval. He has published over 30 journal and conference papers and he also serves as a journal and conference reviewer. Jovan has extensive academic and research experience, coupled with practical expertise in the IT industry. He is currently working as a Senior Technology Consultant at Dell Technologies, covering South Eastern Europe.