IMVIP 2016

Irish Machine Vision
& Image Processing
Conference

August 25/26 2016
National University of
Ireland, Galway

IRISH MACHINE VISION & IMAGE PROCESSING

Conference proceedings 2016

Irish Pattern
Recognition
& Classification
Society

Editor:
Nicholas Devaney
School of Physics
NUI, Galway

Published by the Irish Pattern Recognition & Classification Society

iprcs.org

ISBN 978-0-9934207-1-9

Introduction

The twin fields of Machine Vision and Image Processing continue to develop and find wider and more numerous applications. With the explosion in the number of consumer imaging devices, the world is awash with images and video -- it is estimated that more than 3 billion images are shared online every day. Meanwhile the application of imaging to scientific, medical and engineering problems also continues to grow – from robotic probes on other planets and even comets, to swarms of drones here on Earth. Irish companies and researchers are playing a significant role in developing both the algorithms and the hardware necessary to extract useful information from this data, and the annual Irish Machine Vision and Image Processing conference is the ideal forum to bring together this community.

The 18th IMVIP conference was held in the National University of Ireland, Galway, on 25th and 26th August 2016. It was organised by the School of Physics. A total of 17 papers were selected for either poster or oral presentations; and these papers are brought together in this volume. Thanks are due to the programme committee which reviewed the papers. In addition to these papers, the conference has featured top-class Keynote speakers; Prof. Jenny Read (Newcastle University), Prof. Rudolf Mester (Linköping University, Sweden), Dr. Chris Solomon (University of Kent), Dr. Alexandru Drimbarean (vice-president of Fotonation), Dr. Michael Starr (Valeo), Dr. George Siogkas(Valeo) and Dr. Alireza Dehghani (Movidius).

IMVIP is run in association with the Irish Pattern Recognition & Classification Society (iprcs.org), a member organisation of the International Association for Pattern Recognition (IAPR) and the International Federation of Classification Society (IFCS). I would particularly like to thank the president of IPRCS, Rozen Dahyot, for all her advice and help over the past year.

I am very grateful to our sponsors: Movidius, Fotonation, and the NUIG College of Science. Special thanks also to the graduate students of the Applied Optics group (NUI, Galway) for help with running the conference, and in particular to Colm Lynch who developed the conference web page (http://optics.nuigalway.ie/IMVIP2016/).

Nicholas Devaney
School of Physics
NUI, Galway
Ireland
August 2016

 ISBN 978-0-9934207-1-9

Keynote speakers: Jenny Read

Sponsored by FotoNation

Title: *How many ways are there to solve stereoscopic vision?*

Abstract: Matching up the two perspectives from which we see the world requires the brain to compute when particular locations on the retinas are viewing the same object in space. In principle, there are several ways this could be done. For example, corresponding points should usually have roughly the same contrast, luminance, colour, texture and motion. A stereo vision system could proceed by detecting distinctive features or objects in each eye individually and then finding pairs of features that match. Alternatively, it could simply assess how well regions of the retinas match, without identifying particular features or objects; the output of this computation could then guide scene segmentation and object identification. Both approaches have been employed in machine stereo algorithms. Human stereopsis seems to consist of several distinct modules which use different approaches, and contrast this with a very different form of stereopsis found in an insect, the praying mantis.

About the speaker:

Jenny Read is Professor of Vision Science at the Institute of Neuroscience in Newcastle University. Her particular interest is 3D and stereo depth perception and her research includes detailed psychophysical measurements of depth perception, computational models of the underlying neuronal mechanisms, stereopsis in other species, clinical disorders of vision, and commercial applications of 3D display technologies (http://www.jennyreadresearch.com). Her major projects at the moment include "Man, Mantis and Machine", looking at insect 3D vision, and ASTEROID, developing a new vision test for children.

Keynote speakers: Rudolf Mester

Sponsored by Movidius

Title: *Predictive Visual Perception for Automotive Applications*

Abstract: Understanding the world around us while we are moving means to continuously maintain a dynamically changing representation of the environment, to make predictions about what to see next, and to correctly process those perceptions which were surprising, relative to our predictions. This principle is valid both for animate beings, as well as for technical systems that successfully participate in traffic. At the VSI Lab, we put special emphasis on this recursive / predictive approach to visual perception in ongoing projects for driver assistance and autonomous driving. These processing structures are complemented by statistical modeling of egomotion, environment, and the measurement process. In our opinion, this approach leads to particularly efficient systems, since computational ressources may be focussed on 'surprising' (thus rare) observations, and since this allows for a large reduction of search spaces in typical visual matching and tracking tasks. The talk will present examples for such predictive / recursive processing structures. Furthermore, recent results in the field of monocular, stereo, and multi-monocular (surround vision) applications will be shown.

About the speaker:

Rudolf Mester is guest Professor at the Computer Vision Laboratory at the ElectricalEngineering Dept. (ISY) of Linköping University, Sweden where he leads an initiative for Intelligent Visual Vehicle Sensorics and Visual Surround Sensing. The aim is to generalize state of the art computer vision procedures for typicalvehicle-related tasks such as driver assistance towards multi-view /omnidirectional processing. His research are focused on statistical signal and image processing methods, the construction of robust and reliable vision algorithms and flexible vision systems as well as the theoretical foundations for "seeing machines". (http://www.vsi.cs.uni-frankfurt.de/people/prof-mester/).

 ISBN 978-0-9934207-1-9

Keynote speakers: Chris Solomon

Title: *Making Faces - From concept to commercial product*

Abstract: This paper will describe the scientific basis of work undertaken over a number of years at the University of Kent to find a more effective way to produce facial composites. Facial composites are images produced from an eyewitness' memory which are used by police forces around the world to help identify criminal suspects. We will outline the core concepts and explain how they have led to the E-FIT system, now in routine use in more than 20 countries around the world.

About the speaker:

Chris Solomon is Reader at the University of Kent. He also directs VisionMetric Ltd, a spin-out company which is the UK's leading developer and supplier of facial composite software to the police. His main research activities focus on image processing and evolutionary methods with particular interest in the human face.

 ISBN 978-0-9934207-1-9

Keynote speakers: Alexandru Drimbarean

Title: *Mobile Computational Imaging*

Abstract: From its beginning with the transition from film to digital, mobile computational imaging technologies have evolved, transforming forever the way we experience photography. Even though the latest smartphones can capture images with amazing quality and implement a wide range of features such as panorama, HDR or VIS not possible in the film era, this “magic” is not without challenges. This presentation will outline the key challenges of mobile computational imaging and the evolution of different solutions leading to FotoNation’s IPU (Image Processing Unit) a hardware unit comprising a set of IP cores tightly connected to provide high performance and low power computational imaging.

About the speaker:

Alexandru Drimbarean is the Vice President of Advanced Research team at FotoNation Ireland focusing on developing computer vision and machine learning technologies for mobile, biometrics and automotive applications. He and his group aim at identifying, researching and providing novel innovative techniques and features to enhance, extend and simplify the adoption of imaging technologies. Alexandru received his B.S in Electronic Engineering in Brasov Romania followed by an M.Sc. in Electronic Science at N.U.I Galway in 2002. His interests include image acquisition, processing and understanding as well computer vision and machine learning. Alexandru has authored several journal articles as well as more than 30 patents.

 ISBN 978-0-9934207-1-9

Keynote speakers: Michael Starr, George Siogkas

Title: *Automotive Computer Vision - from ADAS to Autonomous driving*

Abstract:

The automotive world has always been a lucrative application space for cutting edge computer vision research. Nowadays, the hype around driverless vehicles is continuing its upward trend and OEMs and suppliers alike are competing in a race to be the first to deliver a fully autonomous vehicle. The aim of this presentation, is to connect the dots starting at low level supportive computer vision algorithms for ADAS which appeared 10 years ago to the point where fully autonomous vehicles are expected to hit the market in the next 5 years. We will present past, current and future computer vision algorithms that have, or soon will be available on commercial vehicles. Doing so, we will point out challenges, some more obvious than others, that dramatically increase the complexity of the solutions and dictate a cautious approach to developing safety critical algorithms. We will also try to predict what the future has in store for the research community, the OEMs, and the consumers.

 ISBN 978-0-9934207-1-9

About the speakers:

Michael Starr is a computer vision researcher in the Automated Parking Product Group in Valeo Vision Systems, Tuam, Co, Galway. He has worked extensively on automotive computer vision for the delivery of Advanced Driver Assistance Systems (ADAS) including Object Detection, Parking Slot Detection, and Trailer Parking Assist algorithms. Michael also has a strong interest in the application of computer vision algorithms to FPGAs. Prior to his work in Valeo, Michael worked in the design of Electron Multiplying CCD (EMCCD), Short Wave Infrared (SWIR), scientific CMOS and CCD cameras focusing on low noise and high sensitivity for medical and security applications. Michael has patents in the areas of lens soiling detection and ADAS trailer applications.

George Siogkas is a computer vision researcher in the Automated Parking Product Group in Valeo Vision Systems, Tuam, Co, Galway. George has received his PhD in Electrical and Computer Engineering from the University of Patras, Greece in 2013. Before joining Valeo, he worked as a lecturer in a private College in Greece, where he also served as the Head of the Engineering and Informatics Department. His main research interests lie in the areas of computer vision, signal and image processing and machine learning, with a special focus on automotive and biomedical applications. Since he started working for Valeo, George has been contributing to the development of algorithms like park slot marking detection and lane detection. Currently, he leads the Deep Learning related research efforts of the Automated Parking Group.

 ISBN 978-0-9934207-1-9

Programme Committee

- Abdullah Bulbul, Trinity College Dublin
- Ahmed Bouridane, Northumbria University, Newcastle, UK
- Andy Shearer, National University of Ireland Galway
- Antonio Fernández, University of Vigo, Spain
- Bob Fisher, University of Edinburgh
- Bryan Gardiner, Ulster University
- Bryan W. Scotney, Ulster University
- Cem Direkoglu, Middle East Technical University, Cyprus
- Danny Crookes, Queen's University of Belfast
- David Vernon, University of Skövde, Sweden
- Derek Molloy, Dublin City University
- Dermot Kerr, Ulster University
- Donald Bailey, Massey University, New Zealand
- Fionn Murtagh, University of London, UK
- Francesco Bianconi, University of Perugia, Italy
- George Moore, Ulster University
- Hiroshi Sako, Hosei University, Japan
- Jane Courtney, Dublin Institute of Technology
- Joan Condell, Ulster University
- John Barron, The Univerity of Western Ontario, Canada
- John Mc Donald, National University of Ireland Maynooth
- John Winder, Ulster University
- Jonathan Ruttle, SureWash, Dublin
- Kathleen Curran, Univerity College Dublin
- Kevin McGuinness, Dublin City University
- Madonna Herron, Ulster University
- Nicholas Devaney, National University of Ireland, Galway
- Noel O'Connor, Dublin City University
- Paul Mc Kevitt, Ulster University
- Paul Miller, Queen's University of Belfast
- Paul Whelan, Dublin City University
- Philip Morrow, Ulster University
- Reyer Zwiggelaar, Aberystwyth University, UK
- Robert Sadleir, Dublin City University
- Sally McClean, Ulster University
- Sonya Coleman, Ulster University
- Sudeep Sarkar, University of South Florida, USA
- Tom Naughton, National University of Ireland Maynooth

 ISBN 978-0-9934207-1-9

Table of Contents

 ISBN 978-0-9934207-1-9

Background estimation in adaptive optics photoreceptor images

L. Mariotti and N. Devaney

Applied Optics group, School of Physics,
National University of Ireland, Galway

Abstract

With the use of adaptive optics cameras, it is possible to study individual cone photoreceptors in the retina *in vivo*. We are studying the spatial and temporal variations of the reflectance of the individual cones and its relationship with the background intensity.

Keywords: image processing, biomedical imaging, adaptive optics

1 Introduction

Adaptive Optics (AO), a technology that was first developed in astronomy to improve the resolution of images acquired through atmospheric turbulence, has also been implemented in the last decade in the field of retinal imaging. By compensating the eye aberrations, the AO cameras can greatly increase the amount of information available about the retina, allowing to resolve the individual cone photoreceptors [Lombardo et al., 2013].

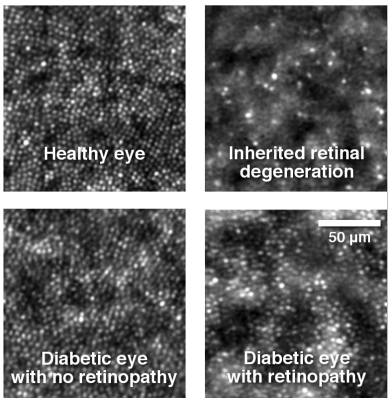


Figure 1: AO images of cone mosaics with different conditions

The investigation of the cone mosaic is providing new insights into the photoreceptor physiology as well as into the early stages of retinal conditions. In addition to their number and position, one of the most noticeable features of the photoreceptors is the spatial and temporal variability of their reflectance, which has been clearly detected with all AO imaging modalities. In some clinical cases the cone reflectance is the only apparent feature of the cones that distinguishes a healthy cone mosaic from a mosaic with altered functionality, but unaltered spatial organization (Figure 1). Therefore, further investigation of the spatial and temporal characteristics of cone reflectance in both healthy and diseased subjects is necessary.

2 Method

We developed and successfully used a processing method that performs a completely automated cone mosaic and cone reflectance analysis. The area to process is selected with a semi-automated algorithm that excludes the blood vessel shadows, requiring manual intervention only in the choice of the numerical parameters of the vessel segmentation [Mariotti et al., 2016]. We are currently working on a procedure that will allow for the estimation of the background contribution to the total cone reflectance.

The cone profiles are separated from the background through a detection and segmentation algorithm [Chiu et al., 2013], which extends the detection by segmenting the cone profiles around the local maxima. A portion of the image surrounding each maximum is transformed into quasi-polar coordinates, and then the contour of the cone is segmented as a layer using graph theory and dynamic programming.

 ISBN 978-0-9934207-1-9

The segmentation is subsequently used to automatically determine the size of a disc structuring element. The structuring element is used to perform a rank leveling operation, which eliminates the bright features (i.e. the cones) replacing them with the surrounding pixel intensity [Russ, 2016]. The image so obtained is then smoothed through convolution with a Gaussian filter with standard deviation equal to the structuring element radius. The background image reveals the low-frequency spatial variations in intensity and is subtracted from the original image (Figure 2).

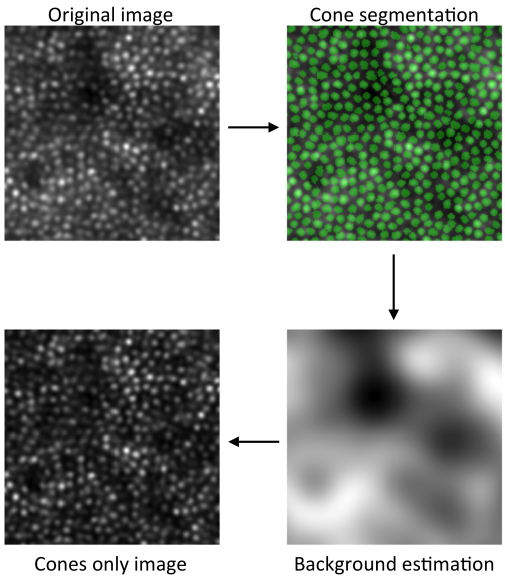


Figure 2: Detail of the cone segmentation and background estimation

3 Conclusions and future work

The separation of background and cone intensity will allow to characterise the spatial and temporal changes in reflectance at different depths in the retina. The reflectance will be studied on a number of subjects, both healthy and diseased, in order to determine if there is a relationship between the spatial pattern and the fraction of light reflected by the background and the presence of retinal conditions (Figure 3). Another objective is to determine if detectable modifications in the reflectance occur in conjunction with the mosaic spatial alterations or possibly before them.

We are making important progress towards the automation of the cone mosaic analysis in AO retinal images and towards the understanding of cone reflectance in the early diagnosis of retinal conditions.

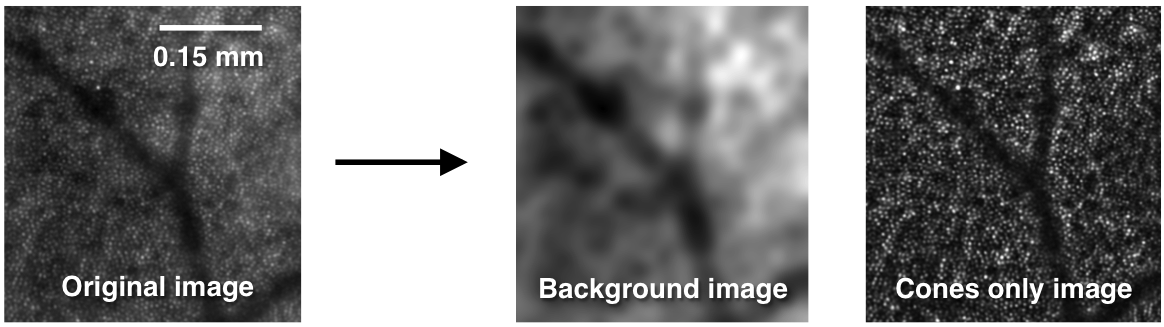


Figure 3: Detail of the cone segmentation and background estimation

References

[Chiu et al., 2013] Chiu, S. J., Lokhnygina, Y., Dubis, A. M., Dubra, A., Carroll, J., Izatt, J. A., and Farsiu, S. (2013). Automatic cone photoreceptor segmentation using graph theory and dynamic programming. *Biomedical Optics Express*, 4(6):924–937.

[Lombardo et al., 2013] Lombardo, M., Serrao, S., Devaney, N., Parravano, M., and Lombardo, G. (2013). Adaptive optics technology for high-resolution retinal imaging. *Sensors (Switzerland)*, 13(1):334–366.

[Mariotti et al., 2016] Mariotti, L., Devaney, N., Lombardo, G., and Lombardo, M. (2016). Understanding the changes of cone reflectance in adaptive optics flood illumination retinal images over three years. *Biomedical Optics Express*, 7(7):2807–2822.

[Russ, 2016] Russ, J. (2016). *The image processing handbook*. CRC Press, 7th edition.

Deep Learning for Biomedical Texture Image Analysis

Vincent Andrearczyk, Paul F. Whelan

Vision Systems Group, School of Electronic Engineering,
Dublin City University, Dublin, Ireland
vincent.andrearczyk3@mail.dcu.ie, paul.whelan@dcu.ie

Abstract

This paper shows promising results in the application of Convolutional Neural Networks (CNN) to biomedical imaging. Texture is often dominant in biomedical imaging and its analysis is essential to automatically obtain meaningful information. Therefore, we introduce a method using a Texture CNN for the classification of biomedical images. We test our approach on three datasets of liver tissues images and significantly improve the state of the art.

Keywords: Texture classification, biomedical imaging, Convolutional Neural Network

1 Introduction

In this paper we show that deep learning, which has established new states of the art in many domains including image classification and segmentation, can be very beneficial to biomedical imaging. Deep learning has recently shown impressive results in texture analysis as a feature extraction tool [1] or in an end-to-end training scheme [2, 3]. The analysis of texture is crucial in medical imaging for applications such as the detection, segmentation and classification of tissues, proteins and lesions. Therefore, we develop a method specifically designed for texture images based on Convolutional Neural Network (CNN) to classify liver tissues. Our neural network approach is of low complexity (memory and computation), which is particularly important in biomedical imaging to allow laboratories or research groups to train the network on a generally limited number of training images (avoiding overfitting) for possibly real-time applications, without requiring extremely powerful Graphical Processing Units (GPUs).

2 Work in progress

The architecture of the Texture-CNN with three convolution layers (T-CNN-3) is described in [2][1]. It uses an energy layer which densely pools (average pooling) the output features from an intermediate convolution layer. This approach discards the overall shape analysis needed for an object recognition task and of negligible importance in texture analysis. The complexity of this network is greatly reduced as compared to classic networks such as AlexNet (nearly three times fewer parameters) while obtaining better results on texture datasets.

Biomedical images are generally large (more than 1000×1000 pixels) and the number of training samples is often limited due to data privacy as opposed to the large training sets for object detection. Since we analyse texture images with repeated patterns, we can split the input images and use a voting score for classification. Thus, we do not need downscaling, which causes a loss of information, and we can increase the size of our training set, which is important for training neural networks.

Our approach is summarized in Figure 1. In this paper we use the IICBU database [4] with images of size 1388×1040. We split these images into 24 sub-images as shown in Figure 1(6 on the horizontal axis and 4 on the vertical axis) and resize them to 227×227. In the training phase, we use all the sub-images as independent samples to finetune the network, while in the testing phase we use a sum voting score among the 24 sub-images to classify the original full-size images. To this end, we sum the classification probabilities given as the softmax outputs of the network of all the 24 sub-images and assign the class with the highest sum to the full-size image.

[1] An implementation of the T-CNN-3 is available here: https://github.com/v-andrearczyk/caffe-TCNN

 ISBN 978-0-9934207-1-9

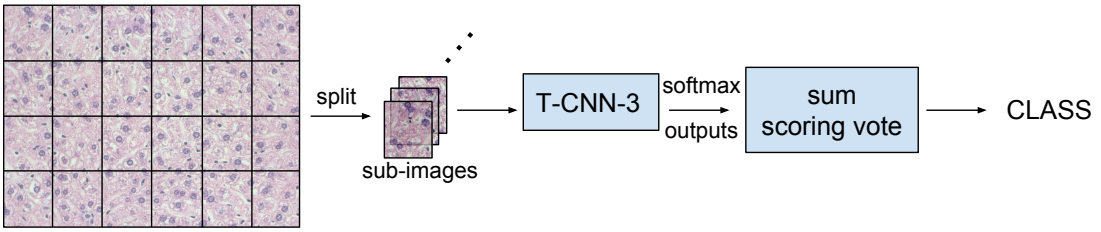


Figure 1: Proposed method including images split, neural network classification and collective classification by sum scoring vote.

The algorithm behaves as an ensemble method which takes a collective decision by summing the classification probabilities of different parts (sub-images) of the image.

3 Experiments

We test the proposed method on three datasets derived from the IICBU database [4]. An example of a tissue image is shown in Figure 1.

In the first experiment, we reproduce the Across-Subject Liver Aging (*AS-LA*) setup from [5] and report the accuracy averaged over 30 runs. The *AS-LA* dataset contains 1027 images obtained from 21 mice grouped into 4 classes. For each mouse, all the images are randomly divided into training (5/6) and test images (1/6).

The second and third experiments, respectively Liver Gender 6 Months on Ad Libitium (*LG6MAL*) and *Lymphoma*, are reproduced from [6]. We report the Mean Average Precision (MAP) measure averaged over 5000 runs as suggested in [6]. The *LG6MAL* experiment contains 265 images grouped into 2 classes (male/female). We report the results on the male class to compare to the state of the art. The *Lymphoma* dataset contains 374 images from 3 types of malignant lymphoma. For both *LG6MAL* and *Lymphoma*, 5% of the data is used for training and the rest for testing.

The results are reported and compared to the state of the art in Table 1. Our method significantly outperforms the state of the art on the three datasets. Even better results can be obtained with deeper networks and/or data augmentation. These results show that our approach could be very beneficial to the field of biomedical imaging. One could extend this approach to the detection and segmentation of tissues, tumors and lesions. However, the number of publicly available biomedical imaging datasets is small which is a significant barrier to the design of better suited and adapted computer vision methods including our approach.

Methods	AS-LA (accuracy)	LG6MAL (MAP)	Lymphoma (MAP)
Our method	**99.1%**	**98.2%**	**65.1%**
SoA	97.01% [5]	97.3% [6]	63.3% [6]

Table 1: Comparison between our method and the state of the art (SoA) in the classification results of tissue images.

References

[1] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, 2016.

[2] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *arXiv preprint arXiv:1601.02919*, 2016.

[3] T.-Y. Lin and S. Maji, "Visualizing and understanding deep texture representations," *arXiv preprint arXiv:1511.05197*, 2015.

[4] L. Shamir, N. Orlov, D. M. Eckley, T. J. Macura, and I. G. Goldberg, "IICBU 2008: a proposed benchmark suite for biological image analysis," *Medical & biological engineering & computing*, vol. 46, no. 9, pp. 943–947, 2008.

[5] H.-L. Huang, M.-H. Hsu, H.-C. Lee, P. Charoenkwan, S.-J. Ho, and S.-Y. Ho, "Prediction of mouse senescence from HE-Stain liver images using an ensemble SVM classifier," in *Intelligent Information and Database Systems*, pp. 325–334, Springer, 2013.

[6] N. Hervé, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid, "Statistical color texture descriptors for histological images analysis," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 724–727, IEEE, 2011.

Super-resolution of Aliased Thermal Imagery

C. Lynch[1,2], N. Devaney[1]

1. Applied Optics Group, School of Physics,
National University of Ireland, Galway
2. Fotonation, Parkmore East Industrial Estate, Galway, Ireland

Abstract

Aliased imagery contains high spatial frequency content encoded in a lower frequency band. Through use of a multiple-image interpolation technique it is possible to extract some of these high spatial frequencies to yield a super-resolved result. We demonstrate this using aliased thermal imagery with potential application for consumer imaging.

Keywords: Super-resolution, image processing, thermal imaging, spatial frequencies

1 Introduction

The resolving power of an optical detector is determined by its pixel pitch. Spatial frequencies that reach the detector above the Nyquist frequency are aliased and high detail is encoded into low spatial frequencies [Bracewell, 2004]. These higher spatial frequencies can potentially be recovered [Brown, 1981]. A Sinc interpolant across several images with aliasing and sub-pixel motion can be used to partially recover some of these aliased frequencies and yield a super-resolved output. As current consumer thermal cameras do not satisfy the Nyquist criterion for sampling, this method has potential application in consumer thermal imaging. While many approaches to super-resolution exist in the literature, the proposed method has low computational cost, making it suited to application on consumer devices [Thapa et al., 2015].

As aliasing is dependent upon spatial frequencies under-sampled by detector elements, small sub-pixel shifts of the camera between images will result in a set of images with differing alias strengths. In our work, pixel-level shifts are first measured using phase-correlation. Sub-pixel shifts are then interpolated using the peak of the phase-correlation output.

A new Sinc-based interpolation method has been developed for super-resolution. Using inter-pixel shifts, $\Delta_{x,y}$, interpolation kernels of the form $K = \alpha \mathrm{Sinc}(x - \Delta_x, y - \Delta_y)$ are calculated, with normalization coefficient α. Unique kernels are formed for each image and combined to form an output with a 2X increase in spatial resolution. The requirement for simultaneous visible imaging has been eliminated [Lynch et al., 2015].

2 Experimentation and Results

A FLIR A5 thermal camera was used, with F number $F\# = 1.25$ and pixel size $\delta = 50\,\mu m$. For a mean wavelength $\lambda = 10\,\mu m$, the diffraction-limited spatial frequency of the optics is $U_C = 80\,lp\,mm^{-1}$, while the pixel Nyquist frequency is $U_N = 10\,lp\,mm^{-1}$ [Dereniak and Boreman, 1996]. As a result, this system is sub-Nyquist sampled and output images are aliased. Shifts of the camera are on the order of millimetres and are introduced by hand jitter during capture.

Upon application of interpolation kernels, a super-resolved output is obtained. Results are shown in figure 1. For performance analysis, input aliased images are also registered to sub-pixel level and upsampled using a Lanczos kernel. The Lanczos kernel is simply a windowed Sinc which satisfies the sampling theorem for lossless upsampling[Bracewell, 2004].

 ISBN 978-0-9934207-1-9

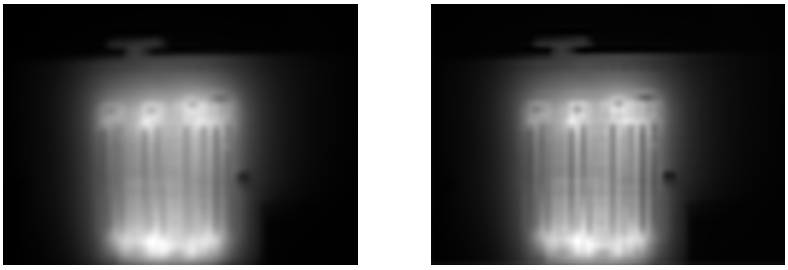


Figure 1: Results from the super-resolution method. Heated wires were used as a resolution target. (Left): images are registered to a sub-pixel level and interpolated using Lanczos interpolation. (Right): resultant image upon application of super-resolution.

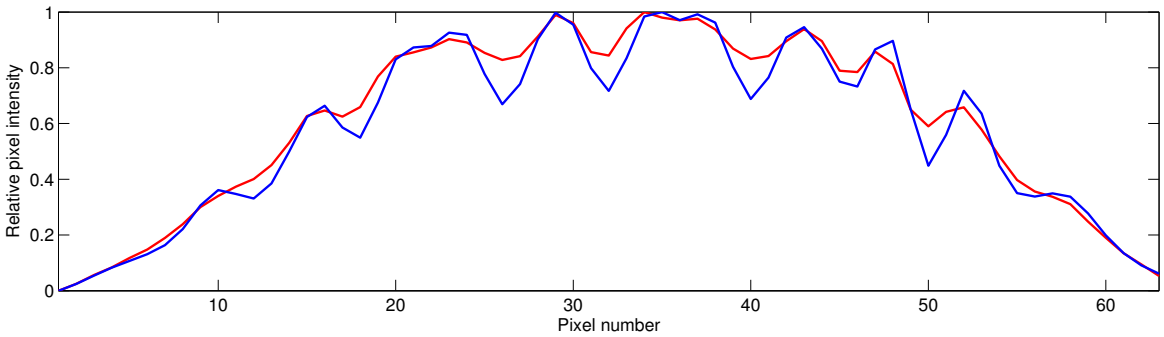


Figure 2: Horizontal profile showing pixel values for the case of Lanczos interpolation (red) and super-resolution (blue). Higher contrast is seen in the super-resolved image, with additional minima where expected.

3 Conclusions

With multiple aliased images, a resolution enhancement is possible using Sinc-based super-resolution kernels. When compared with the optimum case of Lanczos interpolation, additional spatial frequencies are present in the super-resolved output.

Acknowledgements

This work was supported by the Irish Research Council, Fotonation Ireland and the Applied Optics group.

References

[Bracewell, 2004] Bracewell, R. (2004). *Fourier analysis and imaging*. Springer Science & Business Media.

[Brown, 1981] Brown, J. L. (1981). Multi-channel sampling of low-pass signals. *IEEE Trans. Circuits Syst.*, 28(2):101–106.

[Dereniak and Boreman, 1996] Dereniak, E. L. and Boreman, G. D. (1996). *Infrared detectors and systems*. Wiley New York.

[Lynch et al., 2015] Lynch, C. N., Devaney, N., and Drimbarean, A. (2015). Computational methods for improving thermal imaging for consumer devices. In *Proc. SPIE 9485*. SPIE.

[Thapa et al., 2015] Thapa, D., Raahemifar, K., Bobier, W. R., and Lakshminarayanan, V. (2015). A performance comparison among different super-resolution techniques. *Computers & Electrical Engineering*.

About the Acquisition and Processing of Ray Deflection Histograms for Transparent Object Inspection

Johannes Meyer[1], Thomas Längle[2], Jürgen Beyerer[2,1]

[1] *Karlsruhe Institute of Technology KIT*
[2] *Fraunhofer-Institute of Optronics, System Technologies and Image Exploitation IOSB*

Abstract

Objects made from transparent materials are of great importance in our every-day life. In order to work as intended, these objects have to meet high quality criteria. Since their transparency makes many existing machine vision methods for opaque objects inapplicable, novel approaches have to be found. This paper proposes an optical inspection system based on a $4f$ light field camera in concert with parallel illumination, so that local histograms of the deflections of light rays exiting the test object can be acquired. Furthermore, the article presents two different approaches for processing these histograms of ray deflections (HORDs) in order to visualize scattering defects present in the test object. To evaluate the suitability of the proposed acquisition setup and the light field processing methods for the visualization of scattering material defects, different experiments are performed using a physically based rendering framework.

Keywords: machine vision, transparent object inspection, light field processing, image processing, histogram comparison

1 Introduction

Transparent objects and materials are of great importance. They are used in diverse kinds of industries to produce products that have to meet high quality requirements. For example, windshields of automobiles or aircrafts have to be clear, have to protect the passengers from the environmental influences, have to be mechanically durable and must not impair the driver's or pilot's sight. A further example are transparent plastic lenses used in eye-surgery, that have to precisely guide high-power laser beams as intended by the manufacturer in order to help and not harm the patient. Therefore, the respective transparent parts have to be free from material defects like absorbing or scattering particles. There exist elaborated machine vision methods for finding absorbing defects in transparent objects [Meyer, 2014]. However, the detection of scattering impurities in transparent materials with a more complex 3D-geometry is still an open research question. Since scattering defects, e.g., air bubbles, result in a deflection of the incident light rays, the directions of the light rays exiting the test object have to be captured in order to gain information about the defect. However, as both the transparent object's 3D-shape and scattering impurities can deflect light, a potential visual inspection system has to capture spatially resolved information on the direction of light rays exiting the test object. This paper shows, that an appropriately recorded light field contains the necessary information for visualizing scattering defects in transparent objects with a complex 3D-geometry, e.g., a double-convex lens. Furthermore, the present work introduces an optical setup capable of capturing such a light field and presents a mathematical framework for performing the necessary image processing steps on the acquired data.

The paper is organized as follows: Section 2 outlines work on light field processing performed by other researchers. In Sec. 3, a novel optical setup is proposed that acquires light field data suitable for transparent object inspection. Section 4 introduces a mathematical framework that paves the way for conveniently processing the light field data. Furthermore, Sec. 5 describes how scattering material defects are manifested in the light

 ISBN 978-0-9934207-1-9

field and how they can be extracted. The experiments carried out in order to evaluate the presented approach are presented and discussed in Sec. 6. Finally, Sec. 7 provides a summary of all the findings and closes the paper with a short outlook.

2 Related work

The idea of light field cameras is not new. They first were described by Gabriel Lippmann in the early 20th century [Lippmann, 1908]. In the recent years, the technology behind light field cameras emerged greatly, especially because of the research performed by Ng et al. [Ng et al., 2005] and Perwass and Wietzke [Perwass and Wietzke, 2012]. However, the performed research has mainly been targeted on the consumer photography market and not on machine vision applications. This is why there are only a few research groups working on visual inspection applications using light field processing. One of them—the group of [Štolc et al., 2014]—employed multiple line scan cameras to capture the light field reflected by test objects lying on a conveyor belt. Therefore, the cameras were all tilted by a certain angle so that they observed a common line on the conveyor belt from different viewing angles. By this means, they could reconstruct the light field reflected by the objects and use it to calculate all-in-focus images or depth maps. They showed, that the resulting data could successfully be employed for inspecting printed circuit boards for completeness.

In [Soukup et al., 2015], the authors use a light field camera in concert with a variable direction dome illumination in order to partially acquire the bidirectional reflectance distribution function (BRDF) of so-called diffractive optical variable image devices (OVID). OVIDs are frequently used as anti-counterfeiting measures on bank notes since their reflectance highly depends on the illumination's angle of incidence and on the angle of observation. Because of the variable illumination direction and the angular resolution of the sample's reflection achieved by the light field camera, the authors were able to define adequate features suitable for discriminating genuine bank notes from counterfeit ones.

In [Sudhakar et al., 2015], a special Schlieren setup is employed in concert with a compressed sensing approach to acquire light deflection maps of contact lenses. These deflection maps are similar to the deflection histograms introduced in this paper (see Sec. 5). In order to acquire the deflection maps, Sudhakar et al. illuminate the test object with parallel light beams, that are tilted with respect to the optical axis and observe the transmitted light with a telecentric camera system. Since the telecentric lens allows only light rays parallel to the optical axis to reach the sensor and since the angle of the illumination with respect to the optical axis is known, the deflection angles of the transmitted light rays can be calculated. The authors use the deflection maps to calculate the local optical power of the test object, however, they do not use it to test the inspected object for defects.

3 Light field acquisition

As shown in the previous section, there are several methods for capturing light fields. All discussed acquisition systems have different advantages and disadvantages. In order to allow the detection of even very small scattering defects in transparent materials, a potential light field sensor should have a high spatial resolution. Furthermore, there should be a convenient way of calculating the deflection angles corresponding to the captured light rays.

A common light field camera could be used together with a standard lens to capture a full light field during a single exposure [Beyerer et al., 2015]. However, for such an acquisition system, two pixels that belong to different lenslets but have the same relative position with respect to their lenslet, are sensitive for different deflection angles of the captured light rays (see Fig. 1). Therefore, further calibration steps would be needed, what would complicate further processing steps.

To overcome this drawback, this article introduces a novel optical design (see Fig. 2). The sensing part of the setup consists of a lenslet array placed in front of the sensor and of two main lenses. The optical setup can be described as a $4f$ system since the two lenses share a common focal plane and since the distance

 ISBN 978-0-9934207-1-9

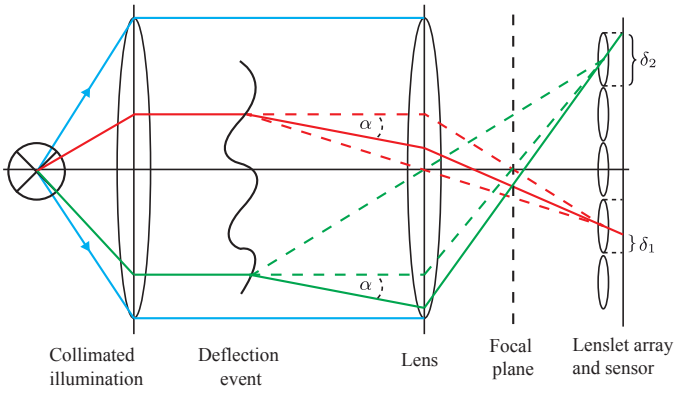


Figure 1: Spatial dependence of angular deflection measurements using a conventional light field camera: two light rays, red and green, get deflected at different spatial positions in the measurement field. Although they have the same deflection angle α, the relative pixel positions δ_1 and δ_2, with respect to the corresponding lenslet, are different.

between the measurement field and the lenslet array is four times a focal length. By this means, pixels that are located underneath different lenslets but have the same relative position with respect to the corresponding lenslet, capture light rays with different origins but with the same range of deflection angles.

In concert with an illumination consisting of parallel rays, all the pixels corresponding to one lenslet represent a local map of the angles by which light rays get deflected inside the respective spatial position of the measuring field.

This optical system has been implemented as a plugin for the physically based rendering framework Mitsuba [Jakob, 2010]. It is used to obtain test images for the experiments described in Sec. 6.

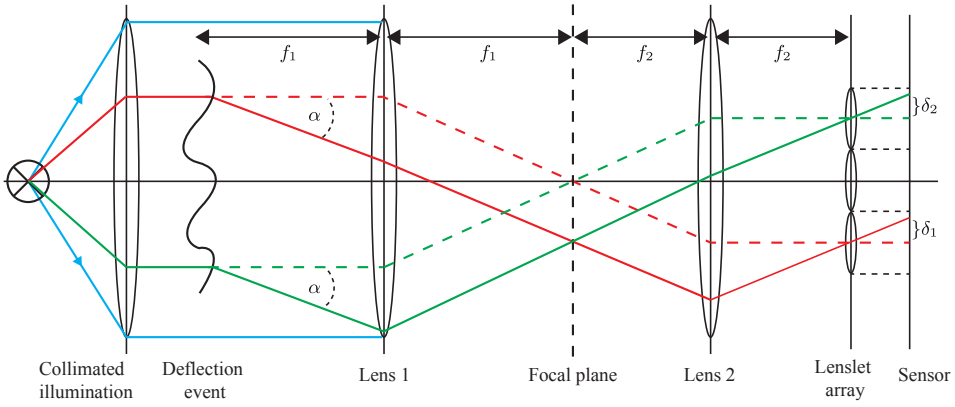


Figure 2: Measuring light deflection angles using a $4f$ system: by means of two lenses with a common focal plane, two light rays, red and green, get deflected by the same angle α at different spatial positions in the measurement field. With respect to the corresponding lenslet, the relative pixel positions δ_1 and δ_2 of the two rays are equal.

4 Mathematical framework

In the following, a framework is introduced that provides a mathematical description of light field data as it could be obtained using the optical setup described in Sec. 3. The light field processing methods presented in the next section are based on this framework.

 ISBN 978-0-9934207-1-9

A light field $L(x,y,\theta,\phi)$ is a function that maps a light ray starting from the position $\mathbf{p} = (x,y)^{\mathrm{T}}$ with a polar angle θ and an azimuthal angle ϕ with respect to a fixed plane in 3D-space to a radiance L. In the acquisition setup from Fig. 2, the lenslet array is responsible for the spatial component, i.e., the (x,y) part of L and the sensor pixels behind each lenslet correspond to the angular information, i.e., the (θ,ϕ) part of L. Therefore, the four-dimensional light field captured using the presented sensor is represented in a 2D-image by spatially multiplexing the spatial and angular component. For demultiplexing and obtaining a more convenient access to the light field data, an alternative formulation is introduced. At first, the discretized pendants m,n,i,j,a of the continuous quantities x,y,θ,ϕ,L are defined by the mappings

$$x \mapsto m\,, \tag{1}$$
$$y \mapsto n\,, \tag{2}$$
$$\theta \mapsto i\,, \tag{3}$$
$$\phi \mapsto j\,, \tag{4}$$
$$L(x,y,\theta,\phi) \mapsto a(m,n,i,j)\,, \tag{5}$$

which depend on the parameters of the optical system and of the sensor. The captured light field can now be expressed via a function S with

$$S : \Omega_S \to (\Omega_A \to [0,255]) : (m,n)^{\mathrm{T}} \mapsto a(m,n,\cdot,\cdot)\,, \tag{6}$$

where Ω_S and Ω_A are the sets of the spatial, respectively, the angular sampling positions and $[0,255]$ denotes the interval of available image values. For a given spatial position $(m,n)^{\mathrm{T}}$, $S(m,n) = a(m,n,\cdot,\cdot)$ is again a function providing the angular information for the respective spatial position. Figure 3 visualizes the concept of the proposed light field representation.

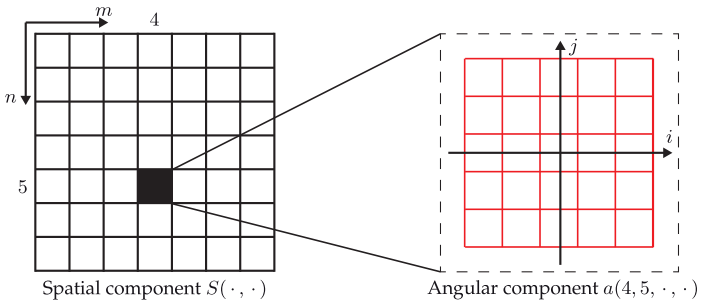


Figure 3: Proposed alternative light field representation by means of a function S, that holds the spatial component of the light field and maps a spatial position $(m,n)^{\mathrm{T}}$ to a corresponding function $a(m,n,\cdot,\cdot)$ holding the respective angular light field component.

5 Light field processing

As stated before, when a defect-free transparent test object is illuminated with collimated light, the object's 3D-geometry will result in deflections of some of the incident light rays. Except for object edges or boundaries, there will be no discontinuities regarding the angular distribution of these deflections with respect to the local neighborhood. However, if the object is affected by a scattering defect (e.g., an enclosed transparent air bubble or a small absorbing particle), the respective light rays will be deflected into multiple directions and result in differences between spatially adjacent angular deflection distributions.
In order to allow a visualization or even detection of scattering defects in acquired light field data S (see Sec. 4), adequate processing steps are necessary that extract features sensitive to the mentioned discontinuities of the angular deflection distribution. For this purpose, this article proposes a chain of processing steps that will be explained in the following paragraphs.

 ISBN 978-0-9934207-1-9

Histogram of ray deflection (HORD) As a first preprocessing step, the input light field data S is normalized to $\tilde{S}$ by calculating histograms out of the single angular deflection distributions:

$$\tilde{S}(m,n) := h(m,n,\cdot,\cdot)\,, \tag{7}$$

$$h(m,n,i,j) := \frac{a(m,n,i,j)}{\sum\limits_{(k,l)\in\Omega_A} a(m,n,k,l)}\,. \tag{8}$$

By this means, influences of varying light intensity are mitigated. The resulting h are called histograms of ray deflections (HORDs). Based on the HORDs, features can be calculated that allow a visualization of material defects resulting in ray deflections. The extraction of two possible features is described in the following sections.

5.1 Discontinuity extraction by histogram comparison

One possible way to find discontinuities between adjacent HORDs is to calculate a special kind of gradient

$$\tilde{\Delta}_{\tilde{S}}(m,n) = \begin{pmatrix} d\left(\tilde{S}(m-1,n), \tilde{S}(m+1,n)\right) \\ d\left(\tilde{S}(m,n-1), \tilde{S}(m,n+1)\right) \end{pmatrix} \tag{9}$$

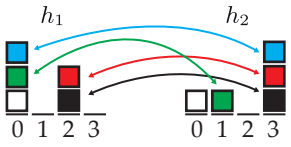
of $\tilde{S}$ in horizontal and vertical direction and to search for peaks in the gradient's norm $\|\tilde{\Delta}_{\tilde{S}}\|$. Now, an adequate distance $d(\cdot,\cdot)$ for two-dimensional histograms has to be defined. Since the histograms corresponding to two rays that have been deflected into different directions should result in a high difference, $d(\cdot,\cdot)$ should consider cross-bin distances.

A distance that is suitable for comparing HORDs in the demanded way is the so-called earth mover's distance $EMD(\cdot,\cdot)$, [Rubner et al., 1998]. The EMD can be imagined as the minimum costs needed for rearranging the probability mass of h_1 to form h_2 (or vice versa). The earth mover's distance is defined by:

$$EMD(h_1,h_2) := \min_{\gamma(k,l)\in\mathcal{M}} \sum_{k=1}^{N}\sum_{l=1}^{N} \gamma(k,l)c(k,l)\,, \tag{10}$$

$$\mathcal{M} = \{\gamma(k,l) : \gamma(k,l) \geq 0, \sum_{l}\gamma(k,l) = h_1(k), \sum_{k}\gamma(k,l) = h_2(l)\}\,, \tag{11}$$

with h_1 and h_2 denoting two histograms with N bins, with $c(k,l)$ denoting the costs of moving one unit probability mass from bin k to bin l and with $\gamma(k,l)$ denoting the amount of probability transfered from bin k of h_1 to bin l of h_2. Figure 4 visualizes the calculation of the one-dimensional EMD for two example histograms.



$$EMD(h_1, h_2) = 1 + 1 + 3 + 1 = 6$$

Figure 4: Example calculation of the one-dimensional earth mover's distance between the two histograms h_1 and h_2: as visualized by the arrows, the black, red and green unit mass each have to be moved by a distance of 1 and the blue unit mass by distance of 3, resulting in an $EMD(h_1,h_2) = 6$.

Experiments carried out by using the earth mover's distance for gradient calculation on simulated images are presented in Sec. 6.

5.2 Discontinuity extraction by vector analysis

Besides histogram distances, methods of vector analysis are also suitable for extracting information on discontinuities of the local angular deflection distribution. For every HORD $\tilde{S}(m,n) = h(m,n,\cdot,\cdot)$, the mean ray

deflection direction $\bar{\mathbf{r}}(m,n)$ can be calculated as follows:

$$\bar{\mathbf{r}}(m,n) = \sum_{(i,j)\in\Omega_A} h(m,n,i,j)\cdot\mathbf{r}(i,j)\,, \tag{12}$$

with $\mathbf{r}(i,j)$ denoting the respective deflection direction of the histogram bins $(\cdot,\cdot,i,j)$. The resulting vector valued function $\bar{\mathbf{r}}$ can now be considered as a vector field and can be further processed using methods of vector analysis.

As mentioned above, scattering material defects result in discontinuities of the angular deflection distribution, which is why it is sensible to look for discontinuities of $\bar{\mathbf{r}}$. Therefore, the Jacobians $\mathbf{J}$ of $\bar{\mathbf{r}}$ can be calculated:

$$\mathbf{J}(m,n) = \begin{pmatrix} \frac{\partial \bar{r}_m}{\partial m}(m,n)\,, & \frac{\partial \bar{r}_m}{\partial n}(m,n) \\ \frac{\partial \bar{r}_n}{\partial m}(m,n)\,, & \frac{\partial \bar{r}_n}{\partial n}(m,n) \end{pmatrix}. \tag{13}$$

For both the m- and n-components of $\bar{\mathbf{r}}(m,n)$, the Jacobians $\mathbf{J}(m,n)$ contain the respective derivatives in m- and n-direction [Horn and Johnson, 2012]. Scattering material defects result in changes of the direction of incident rays and therefore lead to discontinuities of the angular deflection distribution. Since these discontinuities are manifested as high values of the respective components of the Jacobians $\mathbf{J}$, the corresponding defects can be found by applying an adequate matrix norm $\|\cdot\|$ and by looking for peaks in the resulting scalar image $\|\mathbf{J}(m,n)\|$.

Frobenius norm As mentioned above, high values of the components of $\mathbf{J}(m,n)$ might indicate that there is a defect present in the test object at position $(m,n)^{\mathrm{T}}$. Hence, the Frobenius norm [Horn and Johnson, 2012] of $\mathbf{J}$,

$$\|\mathbf{J}(m,n)\|_{\mathrm{F}} = \sqrt{\left|\frac{\partial \bar{r}_m}{\partial m}(m,n)\right|^2 + \left|\frac{\partial \bar{r}_m}{\partial n}(m,n)\right|^2 + \left|\frac{\partial \bar{r}_n}{\partial m}(m,n)\right|^2 + \left|\frac{\partial \bar{r}_n}{\partial n}(m,n)\right|^2}\,, \tag{14}$$

might be a sensible choice, as it results in higher values for higher values of the single components of $\mathbf{J}$. However, the Frobenius norm does not take the directional information into account, that is embedded in $\mathbf{J}$. This characteristic of $\|\cdot\|_{\mathrm{F}}$ could cause Jacobians with a high absolute value in one directional component, i.e., with a high anisotropy, to yield the same norm as Jacobians with moderate values in every directional component, i.e., with a low anisotropy.

6 Experiments

This section describes the experiments[1] conducted for evaluating the proposed light field processing approaches. Therefore, inspection images of virtual test scenes have been rendered using the physically based rendering framework Mitsuba [Jakob, 2010]. The $4f$ light field acquisition system introduced in Sec. 3 and shown in Fig. 2 has been implemented as a sensor plugin. Its spatial resolution has been set to 555×555 pixels and the angular resolution to 9×9 directions. The emitter plugin introduced in [Meyer et al., 2016] serves as the required parallel illumination. Furthermore, inspection images for a conventional machine vision system consisting of a telecentric camera also having a spatial resolution of 555×555 in combination with an area light source have been simulated in order to compare the proposed approach against an existing method.

A double-convex lens has been placed in the virtual scene to serve as a test object. Appropriate manipulations of the test object simulated different kinds of material defects. To all rendered inspection images, the two image processing approaches presented in Sec. 5 have been applied to evaluate their suitability for visualizing the respective types of defects. For every simulated inspection image, a gradient image using the 2D earth mover's distance and the Frobenius norm of the Jacobians of the mean deflection directions have been calculated and converted to pseudocolor images for visualization purposes. Figure 5 shows the results of the experiments.

[1] All data needed for reproducing the presented results are available online at https://www.meyer-research.de .

For a defect-free test object, the resulting pseudocolor images have high intensities mainly at the borders of the test object. This is because the HORDs change abruptly in the respective regions. Furthermore, the sampling noise of the employed rendering framework causes single pixels to show high intensities.
In order to simulate a scattering shape defect, a part of a small sphere has been cut out of the test object in its upper left quarter by means of constructive solid geometry. Although both approaches reveal the defect, its signature seems to be stronger in the image obtained using the earth mover's distance. The conventional telecentric setup also clearly shows this defect.

In the third experiment, a small air bubble, i.e., a small scattering defect, has been placed inside the test object's center. When compared to the images corresponding to the defect-free test object, the defect is clearly visible. For this kind of defect, the earth mover's distance seems to be a particularly adequate choice, since the size of the small dot visible in the test object's center is close to the defect's actual size. In contrast to the proposed approach, the conventional inspection setup is not able to visualize the defect.

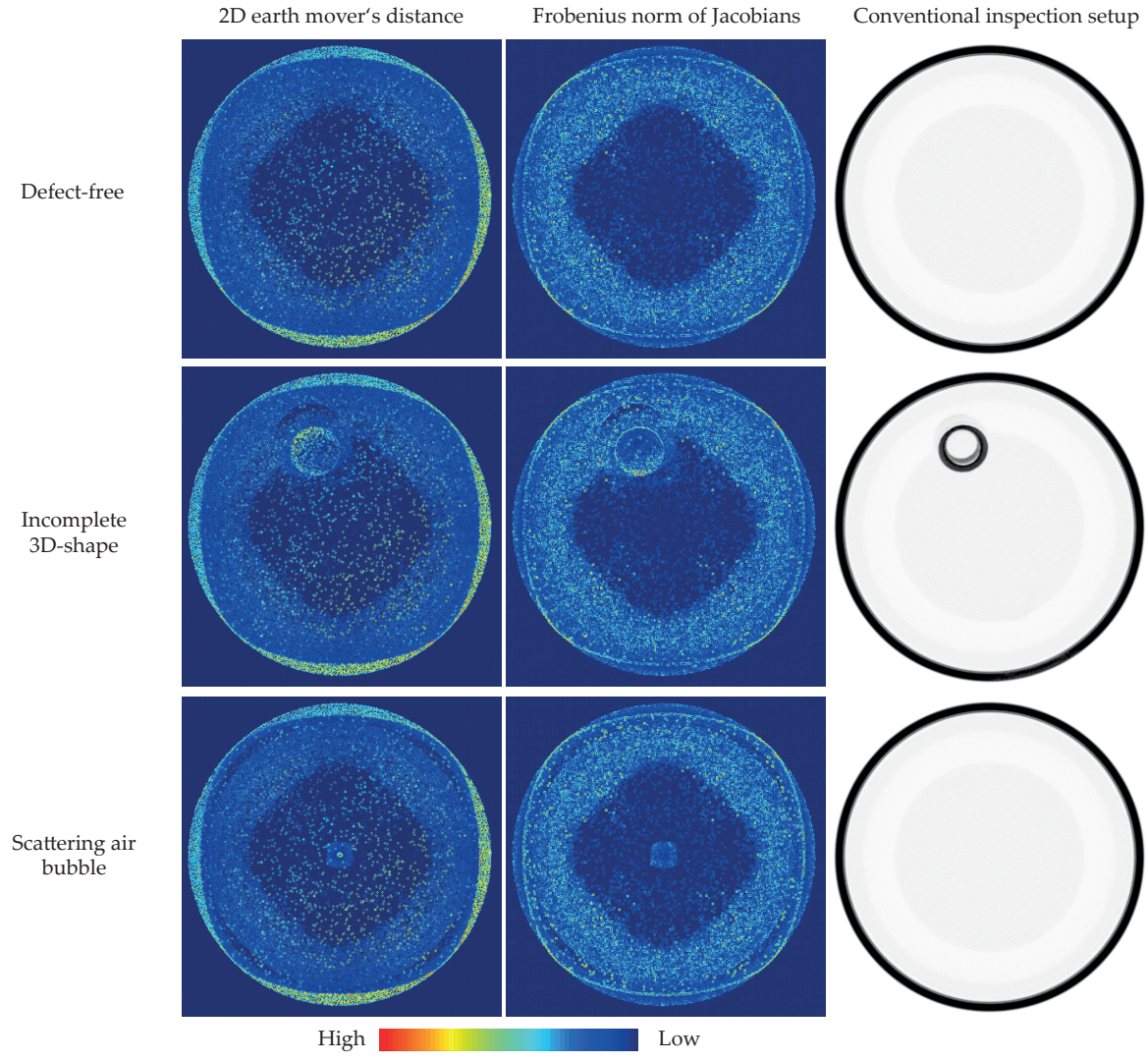


Figure 5: Results of the performed experiments.

In summary, the experiments show that the proposed image processing approaches are suitable for visualizing scattering defects in transparent materials. The earth mover's distance seems to yield images, that contain slightly more information about the defects, however, all simulated kinds of defects are also clearly visible in the images representing the Frobenius norm of the Jacobians. For small scattering inclusions, the proposed approach seems to even outperform conventional inspection methods based on telecentric camera systems.

 ISBN 978-0-9934207-1-9

7 Conclusion

This article introduces a $4f$ light field setup for acquiring spatially resolved histograms of ray deflections (HORDs) that can be employed for testing transparent objects for scattering material defects. Therefore, the paper presents two image processing approaches capable of visualizing these kinds of defects by extracting variations of the light's local deflection directions out of the mentioned histograms. The first approach performs histogram comparisons using the two-dimensional earth mover's distance. The second approach relies on matrix norms that are calculated for the Jacobians of the light's mean deflection direction. Simulations of the proposed acquisition setup using a physically based renderer and further processing of the resulting images show promising results stating the method's suitability for transparent object inspection.

Future work will focus on extending the proposed image processing algorithms, e.g., by supporting other histogram distances like the Bhattacharyya distance. Furthermore, a prototype of the introduced optical system will be set up in order to acquire real measurements allowing to further evaluate the approach.

References

[Beyerer et al., 2015] Beyerer, J., León, F. P., and Frese, C. (2015). *Machine Vision: Automated Visual Inspection: Theory, Practice and Applications*. Springer Berlin Heidelberg.

[Horn and Johnson, 2012] Horn, R. and Johnson, C. (2012). *Matrix analysis*. Cambridge University Press.

[Jakob, 2010] Jakob, W. (2010). Mitsuba renderer. http://www.mitsuba-renderer.org.

[Lippmann, 1908] Lippmann, G. (1908). Epreuves reversibles. photographies integrals. *Comptes-Rendus de l'Académie des Sciences*, 146:446–451.

[Meyer, 2014] Meyer, J. (2014). Visual inspection of transparent objects – physical basics, existing methods and novel ideas. Technical Report IES-2014-04, Karlsruhe Institute of Technology.

[Meyer et al., 2016] Meyer, J., Gruna, R., Längle, T., and Beyerer, J. (2016). Simulation of an inverse schlieren image acquisition system for inspecting transparent objects. *Electronic Imaging*, 2016(19):1–9.

[Ng et al., 2005] Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., and Hanrahan, P. (2005). Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11.

[Perwass and Wietzke, 2012] Perwass, C. and Wietzke, L. (2012). Single lens 3d-camera with extended depth-of-field. In *IS&T/SPIE Electronic Imaging*, pages 829108–829108. International Society for Optics and Photonics.

[Rubner et al., 1998] Rubner, Y., Tomasi, C., and Guibas, L. (1998). A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE.

[Soukup et al., 2015] Soukup, D., Štolc, S., and Huber-Mörk, R. (2015). Analysis of optically variable devices using a photometric light-field approach. In *IS&T/SPIE Electronic Imaging*, pages 94090R–94090R. International Society for Optics and Photonics.

[Štolc et al., 2014] Štolc, S., Huber-Mörk, R., Holländer, B., and Soukup, D. (2014). Depth and all-in-focus images obtained by multi-line-scan light-field approach. In *IS&T/SPIE Electronic Imaging*, pages 902407–902407. International Society for Optics and Photonics.

[Sudhakar et al., 2015] Sudhakar, P., Jacques, L., Dubois, X., Antoine, P., and Joannes, L. (2015). Compressive imaging and characterization of sparse light deflection maps. *SIAM Journal on Imaging Sciences*, 8(3):1824–1856.

Abnormal pedestrian trajectory analysis based on arbitrary-length clustering

Diane Murdock and Jesus Martinez del Rincon

The Centre for Secure Information Technologies (CSIT), Queen's University Belfast, UK

Abstract

This papers examines the use of trajectory distance measures and clustering techniques to define normal and abnormal trajectories in the context of pedestrian tracking in public spaces. In order to detect abnormal trajectories, what is meant by a normal trajectory in a given scene is firstly defined. Then every trajectory that deviates from this normality is classified as abnormal. By combining Dynamic Time Warping and a modified K-Means algorithms for arbitrary-length data series, we have developed an algorithm for trajectory clustering and abnormality detection. The final system performs with an overall accuracy of 83% and 75% when tested in two different standard datasets.

Keywords: Trajectory Analysis, Dynamic Time Warping, Clustering, Abnormal detection

1 Introduction

Video surveillance of public spaces in order to study activities and behaviours is becoming increasingly popular in today's society. The benefit of processing and analysing surveillance data automatically is becoming increasingly obvious since allows activity recognition and anomaly identification. This is due to the need for heightened public safety and crime prevention by law enforcement agencies in vast areas and camera networks. In particular, the study of trajectory patterns corresponding to erratic or obscure movements and activities, such as wandering or loitering, may suggest suspicious and abnormal behaviour in public spaces.

A new generation of pedestrian detectors and trackers, able to provide an unprecedented accuracy even in dense and moderately crowded sequences, allows the development of human behaviour analysis systems that can cater from their outputs. Thus, abnormal behaviour detection based on the analysis of accurate trajectories provided by automated systems has emerged in the literature [Morris and Trivedi, 2009, Zhang et al., 2006]. While incorporating other sources of information in addition to the simple raw trajectory data, such as local motion descriptors [Datta et al., 2002] or spatio-temproal data [Robertson and Reid, 2006] may provide more precise information about the types of actions and give better results in the detection of abnormal behaviour, there are also disadvantages and they make the system more dependent on the environment and camera setup.

In this context, the present paper focuses on trajectory analysis of pedestrians in public spaces. To distinguish between normal and abnormal trajectories, two main processes are applied. Our proposed framework first defines what is meant by a normal trajectory in a given scene, and then classifies as abnormal every trajectory that deviates from this normality. Given a set of 'normal' trajectories, these are grouped into clusters according to information encapsulated into the trajectories elements. A distance able to take into consideration the specific characteristic of human motion patterns, such as different lengths, speeds and jittering, is crucial in such process. Trajectories that are close to the cluster mean are considered normal, while all trajectories that lie further away are considered abnormal.

 ISBN 978-0-9934207-1-9

1.1 State of the art

In trajectory analysis, unsupervised learning provides a versatile and effective approach given the lack of prior information regarding the different types of possible trajectories for each given scenario. In this context, clustering has been proved as a standard approach to group and classify trajectories. Among the different clustering options, diverse approaches such as k-means, fuzzy k-means, graph mining, Expectation-Maximization, Self-organised maps and Hidden Markov models have been evaluated with no significant differences between them [Morris and Trivedi, 2009]. On the contrary, the choice of the distance used to compare trajectories shows a bigger relevance and it relates directly with the resulting clustering [Zhang et al., 2006]. For instance, Euclidean distance allows measuring the dissimilairty between two trajectories but it is limited to trajectories with exactly the same number of elements, which is unusual in pedestrian analysis.

As a consequence, a preprocessing step is usually required prior to clustering. Since human trajectories are diverse in their execution, a normalization process is required to allow clustering algorithms to properly group trajectories based on their motion patterns and shape rather than by speed variations, global direction, fragmentation or punctual differences.

In order to address this limitation, the simplest solution consists in applying a geometric transformation to raw trajectory data prior to its classification. In [Sillito and Fisher, 2008, Majecka, 2009], raw trajectories are approximated by cubic spline curves with a fixed number of points. In [Johnson and Hogg, 1996], a slightly different approach was proposed, where each trajectory is converted into a sequence of 4-dimensional flow vectors composed of 2D position coordinates and velocity of the tracked object, followed by a clustering algorithm based on neural networks to obtain a finite set of prototype vectors. These approaches allow trajectories to be represented by the same number of attributes in order to facilitate the posterior analysis. However, they entail a set of assumptions, such as the required number of points or the order of the spline approximation, which is not constant for different motion patterns and largely unknown a priori for a give scene.

Other approaches [Bashir et al., 2007] made use of dimensionality reduction techniques such as principal component analysis (PCA) to select automatically the most relevant elements before clustering. This permits to initially overestimate the number of elements in a trajectory that is then projected in lower dimensional space. However, a fixed number of points should be still decided initially and an interpolation/subsampling may be required for every trajectory. In general, they suffer to distinguish speed variations [Morris and Trivedi, 2009].

As alternative, the use of more complex distance metrics that allow comparing two set of data of arbitrary number of elements can be applied. Hausdorff distance has been proposed [Junejo et al., 2004], but since only captures the minimum distance between two shapes, the subtle information encapsulated in human trajectories and required for its classification may be lost [Zhang et al., 2006] leading to poor classification rates. Better results have been achieved with the use of Dynamic Time Warping (DTW) [Ratanamahatana and Keogh, 2004] and Longest Common Subsequence (LCSS) [Buzan et al., 2004]. The extra complexity of these distance metrics allow simpler clustering algorithms to be employed in the later stage. However, while the aforementioned distances allow comparing trajectories of different number of elements they may bring new problems such as the removal of contextual information, such as global direction, entrance and exit in the scene, etc., helpful for the classification of trajectories.

Moreover, only a minority of approaches in the literature have addressed the detection of abnormal behaviour, but limited to classify trajectories into known classes.

2 System Architecture

The aim of our proposed system is to detect abnormal trajectories. In order to achieve this goal, the meaning of what a normal trajectory in a given scene is should be first defined. Figure 1 depicts our proposed framework. Our system is composed of a k-means-based clustering algorithm, which has been modified in order to allow arbitrary length trajectories to be compared and grouped using DTW as a distance metric. The proposed modifications also allowed trajectories of different element numbers to be averaged and their variability to be measured, which is a novel contribution regarding other implementations. The resulting cluster means will

 ISBN 978-0-9934207-1-9

represent the normal trajectory prototypes and their learned variation, in the form of standard deviation, will be finally used in testing to classify and separate normal and abnormal trajectories.

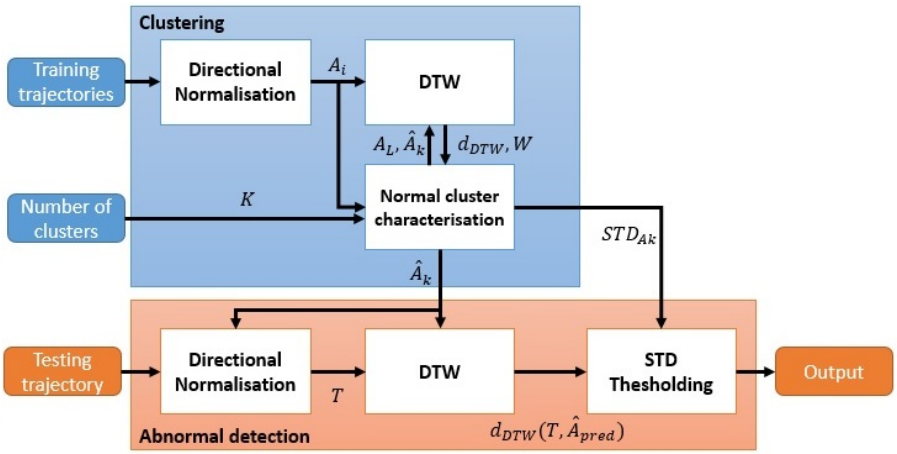


Figure 1: *Framework overview*

The input data to our system are the raw trajectories provided by any automated human pedestrian detector and tracking system. We define a trajectory as a nxd-dimensional time series composed of a sequence of n elements of d-dimensions associated with a pedestrian or an object moving through a scene. In order to fully exploit the information provided by the sourcing automated system, each element is composed by the 2D spatial coordinates and their corresponding instant velocities as attributes. Thus a trajectory A is defined as follows:

$$A = (a^1, a^2, ...a^n) = \left((x^1, y^1, v_x^1, v_y^1), (x^2, y^2, v_x^2, v_y^2), \ldots, (x^n, y^n, v_x^n, v_y^n)\right) \tag{1}$$

where n is the number of elements, (x, y) is each of the 2D points composing the trajectory and (v_x, v_y) their corresponding instant velocities.

2.1 Dynamic Time Warping

DTW is a time series alignment algorithm which aims to align two time series of coordinates by warping the time axis iteratively until an optimal match between the two time-series has been found [Ratanamahatana and Keogh, 2004]. In this manner, two time series that are similar but locally out of sync can be aligned in a nonlinear manner. The d_{DTW} distance between two trajectories A_1 and A_2 of lengths n_1 and n_2 respectively is described as

$$d_{DTW}(A_1, A_2) = \min\left\{\sqrt{\sum_{k=1}^{\min(n_1,n_2)} w_k}\right\} \tag{2}$$

where $w_k \in W_{A_1,A_2}$ is the matrix element $C(i,j)_k$ that belongs to the k^th element of the warping path W_{A_1,A_2}, a continuous set of matrix elements that represent the mapping -or minimum path- between A_1 and A_2. Each matrix element is calculated recursively following:

$$C(i,j) = d(a_1^i, a_2^j) + \min\{C(i-1,j-1), C(i-1,j), C(i,j-1)\} \tag{3}$$

being $d(a_1^i, a_2^j)$ the Euclidean distance between two trajectory elements.

In our framework, DTW was chosen above other alternatives such as LCSS due to DTW sensitiveness to outliers. Since our aims is to detect anomalies, this sensitivity would draw attention to abnormalities in the data.

 ISBN 978-0-9934207-1-9

2.1.1 Directional normalisation

Since we aim to define normality versus abnormality rather than thoroughly classify normal trajectories, trajectories from a point X to a point Y will be considered identical to those from Y to X for the sake of compactness and better resource management. However, given that in DTW compares sequentially each element in the trajectory without having and overall look at the trajectory, DTW will return a large distance even when two trajectories are identical element by element. This means that both trajectories could be assigned to different clusters. To avoid this undesired effect for our application, the global direction of the trajectories is aligned before applying DTW by flipping each trajectory if $d(a_1^1, a_2^{n_2}) < d(a_1^1, a_2^1)$ so their closest extreme points are located at the initial position of the data series.

2.2 Trajectory Clustering

We base our system on a modification of the k-means algorithm that allows times series of different lengths to be clustered. Given a number K of expected clusters, the algorithm first assigns each trajectory to a cluster using DTW such that the distance between the trajectory and the cluster mean $\hat{A}_k$ is the minimum distance.

$$k_i = \arg\min_k \{d_{DTW}(A_i, \hat{A}_k)\} \tag{4}$$

Then, each cluster average is recalculated using all the correspondingly assigned trajectories. This algorithm iterates until convergence, i.e. cluster means do not change, or a maximum number of iterations is reached.

However, while the cluster means $\hat{A}_k$ are easily initialised by randomly selecting K trajectories in the training set, the computation of a trajectory average given sets of elements of different lengths is not trivial. In order to recalculate the new cluster means, our method uses the mappings W provided by DTW between the longest trajectory A_L and every other trajectory A_i in its cluster. Once the correspondence between the longest reference trajectory elements to every other element within the trajectories in the cluster is known, the cluster mean can be computed as the concatenation of element averages with different number of elements using eq. 6. Figure 2 illustrates this procedure with a graphical representation.

$$\hat{A}_k = (\hat{a}_k^1, \hat{a}_k^2, ..., \hat{a}_k^{n_L}) \tag{5}$$

$$\hat{a}_k^n \propto \sum_{i \in k} \left(\sum_{w^n \in W_{A_L, A_i}} a_i^{w^n} \right) \tag{6}$$

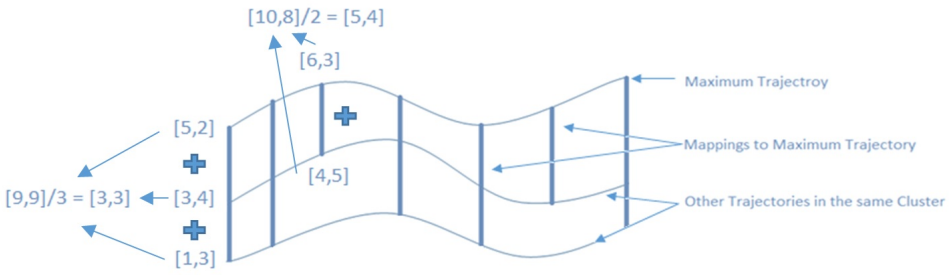


Figure 2: *Mapping correspondence between three trajectories and averaging procedure*

Two alternatives were tested for this paper to define the longest trajectory. First, the longest trajectory in a cluster was defined as the trajectory that possess the largest number of elements n_i so

$$L = \arg\max_i \{n_i\} \tag{7}$$

Second, the longest trajectory was defined as the trajectory with the largest physical distance between its start and end point

$$L = \arg\max_i \{d(a_i^1, a_i^{n_i})\} \tag{8}$$

The comparison between these two approaches will be analysed in the experimental section.

2.3 Defining normality and detecting abnormality

Once the trajectory clustering has converged to a solution, each cluster mean $\hat{A}_k$ can be used as a prototype for each normal trajectory type. By selecting the minimum distance, a new trajectory T can be assigned and classified to a specific type or normal trajectory, so

$$pred = \arg\min_k \{d_{DTW}(T, \hat{A}_k)\} \tag{9}$$

While the previous equation predicts the most likely classification assuming normality, abnormal trajectories will still be considered as part of a normal cluster. To avoid it, a trajectory will be classified as abnormal if it lies outside a given distance or threshold from its assigned cluster. We propose a threshold which is different for each of the normal clusters and proportional to the expected variance learned during training. This variance is calculated in a similar manner to the cluster averages, where the previously generated mappings W are used to calculate the standard deviation around the means of each of the clusters, using the equation:

$$STD_{A_k} = \sqrt{\frac{1}{n_L} \sum_{n=1}^{n_L} std^2_{a_k^n}} \tag{10}$$

$$std^2_{a_k^n} \propto \sum_{i \in k} \left(\sum_{w^n \in W_{A_L, A_i}} (a_i^{w^n} - \hat{a}_k^n)^2 \right) \tag{11}$$

Finally, abnormal trajectories are detected by applying thresholding:

$$Abnormal = \begin{cases} 1 & \text{if } d_{DTW}(T, \hat{A}_{pred}) > \alpha \cdot STD^2_{A_{pred}} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where α is a constant chosen empirically.

3 Experimental Results

3.1 Datasets and experimental setup

Two different datasets have been used to validate the experiments and ensure that the conclusions are not depending on scenario or camera setup. The first dataset is the Edinburgh Informatics Forum Pedestrian Database [Edinburgh, 2016] which contains trajectories of detected targets of people walking observed from a zenital view. Since manual annotation is needed to validate our experiments, a subset with the first 150 trajectories in the file tracks.24Aug were selected and classified as 'normal' or 'abnormal', obtaining an split of 66% normal and 34% abnormal trajectories. As annotation criterion, for a trajectory to be considered normal, it must start and end at an entry and exit in the scene with little diversion. Any trajectory that does not meet these criteria is considered abnormal.

The second dataset is the Oxford Real-Time Surveillance Town Centre Dataset [Oxford, 2016], which consists of a video surveillance sequence of in a town centre on a busy pedestrianised shopping street with several points of interest and shop entrances. The first 150 trajectories were selected and annotated for our evaluation, obtaining a 76.7% - 23.3% normal versus abnormal split.

3.2 Results

Different variations of our methodology were tested: using only 2D coordinates as attributes for each element (version 1), using 4D attributes for each element of the trajectory (2D coordinated + instant velocity) (version 2), using the directional alignment explained in section 2.1.1 (version 3 and 4). Version 3 and 4 differ in their

 ISBN 978-0-9934207-1-9

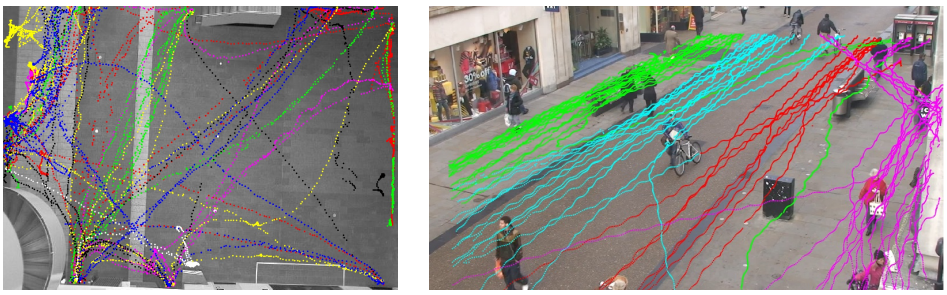


Figure 3: *a) Edinburgh dataset and b) Oxford dataset clustering obtained with our system (v4).* Different colours represent different clusters

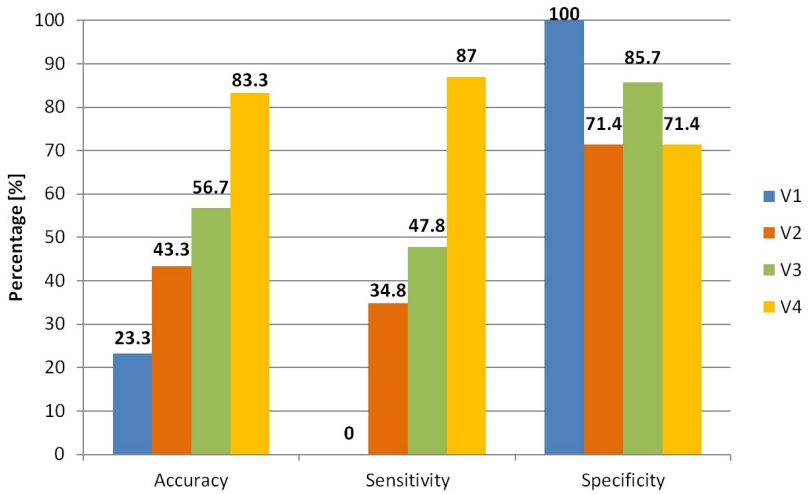


Figure 4: *Comparison between the 4 different versions of our system on the Edinburgh dataset*

definition of longest trajectory used as reference for each cluster: version 3 defines it as the trajectory with most elements while version 4 uses the trajectory with the largest physical distance (see eq. 7 and 8).

Results are shown in Figure 4 for the Edinburgh dataset, using an 80%-20% split between training and testing (similar to [Majecka, 2009]) but respecting the original normal-abnormal ration between them. Parameter α was set to 180 while the number of normal cluster K was equal to 8. It can be seen how each version improves the previous one, demonstrating the individual value of each contribution. Thus, adding velocity to the comparison of each individual element improved the overall d_DTW and the mapping W between trajectories. This is due to the fact that DTW algorithm fully removes the global structure and position of the trajectories which results on a loss of the trajectory structure. On the contrary, the explicit addition of velocity information preserves better the relations between trajectory elements. Direction normalisation also improves the result by reducing the number of clusters needed for defining normality and avoiding a multiplicity of cluster with similar trajectories. Finally, the best results are obtained when using as a reference for each cluster the longest physical trajectory rather than the trajectory with the largest number of elements. This is because humans tend to wander around points of interest in the scene for a long time which may result on extremely long outliers which should not be taken as reference to avoid corruption. Qualitative effect of each of this modifications are depicted in Fig. 5.a, b and c.

Previous results were obtained with manually-tuned parameter values α and K, so certain degree of overfitting to teh scenario is expected. To evaluate the sensitivity of our method to those parameters, parameter values were varied over a range for our best framework (version 4). Results are reported in Fig. 6. While the final performance varied according to the particular values, it can be observed how good performance is still achieved over a large range of values, which makes the algorithm easy to tune.

In order to evaluate the generality of our framework, our system (version 4) was evaluated on the Oxford dataset. A 50%-50% split between training and testing was used and parameters were set to $\alpha = 600$ and $K = 4$. Lower performance is achieved in this second dataset, which may be explained by the more limited definition

 ISBN 978-0-9934207-1-9

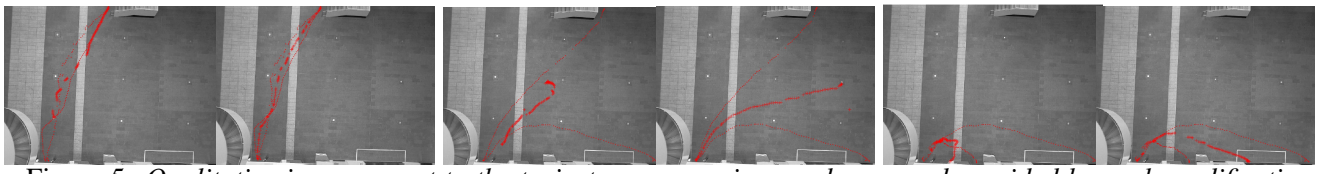


Figure 5: *Qualitative improvement to the trajectory comparison and averaged provided by each modification of the clustering algorithm, from left to right: a) Before and b) after adding instant velocity to the trajectory series, c) before and d) after directional normalisation, using e) maximum number of elements or f) longest physical distance to define longest cluster trajectory. Dots represent raw trajectory points while x represents the calculated average*

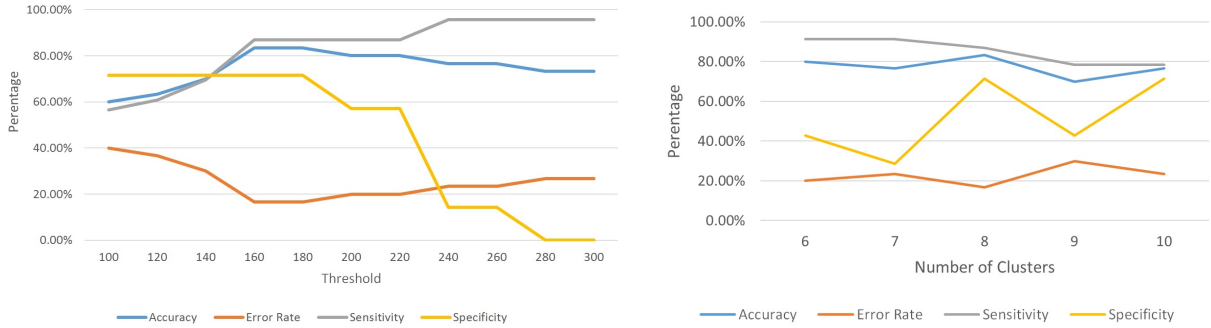


Figure 6: *Performance results varying the parameters, from left to right: a) threshold constant α , b) number of cluster K.*

of this scenario, where most normal trajectories are linear motions along the street. This also justify teh use of a larger α value, to ensure that normal trajectories are still included as such.

Finally, our system was compared against [Majecka, 2009], one of the scarce methods in the literature that targets abnormal trajectory detection rather than trajectory classification. The same testing set (20% of the Edinburgh dataset) was used in the comparison. Results could not be generated with this system in the Oxford dataset due to the very different scenario and the lack of a training model. These results are summarised in Table 1. Our method achieved the same performance than [Majecka, 2009] but with a much simpler and reduced training set (only 120 training trajectories versus a few thousand). Opposite to [Majecka, 2009], our methodology does not require supervised learning and abnormal trajectories must be removed. Furthermore, since our system does not apply any geometric transformation to the trajectories, it can also be easily applied to other scenarios with different camera perspectives.

4 Conclusion

In this paper, an unsupervised framework for abnormal trajectory analysis has been proposed. The system proposed an extension of k-means clustering that allows dealing with human trajectories of varied and arbitrary sizes. Our system has been evaluated in two different scenarios obtaining good performance in both, being the results at state-of-art level in the standard Edinburgh dataset, and showing advantages regarding previous methodologies in terms of simplest and fully unsupervised training as well as easier extension to different environments and camera perspectives.

As future work, an extension to automatically determine the number of cluster using Expectation maximization will be proposed.

References

[Bashir et al., 2007] Bashir, F. I., Khokhar, A. A., and Schonfeld, D. (2007). Object trajectory-based activity classification and recognition using hidden markov models. *EEE Trans. Image Processing*, 16:1912Ű1919.

Table 1: *Result comparison over the 2 datasets.*

	Edinburgh				Oxford			
Method	Acc	Err	Sens	Spec	Acc	Err	Sens	Spec
Ours (V4)	0.83	0.17	0.87	0.71	0.75	0.25	0.78	0.63
[Majecka, 2009]	0.83	0.17	0.87	0.71	-	-	-	-

[Buzan et al., 2004] Buzan, D., Sclaroff, S., and Kollios, G. (2004). Extraction and clustering of motion trajectories in video. In *Intl. Conf. on Pattern Recognition (ICPRŠ04)*, volume 2.

[Datta et al., 2002] Datta, A., Shah, M., and Lobo, N. D. V. (2002). Person-on-person violence detection in video data. In *International Conference on Pattern Recognition (ICPR'02)*, volume 1, page 433Ű438.

[Edinburgh, 2016] Edinburgh (2016). The edinburgh informatics forum pedestrian database. http://homepages.inf.ed.ac.uk//rbf//FORUMTRACKING//.

[Johnson and Hogg, 1996] Johnson, N. and Hogg, D. (1996). Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14:583Ű592.

[Junejo et al., 2004] Junejo, I., Javed, O., and Shah, M. (2004). Multi feature path modeling for video surveillance. In *17th Intl. Conf on Pattern Recognition (ICPRŠ04)*, volume 2, pages 716–719.

[Majecka, 2009] Majecka, B. (2009). Statistical models of pedestrian behaviour in the forum. Master's thesis, University of Edinburgh.

[Morris and Trivedi, 2009] Morris, B. and Trivedi, M. (2009). Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR'09)*.

[Oxford, 2016] Oxford (2016). Oxford stable multi-target tracking in real-time surveillance dataset. http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html.

[Ratanamahatana and Keogh, 2004] Ratanamahatana, C. and Keogh, E. (2004). Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*.

[Robertson and Reid, 2006] Robertson, N. and Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104:232Ű248.

[Sillito and Fisher, 2008] Sillito, R. and Fisher, R. (2008). Semi-supervised learning for anomalous trajectory detection. In *British Machine Vision Conference (BMVC'08)*.

[Zhang et al., 2006] Zhang, Z., Huang, K., and Tan, T. (2006). Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1135–1138.

Comparison of approaches to landmark identification on 3D torso surface meshes for breast reconstruction

S. Foster, P.J. Morrow, B.W. Scotney, R.J. Winder, S.A. McIntosh

Faculty of Computing & Engineering,
Coleraine Campus, Ulster University

Belfast Health and Social Care Trust

Abstract

Breast reconstruction is a vital part of breast cancer treatment for many women and can contribute to maximizing their quality of life by reducing the impact of breast cancer on their physical appearance. The limited reproducibility of subjective outcome evaluation techniques including panel evaluation has indicated a need for objective methods. Anthropometry requires fiducial points to be directly marked on the torso of subjects being assessed or indirectly on photographs before measurements are carried out, however subjectivity still exists with anthropomorphic techniques. Automating the identification of fiducial points such as the nipples will permit more consistent and reproducible quantitative measures of breast morphology. This paper investigates algorithms for automatic detection of nipples on 3D surface images and the impact that applying various thresholding and clustering operations on 2D texture data has on the automated placement of the nipple compared to the ground truth manually marked location.

Keywords: Image Processing, landmark detection, 3D surface mesh, segmentation, objective evaluation

1 Introduction

During 2010-2014 in Northern Ireland, 1,283 females were diagnosed with breast cancer per year and 306 per year died from the disease, with the lifetime risk of women developing breast cancer being 1 in 11 [1]. In the last 10 years female death rates have fallen by around a fifth [2], which may be due to earlier diagnosis through the introduction of breast cancer screening and advancements in treatments including chemotherapy, radiotherapy and surgery. Breast cancer surgery may involve partial removal of the breast or a mastectomy procedure during which the entire breast is removed. Breast reconstruction is the rebuilding of the breast mound using prosthetic implants or tissue taken from other parts of the body to create a natural breast shape [3]. Sabczynski *et* al. [4] developed a system to support surgeons in surgical planning by permitting visualization of the foreseen breast surgery results, this provides patients with opportunity to participate in a Shared Decision making process with their clinicians.

Evaluating the cosmetic outcome of breast reconstruction is essential in order to make improvements in current strategies by identifying variables which affect breast aesthetics [5]. Visual assessment by a panel of observers is the most frequently used method for evaluating patient cosmetic outcome [6]–[8]. The panel may comprise of independent clinicians and lay persons, where each panel member independently scores a range of aspects using photographs of the breasts, taking into account breast symmetry, scars and skin changes [9]. The overall aesthetic outcome of the patient being assessed is scored using a rating scale which ranks comparisons between their treated and untreated breast [10]. Results obtained through panel evaluation have been shown to lack accuracy and reproducibility through low intra and inter-rater agreement [11], [12]. The process is time consuming and impractical as requiring the participation of multiple health professionals is a hindrance when there are large volumes of patients to assess. Effective evaluation of

breast reconstruction surgery requires an objective, consistent and efficient processing technique capable of providing three-dimensional measures of breast aesthetics.

Three-dimensional (3D) imaging has gained credibility in the research environment and has the potential to provide efficient, accurate and repeatable objective outcome measures [13]–[17], incorporating parameters which are not available from two-dimensional images such as volume, surface area, projection, contour and symmetry [10]. Distances between fiducial points on the female torso have been used to objectively quantify aspects of the breast such as ptosis [18]. Current objective breast assessment software BCCT.core [19] and AxisThree [20] require manual placing of landmarks on frontal torso images prior to performing symmetry calculations. Knowledge of surgical terminology is required to accurately locate a number of landmarks on the female torso, therefore these applications require a clinician to carry out the evaluation. The manual landmark procedure is time consuming and has been shown to introduce inter- and intra-observer variability [21]. Automating the detection of fiducial points will achieve efficient, productive, accurate and reliable breast evaluations. This paper investigates a number of algorithms for automated nipple identification on 3D torso surface meshes. Some preliminary results are presented for various thresholding and segmentation techniques by comparing the location of the automatically detected nipple to the manually detected ground truth position. The remainder of this paper is organized as follows. Section 2 highlights work related to this study. Section 3 describes the study methodology including the segmentation approaches assessed and methods of data analysis. Section 4 presents and discusses preliminary results and the study is concluded in Section 5.

2 Related work

Objective outcome measures incorporating algorithms capable of performing automated volumetric, symmetry, shape, scar visibility and projection calculations of the breast region for breast reconstruction would provide clinicians with an efficient, comprehensive and consistent method of evaluation. Implementation of automated detection of prominent points on 3D torso surface images will enable measurements carried out by software to be more usable, efficient and applicable for clinical use. Identifying the location of nipples on a 3D surface mesh will permit symmetry measurements to be performed such as the distance between the left and right nipple, which can be used to objectively assess the cosmetic outcome of a patient. Automatically identifying landmarks on 3D surface images has the potential to enable a patient-friendly method of gathering reliable and accurate anthropometric measurements by reducing intra and inter-observer variability.

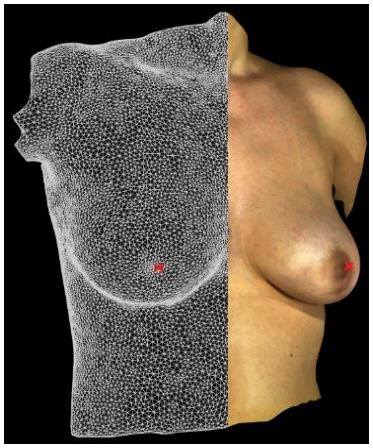


Figure 1: Nipples identified on 3D mesh wireframe and texture.

Merchant *et* al. [22] proposed algorithms to identify the nipples, sternal notch and umbilicus in 3D surface images using surface curvature and 2D texture data. Automatic segmentation of the 3D torso surface image into relative regions of interest regarding the typical location of each anatomical landmark was applied prior to curvature analysis. Gaussian and mean curvature information of the 3D surface enabled landmarks to be detected by searching the 3D image for curvature measurements correlating to the typical features they exhibit. Nipples are commonly located at the peak of the breast mounds, which are convex in shape and exhibit high elliptic Gaussian curvature. To determine an initial estimate for the location of the nipple Merchant *et* al. [22] computed a sum of the Gaussian and mean curvature values and corresponding z-value for each vertex on the 3D surface mesh. The largest value calculated was selected as the initial estimate as nipples are typically regions of high ellipticity and convexity with a high value along the z-axis due to being outwardly projected. 400 vertices surrounding the initial estimate were identified by traversing the mesh and selecting vertices in 1-ring neighborhoods. The colour map of the selected vertices was converted to greyscale before a thresholding operation produced binary pixels. The maximum and minimum intensity values in the greyscale map were retrieved pixels with intensity values less than minimum + 0.1 (maximum – minimum) were

 ISBN 978-0-9934207-1-9

assigned a binary value of 1, and remaining pixels assigned a value of 0. This thresholding procedure permits the selection of points within 10% of the total contrast of the selected sub-region as the nipple areola complex has been found to contain a significantly higher percentage of melanin than the surrounding breast skin making it more pigmented [23]. The centroid of the resulting binarized region can then be computed and mapped back onto the 3D mesh surface, determining the final nipple location [22]. The algorithms were validated by comparing the automatically detected co-ordinates with those manually detected. The study concluded that the landmarks outlined were reliably identified and curvature analysis of 3D surface images is an appropriate technique for determining properties of the breasts such as symmetry and projection which contribute to evaluating breast aesthetics [22].

Detection of nipples on two-dimensional images of female breast was investigated as an approach for adult content recognition by Wang *et* al. [24]. This study opted to use R, G, B colour information in their nipple detection procedure as the nipple skin regularly contains more R (Red) component and less G (Green) component compared to the non-nipple skin which regularly contains less R and more G components. This prior knowledge of colour model composition was used in the two-stage nipple detection algorithm after a canny edge filter was applied. The method proposed by Wang *et* al. [24] was tested on a database of 980 images and the experimental results shown the algorithm to be efficient and accurate when detecting nipples on 2D images of various subject positioning.

Related work has previously focused on implementing automated quantitative analysis of breast morphology in BCCT.core objective software. A semi-automated 2D breast contour detection technique initially developed by Cardoso and Cardoso [25] required manual selection of two breast contour endpoints before the algorithm automatically detected the contour in-between using a shortest path approach. Cardoso *et* al. [26] presented an improved algorithm capable of automatically detecting the endpoints, achieving a fully automatic method of breast contour detection. However the quality of the contour detected was shown to be dependent on position of the subject as the endpoints are located where the contour of the arm intersects the trunk contour and if these are overlapped for instance if the subject places their arms down by their sides then accurate positioning of the endpoints is hindered [26].

3 Methodology

This study investigated modified versions of the approach introduced by Merchant *et* al. [22] for automated detection of the nipples on 3D torso images. In this paper various modified thresholding and additional clustering techniques will be applied to both greyscale and colour 2D texture data of the 3D surface mesh after the initial estimate of the left and right nipple and neighboring vertices have been selected. The position of each automatically identified nipple will be compared to the position of the ground truth, which is manually selected on the 3D torso mesh.

3.1 Experiments

OpenFlipper is an open source, multi-platform application and programming framework which was used for development of algorithms [27].

As the nipples are located at the peak of the breast mound on each breast, the 3D mesh is divided into regions of interest (ROI) in order to perform the search in the area of the mesh where the nipples are likely to be located. Vertices of the mesh are firstly iterated to find the maximum and minimum x and y co-ordinate values on the mesh and using these values the mesh is then split into two halves, left and right each containing one breast. The ROI also takes into account the location of the breasts on the y-axis and focuses on the middle region of the torso. After the ROI's are defined, the vertices in each are searched to find the vertex with the greatest z coordinate which is selected as the initial estimate for the nipple. The uv 2D texture coordinates that are used to map the texture onto the 3D mesh are retrieved for neighboring vertices of the initial estimate and used to calculate the x and y position of each selected vertex on the 2D texture image using Equations (1) and (2).

$$x = u * texture\_width \quad (1)$$

$$y = texture\_height - (v * texture\_height) \quad (2)$$

 ISBN 978-0-9934207-1-9

Various thresholding and clustering techniques are then applied to both greyscale and colour 2D texture data of the selected vertices on the 3D surface mesh in effort to segment the nipple from the surrounding areola and skin. The centroid of the segmented region or cluster is computed as an x, y location on the 2D texture, which is then remapped onto the 3D mesh and selected as the final nipple location.

The following sections will describe in detail the procedure used in each experiment to segment the nipple from surrounding skin using 2D texture data to locate the position of the nipple on the 3D surface mesh.

3.1.1 10% thresholding

Thresholding is an approach used to segment an image by taking a greyscale image as input and setting pixels with intensity values higher than a threshold to 0 and all remaining pixels to 1; producing a binary image. This approach to thresholding used in the study by Merchant *et* al. [22] allows the selection of points that are within 10% of the total contrast of the sub-region containing neighboring vertices of the initial nipple estimate. The image texture file belonging to the 3D mesh currently loaded is retrieved by the software and converted to greyscale. The pixel intensity values of selected vertices are searched to find the maximum and minimum. These values are then used in Equations (3) and (4) to calculate a threshold value.

$$threshold\_left = min\_left + 0.1\ (max\_left - min\_left) \quad (3)$$

$$threshold\_right = min\_right + 0.1\ (max\_right - min\_right) \quad (4)$$

Pixels less than the threshold value are assigned a value of 1 and the remaining pixels a value of 0. The centroid of the pixels assigned a value of 1 is then calculated and remapped onto the 3D mesh as the final nipple location.

3.1.2 20% thresholding

The thresholding applied here follows the same approach as above, however allows the selection of points within 20% of the total intensity of the sub-region. The pixel intensity values of selected vertices are searched to find the maximum and minimum. These values are then used in Equations (5) and (6) to calculate a threshold value.

$$threshold\_left = min\_left + 0.2\ (max\_left - min\_left) \quad (5)$$

$$threshold\_right = min\_right + 0.2\ (max\_right - min\_right) \quad (6)$$

Once again pixels less than the threshold value are assigned a value of 1 and the remaining pixels a value of 0 and the centroid of the pixels assigned a value of 1 position is determined before being remapped back onto the 3D mesh.

3.1.3 Otsu automated thresholding

Rather than simply setting a threshold based on pixels with intensities within a certain percentage of the total contrast as above an automated approach to thresholding can be used. Otsu's automated thresholding method assumes that the extracted texture data of the selected sub-region of neighboring vertices contains two classes of pixels, the nipple and surrounding areola. The image data is first converted to greyscale and a histogram is created of the pixel intensities. The optimum threshold of the bimodal histogram is calculated by Otsu and the pixel intensities of selected vertices are checked. If the intensity is less than the optimum threshold the pixel is assigned a value of 1 and if the intensity is greater a value of 0. The centroid is located and remapped back onto the mesh as before.

3.1.4 K-Means clustering RGB colour model

Figure 2: Clustered RGB image.

K-Means clustering is one of the simplest unsupervised learning algorithms [28]. Image segmentation using k-means assigns labels to each pixel in the 2D image texture based on their RGB colour values. Pixels of similar colour formation are assigned the same label and therefore belong to the same cluster. With reference to our data it is hoped that pixels belonging to the nipple will belong to a different cluster than pixels from the surrounding nipple areola complex. The image texture file belonging to the 3D mesh currently loaded is retrieved by the software and mapped into samples, each dataset of the sample consists of a RGB pixel group. The number of clusters required has been empirically selected as 5 to permit adequate segmentation of the nipple from surrounding areola and skin. K-Means is then executed on the texture image and the centers of the computed clusters mapped onto a clustered image as shown in Figure 2. The texture location of the initial nipple estimate's neighboring vertices on the clustered image is accessed and the smallest cluster determined. The centroid of the smallest cluster is remapped onto the mesh as the final nipple position.

3.1.5 K-Means clustering HSV colour model

K-means clustering is carried out using the HSV colour space for this experiment. The HSV colour space is often preferred over the RGB colour model in image segmentation as it separates colour information and intensity [29]. The image texture file is converted into the HSV colour space and mapped into samples, each dataset of the sample consists of a HSV pixel group. The number of clusters required has been empirically selected as 5 to permit adequate segmentation of the nipple from surrounding areola and skin. K-Means is then executed on the texture image and the centers of the computed clusters mapped to onto a clustered image. Once again the smallest cluster in the sub-region is determined and it's centroid of the smallest cluster is remapped onto the mesh as the final nipple position.

3.2 Data analysis

The methodology proposed in this paper for this initial study was assessed on a dataset of four 3D frontal torso surface meshes to gather preliminary results. The performance of each experiment was evaluated by comparing the 3D coordinates of the automatically detected nipples and the manual placement of each nipple in the software, which was used as the ground truth. To compare the position of the nipples located automatically and manually, the euclidean distance (mm) between the automatically detected coordinates and those manually marked on software was calculated using Equation (7)

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{7}$$

where $x_1, y_1, z_1$are the vertex coordinates of the automatically detected nipple and x_2, y_2, z_2 are the vertex coordinates of the manually located nipple.

The ground truth nipple position was determined for each mesh by manually marking the left and right nipple on the surface mesh. The vertex position of each manually located nipple is then stored to be used as in each experiment to calculate the distance between the ground truth and the automated position of each nipple.

 ISBN 978-0-9934207-1-9

4 Preliminary results and discussion

Left and right nipples were automatically detected on four 3D surface meshes by performing the procedures outlined in Section 3. The euclidean distance (mm) between the automatically detected nipple and the manual ground truth was determined for each procedure. The distances calculated for each of the 4 meshes and average for each approach are presented in Table 1.

Table 1: Distances (mm) between the automated and ground truth positions for the left (NL) and right nipple (NR)

Approach	Mesh #1		Mesh #2		Mesh #3		Mesh #4		Average of 4 Meshes (mm)	
	NL	NR	NL	NR	NL	NR	NL	NR	NL	NR
10% threshold	4.84	4.79	6.68	0.00	7.07	5.00	6.38	3.17	6.24	3.24
20% threshold	2.28	5.80	4.63	0.00	7.07	2.91	2.68	2.87	4.16	2.90
K-Means (RGB)	2.28	4.79	4.63	0.00	7.07	2.91	2.68	0.00	4.16	1.92
K-Means (HSV)	4.84	3.88	18.40	12.30	25.62	21.54	5.59	2.87	13.61	10.15
Otsu threshold	4.84	23.51	15.76	0.00	25.87	2.91	26.62	24.75	18.27	12.79

The initial results for K-Means (RGB) clustering were promising with the algorithm performing more accurate nipple detection than the 10% thresholding method initially proposed by Merchant *et* al. [22] on five out of the total eight occasions and matching the accuracy of 10% thresholding on the remaining three. Position of the automatically detected right nipple for Mesh #2 and #4 (Figure 3(c)) equaled the manual ground truth position using the K-Means (RGB) approach, where the 10% thresholding algorithm resulted in less accurate detection of the right nipple for Mesh #4. The proposed K-Means (RGB) clustering approach may have produced more accurate preliminary results than the method suggested by Merchant *et* al. [22] because of the extra information at pixel level contained in the RGB colour model compared to the greyscale image data used in the 10% thresholding algorithm. The 20% thresholding results were encouraging for the left nipple, however didn't match the accuracy of the K-Means (RGB) clustering method for detecting the right nipple which produced the lowest average of 1.92mm. K-Means (HSV) performed well on Mesh #1 and #4, generating lower euclidean distances than the 10% thresholding approach on 4 occasions, however poor performance on Mesh #2 and #3 resulted in a high average. The Otsu automated thresholding approach proved to be the least accurate at detecting both nipples for all meshes in our initial dataset. This may due to Otsu's method of thresholding relying on bimodal histogram to select an optimum threshold between two peaks representing the foreground and background, which may not have been the case for the texture image data of the initial dataset [30]. Variation between distances calculated for the left and right nipple was present in all 4 meshes, this may have been caused by changes in lighting conditions over the torso region impairing the quality of the texture image. The variation in lighting conditions of surface meshes is to be expected therefore further development of algorithms should aim to perform accurate nipple detection taking this into consideration.

Figure 3: Automatically detected nipples (red cross) and manual ground truth (blue square) after executing (a) 10% thresholding (b) Otsu (c) K-Means (RGB) on Mesh #4.

 ISBN 978-0-9934207-1-9

5 Conclusion

This paper presents various segmentation techniques using both greyscale and colour image data combined with 3D surface meshes to distinguish the nipple from surrounding areola and skin on 3D torso surface images. Segmentation using colour information from the 2D texture data of the 3D mesh presented promising initial results with K-Means (RGB) performing the most accurate detection on average for the right nipple when compared to the other approaches, however as this study was carried out using a small initial dataset the robustness of all algorithms could not be assessed completely. There is clear potential for the 20% thresholding which matched the accuracy of the K-Means (RGB) algorithm when detecting the left nipple, with both approaches obtaining the lowest average. The K-Means (HSV) clustering method produced very good results for some meshes proving to be more accurate than Otsu automated thresholding method. Future work will focus on improving segmentation algorithms by performing clustering on the sub-region of the initial estimate instead of the entire texture and ensuring algorithms are able to perform accurate nipple detection on meshes with varying lighting conditions.

References

[1] Northern Ireland Cancer Registry (NICR), "Breast cancer," 2013. [Online]. Available: http://www.qub.ac.uk/research-centres/nicr/FileStore/PDF/FactSheets/Filetoupload,629853,en.pdf. [Accessed: 22-Mar-2016].

[2] Cancer Research UK, "Breast Cancer Mortality." [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Two. [Accessed: 01-Nov-2015].

[3] D. A. Hudson, "Factors determining shape and symmetry in immediate breast reconstruction.," *Ann. Plast. Surg.*, vol. 52, no. 1, pp. 15–21, Jan. 2004.

[4] J. Sabczynski, H. Barschdorf, T. Bülow, M. J. Cardoso, J. S. Cardoso, A. Clarke, B. Eiben, P. Gouveia, D. Hawkes, J. Hipwell, M. Keshtgar, R. Lacher, D. Kutra, G.-J. Liefers, K. Meetz, B. Molenkamp, J. P. Monteiro, A. Mosahebi, H. P. Oliveira, R. Sinkus, D. Stoyanov, V. Vavourakis, C. Jh Van De Velde, N. Williams, S. Young, and H. Zolfagharnasab, "PICTURE: Predicting the cosmetic outcome of breast cancer surgery."

[5] M. J. Cardoso, H. Oliveira, and J. Cardoso, "Assessing cosmetic results after breast conserving surgery.," *J. Surg. Oncol.*, vol. 110, no. 1, pp. 37–44, 2014.

[6] D. R. H. Christie, M. Y. O'Brien, J. A. Christie, T. Kron, S. A. Ferguson, C. S. Hamilton, and J. W. Denham, "A comparison of methods of cosmetic assessment in breast conservation treatment," *The Breast*, vol. 5, no. 5, pp. 358–367, Oct. 1996.

[7] K. C. A. Sneeuw, N. K. Aaronson, J. R. Yarnold, M. Broderick, J. Regan, G. Ross, and A. Goddard, "Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 1. Comparison of patients' ratings, observers' ratings and objective assessments," *Radiother. Oncol.*, vol. 25, no. 3, pp. 153–159, Nov. 1992.

[8] C. Vrieling, L. Collette, E. Bartelink, J. H. Borger, S. J. Brenninkmeyer, J. C. Horiot, M. Pierart, P. M. Poortmans, H. Struikmans, E. Van Der Schueren, J. a. Van Dongen, E. Van Limbergen, and H. Bartelink, "Validation of the methods of cosmetic assessment after breast-conserving therapy in the EORTC 'boost versus no boost' trial," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 45, no. 3, pp. 667–676, 1999.

[9] M. H. Haloua, N. M. A. Krekel, G. J. A. Jacobs, B. Zonderhuis, M. Bouman, M. E. Buncamper, F. B. Niessen, H. A. H. Winters, C. Terwee, S. Meijer, and M. P. van den Tol, "Cosmetic Outcome Assessment following Breast-Conserving Therapy: A Comparison between BCCT.core Software and Panel Evaluation," *Int. J. Breast Cancer*, vol. 2014, pp. 1–7, 2014.

[10] M. S. Kim, J. C. Sbalchiero, G. P. Reece, M. J. Miller, E. K. Beahm, and M. K. Markey, "Assessment of breast aesthetics.," *Plast. Reconstr. Surg.*, vol. 121, no. 4, p. 186e–94e, 2008.

[11] M. J. Cardoso, J. Cardoso, A. C. Santos, H. Barros, and M. C. de Oliveira, "Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer," *The Breast*, vol. 15, no. 1, pp. 52–57, 2006.

[12] H. Henseler, J. Smith, A. Bowman, B. S. Khambay, X. Ju, A. Ayoub, and A. K. Ray, "Subjective versus objective assessment of breast reconstruction," *J. Plast. Reconstr. Aesthetic Surg.*, vol. 66, no. 5, pp. 634–639, 2013.

[13] K. Aldridge, S. A. Boyadjiev, G. T. Capone, V. B. DeLeon, and J. T. Richtsmeier, "Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogrammetric images," *Am. J. Med. Genet. Part A*, vol. 138A, no. 3, pp. 247–253, 2005.

[14] T. Catherwood, E. McCaughan, E. Greer, R. A. J. Spence, S. A. McIntosh, and R. J. Winder, "Validation of a passive stereophotogrammetry system for imaging of the breast: A geometric analysis," *Med. Eng. Phys.*, vol. 33, no. 8, pp. 900–905, 2011.

[15] D. L. Esme, A. Bucksch, and W. H. Beekman, "Three-Dimensional Laser Imaging as a Valuable Tool for Specifying Changes in Breast Shape After Augmentation Mammaplasty," *Aesthetic Plast. Surg.*, vol. 33, no. 2, pp. 191–195, 2009.

[16] H. Henseler, J. Smith, A. Bowman, B. S. Khambay, X. Ju, A. Ayoub, and A. K. Ray, "Investigation into variation and errors of a three-dimensional breast imaging system using multiple stereo cameras," *J. Plast. Reconstr. Aesthetic Surg.*, vol. 65, no. 12, pp. e332–e337, 2012.

[17] L. Kovacs, M. Eder, R. Hollweck, A. Zimmermann, M. Settles, A. Schneider, M. Endlich, A. Mueller, K. Schwenzer-Zimmerer, N. a. Papadopulos, and E. Biemer, "Comparison between breast volume measurement using 3D surface imaging and classical techniques," *Breast*, vol. 16, no. 2, pp. 137–145, 2007.

[18] M. S. Kim, G. P. Reece, E. K. Beahm, M. J. Miller, E. Neely Atkinson, and M. K. Markey, "Objective assessment of aesthetic outcomes of breast cancer treatment: Measuring ptosis from clinical photographs," *Comput. Biol. Med.*, vol. 37, no. 1, pp. 49–59, 2007.

[19] M. J. Cardoso, J. Cardoso, N. Amaral, I. Azevedo, L. Barreau, M. Bernardo, D. Christie, S. Costa, F. Fitzal, J. L. Fougo, J. Johansen, D. Macmillan, M. P. Mano, L. Regolo, J. Rosa, L. Teixeira, and C. Vrieling, "Turning subjective into objective: The BCCT.core software for evaluation of cosmetic results in breast cancer conservative treatment," *The Breast*, vol. 16, no. 5, pp. 456–461, 2007.

[20] "Axis Three for Breast Simulations," 2016. [Online]. Available: http://www.axisthree.com/products/breast-surgery-simulation. [Accessed: 24-May-2016].

[21] M. Kawale, J. Lee, S. Y. Leung, M. C. Fingeret, G. P. Reece, M. A. Crosby, E. K. Beahm, M. K. Markey, and F. A. Merchant, "3D Symmetry measure invariant to subject pose during image acquisition," *Breast Cancer Basic Clin. Res.*, vol. 5, no. 1, pp. 131–142, 2011.

[22] F. Merchant, M. Kawale, G. Reece, M. Crosby, E. Beahm, M. Fingeret, and M. Markey, "Automated Identification of Fiducial Points on 3D Torso Images," *Biomed. Eng. Comput. Biol.*, p. 57, 2013.

[23] N. Dean, J. Haynes, J. Brennan, T. Neild, C. Goddard, B. Dearman, and R. Cooter, "Nipple-areolar pigmentation: Histology and potential for reconstitution in breast reconstruction," *Br. J. Plast. Surg.*, vol. 58, no. 2, pp. 202–208, 2005.

[24] Y. Wang, J. Li, H. Wang, and Z. Hou, "Automatic nipple detection using shape and statistical skin color information," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5916 LNCS, pp. 644–649, 2009.

[25] J. S. Cardoso and M. J. Cardoso, "Breast contour detection for the aesthetic evaluation of breast cancer conservative treatment," *Adv. Soft Comput.*, vol. 45, pp. 518–525, 2007.

[26] J. S. Cardoso and L. F. Teixeira, "Automatic breast contour detection in digital photographs," *Int. Conf. Heal. Informatics*, pp. 91–98, 2008.

[27] "OpenFlipper," 2016. [Online]. Available: www.openflipper.org. [Accessed: 24-May-2016].

[28] A. Mohanty, "Analysis of Color Images using Cluster based Segmentation Techniques," vol. 79, no. 2, pp. 42–47, 2013.

[29] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval," *Int. Conf. Image Process.*, vol. 2, pp. II–589–II–592, 2002.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

Recent techniques for (re)colouring

M. Grogan, J. V. E. Carvalho & R. Dahyot

School of Computer Science and Statistics
Trinity College Dublin
Ireland

Abstract

This paper investigates how several techniques can be used together for colouring frames in grey level sequences. A trained deep neural network is used to colour a grey level image coherently [Iizuka et al., 2016], and this colour image can be recoloured further to change its feel [Grogan et al., 2015]. When considering videos however, artifacts are created in the first step when the same semantic object can occasionally be given different colours from frame to frame in the sequence creating a flicker in the resulting coloured sequence.

Keywords: colour transfer, colouring, deep learning, flicker

We have recently proposed a new parametric framework for colour transfer to transform the colour feel of images and videos using a chosen colour palette image [Grogan et al., 2015, Grogan and Dahyot, 2015]. The parametric modelling of transfer functions allows for easy storage and interpolation between several transfer functions, for creating new colouring schemes, making the approach easily tunable by artists. Moreover using binary or fuzzy masks, spatio-temporal variations between several palettes can be applied. Figure 1 shows for instance two colour palettes images (lower right) that are used for recolouring a target image (middle left) applying a spatial mask (lower left) for mixing both colouring effects in the resulting recoloured target image (top). When available, semantic based binary masks can efficiently be used to create effects on each object in the scene, and likewise colour transfer function can be estimated between local regions as opposed to capturing global colour information in the whole target and palette images.

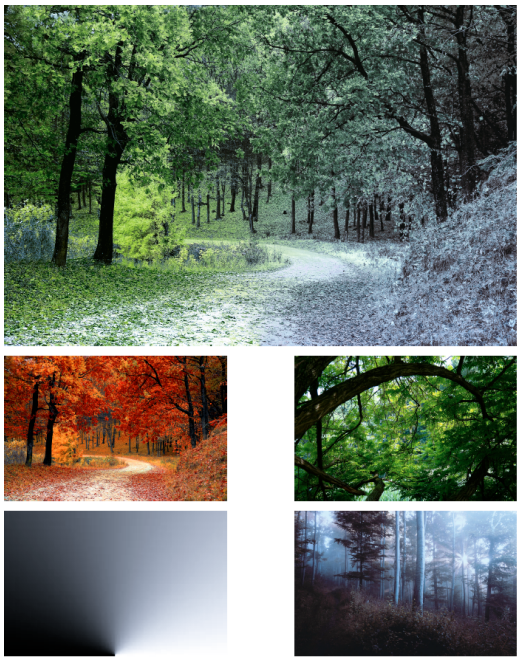


Figure 1: Colour transfer [Grogan et al., 2015, Grogan and Dahyot, 2015]: recoloured target image (top), target image (middle row left), palette images (midlle and bottom right) and mask (bottom left).

Several recent techniques based on deep learning architecture have been proposed to transfer style and colour between digital image material [Yan et al., 2016, Iizuka et al., 2016]. For instance Yan et al trained a deep neural network to replicate artist skills for editing images (e.g. effect *Foreground Pop-Out* that increases contrast and saturation of foreground salient objects, and *Watercolor* effect that mimics waterpainting style) [Yan et al., 2016]. This learning based approach allows to take into account the semantic content locally in digital images as well as their corresponding perceptual importance for manipulation by artists.

 ISBN 978-0-9934207-1-9

Similarly, Iizuka et al. learn semantic content of images and their relations to their colour information to then automatically colourize greyscale digital photographs using convolutional neural network [Iizuka et al., 2016]. Figure 2 shows some results of colouring a greyscale video using Iizuka et al.: when applied to a sequence, colouring each frame independently create a flickering effect: some area or object in the video are changing colour from frame to frame. This flickering artifact can worsen by applying other post processing steps such as colour transfer. Our current efforts aims at adapting Grogan et al [Grogan et al., 2015] colour transfer techniques for removing the occurrences of this flicker taking example of the filtering approach proposed by Pitié et al to correct flicker in old films [Pitié et al., 2004].

Figure 2: Several frames from the sequence *traffic* from CDV2014 dataset [Nuutinen et al., 2016] recolored using Iizuka algorithm [Iizuka et al., 2016] (second row), and recolored with two different palettes [Grogan et al., 2015] (rows 3 and 4).

Acknowledgments. This work has been supported by a Ussher scholarship from Trinity College Dublin (Ireland), by the Science Without Borders Programme (CAPES, Personal Improvement Coordination of Higher Education, Brazil) and partially supported by EU FP7-PEOPLE-2013- IAPP GRAISearch grant (612334).

References

[Grogan and Dahyot, 2015] Grogan, M. and Dahyot, R. (2015). L2 registration for colour transfer in videos. In *Conference on Visual Media Production*, London, UK.

[Grogan et al., 2015] Grogan, M., Prasad, M., and Dahyot, R. (2015). L2 registration for colour transfer. In *European Signal Processing Conference (Eusipco)*, Nice France.

[Iizuka et al., 2016] Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2016)*, 35(4).

[Nuutinen et al., 2016] Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., and Häkkinen, J. (2016). Cvd2014 a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086.

[Pitié et al., 2004] Pitié, F., Dahyot, R., Kelly, F., and Kokaram, A. (2004). A new robust technique for stabilizing brightness fluctuations in image sequences. In *Workshop on Statistical Methods in Video Processing (in conjunction to ECCV)*, Prague, Czech Republic. Springer, ISBN: 3-540-23989-8.

[Yan et al., 2016] Yan, Z., Zhang, H., Wang, B., Paris, S., and Yu, Y. (2016). Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 35(2):11:1–11:15.

Analysis of variable-order interacting multiple model algorithms for cell tracking

K. Lomanov[1,2], J. Martínez del Rincón[1], P. Miller[1], and H. Gribben[2]

[1]*The Institute of Electronics, Communications and Information Technology (ECIT) , Queen's University of Belfast*
[2]*Andor Technology, 7 Millennium Way, Belfast*

Abstract

In this paper we propose a modification of the Interacting Multiple Model (IMM) algorithm to effectively track complex dynamics in cell images. Our solution proposes a more efficient use and combination of the multiple Kalman filter estimations that lead to a performance improvement in multi-cell sequences, with an increase of up to 10% in the recall value, compared to the classic IMM. First and second order models are evaluated in the scope of cell migration. The system is evaluated and compared against a baseline using 3D synthetic confocal microscopy images, where cells behave realistically according to actual cell trajectories extracted from real sequences in biology.

Keywords: Live Cell Tracking, Particle Tracking, IMM Algorithm, Second-Order Markov Chains

1 Introduction

Progress in medicine significantly depends on developing a deeper understanding of cell movements and cell interactions. Recent developments in stochastic and deterministic super-resolution microscopy techniques yield images with a resolution below the diffraction limit [Huang et al., 2009], giving biologists the technology to observe processes at the nanometre scale in real time, which was not possible before. In spite of all the advantages that new technologies have brought, some challenges associated with them have also arisen, such as the large amount of raw data which needs to be processed to extract precise information, obtain quantitative characterizations of the observed phenomena, and draw meaningful conclusions [Meijering et al., 2006]. Thus, automated acquisition and analysis of cell images has become more and more essential for biomedical research over the past 15 years [Peng, 2008].

Among the features that can be extracted and analysed automatically, the migration and mobility of a cell in a 3D environment is crucial for understanding an immune response and wound healing [Pivarcsi et al., 2004, Gurtner et al., 2008]. Automatic tracking of moving cells can also be used to perform the required continuous adjustments needed to keep the objects of interest within the imaging field and in focus. However, multiple cell tracking is a complex task that is far from being solved, given the variety of behaviours and interactions that cells can express in different sequences, including migration, crawling, splitting, and phagocytosis, to name a few. This diversity requires the use of one or multiple complex motion models that must be selected and combined to ensure an accurate and robust tracking.

In this paper, we investigate the use of tracking algorithms to extract trajectories from multiple cells moving in a 3D environment. In particular, the use of the Interacting Multiple Model (IMM) algorithm is proposed given its ability to combine different motion models at every given time. A novel approach to the combination of multiple model estimations is proposed, which surpasses the traditional approach.

 ISBN 978-0-9934207-1-9

Given the limited amount of real data obtained via confocal microscopy, the system is evaluated using realistic 3D synthetic images. The trajectories of the moving objects correspond to cell trajectories extracted from real sequences in biology [de Solórzano et al., 2015] to ensure natural behaviours and interactions [Wilson et al., 2016].

2 State of the Art

Traditionally, cell detection, segmentation, and tracking used to be performed manually (by pointing and clicking the objects of interest on each frame). However, several reasons make such a task tedious, or even impossible. First, the datasets are now so large [Meijering et al., 2006, Peng, 2008], that manually processing them would take days, and selecting smaller subsets means losing relevant information and taking biased decisions. Second, manually determining the centroids of the cells is a user-dependent measurement, and as such is particularly error prone. Over the last years, a large research effort in computer science has been directed at developing effective automatic tracking algorithms.

Until recently, deterministic approaches were mostly used for cell tracking [Meijering et al., 2006]. These approaches consist of detecting the cells on each frame, and then linking them to form tracks through data association. As a consequence, these methods highly depend on the performance of the segmentation algorithm. While tracking by detection works effectively in other related fields, such as video surveillance, it struggles in biological applications due to the diverse and generally poor quality of biomedical and cell images and their low signal to noise ratio. These approaches are also impacted by the lack of specific and reliable cell detectors, therefore relying on general purpose segmentation such as simple thresholding or slightly more complex methods such as the watershed transform [Meijering et al., 2006] or wavelet transform [Genovesio et al., 2006].

As an alternative, a significant number of algorithms using a probabilistic approach, known as Bayesian tracking, have been proposed. The basic principle is to infer the current state using the observation and the previous states. These probabilistic approaches show better results [Jaqaman et al., 2008], especially when frequent segmentation errors are expected. The Kalman filter [Kalman, 1960] is a common approach, which is optimal in the case of Gaussian distributions. However, this assumption is not correct in cell motion, in particular for multi-target problems. The particle filter [Doucet and Johansen, 2011], which is based on the same principle, deals with non Gaussian and non linear cases, that are biologically relevant. However, it suffers from high complexity and particle degeneration that also imply the use of specific assumptions. As an intermediate solution, the IMM algorithm was presented in [Genovesio et al., 2006] as a novel method for tracking multiple microscopic objects in 3D space, in real-time. This solution combines several Kalman filters with different dynamic models to quickly adapt to changes of state. It is nonetheless difficult to choose relevant models for biological processes as well as determining the optimal combination of those models to deal with complex dynamics in multi-target interacting scenarios. Although Brownian motion is usually chosen for describing cell motion, this is an over-simplification of their motion patterns [Codling et al., 2008, Selmeczi et al., 2008].

3 Interacting Multiple Model algorithm

The IMM algorithm is a Bayesian iterative algorithm that uses a combination of several Kalman filters, which allow estimating the state X_k of a tracked object at a given time step k as a combination of several dynamical models' estimations.

Assuming N dynamic models are being considered to model the most frequent behaviours of the tracked objects, each corresponding dynamic model can be expressed as a transition matrix D^j for $j \in [\![1, N]\!]$, following conventional notation. Thus, given the estimated state $\hat{X}_k$ and uncertainty $\hat{C}_k$ for each tracked object at the previous time step, N predictions for the new time step $k+1$, as well as their corresponding uncertainties C^j, are calculated using the equations:

$$X^j_{\text{pred},k+1} = D^j \tilde{X}^j_k \tag{1}$$

 ISBN 978-0-9934207-1-9

$$C^j_{\text{pred},k+1} = D^j \tilde{C}^j_k {D^j}^T + Q_j \tag{2}$$

where Q_j is the process noise covariance, and $\tilde{X}^j_k$ and $\tilde{C}^j_k$ are the mixed state and covariance, such that:

$$\tilde{X}^j_k = \sum_{i=1}^{N} u^{i|j}_k \hat{X}^i_k \tag{3}$$

$$\tilde{C}^j_k = \sum_{i=1}^{N} u^{i|j}_k \left[\hat{C}^i_k + \left(\hat{X}^i_k - \tilde{X}^i_k \right) \left(\hat{X}^i_k - \tilde{X}^i_k \right)^T \right] \tag{4}$$

where $u^{i|j}_k = \dfrac{1}{u^j_{\text{pred},k+1}} p_{ij} u^i_k$ is the conditional model probability, with $u^j_{\text{pred},k+1} = \sum_{i=1}^{N} p_{ij} u^i_k$ the predicted model probability, and p_{ij} the probability to switch from the i^{th} to the j^{th} model between k and $k+1$.

Each of these predictions should be compared at present time $k+1$ against the observation Z_{k+1} provided by the segmentation algorithm to validate the adequacy of the chosen motion model to the current target motion.

$$\hat{X}^j_{k+1} = X^j_{\text{pred},k+1} + G^j_{k+1} \left(Z_{k+1} - H X^j_{\text{pred},k+1} \right) \tag{5}$$

$$\hat{C}^j_{k+1} = C^j_{\text{pred},k+1} \left(I - G^j_{k+1} H \right) \tag{6}$$

I being the identity matrix, H the observation matrix and G the Kalman gain such that

$$G^j_{k+1} = C^j_{\text{pred},k+1} H^T \left(H C^j_{\text{pred},k+1} H^T + R \right)^{-1} \tag{7}$$

where R is the measurement noise covariance.

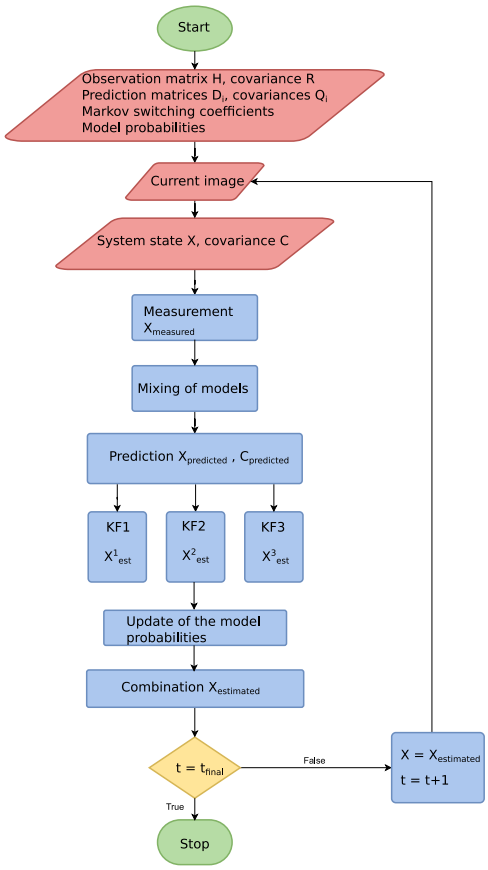


Figure 1: The IMM algorithm.

Since none of the used models is likely to provide a perfect match and the behaviour of the tracked object can be better explained as a combination of them all, the IMM computes the final combined estimated state and covariance as a weighted average of the individual estimates:

$$\hat{X}_{k+1} = \sum_{j=1}^{N} u^j_{k+1} \hat{X}^j_{k+1} \tag{8}$$

$$\hat{C}_{k+1} = \sum_{j=1}^{N} u^j_{k+1} \left[\hat{C}^j_{k+1} + \left(\hat{X}^j_{k+1} - \hat{X}_{k+1} \right) \left(\hat{X}^j_{k+1} - \hat{X}_{k+1} \right)^T \right] \tag{9}$$

A flow diagram of the algorithm is shown in Figure 1.

3.1 Model probabilities

The weights u^j_{k+1} used in equations 8 and 9 are normalised factors proportional to the likelihood λ^j_k of how well each dynamic model j fits the observation at each time step k:

$$u^j_{k+1} = \frac{1}{\sum_{i=1}^{N} u^i_{\text{pred},k+1}\lambda^i_{k+1}} u^j_{\text{pred},k+1}\lambda^j_{k+1} \tag{10}$$

where λ^j_{k+1} is the likelihood of the filter j matched to the model j and is computed as follows:

$$\lambda^j_{k+1} = \frac{1}{\sqrt{\det\left(2\pi S^j_{k+1}\right)}} \exp\left[-\frac{1}{2}\left(Z_{k+1} - HX^j_{\text{pred},k+1}\right)^T \left(S^j_{k+1}\right)^{-1}\left(Z_{k+1} - HX^j_{\text{pred},k+1}\right)\right] \tag{11}$$

with S^j_{k+1} defined as the covariance of the innovation of the j-th Kalman filter.

The use of these model probabilities allows the IMM to effectively track an object if its movement approximately matched any of the dynamic models or combination of them. Since these weights are recalculated at each time step k, the behaviour of the tracked object can vary over time from one dynamic model or behaviour to another while still being effectively tracked by the IMM.

3.2 Second-order Markov chain based IMM

While the weighting process described in the previous section is effective, non-accurate estimations may happen due to noisy observations and wrong model choices. In order to filter some of those errors, in this section we employ an improved calculation of the model probabilities based on the assumption that the current model combination depends on the two previous time steps. This process is coherent with cell migration, whose motion and behaviour tend to be consistent over short periods of time, once the chosen cell mechanism has been initiated.

An efficient way to integrate second-order information was designed in [Lan et al., 2013] for manoeuvring target tracking applications. We applied this method, called SIMM (Second-order Markov chain based IMM) algorithm, to biological data for the first time. The difference with the classic IMM, based on a first order Markov chain, is that before the interaction step given by the equations 3 and 4, there is an additional step that consists of updating the switching probabilities p^k_{ij}, following the equations below:

$$p^k_{ij} = \sum_{l=1}^{N} p_{j|l,i} \times \mu^k_{i,l} \tag{12}$$

where $p_{j|l,i}$ are the transition probabilities of the second-order Markov chains and

$$\mu^k_{i,l} = \frac{1}{\sum_{q=1}^{N} u^{q|i}_{k-1}\lambda^{i|q}_k} u^{l|i}_{k-1}\lambda^{i|l}_k \tag{13}$$

where $\lambda^{i|l}_k$ is the past likelihood of the filter l matched to the model i:

$$\lambda^{i|l}_k = \frac{1}{\sqrt{\det\left(2\pi S^{i|l}_k\right)}} \exp\left[-\frac{1}{2}\left(Z_k - HD^i\widehat{X}^l_{k-1}\right)^T \left(S^{i|l}_k\right)^{-1}\left(Z_k - HD^i\widehat{X}^l_{k-1}\right)\right] \tag{14}$$

with $S^{i|l}_k$ defined as the covariance of the innovation of the i-th Kalman filter using the estimation from the l-th Kalman filter.

 ISBN 978-0-9934207-1-9

3.3 Hard estimation of combined state

Finally we propose a novel modification to both the IMM and SIMM. While previous approaches estimate the best possible combination of model at each time step, this estimated state $\hat{X}_k$ and covariance $\hat{C}_k$ are not directly used in the next prediction but replaced by the mixed state and covariance $\tilde{X}_k$ and $\tilde{C}_k$. We hypothesize that since this mixed variable relies on p_{ij} and $p_{l,i,j}$, which are manually or empirically chosen and fixed for every sequence, this decision may not provide a better reference for each individual predictions than the agreed previous estimation. Therefore, in our modified versions, we replace equations 1 and 2 by:

$$X^i_{\text{pred},k+1} = D^i \widehat{X}_k \tag{15}$$

$$C^i_{\text{pred},k+1} = D^i \widehat{C}_k {D^i}^T + Q_i \tag{16}$$

We will refer to these algorithms as the modified IMM and modified SIMM.

3.4 Data association

Since our application aims to track multiple objects, data association between each of the M tracked objects $\{\widehat{X}_{k+1}\}_m$ for $m \in [\![1, M]\!]$ and the observations $\{Z_{k+1}\}_m$ must be solved to ensure a correct allocation of observation and predictions and a correct object tracking without identity swapping.

Following a common strategy in biology [Genovesio et al., 2006], we make use of a greedy linear assignment association in this paper.

4 Experiments

4.1 Algorithm setup

In order to track cells in 3D confocal microscopy, the (x, y, z) positions of their centroids must be included as variables in the state vector. Given the need of modelling first and second order motion models to explain cell migration, velocity and acceleration vectors are also coded, resulting in the state vector:

$$X_k = (x_k, y_k, z_k, x_{k-1}, y_{k-1}, z_{k-1}, x_{k-2}, y_{k-2}, z_{k-2})^T$$

Three different dynamical models are chosen for our IMM implementation: constant velocity (CV), constant acceleration (CA) and Brownian motion (BM). All three models are common behaviours for biological objects [Genovesio et al., 2006], such as cells or even smaller particles (organelles, viruses, etc.). We represent the CV model by a linear extrapolation of the locations, the CA by a linear extrapolation of the velocities, and the BM by adding Gaussian noise to the previous state. They are coded as the following transition matrices:

$$D_{\text{CV}} = \begin{pmatrix} 2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, D_{\text{CA}} = \begin{pmatrix} 3 & 0 & 0 & -3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 & -3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & -3 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, D_{\text{BM}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Process and measurement noise covariances Q_i and R are chosen with the following values in all our experiments: $Q_{\text{CV}} = I_9$, $Q_{\text{CA}} = I_9$, $Q_{\text{BM}} = \text{diag}_9(1000)$, $R = \text{diag}_3(400)$.

 ISBN 978-0-9934207-1-9

4.2 Dataset

Given the limited amount of real data from confocal microscopy, our proposed systems and baseline are evaluated using 3D synthetic images. The images are generated using a simulator where cells are represented by spheres. For a realistic result, the image is first convolved with a point spread function (PSF) and then Poisson noise is applied by generating each output pixel from a Poisson distribution with a mean value equal to the input pixel intensity. Using such a simulator allows isolating the tracking from the segmentation process in order to better evaluate each part separately, as well as to evaluate the tracking robustness against different levels of signal to noise ratio. The segmentation is a simple thresholding of value $t = 0.5$. An example of the generated sequences is depicted in Figure 2.

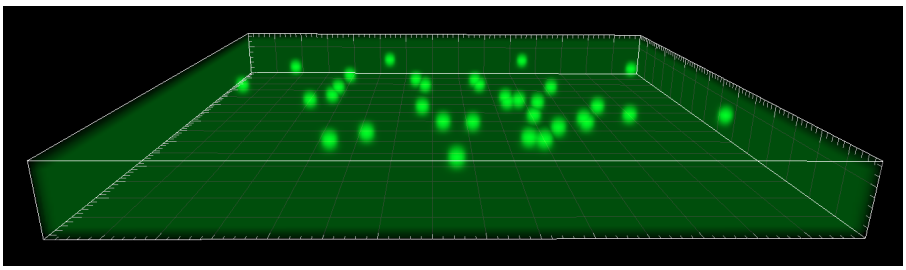


Figure 2: 3D synthetic image of 33 cells, rendered in Imaris

Given the importance of evaluating our tracking algorithms against realistic cell migration behaviours, the trajectories of moving HeLA cells are imported from real sequences [de Solórzano et al., 2015] into our generator.

Twelve sequences of increasing complexities, with the number of targets increasing from 8 to 50, are used for testing. Imported trajectories are used as ground truth to calculate the system's performance. Recall and precision, respectively defined in 17 and 18, are used as metrics. A true positive TP is defined for a minimum overlap $b = 0.5$ between estimation and ground truth. This overlap is computed as the ratio $b = \frac{A_{\text{overlap}}}{A_{\text{disc}}}$, where A_{overlap} is the intersection area of the two bounding circles, and A_{disc} the surface of the disc defined by one bounding circle. The bounding circles are defined as circles of the same radius r as the simulated cells, with one centred on the tracker estimated location and the other on the ground truth location. A false negative FN occurs when the overlap between estimation and ground truth is inferior to b, and a false positive FP is when there is an estimation but no corresponding ground truth. TP, FP and FN are computed as the total values over all the sequences.

$$\text{recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{18}$$

4.3 Experiments

We compute the recall and precision of all proposed algorithms for different densities (Table 1), and different noise levels (Table 2). The packing density η is the ratio of the total volume occupied by the cells to the total volume considered. For instance, a packing density of 8.0×10^{-4} corresponds here to 8 cells in the volume shown in Figure 2.

4.4 Analysis

Table 1 shows that the modification we introduced improves both the performance of the IMM and the SIMM, with a significant increase in the recall, for every packing density value. The precision is always close to 1 because the only source of false positives or distractors in our generator is due to fragmented detections, which are unlikely to happen in data without noise.

 ISBN 978-0-9934207-1-9

	IMM		Modified IMM		SIMM		Modified SIMM	
Packing density η	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
8.0×10^{-4}	0.85	0.99	0.91	1.00	0.68	1.00	0.95	1.00
1.5×10^{-3}	0.79	0.99	0.85	0.99	0.66	0.99	0.88	0.99
2.5×10^{-3}	0.79	0.98	0.85	0.98	0.68	0.98	0.89	0.98
5.0×10^{-3}	0.62	0.94	0.70	0.95	0.55	0.95	0.75	0.94

Table 1: Results of the experiment without noise

	IMM		Mod IMM		SIMM		Mod SIMM	
Max. value of noise distribution	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
50	0.78	0.53	0.84	0.51	0.63	0.55	0.91	0.46
30	0.54	0.61	0.58	0.57	0.51	0.63	0.65	0.49
10	0.03	1.00	0.03	1.00	0.03	1.00	0.03	1.00

Table 2: Results of the experiment for different noise levels, with $\eta = 8.0 \times 10^{-4}$

Compared to the IMM, the initial version of the SIMM fails to show the results that could be expected according to [Lan et al., 2013]. While this may suggest that second order combination is not suitable for the behaviour displayed by HeLa cells, which is very different from the behaviour of a manoeuvring target, it improves the results when combined with our proposed modification.

The algorithm that yields the best recall is the modified version of the SIMM. Our modification allows the algorithm to update the model probabilities quicker and take advantage of the second order chain to generate a more accurate estimated state while filtering the mistakes towards the next step predictions. The two SIMM type algorithms are also the ones that resist best when the noise increases, as we can see on Table 2, although none of the methods give acceptable results when the noise level is the highest.

5 Conclusion

In this paper, we have described a novel algorithm for multiple cell tracking, which uses an SIMM method, for the first time in biology, in combination with a modified dynamic model prediction. The data association is performed using the greedy linear assignment type. The performance of the algorithm has been tested and compared to a baseline and incremental improved versions, using synthetic data sequences generated from real cells under different density and noise level conditions. Our results show an improvement in terms of recall and precision, compared to the classic IMM and SIMM algorithms.

As future work, we aim to investigate more efficient data association methods, in order to improve the algorithm's performance in high density situations, and we will evaluate the tracker against real sequences.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642866.

References

[Codling et al., 2008] Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society Interface*, 5(25):813–834.

[de Solórzano et al., 2015] de Solórzano, C. O., Kozubek, M., Meijering, E., and Barrutia, A. M. (2015). ISBI cell tracking challenge. `http://www.codesolorzano.com/celltrackingchallenge`. Accessed: 26/05/2016.

[Doucet and Johansen, 2011] Doucet, A. and Johansen, A. M. (2011). *A tutorial on particle filtering and smoothing: fifteen years later*, chapter 24, pages 656–704. Oxford Handbook of Nonlinear Filtering. Oxford University Press.

[Genovesio et al., 2006] Genovesio, A., Liedl, T., Emiliani, V., Parak, W. J., Coppey-Moisan, M., and Olivo-Marin, J.-C. (2006). Multiple particle tracking in 3-d+t microscopy: Method and application to the tracking of endocytosed quantum dots. *IEEE Transactions on Image Processing*, 15(5):1062–1070.

[Gurtner et al., 2008] Gurtner, G. C., Werner, S., Barrandon, Y., and Longaker, M. T. (2008). Wound repair and regeneration. *Nature*, 453(7193):314–321.

[Huang et al., 2009] Huang, B., Bates, M., and Zhuang, X. (2009). Super resolution fluorescence microscopy. *Annual Review of Biochemistry*, 78:993–1016.

[Jaqaman et al., 2008] Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L., and Danuser, G. (2008). Robust single particle tracking in live cell time-lapse sequences. *Nature methods*, 5(8):695–702.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

[Lan et al., 2013] Lan, J., Li, X. R., Jilkov, V. P., and Mu, C. (2013). Second-order markov chain based multiple-model algorithm for maneuvering target tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 49(1):3–19.

[Meijering et al., 2006] Meijering, E., Smal, I., and Danuser, G. (2006). Tracking in molecular bioimaging. *IEEE Signal Processing Magazine*, 23:46–53.

[Peng, 2008] Peng, H. (2008). Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827–1836.

[Pivarcsi et al., 2004] Pivarcsi, A., Kemény, L., and Dobozy, A. (2004). Innate immune functions of the keratinocytes. *Acta Microbiologica et Immunologica Hungarica*, 51(3):303–310.

[Selmeczi et al., 2008] Selmeczi, D., Li, L., Pedersen, L. I., Nrrelykke, S., Hagedorn, P. H., Mosler, S., Larsen, N. B., Cox, E. C., and Flyvbjerg, H. (2008). Cell motility as random motion: a review. *The European Physical Journal Special Topics*, 157(1):1–15.

[Wilson et al., 2016] Wilson, R. S., Yang, L., Dun, A., Smyth, A. M., Duncan, R. R., Rickman, C., and Lu, W. (2016). Automated single particle detection and tracking for large microscopy datasets. *Royal Society Open Science*, 3(5).

 ISBN 978-0-9934207-1-9

Visual Speech Encoding based on Facial Landmark Registration

Ram P. Krish, Paul F. Whelan

Vision Systems Group, School of Electronic Engineering,
Dublin City University, Dublin, Ireland.
ram.krish@dcu.ie, paul.whelan@dcu.ie

Abstract

Visual Speech Recognition (VSR) related studies largely ignore the use of state of the art approaches in facial landmark localization, and are also deficit of robust visual features and its temporal encoding. In this work, we propose a visual speech temporal encoding by integrating state of the art fast and accurate facial landmark detection based on ensemble of regression trees learned using gradient boosting. The main contribution of this work is in proposing a fast and simple encoding of visual speech features derived from vertically symmetric point pairs (VeSPP) of facial landmarks corresponding to lip regions, and demonstrating their usefulness in temporal sequence comparisons using Dynamic Time Warping. VSR can be either speaker dependent (SD) or speaker independent (SI), and each of them poses different kind of challenges. In this work, we consider the SD scenario, and obtain 82.65% recognition accuracy on OuluVS database. Unlike recent research in VSR which makes use of auxiliary information such as audio, depth and color channels, our approach does not impose such constraints.

Keywords: Visual speech, temporal encoding, facial landmarks, dynamic time warping.

1 Introduction

Speech perception is the process by which the sounds of language is *heard, interpreted* and *understood*. The interpreting aspect also includes focusing on visual cues of the speech. The interactions between acoustic and visual information in speech perception was shown by McGurk, the phenomenon being popularly known as *McGurk Effect* [McGurk and MacDonald, 1976]. People with better sensory integration are more susceptible to McGurk effect. Visual cues generally used by humans for speech perception constitute *lip-motion, head movements, facial expressions, body gestures, language structures, contexts, etc.* Such a process is referred to as *speech reading* [Newman et al., 2010].

From an automated computational point of view in Visual Speech Recognition (VSR) where only visual cues are derived as features for recognition, lip-motion is considered more feasible compared to other visual cues. So, lip-motions encoded as visual features contributes towards VSR. When only acoustic

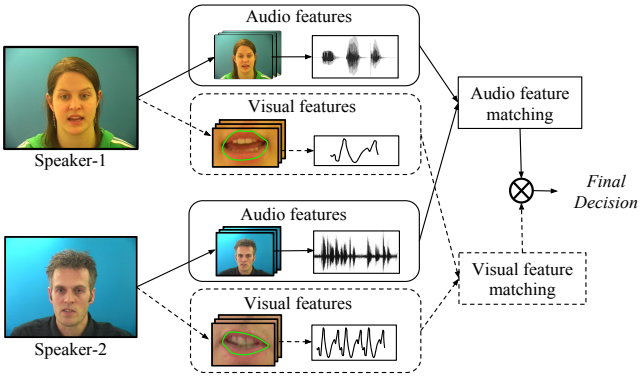


Figure 1: An Audio-Visual Speech Recognition (AVSR) system where the audio and visual (*lip-motion*) features of two speakers are compared to evaluate if they are similar or not. The images of speakers are taken from GRID audio-visual corpus [Cooke et al., 2006].

 ISBN 978-0-9934207-1-9

features are used in recognition, the system is referred to as Automatic Speech Recognition (ASR), and when both acoustic and visual features are used, the system is referred to as Audio-Visual Speech Recognition (AVSR). Figure 1 represents a general AVSR system comparing the audio-visual features of two speakers to evaluate the similarity in speech.

One of the major challenges in lip-motion analysis is due to *phonemes* and *visemes* not sharing one-to-one correspondence. A phoneme is one of the units of sound that distinguishes one word from another in a particular language. A viseme is defined as a visually distinguishable unit of speech in visual domain, the equivalent of phoneme in audio domain. Often, several phonemes correspond to single viseme [Cappelletta and Harte, 2012]. For example, words *pet, bell* and *men* are difficult to distinguish based on lip-motion because they have similar visemes while phonemes are different. The current *ARPAbet* phoneme set for standard English pronunciation maintained by Carnegie Mellon University Pronunciation Dictionary has 39 phonemes [Arp, CMU]. There is no standard viseme set similar to that of phonemes.

The detailed review on recent advances in the area of visual speech decoding [Zhou et al., 2014] points out that state of the art approaches to facial landmark localization is largely ignored in the development of VSR. The review also emphasized about the need for a better visual feature representation encoding the temporal information so as to improve the robustness of VSR. We address these two challenges in our work, and propose a methodology to improve VSR based on lip-motion analysis by incorporating a state of the art fast and accurate facial landmark detection which incorporates ensemble of regression trees learned using gradient boosting [Kazemi and Sullivan, 2014], as well as a simple temporal sequence encoding of visual features which can be verified using Dynamic Time Warping (DTW) algorithm.

The remaining part of this paper is organized as follows: a brief review of the related works in VSR, the OuluVS database used in our work, the proposed algorithm for temporal encoding of features corresponding to lip-motion (VeSPP), experiments demonstrating the robustness of the proposed method in speaker dependent scenario where the encoded features are compared using DTW, followed by conclusion and future work.

2 Related Works

Development of VSR in general involves visual feature extraction, its representation and classification. Extensive works was done on speech recognition based on audio signal alone, or on integrating audio and visual signals. Very little work has been reported in the literature for VSR alone. Various models for lip-motion analysis were studied by involving techniques such as Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Active Appearance Model (AAM), Hidden Markov Model (HMM), Local Binary Patterns (LBP), Support Vector Machines (SVM), etc. A detailed review of these techniques applied for lip-motion analysis can be found in [Zhou et al., 2014].

Of all the models, HMM is the most widely used technique in the domain of VSR [Liu and Cheung, 2014, Yu et al., 2009]. This is mainly due to the fact that HMM can incorporate strong temporal correlations between observed frames. However, the main challenges faced by HMM based VSR systems are: 1) the visual features obtained are not discriminant enough for lip-motion analysis and similarity computation, 2) the learned models are not sufficient to discriminate and characterize different lip-motion activities [Liu and Cheung, 2014].

In acoustic speech domain, there are well established features (for example, Mel-frequency cepstral coefficients (MFCC)), but in VSR, there are no standard accepted visual features. In general, visual features are broadly classified as *image-based, motion-based, geometric-based* and *model-based* [Zhou et al., 2014]. Many VSR based works in recent literature use auxiliary information such as audio corresponding to frames for preprocessing [Zhou et al., 2011], depth and color channel information [Pei et al., 2013] to accompany visual data. Though the systems which use such extra information report improved recognition accuracy, they tend to be more restricted in VSR domain, and these methods cannot be generalized. Also, usually in speaker dependent (SD) scenario, the number of training data available will be less than that of speaker independent (SI) scenario. This scarcity in training data is a major challenge for SD scenario.

In this work, we focus on the SD scenario and represent visual features using geometric-based attributes derived from facial landmarks corresponding to the lip region. These landmarks are obtained using an ensemble

 ISBN 978-0-9934207-1-9

of regression trees learned using gradient boosting as explained in [Kazemi and Sullivan, 2014]. The derived geometric features as well as their temporal encoding is described in the algorithm section, which is the main contribution of this paper. To these geometric features, we apply DTW to obtain a similarity score between any two given lip-motions. Our approach do not need any auxiliary information such as audio, depth or color channels. A similar purely visual only study was proposed by Zhao et al., where spatiotemporal local texture descriptors (LBP-TOP) are used for VSR [Zhao et al., 2009]. We will be following the experimental protocol and compare our results in SD scenario to the results reported in [Zhao et al., 2009].

3 Database

Although there are abundant audio-only databases for ASR, there exist only a few databases suitable for visual-only or audio-visual research. Among the audio-visual databases, many of them contain only recording of one subject, or are limited to isolated digits, letters or short list of fixed phrases not suitable for our experiments [Zhou et al., 2014]. There are few databases providing phrase data, but in many of them either the number of speakers is small or the speakers utter different phrases. For example, in GRID database [Cooke et al., 2006], all the phrases are different.

In this work we chose the OuluVS database which is publicly available and is a benchmark database in visual speech domain [Zhao et al., 2009]. It is a database containing the video and audio data for 20 subjects uttering 10 daily-use short phrases repeated up to 5 times making it suitable for visual speech lip reading experiments. The 10 phrases contained in OuluVS are: *Hello, Excuse me, I am sorry, Thank you, Good bye, See you, Nice to meet you, You are welcome, How are you, Have a good time*. The speakers were from 4 different countries with different accents and speaking rates which makes the dataset challenging. The videos were recorded in an indoor controlled environment. The frame rate was set as 25 fps and the image resolution was 720×576 pixels.

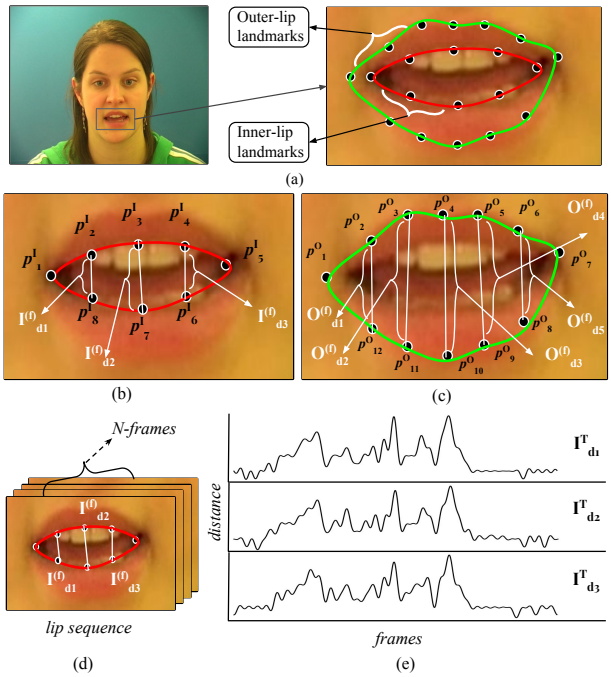


Figure 2: Various stages of the algorithm. (a) the facial landmark obtained for both inner and outer lip regions, (b) three vertical distances derived from the inner-lip landmarks, (c) five vertical distances derived from the outer-lip landmarks, (d) lip sequence consisting of N-frames, (e) temporal encoding of three vertical distances derived from the inner-lip landmarks for the video sequence consisting of N frames. The image of speaker is taken from GRID audio-visual corpus [Cooke et al., 2006] for demonstration purpose only.

4 Algorithm

The algorithm consists of four major stages: lip region landmark detection, visual feature extraction, visual feature encoding, and temporal sequence matching based on DTW for verification. Some of the recent work in facial landmark detection and lip-motion analysis can be found in [Kazemi and Sullivan, 2014, Katina et al., 2015, Sukno et al., 2015, Cao et al., 2014, Liu et al., 2015]. In this work, we used the algorithm proposed in [Kazemi and Sullivan, 2014] for facial landmark detection which used an ensemble of regression trees learned using gradient boosting. The Dlib C++ library [King, 2009] was used to train and obtain these landmarks for lip regions. Face detection was performed using Histogram of Oriented Gradients (HOG) as implemented in [King, 2009].

 ISBN 978-0-9934207-1-9

Stage 1: Landmark detection

Step 1: Using the algorithm proposed in [Kazemi and Sullivan, 2014], estimate the landmarks corresponding to the face.

Step 2: We detect 8 and 12 landmark points corresponding to inner-lip region and outer-lip region respectively (see Figure 2(a)). Depending on how the system is trained to obtain the landmarks, the number of points representing the lip region may vary. Let $p_i^I = (x_i, y_i)$ and $p_j^O = (x_j, y_j)$ represent the i^{th} and j^{th} landmark points for inner-lip and outer-lip regions respectively.

Stage 2: Visual feature extraction

Step 3: Locate the vertically symmetric landmark point pairs corresponding to inner-lip and outer-lip regions. In Figure 2(b), the vertically symmetric point pairs for inner-lip regions are (p_2^I, p_8^I), (p_3^I, p_7^I) and (p_4^I, p_6^I).

Similarly, for outer-lip region, the vertically symmetric point pairs are (p_2^O, p_{12}^O), (p_3^O, p_{11}^O), (p_4^O, p_{10}^O), (p_5^O, p_9^O), and (p_6^O, p_8^O).

Step 4: The distance $dist(p_k, p_l)$ between the vertically symmetric points (VeSPP) p_k and p_l represents a visual feature for a given frame. The vertical distance can be the *euclidean-distance* or the *absolute difference* between the symmetric points. In Figure 2(b), $I_{d1}^{(f)} = dist(p_2^I, p_8^I)$, $I_{d2}^{(f)} = dist(p_3^I, p_7^I)$ and $I_{d3}^{(f)} = dist(p_4^I, p_6^I)$ are the VeSPP features representing the inner-lip region for a given frame f.

Similarly, $O_{d1}^{(f)}$, $O_{d2}^{(f)}$, $O_{d3}^{(f)}$, $O_{d4}^{(f)}$, $O_{d5}^{(f)}$ are estimated from vertically symmetric points for the outer-lip region, as shown in Figure 2(c).

Stage 3: Visual feature encoding

Step 5: Assuming the lip sequence consists of N-frames (Figure 2(d)), repeat Steps $1-4$ for each frame and temporally concatenate the vertical distances corresponding to each vertically symmetric pairs to obtain its VeSPP temporal encoding. For example, $I_{d1}^T = \langle I_{d1}^{(1)}, I_{d1}^{(2)}, ..., I_{d1}^{(N)} \rangle$ represents the VeSPP temporal encoding for the vertically symmetric pair (p_2^I, p_8^I) corresponding to the inner-lip region for the lip sequence consisting of N-frames. Similarly, obtain the VeSPP feature temporal encoding of all vertically symmetric points corresponding to inner and outer lip regions.

Figure 2(e) shows the plot of the temporal encodings of I_{d1}^T, I_{d2}^T and I_{d3}^T for the inner-lip region thus obtained corresponding N frames (X-axis corresponds to frame number and Y-axis corresponds to vertical distance).

Stage 4: Similarity computation for verification

Step 6: Lip-motions representing a particular phrase by the same person at different instances may vary in both time and speed. The VeSPP features proposed in the previous steps encode the lip-motion as a temporal sequence. DTW can be used to find an optimal match between any two such encoded lip-motions. DTW is a dynamic programming based distance measure which allows a non-linear mapping of one temporal sequence onto another by minimizing the distance between them.

Suppose qI_{d1}^T and rI_{d1}^T represents the VeSPP temporal encoding of vertically symmetrical pairs (p_2^I, p_8^I) of two instances of lip-sequence videos consisting of M and N frames respectively where,

$$qI_{d1}^T = \langle q_1, q_2, q_3, ..., q_i, ..., q_M \rangle, \tag{1}$$

$$rI_{d1}^T = \langle r_1, r_2, r_3, ..., r_j, ..., r_N \rangle \tag{2}$$

 ISBN 978-0-9934207-1-9

where $q_i = I_{d1}^{(i)} = dist(p_2^I, p_8^I)$ of the i^{th} frame of the query video, and $r_j = I_{d1}^{(j)} = dist(p_2^I, p_8^I)$ of the j^{th} frame of the reference video.

These VeSPP features may correspond to lip-motions of the same speaker or different speakers.

To perform a non-linear alignment between qI_{d1}^T and rI_{d1}^T using DTW, we construct an $M \times N$ matrix where the $(i,j)^{th}$ entry of the matrix corresponds to squared distance $d(q_i, r_j) = (q_i - r_j)^2$ which is the alignment between q_i and r_j. The best match between qI_{d1}^T and rI_{d1}^T is found by retrieving a path through this matrix that minimizes the total cumulative distance between them. Essentially, the optimal path is the path that minimizes the warping cost

$$DTW(qI_{d1}^T, rI_{d1}^T) = min\Big(\sum_{k=1}^{K} w_k\Big) \tag{3}$$

where w_k is the matrix element $(i,j)_k$ that also belongs to the k^{th} element of a warping path W, a contiguous set of matrix elements that represents an optimal mapping between qI_{d1}^T and rI_{d1}^T. The warping path can be found using the dynamic programming to evaluate the recurrence

$$C(i,j) = d(i,j) + min\left\{\begin{array}{l} C(i,j-1) \\ C(i-1,j) \\ C(i-1,j-1) \end{array}\right\} \tag{4}$$

where $d(i,j)$ is the distance calculated for the current cell, and $C(i,j)$ is the cumulative distance of $d(i,j)$ and the minimum cumulative distance from the three adjacent cells.

Step 7: Let $c = DTW(qI_{d1}^T, rI_{d1}^T)$ be the minimum warping cost obtained, which is a dissimilarity measure, i.e., the lower the warping cost, the lower their dissimilarity which implies both temporal sequences are similar. We can transform the dissimilarity score to a similarity score S by

$$S = exp(-c) \tag{5}$$

Now, S close to 0 implies the temporal sequence compared are dissimilar, and when S is close to 1, the temporal sequences are similar. The ideal case is when same copies of signals are compared, which leads to a DTW value of 0 which in turn upper bounds to a similarity value of 1. The similarity scores thus obtained are normalized in the range $[0,1]$. So, our results can be directly used for multi-modal score fusions in AVSR applications.

5 Experiments

5.1 Experiment protocol

We used OuluVS video database in our experiments. The details about this database is briefly described in Section 3. We tested our proposed methodology for building a speaker dependent lip reading system. For each of the 20 speakers, the leave-one-video-out cross validation was carried out, i.e., one video is used for testing as a query template, and the rest were used as reference template. Since each of the 20 speakers uttered 10 different phrases repeated at least 5 times, there should be in total $20 \times 10 \times 5$ test comparisons in the cross validation scenario. In OuluVS video database, three video files corresponding to the repetition of three different phrases are not available. So, in our experiments, we have 997 test comparisons in total. For each testing, there are at most 4 match (*genuine*) scores and 19×4 non-match (*impostor*) scores. We determined whether the given comparison is a match or non-match based on the maximum similarity score obtained in the comparisons. We report the overall results of the cross validation in terms of recognition rate, obtained using M/N (M is the total number of correctly recognized sequence and N is the total number of testing sequence). Together with the recognition rate, we also generate the confusion matrix to see the clustering ability of the proposed method.

 ISBN 978-0-9934207-1-9

Configuration *(label)*	VeSPP feature configuration *(raw feature ⊕ first derivative)*	Recognition rate *(in %)*
C1	$(I^T_{d2}) \oplus (I'^T_{d2})$	71.61
C2	$(I^T_{d2} + O^T_{d2}) \oplus (I'^T_{d2} + O'^T_{d2})$	80.14
C3	$(I^T_{d2} + O^T_{d2} + O^T_{d3}) \oplus (I'^T_{d2} + O'^T_{d2} + O'^T_{d3})$	82.24
C4	$(I^T_{d2} + O^T_{d2} + O^T_{d3} + O^T_{d4}) \oplus (I'^T_{d2} + O'^T_{d2} + O'^T_{d3} + O'^T_{d4})$	**82.65**

Table 1: Recognition accuracy of speaker dependent experiments on OuluVS database for different configurations of VeSPP features.

Phrase	Excuse	Goodbye	Hello	How	Nice	Seeyou	Sorry	Thank	Time	Welcome
Excuse	**85.9**	4.0	0.0	0.0	7.1	1.0	2.0	0.0	0.0	0.0
Goodbye	3.0	**81.0**	1.0	2.0	5.0	1.0	0.0	0.0	0.0	7.0
Hello	2.0	0.0	**67.0**	0.0	1.0	15.0	1.0	14.0	0.0	0.0
How	0.0	0.0	2.0	**86.0**	2.0	2.0	2.0	3.0	1.0	2.0
Nice	2.0	0.0	0.0	4.0	**91.9**	1.0	0.0	1.0	0.0	0.0
Seeyou	1.0	0.0	9.1	1.0	2.0	**72.7**	0.0	13.1	0.0	1.0
Sorry	1.0	0.0	0.0	3.0	3.0	0.0	**88.0**	1.0	3.0	1.0
Thank	0.0	0.0	8.0	1.0	1.0	15.0	0.0	**74.0**	0.0	1.0
Time	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	**91.0**	6.0
Welcome	0.0	2.0	1.0	2.0	5.0	0.0	0.0	1.0	0.0	**89.0**

Table 2: Confusion matrix showing the recognition accuracy in percentage for the cross validation of 10 phrases uttered by 20 speakers of OuluVS database for the configuration C4 in Table 1.

5.2 Speaker dependent system

We used various configurations of VeSPP features and their first derivatives for testing the performance of the proposed visual speech encoding. In this experiment, we highlight only those VeSPP features which lead to best performance. So, the features discussed in this experiment are $I^T_{d2}, O^T_{d2}, O^T_{d3}, O^T_{d4}$, and we discard discussions about other VeSPP features. The first derivative of these sequences are also used which will be denoted as $I'^T_{d2}, O'^T_{d2}, O'^T_{d3}, O'^T_{d4}$. The first derivative is obtained by taking the difference between two consecutive values. We also generated a concatenated version of the features. Such concatenations of visual features were studied previously in [Zhou et al., 2011] and has shown performance improvements. We denote the concatenate version of I^T_{d2}, O^T_{d2} as $(I^T_{d2} + O^T_{d2})$. When we talk about match score for concatenated version $(I^T_{d2} + O^T_{d2})$, it is obtained by $DTW(qI^T_{d2} + qO^T_{d2}, rI^T_{d2} + rO^T_{d2})$ as mentioned in Eq.(3). When raw features and first derivative features are combined to obtained the match score, we denote it as $(I^T_{d2}) \oplus (I'^T_{d2})$, and the final score is obtained by adding the individual scores: $DTW(qI^T_{d2} + rI^T_{d2}) + DTW(qI'^T_{d2} + rI'^T_{d2})$.

Table 1 lists the recognition accuracy for various configurations of the VeSPP features. We used the *euclidean-distance* for VeSPP features. We noticed that the concatenated version of four features (configuration C4 in Table 1) taken from inner and outer lip regions together with their first derivative achieves the best result of 82.65% recognition accuracy. We also notice that for configuration C2, the performance is only slightly lower than that of C4. So, the proposed encoding scheme can be utilized for a faster implementation with a very small trade-off between the speed and accuracy.

We also report in Table 2 the confusion matrix to understand the clustering ability of our proposed method. The confusion matrix is for the 10 phrases in the OuluVS database by 20 speakers for the configuration C4 in Table 1 for the leave-one-video-out cross validation. The number at the i^{th} row and j^{th} column gives the percentage of i^{th} phrase being classified as j^{th} by our method.

 ISBN 978-0-9934207-1-9

Method	Recognition rate
[Zhao et al., 2009] (Automatic)	64.20%
[Zhao et al., 2009] (Manual)	70.20%
VeSPP method (Configuration C4)	**82.65%**

Table 3: Comparison with other visual only recognition accuracy for speaker dependent results for OuluVS database.

In Table 3, we compare our result with that of [Zhao et al., 2009] which followed the same experimental protocol as ours. In [Zhao et al., 2009], they propose two different experiments where the mouth region is manually located as well as automatically detected. In our experiment, mouth region is automatically detected, and we obtained 82.65% recognition accuracy compared to 64.20% obtained in [Zhao et al., 2009]. Also, our method outperforms the manually processed method which achieved 70.20% recognition accuracy.

In [Zhou et al., 2011], they report a much better result for automatic and manual processing for speaker dependent scenario, but those results cannot be compared to ours because, they used audio information to locate speaking and non-speaking frames, and then removed the non-speaking frames from the training video. So, their experiment is not purely visual only scenario and makes the audio information necessary to generate improved results. In [Pei et al., 2013], they proposed a method which uses depth information and color channels in VSR experiments, and their protocols were different. So, we discarded comparing our results to [Zhou et al., 2011] and [Pei et al., 2013].

6 Runtime analysis

The runtime complexity for facial landmark detection using ensemble of regression trees is a constant $O(TKF)$ where T, K and F are number of strong regressors, number of weak regressors and depth of trees respectively [Kazemi and Sullivan, 2014]. Deriving visual features corresponding to vertically symmetric pairs is a constant time operation, and is just taking absolute difference which is of $O(1)$. Once we have the temporally encoded VeSPP features, the verification can be performed using a fast-DTW comparison which can be performed in $O(N)$ time complexity, where N is the length of the temporal sequences [Salvador and Chan, 2004]. So, the proposed lip-motion verification can be achieved in linear time complexity upon detecting the face.

7 Conclusion and Future Work

We have proposed a robust temporal encoding of visual features (VeSPP) for lip-motion sequences based on distance computed from vertically symmetric points corresponding to lip regions. We used state of the art facial landmarks detection, and demonstrated its usefulness in lip-motion based verification using DTW comparison on a challenging database where the phrases are of different accents and speaking rates. Our experiments justify that concatenation of VeSPP visual features corresponding to inner-lip and outer-lip region provide better recognition accuracy and obtained 82.65% recognition rate. The fact that the proposed VeSPP features can be compared using DTW demonstrates its robustness in terms of negating the need for any training unlike HMM where more data samples are needed to train its model. In many real-time VSR applications, we cannot always expect to acquire more training samples, especially in case of speaker dependent scenario. Also, our method does not mandate any auxiliary information such as audio, depth or color channels, and is feasible on visual-only 2D data. We will be extending this work to build a speaker independent system based on the visual feature encoding developed in this work, as well as testing the system in real-time scenario.

Acknowledgments

This research was supported by funding from the charity RESPECT and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no: PCOFUND-GA-2013-608728.

References

[Arp, CMU] (CMU). Carnegie Mellon University Pronunciation Dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[Cao et al., 2014] Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

[Cappelletta and Harte, 2012] Cappelletta, L. and Harte, N. (2012). Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)*, pages 322–329.

[Cooke et al., 2006] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

[Katina et al., 2015] Katina, S., McNeil, K., Ayoub, A., Guilfoyle, B., Khambay, B., Siebert, P., Sukno, F., Rojas, M., Vittert, L., Waddington, J., et al. (2015). The definitions of three-dimensional landmarks on the human face: an interdisciplinary view. *Journal of anatomy*.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1867–1874. IEEE.

[King, 2009] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

[Liu et al., 2015] Liu, H., Zhang, X., and Wu, P. (2015). Regression based landmark estimation and multi-feature fusion for visual speech recognition. In *IEEE International Conference on Image Processing*, pages 808–812.

[Liu and Cheung, 2014] Liu, X. and Cheung, Y.-M. (2014). Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification. *Information Forensics and Security, IEEE Transactions on*, 9(2):233–246.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*.

[Newman et al., 2010] Newman, J. L., Theobald, B.-J., and Cox, S. J. (2010). Limitations of visual speech recognition. In *AVSP*, page 1.

[Pei et al., 2013] Pei, Y., Kim, T.-K., and Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 129–136.

[Salvador and Chan, 2004] Salvador, S. and Chan, P. (2004). Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD workshop on mining temporal and sequential data*, pages 70–80.

[Sukno et al., 2015] Sukno, F. M., Waddington, J. L., and Whelan, P. F. (2015). 3D Facial Landmark Localization With Asymmetry Patterns and Shape Regression from Incomplete Local Features.

[Yu et al., 2009] Yu, D., Ghita, O., Sutherland, A., and Whelan, P. F. (2009). A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition. In *Advances in Image and Video Technology*, pages 398–409. Springer Berlin Heidelberg.

[Zhao et al., 2009] Zhao, G., Barnard, M., and Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265.

[Zhou et al., 2014] Zhou, Z., Zhao, G., Hong, X., and Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605.

[Zhou et al., 2011] Zhou, Z., Zhao, G., and Pietikäinen, M. (2011). Towards a practical lipreading system. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 137–144. IEEE.

Fast Corner Detection Using a Spiral Architecture

J. Fegan,[1] S.A., Coleman,[1] D. Kerr,[1] B.W., Scotney[2]

[1]*School of Computing and Intelligent Systems,*
[2]*School of Computing and Information Engineering,*
Ulster University, Northern Ireland

Abstract

Fast image processing is a key element in achieving real-time image and video analysis. Here, a novel framework based on a spiral architecture is used to facilitate fast image processing, in particular, fast corner detection. Unlike a conventional image addressing scheme where the picture elements are indexed using two-dimensional Cartesian coordinates, a spiral addressing scheme enables the image to be stored and indexed as a one-dimensional vector. Image processing is hastened through the combined use of the one-dimensional structure and a lookup table. The performance is evaluated by the application of a corner detector based on the Harris corner detection algorithm. The results demonstrate the efficiency of the proposed approach compared with a typical two-dimensional implementation.

Keywords: Fast Image Processing, Spiral Architecture, Corner Detection, Lookup Table, Eye Tremor

1 Introduction

Since its inception, digital image processing has largely leaned upon the intuitive notion that two-dimensional (2D) visual data can be sampled as a matrix of picture elements (pixels) using a rectangular lattice of sensors. This has resulted in a Cartesian coordinate system where each pixel is referenced by an index in the horizontal and vertical directions. This approach has worked well for tasks where the time taken to process an image is not a primary concern. However, research has shown that, compared with a one-dimensional (1D) approach, operating on a matrix requires additional computation to locate the pixels in two directions [1]–[5]. This has implications for activities such as video processing where the system in question is expected to operate on a stream of consecutive image frames under strict time restraints. Subsequently there has been a growing interest among researchers to explore alternative image representations. Strategies such as Hexagonal Image Processing (HIP) have demonstrated computational performance improvements resulting in a subsequent runtime advantage over a conventional 2D approach [1], [2]. Despite this fact, existing image capture and processing hardware is predominately based on a rectangular architecture and this has limited the research and practical applications of hexagonal imaging methods [3]–[6]. In response to these concerns a new framework has been proposed, one that attempts to reconcile the prevalent rectangular framework with the strengths of the HIP framework. This new approach uses a newly developed sampling method based on a square spiral (squiral) address scheme that is similar to the hexagonal address scheme of the HIP framework [6]. In this paper, corner detection is used to evaluate the performance and effectiveness of this approach.

 ISBN 978-0-9934207-1-9

2 Squiral Image Processing Framework

Previous research has shown that the Squiral Image Processing (SIP) framework is capable of delivering fast results [3]–[5]. This is partly attributed to the SIP address scheme, which avoids using 2D Cartesian coordinates in favour of a 1D index sequence. More precisely, an image in the SIP framework is sized according to a template called a layer. The first layer (0) is a single pixel located at the centre of the image. The next layer (1) encompasses layer 0 and its 8 surrounding pixels. Thereafter, each subsequent layer encompasses its preceding layer and 8 regions of the same size that surround it. The 9 element makeup of each layer promotes a recursive base 9 address scheme. In this scheme the indexing begins at the centre element of each layer and continues outwards in a clockwise spiral (Figure 1). Ultimately each layer is denoted by the position of a digit in a pixel's index and an element is denoted by the value of that digit. In accordance with this scheme the image is unravelled onto a vector. This means that only a single loop is required to traverse the image for subsequent processing. Despite this benefit, the vectorised nature of a SIP image makes it difficult to locate the neighbours of pixels that are not the centre of a squiral. There are currently two solutions that solve this problem: create and reference a table that stores each pixels Cartesian neighbour locations as SIP indices [1], [2]; find each pixels neighbours by shifting the pixels in various directions, an approach based on a biological process of involuntary eye movements called tremors [7].

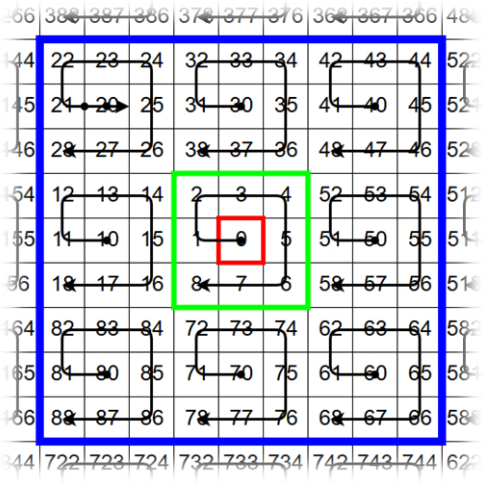


Figure 1: SIP Address Scheme

2.1 SIP Neighbour Lookup Table

	1	2	3	4	5	6	7	8
1	15	14	2	3	0	7	8	16
2	14	26	38	37	3	0	1	15
3	2	38	37	36	4	5	0	1
4	3	37	36	48	52	51	5	0
5	0	3	4	52	51	58	6	7
6	7	0	5	51	58	62	74	73
7	8	1	0	5	6	74	73	72
8	16	15	1	0	7	73	72	84

Figure 2: Neighbour LUT

The first of the two aforementioned solutions relays on a pre-calculated lookup table (LUT) to find the locations of each pixel's Cartesian neighbours in a SIP vector. In this situation each pixel has its own record of neighbour indices in the LUT. Accordingly, a pixel's index is used to access a record and find its neighbour pixels. For example, the LUT in Figure 2 demonstrates that the 8 immediate neighbours of pixel 1 can be found at positions 15, 14, 2, 3, 0, 7, 8, and 16 in the SIP vector. It should be noted that creating the LUT is a one-time procedure and it can be saved and reloaded as required. Incidentally, a single LUT can be used with differently sized SIP images. A notable benefit of this solution is that neighbourhood operations require only two loops. This is opposed to four loops that are common in a typical 2D approach [3], [4]. In the case of neighbourhood navigation, the two neighbourhood loops that are normally used are replaced with one loop that is used to fetch a neighbour index from the LUT. In the case of convolution, the filter in question will need to be vectorised in accordance with the SIP address scheme. This permits the use of a single loop where the index of each filter element is used to retrieve a neighbour index from the LUT. In both cases the use of the LUT presents another advantage by avoiding computation that is otherwise needed to locate a pixel's neighbours.

 ISBN 978-0-9934207-1-9

2.2 SIP Eye Tremor Image Processing

As previously noted, the biological behaviour of eye tremor can be mimicked to find a pixel's Cartesian neighbours in a SIP vector. In this solution the image is sampled once initially before it is offset and resampled several times [1]–[5]. More specifically, in the SIP framework, the initial 2D image is treated as a base which is sampled according to the SIP address scheme. After this the image is offset so that the next element in the SIP address sequence is centred on the rectangular lattice. The offset image is sampled and the process is repeated several times, (Figure 3.1). The output of this procedure is a matrix composed of several SIP vectors [3]–[5]. The diagonal symmetry of this matrix (Figure 3.2.) means that it can be navigated in two different ways: the first way is to loop through each pixel in the base image and refer to the vertically adjacent neighbours in the offset images; the second way is to sparsely process a fraction of the pixels in each image with reference to their horizontally adjacent neighbours. Previous research on edge detection has indicated that this approach is faster than 2D edge detection and SIP edge detection using a neighbour LUT [1], [2]. For this reason, it was the first strategy adopted in the development of a SIP corner detector.

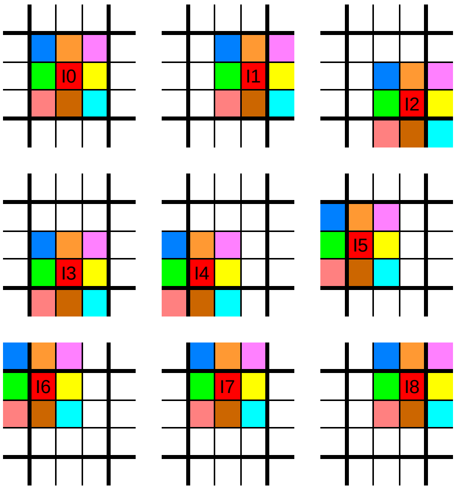


Figure 3.1: Eye Tremor Sampling Scheme

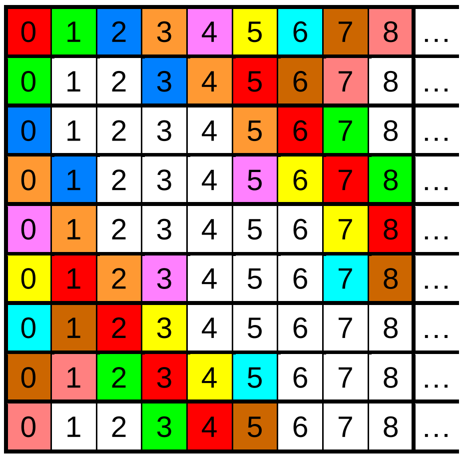


Figure 3.2: Eye Tremor Image

3 SIP Corner Detection

The goal of this research was to investigate the performance of the SIP framework for the application of corner detection. To this end the Harris corner detector was used as a base in the development of the corner detector presented here. As part of this corner detection procedure, the image gradient is computed in two directions. A common way to do this is to apply a gradient filter by convolution [8]. In most cases two filters are used: the first filter computes the gradient in the horizontal direction; the second filter computes the gradient in the vertical direction. Once the gradients are computed they are manipulated and smoothed. This smoothing is typically achieved through another convolution with a Gaussian smoothing filter [8]. However, this presents a problem for the eye tremor approach because a convolution with the gradient filter generates an output where most of the gradient pixels are not adjacent to their Cartesian gradient neighbours (Figure 4). Therefore, the gradient neighbour pixels must be located using base 9 arithmetic, or the eye tremor gradient image must be combined to form a complete 2D gradient image and sampled using the eye tremor scheme. To avoid the computational cost of these solutions, in the work presented here a neighbour LUT was used.

ISBN 978-0-9934207-1-9

0	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16	…
1	15	14	2	3	0	7	8	16	11	155	154	12	13	10	17	…
2	14	26	38	37	3	0	1	15	12	154	146	28	27	13	10	…
3	2	38	37	36	4	5	0	1	13	12	28	27	26	14	15	…
4	3	37	36	48	52	51	5	0	14	13	27	26	38	2	1	…
5	0	3	4	52	51	58	6	7	15	10	13	14	2	1	8	…
6	7	0	5	51	58	62	74	73	16	17	10	15	1	8	72	…
7	8	1	0	5	6	74	73	72	17	18	11	10	15	16	84	…
8	16	15	1	0	7	73	72	84	18	156	155	11	10	17	83	…

0	1	2	3	4	5	6	7	8	10	…
1	15	14	2	3	0	7	8	16	11	…
2	14	26	38	37	3	0	1	15	12	…
3	2	38	37	36	4	5	0	1	13	…
4	3	37	36	48	52	51	5	0	14	…
5	0	3	4	52	51	58	6	7	15	…
6	7	0	5	51	58	62	74	73	16	…
7	8	1	0	5	6	74	73	72	17	…
8	16	15	1	0	7	73	72	84	18	…

Figure 4: Post Neighbourhood Processing Problem

After the convolution with the smoothing filter, the outputs are used in an auto-correlation function to calculate a corner score for each pixel [8]. After this a threshold is applied to the corner scores to reveal prominent corner pixels. Optimal corner points can then be selected by suppressing all non-maximum corner pixels within a specified neighbourhood region. In the previous steps of the Harris algorithm an 8 neighbour LUT was sufficient to process the image because the operations only concerned a pixel's 8 immediate neighbours. However, to perform effective non-maximum suppression (NMS), it is usually necessary to evaluate a larger neighbourhood. It was thought, at first, that this neighbourhood could be navigated using the 8 neighbour LUT, a pixel's own neighbour record and the neighbour records of other pixels. In practice, problems are caused by the order of the pixels in the SIP vector. For example, in Figure 5a the 9x9 (layer 2) neighbourhood of pixel 10 is found using its own neighbour record and the neighbour records of the 8 corresponding pixels that surround it. However, it is shown in Figure 5b that the 8 corresponding pixels that succeed pixel 10 in the SIP vector are not the pixels that surround it in the 2D plane. This means that computation is needed to find these corresponding pixels and their neighbour records. In this paper this problem was overcome by using a larger, 728 neighbour LUT.

276	268	267	266	388	387	386	378	377	376	368
134	142	143	144	22	23	24	32	33	34	42
135	141	140	145	21	20	25	31	30	35	41
136	148	147	146	28	27	26	38	37	36	48
104	152	153	154	12	13	14	2	3	4	52
105	151	150	155	11	10	15	1	0	5	51
106	158	157	156	18	17	16	8	7	6	58
174	162	163	164	82	83	84	72	73	74	62
175	161	160	165	81	80	85	71	70	75	61
176	168	167	166	88	87	86	78	77	76	68
834	842	843	844	722	723	724	732	733	734	742

(a)

286	278	277	276	268	267	266	388	387	386	378	377	376	368	367	366	488
124	132	133	134	142	143	144	22	23	24	32	33	34	42	43	44	522
125	131	130	135	141	140	145	21	20	25	31	30	35	41	40	45	521
126	138	137	136	148	147	146	28	27	26	38	37	36	48	47	46	528
114	102	103	104	152	153	154	12	13	14	2	3	4	52	53	54	512
115	101	100	105	151	150	155	11	10	15	1	0	5	51	50	55	511
116	108	107	106	158	157	156	18	17	16	8	7	6	58	57	56	518
184	172	173	174	162	163	164	82	83	84	72	73	74	62	63	64	582
185	171	170	175	161	160	165	81	80	85	71	70	75	61	60	65	581
186	178	177	176	168	167	166	88	87	86	78	77	76	68	67	66	588
824	832	833	834	842	843	844	722	723	724	732	733	734	742	743	744	622

(b)

Figure 5: NMS Problem

 ISBN 978-0-9934207-1-9

4 Performance Evaluation

In preparation for testing, several steps were taken to ensure the outcomes of this research would be reliable and fair. The corner detector developed for the SIP framework was adapted from a corresponding 2D counterpart. This was done to minimise syntactical differences that could affect an implementation's runtime performance. A 243x243 pixel (layer 5) image was used as a test case in all the experiments presented here. The Sobel filter was selected to compute the image gradient, and a 3x3 pixel (layer 1) Gaussian filter with a standard deviation of 1 was used for smoothing. A threshold of 70,000 was applied to the Harris response matrix, and a 27x27 pixel (layer 3) neighbourhood region was used for NMS. The convolution was non-separable and three modes were used for all neighbourhood operations: ***Ignore*** pixels with any neighbour indices outside the image bounds; ***Discard*** neighbour indices outside the image bounds; ***Wrap*** neighbour indices that are outside an image border so that they assume an index at the opposite border. For the first two modes a neighbourhood LUT with out of bound neighbour indices was used. For the last mode, a LUT with wrapped index values was used. The corner maps in Figure 5 were produced using the configuration outlined above.

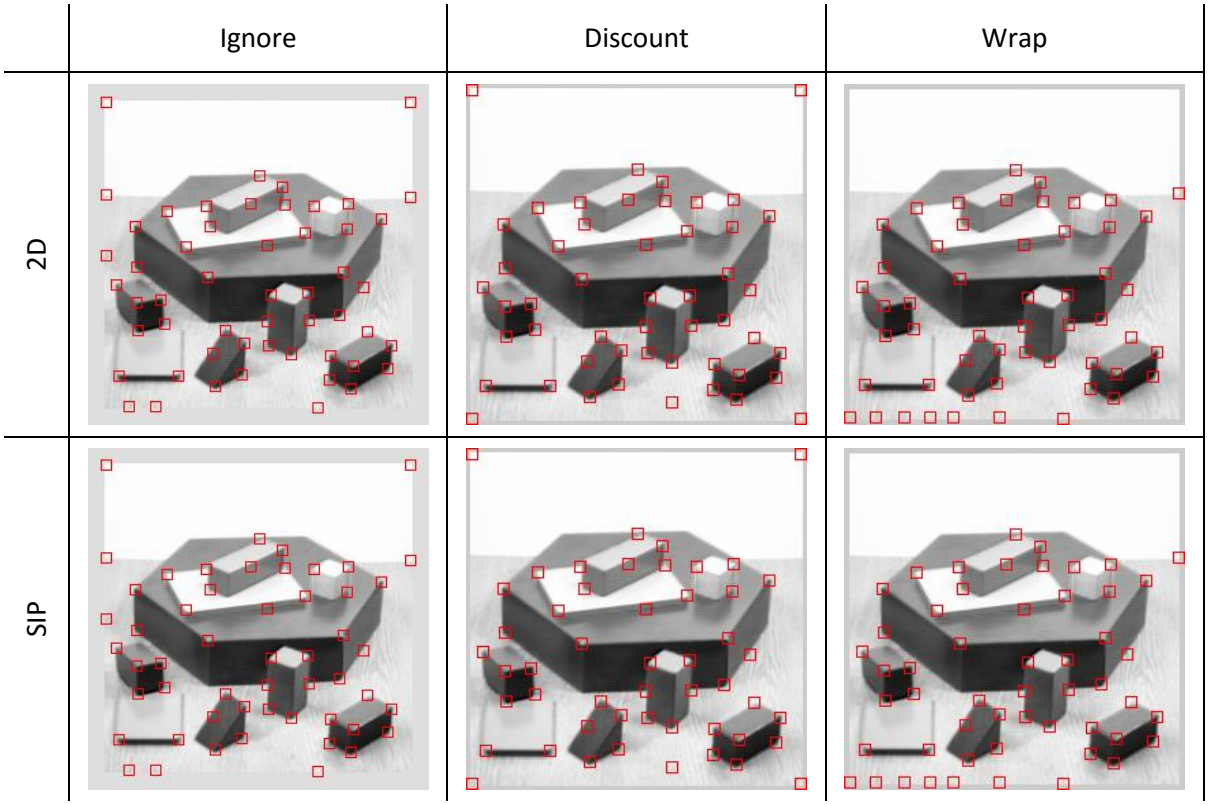


Figure 6: Corner Maps

Note that a few corners have been detected around the image borders; this is a typical response for neighbourhood operations at these regions. Regardless, any issues that these anomalies present can usually be overcome by ignoring or cropping them. As a note for future research, the question is open as to how border pixels in a 1D SIP vector will be handled. For now though, it can be seen that the corner maps of both frameworks are identical. To add further credence to this, the results have been verified as numeric equals in Matlab. This, like the findings in previous SIP research asserts that the SIP framework is capable of delivering outputs that are identical to those produced by a 2D framework [3], [4].

 ISBN 978-0-9934207-1-9

4.1 Runtime Evaluation

The tables below show the times it took to perform SIP conversion, as well as corner detection, in both frameworks with the three modes of neighbourhood operation. For these experiments the runtime assessments were conducted on an Intel Core i7 4790 CPU with 16GB of RAM. The times are based on the average wall-clock times over 1000 runs and were measured using the Matlab functions, *tic* and *toc*. These are the functions recommended by the official Matlab documentation for measuring time reliably [9]. Table 1 shows the runtimes of the conversions to and from a layer 5 SIP image. Table 2 shows the runtimes for the corner detection procedure as measured from the first convolution with the Sobel filter and ending with NMS. The runtime costs for loading the image, setting up filters and displaying the corner maps are not accounted for because they have no bearing on the corner detection process.

2D -> SIP	0.0000095s
SIP -> 2D	0.0006358s

Table 1: Layer 5 Framework Conversion Times

	Runtimes	
	2D	SIP
Ignore	0.3851772s	0.2128920s
Discard	0.5171593s	0.2280309s
Wrap	0.5433419s	0.1678514 s

Table 2: Corner Detection Runtime Results

For the 2D framework with mode Ignore, a 13 pixel border was added to the image to expand it to 269x269 pixels. This was done to account for the 13 neighbour pixels that would extend outside the image during layer 3 NMS. For the same reason, a layer 6 border was added to the SIP image. In both cases, convolution and NMS were restricted to the central layer 5 region. This was done to permit an inbounds convolution on the same data set used by the other modes. Likewise, it keeps the operating conditions across both frameworks as similar as possible. In the case of Discard, a simple boundary check was used to remove neighbour indices that were out of bounds: the 2D framework required four checks to ensure that a neighbour index was within the four image borders; the SIP framework required only one check to ensure that a neighbour index was less than the upper bounds of the SIP vector.

The results show that, compared to a 2D framework, corner detection can be performed much faster on the SIP framework if it is used in conjunction with a neighbourhood LUT. This is observed even if the framework conversion times (Table 1) are summed with the SIP corner detection runtimes. In agreement with previous research it is believed that this performance gain is due, in part, to the noted characteristics of SIP. Namely, that less loops are needed to navigate a vectorised SIP image [3], [4]. The other part of this performance gain is due to the neighbourhood LUT which avoids runtime costs that are normally needed to calculate a pixel's neighbour indices. This is especially notable for neighbourhood operations that use a Wrap mode where the neighbour indices would normally undergo additional calculations to find their circular value.

 ISBN 978-0-9934207-1-9

5 Conclusion

It has been shown that the SIP framework is capable of detecting corners in a way that is equal to a 2D approach, but with significant improvements in algorithmic run-times for non-separable operators. Furthermore, it has been demonstrated that a framework based on a spiral scheme and utilising a neighbourhood LUT is capable of producing fast results. The issues raised in this paper highlight the need to further investigate the eye tremor approach and overcome the current implementation issues and potentially speed up this process further. It could also be useful, in terms of memory allocation, to investigate other methods for addressing SIP neighbourhoods using an 8 neighbour LUT. Future research will extend on the work presented here by developing SIP based Interest Point detectors for applications on video data and ultimately high speed robotic vision challenges.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 607691, SLANDAIL (Security System for Language and Image Analysis). This work was completed under a PhD studentship supported by the Department of Education and Learning (DEL).

The materials presented and views expressed here are the responsibility of the author(s) only. The EU Commission takes no responsibility for any use made of the information set out.

References

[1] B. Scotney, S. Coleman, and B. Gardiner, "Biologically Motivated Feature Extraction Using the Spiral Architecture," in *International Conference on Image Processing*, 2011, pp. 221 – 224.

[2] S. Coleman, S. Bryan, and G. Bryan, "A Biologically Inspired Approach for Fast Image Processing," in *International Conference on Machine Vision Applications*, 2013, pp. 129 – 132.

[3] M. Jing, B. Scotney, S. Coleman, and M. McGinnity, "A Novel Spiral Addressing Scheme for Rectangular Images," in *International Conference on Machine Vision Applications*, 2015, pp. 102 – 105.

[4] M. Jing, S. Coleman, B. Scotney, and M. McGinnity, "Multiscale 'Squiral' (Square-Spiral) Image Processing," in *Irish Machine Vision and Image Processing (IMVIP)*, 2015.

[5] M. Jing, S. Coleman, and B. Scotney, "Biologically Motivated Spiral Architecture for Fast Video Processing," in *International Conference on Image Processing*, 2015, pp. 2040 – 2044.

[6] L. Middleton and J. Sivaswamy, *Hexagonal Image Processing: A Practical Approach*, vol. 224, no. 4. 2005.

[7] A. Róka, Á. Csapó, B. Reskó, and P. Baranyi, "Edge Detection Model Based on Involuntary Eye Movements of the Eye Retina System," *Acta Polytech. Hungarica*, vol. 4, no. 1, pp. 31–46, 2007.

[8] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Procedings of the Alvey Vision Conference*, 1988, pp. 147–151.

[9] Mathworks, "Measure Performance of Your Program," *R2016a Documentation*, 2016. [Online]. Available: https://uk.mathworks.com/help/matlab/matlab_prog/measure-performance-of-your-program.html?requestedDomain=www.mathworks.com. [Accessed: 16-Mar-2016].

Field investigation of contactless displacement measurement using computer vision systems for civil engineering applications

D. Lydon[1], S.E Taylor[1], J. Martinez-Del-Rincon[2], D. Hester[1], M. Lydon[1], and D.Robinson[1]

[1] *School of Planning Architecture and Civil Engineering, Queens University Belfast, BT9 5AG, Northern Ireland*
[2]*School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, BT9 5BN, Northern Ireland*

Abstract

Much of the bridge stock on major transport links in North America and Europe was constructed in the 1950s and 1960s and has since deteriorated or is carrying loads far in excess of the original design loads. Structural Health Monitoring Systems (SHM) can provide valuable information on the bridge capacity but the application of such systems is currently limited by access and bridge type. This paper investigates the use of computer vision systems for SHM. A series of field tests have been carried out to test the accuracy of displacement measurements using contactless methods. A video image of each test was processed using a modified version of the optical flow tracking method to track displacement. These results have been validated with an established measurement method using linear variable differential transformers (LVDTs). The results obtained from the algorithm provided an accurate comparison with the validation measurements. The calculated displacements agree within 2% of the verified LVDT measurements, a number of post processing methods were then applied to attempt to reduce this error.

Keywords: Digital Image Correlation, Camera based Monitoring, Structural Health Monitoring

1 Introduction

This paper investigates the use of computer vision systems for Civil Engineering applications, in particular Structural Health Monitoring (SHM). In essence the goal of SHM is to infer information about the condition or health of a structure by analysing data (often displacement or its time domain derivatives) collected on the structure, and where necessary make appropriate repairs. In the long term, monitoring with cameras is expected to be more broadly utilised for structural engineering purposes because of its potential for inexpensive deployment in real life bridges. Essentially installing sensors on active bridges is logistically difficult and expensive, therefore contactless camera based monitoring is potentially a very useful tool.

A recent example of vision based systems being applied to bridge monitoring is from [Ojio et al., 2016]. In their study a series of vehicles of known weight passing over a bridge were used to determine that changes to deflection can be used a measure of changes to bridge stiffness. In this paper, a set of field trials were carried out to determine the level of accuracy in deflection measurement using a non-contact camera monitoring system. A digital camera was set up to monitor this test and the resulting video images were post-processed to calculate displacements. A modified version of the Kanade-Lucas-Tomasi (KLT) optical flow algorithm[Tomasi and Kanade, 1991] was used to track deflections of the specimen, with these results validated using conventional displacement measurement techniques, such as linear variable displacement transformers (LVDT). The results from directly above the LVDT were compared to readings taken at locations on either side of the LVDT and plotted on a quadratic curve function. In future similar procedures could potentially be used to monitor structural changes in existing buildings and monuments without the need for physical

contact. This novel technique provides simplicity in deployment compared to traditional structural assessment methods which require access and sensor installation. This method is more cost-effective as the response will be measured without any need for sensors attached to the structure overcoming access problems.

2 State of the Art

The methods used for tracking features in a series of images can be categorised under two different headings, optical and normalised cross correlation.

2.1 Normalised Cross Correlation

For this approach, the region of interest (ROI) was defined and then treated as a sub image of the principal image in the sequence. This method is a coarse fine approach to obtain a pixel level displacement. It is performed by mapping the reference sub image on the deformed sub-image. The normalised cross correlation matrix of the two images is then calculated by the use of the mormxcorr2 [Lewis, 1995] function in Matlab. The peak of this matrix occurs where the sub images are best correlated. If there is a difference in peak locations, the ROI has been displaced and a graph tracking this change can be plotted.

2.2 Optical Flow

The Optical flow (OF) technique is used to calculate an approximation of 3D velocities onto the imaging surface, 2D motion field, using spatiotemporal patterns of images. The surface deflection is extracted by identifying features on the test specimen then tracking them through successive image frames. For the computational time be practical for field usage it is recommended that a region of interest (ROI) is specified in which to extract and track features[Fukuda et al., 2010]. In this study, it was decided that a modified version of the optical flow algorithm would be used to calculate displacements of the test specimen due to its enhanced running time and similar level of accuracy to normalised cross correlation. The method used to extract the features for tracking in this test series was the Harris-Stephens corner and edge detector.

2.2.1 Harris Features

This method uses a combination of edge and corner features to determine points of interest. It measures changes in pixel intensity across an image using an autocorrelation function, and determines the quality and number of edge-corner features in an image[Harris and Stephens, 1988].

The Kanade-Lucas-Tomasi (KLT) algorithm takes the features extracted by the above approaches and constructs a window (W) based on the points. It then compares each image in a sequence, and tracks the displacement of W through these images and maps it to an affine transformation T. This transformation can then be plotted using Matlab and the displacement of our ROI determined.

2.3 Computer Vision Algorithms Commonly used in Civil Engineering

Approaches for Computer vison in SHM can be broken down into several methodologies, which are detailed in the sections below.

2.3.1 Hybrid Camera-Sensor Approaches

Previous research has demonstrated the feasibility of integrating imaging devices with traditional SHM technology [Zaurin and Catbas, 2010]. This method involves using the computer vision to monitor bridge traffic, while an LVDT is mounted to the bottom of the bridge to measure deflection readings. Another system was laid out in [Yan et al., 2008] where readings from a strain gauge were linked video images of vehicles passing over a test bed setup in order to classify the vehicles into 7 different classes.

 ISBN 978-0-9934207-1-9

2.3.2 Target based Camera only approaches

Replacing the traditional sensors such as LVDT/accelerometers with cameras for measuring displacement is the logical next step in this area. Early work in the field involved the use of a target based system for locating features to be tracked on a bridge as it was not possible to extract targets for tracking from natural features. The study detailed in [Shih and Sung, 2013] compared Digital Image Correlation (DIC) readings to verified measurements from accelerometers that had been attached to the test specimen. Additional work was carried out in the field by [Lee and Shinozuka, 2006] with comparable results to LVDT.

2.3.3 Natural Features based Camera only approaches

With dramatic improvements in commercially available digital cameras, it is becoming increasingly possible to develop computer vision systems for deflection monitoring using natural features of the bridge structure. This would enable DIC to be a non-contact full field measurement of displacement system, hence overcoming the access limitations of existing SHM systems. Work has been done in this area by [Feng et al., 2015, Malesa et al., 2010] using differing methods of deflection calculation, with no clear optimal method as of yet being established.

3 EXPERIMENTAL STUDIES

This section describes the methodology employed to experimentally assess the performance, i.e. the sensitivity and accuracy of the optical method in measuring displacement.

3.1 Test Setup

A full scale experimental investigation was carried out at Banagher Precast Concrete Ltd, Co Offaly on a 10m long concrete floor slab which was prestressed with basalt reinforcing bars. The slab was cast on the 22nd of January 2016 and allowed to cure for over 40 days. The slab was then tested on the 16th of March 2016. The slab member was simply supported and the load is applied at mid-span by a +300 mm stroke mono-directional hydraulic jack counteracting on a strong steel reaction frame. A steel repartition box beam was placed in between the jack and the slab.

Figure 1: Testing Rig

Figure 2: Camera Location

Figure 1 shows a picture of the test rig. Three linear variable displacement transducers (LVDT) were placed at the mid-span section, one in the centre and two at the edges. Two additional analogic displacement transducers were placed at about 800 mm from the edges to be able to clear the mid-span deflection from the shortening of the timber slats. A manually controlled hydraulic pump provided with an analogic pressure gauge supplied the pressurised oil to the ram. A Nikon 810 camera was set up 6m from the beam to provide a means of determining the slab deflection using fully contactless methods, as shown in Figure 2.

3.2 Testing Plan

A series of loading cycles were carried out on the slab, the details have been presented in table 1. A load of 10bar was applied to investigate the precracking behaviour; the load was then increased to induce cracking, and finally the load was then increased to failure of the specimen.

Test Number	Pressure applied(bars)	Notes	LVDT
1	0-10-0	Video and photo	YES
2	0-10-0	Video	YES
3	0-30-0	video	YES
4	0-30-0	photo	YES
5	0-30-0	photo	YES
6	0-35-0	video	YES
7	0-40-0	photo	YES
8	0-40-0	photo	YES
9	0-50-0	photo/video	YES
10	0-50-0	photo	YES
11	0-58-0	photo	NO
12	0-58-0	photo	NO
13	0-62-0	video	NO

Table 1: Test Details.

3.3 Algorithm Development

The code used for the calculation of the deflection is based on a modified version of a face tracking algorithm contained in the Image Processing toolbox of Matlab. The Harris feature extraction method was used to find the features to be tracked through the video. A ruler was attached to the test specimen in order to give a pixel-mm calibration for the deflection plots. Affixing the ruler will give an accurate calibration factor for pixel-mm conversion, however a system for determining pixel values based on focal length of camera/distance from target is currently being developed in order to present a truly contactless approach. An area directly above the LVDT was chosen for feature extraction, and the 40 strongest features were chosen to be followed. Once these values had been tracked throughout the series of frames, the Euclidean distance between the features in the original reference frame and each subsequent frame was calculated and averaged over the number of points, with the resulting values plotted in the graphs below.

4 Results

The results for Test 3 are shown in Figure 3, both the LVDT and camera based data have been included. The graph validates that the camera based monitoring corresponds to the established LVDT measurements. The result from the camera based monitoring are continuous for the test duration, however the LVDT readings have been taken at discrete times during the test, the markers indicate the actual readings. An optical flow algorithm has been used for the post processing of the test videos; the graph shows the LVDT recorded a maximum deflection of 3.044mm compared to a calculated deflection of 2.98mm from the vision based method. As the error was less than 2% the vision based method has been validated as a viable method of measuring deflection and the results from the other tests confirm this correlation. In an attempt to reduce the error below 2% a number of different post processing methods were then applied to the video image from test 3. As previously stated the optical flow algorithm requires a ROI to be chosen for analysis, the results presented in Figure 3 were determined for a ROI directly above the LVDT location but the following results have been determined by various methods from the ROI highlighted in Figure 4 with ROI3 being located directly over the LVDT. Initially the results from ROI 1,2,4 & 5 were averaged, they were subsequently plotted on a graph and a 3rd

 ISBN 978-0-9934207-1-9

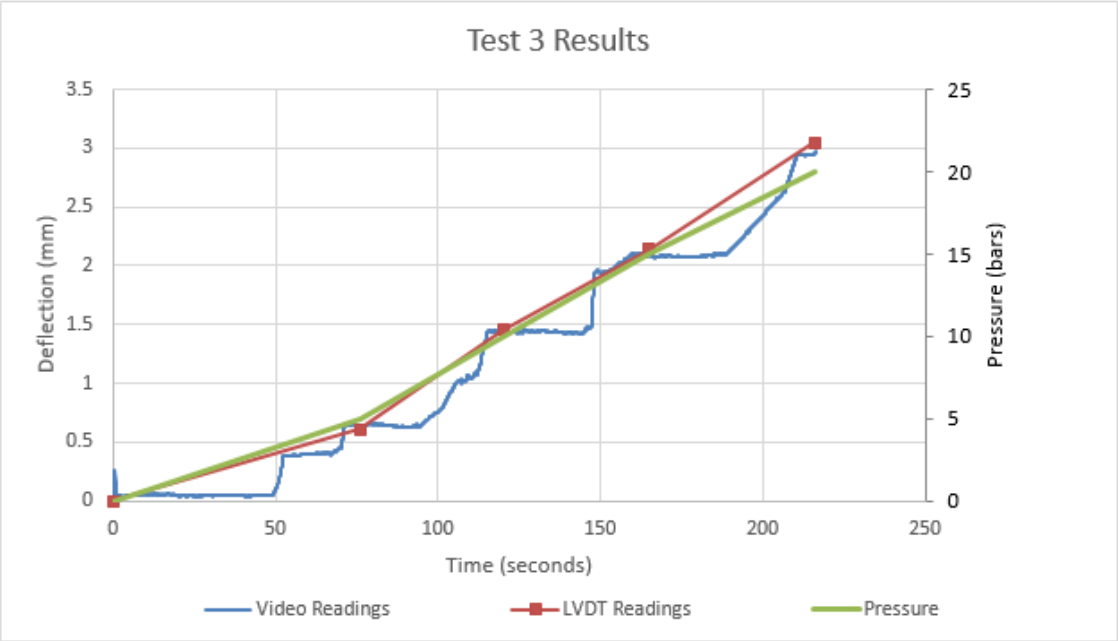


Figure 3: Test 3 Results

order polynomial trend line was added to determine the midpoint deflection. The results for the maximum deflection are presented in Table 2.

ROI	Vision	LVDT	Error
Above LVDT	2.99	3.04	2%
Average 1,2,4 & 5	3.00	3.04	2%
Polynomial	2.99	3.04	2%

Table 2: Calculated Deflections(mm).

The data in Table 2 shows that an error still exists between the LVDT and camera based deflections, the most accurate calculated deflection was determined from ROI 3(the region directly above the LVDT), where a vision based deflection of 3.03mm was calculated eliminating the error. Based on this it was assumed that since the camera was not perpendicular to the test specimen the accuracy of pixel to mm calibration was inversely proportional to the distance between the ROI and the ruler. Further modifications were then carried out to attempt to minimise this error, the post processing algorithm was then adjusted to compensate for this by applying a normalisation factor to the x coordinates of the measured deflection points. Deflections from the 5 ROI were then recalculated and the results are presented in Table 3. The data in Tables 2 and 3 confirm that the tested method did not

Figure 4: ROI Locations.

ROI	Vision	LVDT	Error
Above LVDT	2.95	3.04	3%
Average 1,2,4 & 5	2.96	3.04	3%
Polynomial	2.95	3.04	3%

Table 3: Calculated Deflections after Perspective Correction(mm).

improve the accuracy of the deflection calculation, based on all of the findings included in this paper it has been

 ISBN 978-0-9934207-1-9

determined that the most accurate deflection calculation can be detected by locating the reference measurement scale close to the desired ROI. Future work is now underway to quantify the effect of varying the camera angle in relation to the test specimen. A series of lab trials will be used and the findings will then be applied to the monitoring of a real bridge structure on our regional road network. These initial trials have indicated that camera based monitoring has the potential to provide accurate deflection measurements and can be used as a suitable alternative to LVDTs. The applicability of this can be seen in the data obtained from the two fail tests in this experimental program. During tests 12 & 13 the LVDTs were removed from the test set up as there was a risk of total failure which would result in critical damage to the equipment. However as the deflection measurements were still of significant interest the vision based monitoring was carried out during these tests. The results for test 12 have been presented in Figure 5.

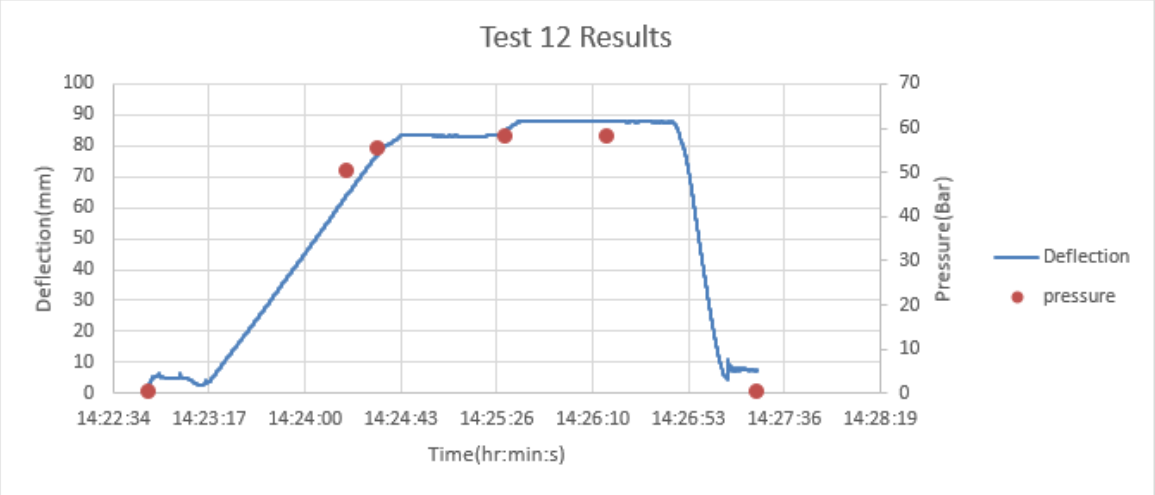


Figure 5: Test 12 Results

5 Conclusions

The results presented in this report confirm vision based monitoring to be a viable method of tracking deflection. This approach has been validated and provided results in a testing situation which would not have been possible using LVDTs. Based on the application of vision based measurement across other engineering disciplines significant further work is now required to realise the full potential of vision based monitoring for civil and structural applications.

Acknowledgments

The experimental activity has been performed within the objectives of a US-Ireland research project, funded by the Invest Northern Ireland, Science foundation Ireland and the National Science Foundation. The technicians of Banagher Precast Concrete and the Eirocrete research project that developed the test specimen are acknowledged, especially Bruno Dal Lago, Peter Deegan and Philip Crossett.

References

[Feng et al., 2015] Feng, M. Q., Fukuda, Y., Feng, D., and Mizuta, M. (2015). Nontarget Vision Sensor for Remote Measurement of Bridge Dynamic Response. *Journal of Bridge Engineering*, 20(12).

[Fukuda et al., 2010] Fukuda, Y., Feng, M. Q., and Shinozuka, M. (2010). Cost-effective vision-based system for monitoring dynamic response of civil engineering structures. *Structural Control and Health Monitoring*, 17(8):918–936.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector.

[Lee and Shinozuka, 2006] Lee, J. J. and Shinozuka, M. (2006). Real-Time Displacement Measurement of a Flexible Bridge Using Digital Image Processing Techniques. *Experimental Mechanics*, 46(1):105–114.

[Lewis, 1995] Lewis, J. (1995). Fast Template Matching.

[Malesa et al., 2010] Malesa, M., Szczepanek, D., Kujawińska, M., Świercz, A., and Kołakowski, P. (2010). Monitoring of civil engineering structures using Digital Image Correlation technique. In *EPJ Web of Conferences*, volume 6, page 31014. EDP Sciences.

[Ojio et al., 2016] Ojio, T., Carey, C., OBrien, E., Doherty, C., and Taylor, S. (2016). Contactless Bridge Weigh-in-Motion. *Journal of Bridge Engineering*, page 4016032.

[Shih and Sung, 2013] Shih, M.-H. and Sung, W.-P. (2013). Developing Dynamic Digital Image Techniques with Continuous Parameters to Detect Structural Damage. *The Scientific World Journal*, 2013:453468.

[Tomasi and Kanade, 1991] Tomasi, C. and Kanade, T. (1991). Detection and Tracking of Point Features.

[Yan et al., 2008] Yan, L., Fraser, M., Elgamal, A., Fountain, T., and Oliver, K. (2008). Neural Networks and Principal Components Analysis for Strain-Based Vehicle Classification. *Journal of Computing in Civil Engineering*, 22(2):123–132.

[Zaurin and Catbas, 2010] Zaurin, R. and Catbas, F. N. (2010). Integration of computer imaging and sensor data for structural health monitoring of bridges. *Smart Materials and Structures*, 19(1):015019.

An eye movement study on visual perception of holographic, stereoscopic, and 2D images

T.M. Lehtimäki,[1] M. Niemelä,[2] R. Näsänen,[3] R.G. Reilly,[1] and T.J. Naughton[1]

[1]*Department of Computer Science, Maynooth University, Maynooth, Co. Kildare, Ireland*
[2]*Former affiliation: University of Oulu, Oulu Southern Institute, Nivala, Finland*
[3]*Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland*

Abstract

We study observers looking at a three-dimensional (3D) scene captured in a traditional glass plate hologram using eye-tracking and compare this with stereoscopic and 2D images. Our results can guide development of future digital holographic displays. We have studied how five participants move their eyes when viewing particular features at different depths in holograms, and how the eye movement patterns differ from viewing conventional stereo image stimuli.

Keywords: Holographic display, visual perception, eye movements, binocular, stereopsis

1 Introduction

Digital holography [Goodman and Lawrence, 1967] has always sought higher resolution devices. Recent advances in 3D digital holographic displays attempt to combine multiple display devices to improve the field of view, stereo parallax, and motion parallax [Kujawinska et al., 2014]. However, digital cameras and digital display devices still do not match the resolution and physical size of traditional glass plate holograms. This implies that one must make compromises when designing such a system. In digital holographic video transmission and optical display, the technical choices, compromises, and optimizations should be driven by our understanding of visual perception [Yaroslavsky, 2008]. There have been perceptual studies on digital holographic data displayed on stereoscopic displays [Näsänen et al., 2010]. However, visual perception studies with optical holographic displays have been mostly concerned with subjective analysis [Barabas and Bove Jr, 2013] or mathematical analysis of their capabilities [Finke et al., 2015]. One reason for this is the small object reconstruction lateral size with current holographic displays causing a small field of view, reduced stereo parallax, and reduced motion parallax. A laser on the display side brings with it additional speckle and eye safety challenges.

We wish to bridge this gap by studying visual perception of traditional glass plate holograms which fulfill all properties of a holographic display for still images. By using glass plate holograms we can have a visually rich scene at a size which gives possibilities to study binocular parallax and measure eye movements. The purpose of this study is to quantitatively analyse how people look at a 3D scene reconstructed by a holographic display. We measure how people move their eyes when viewing holograms and how the binocular eye movement patterns differ from viewing 2D and stereo stimuli. We also investigate how the two eyes work together when viewing particular details at different depths and what differences and similarities there are when viewing stereo images and holograms.

2 Eye tracking experiment

The experiment was performed with 5 subjects. 2D and stereo stimuli were presented on a stereoscopic display (24 in. Hyundai W240S), which was viewed with circular polarizing glasses when used in stereo mode. In

 ISBN 978-0-9934207-1-9

order to collect the eye movement data, a binocular eye tracker (Eye Link II) was used. For evaluating the glass plate holograms, the set up shown in Figure 1(a) was constructed. Seven different holograms were used as hologram stimuli. An example is shown in Figure 1(b). 2D and stereo image stimuli were created by taking photographs of the holograms from two appropriate viewing angles.

3 Results and discussion

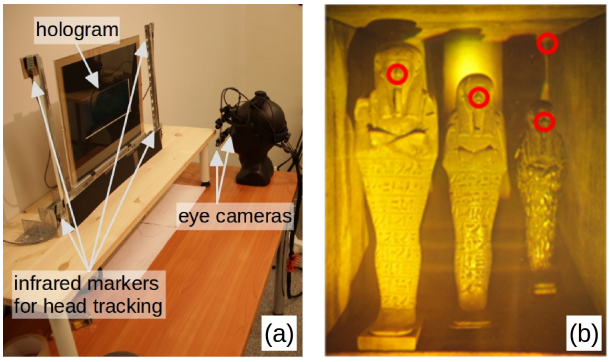


Fig. 1: (a) Eye tracking set up for glass plate holograms. (b) Example stimulus showing four interest areas.

We found that the convergence variation between eyes was greater when viewing the hologram stimuli than equivalent stereo or 2D stimuli. The subjectively evaluated perceived depth variation was also greater for hologram stimuli than stereo or 2D stimuli. As an example result, the range of disparity values for Subject 3 for the "Airplane" stimulus for 2D were 117 and -5, for stereo 121 and -14, and for hologram 201 and -147. These values denotes the difference between right and left eye focus points, respectively, measured in pixels in the calibration plane. The negative and positive values perceptually mean that an object appeared in front of, and behind, respectively, the calibration plane (the screen). Our long term aim is to contribute to the understanding of what are the visual perception requirements for the new generation of 3D holographic displays. Technology has not advanced sufficiently yet to allow a high quality holographic display for unrestricted viewing of dynamic real world scenes with a wide viewing angle. By studying the human visual system's response to glass plate holograms, we can determine what technical compromises have the least impact on unrestricted viewing on future digital holographic displays. Further considerations for holographic displays include the inherent properties of multiple accommodation, monocular parallax, and monocular cues such as blur, perspective, and occlusion.

Acknowledgements. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant no. 13/CDA/2224 and the European Community's Seventh Framework Programme FP7/2007–2013 under grant agreement no. 216105.

References

[Barabas and Bove Jr, 2013] Barabas, J. and Bove Jr, V. M. (2013). Visual perception and holographic displays. In *Journal of Physics: Conference Series*, volume 415, page 012056. IOP Publishing.

[Finke et al., 2015] Finke, G., Kujawińska, M., and Kozacki, T. (2015). Visual perception in multi slm holographic displays. *Applied Optics*, 54(12):3560–3568.

[Goodman and Lawrence, 1967] Goodman, J. W. and Lawrence, R. (1967). Digital image formation from electronically detected holograms. *Applied physics letters*, 11(3):77–79.

[Kujawinska et al., 2014] Kujawinska, M. et al. (2014). Multiwavefront digital holographic television. *Optics Express*, 22(3):2324–2336.

[Näsänen et al., 2010] Näsänen, R. et al. (2010). Presentation and perception of digital hologram reconstructions of real-world objects on conventional stereoscopic displays. In *Information Optics and Photonics*, pages 129–142. Springer.

[Yaroslavsky, 2008] Yaroslavsky, L. P. (2008). Computer-generated holograms and 3-d visual communication. *J. Holography Speckle*, 5:1–6.

Classification of images using semi-supervised learning and structural similarity measure

H. Cecotti and B. Gardiner

Intelligent Systems Research Centre, Ulster University, Londonderry, UK

Abstract

In this paper, we evaluate the performance of graph-based semi supervised learning (SSL) for the classification of images, by using the structural similarity index measure (SSIM) to build the adjacency matrix of the graph. Performance evaluation was carried out with the TID2013 database. The results support the conclusion that SSIM can be efficiently used with graph-based SSL to retrieve images that are similar.

Keywords: Image processing, Semi-supervised learning, Structural Similarity Measure.

1 Introduction

Image quality assessment (IQA) has become an important issue in applications dealing with large number of images. IQA aims to use computational models to measure the image quality consistently with subjective evaluations. The most efficient IQA methods take into account the human visual system (HVS) and image features based on the luminance, the contrast, and the frequency content. The most famous IQA is the structural similarity (SSIM) index. This technique is motivated by the need to capture the loss of structure in the image. The main assumption in SSIM is that HVS is highly adapted to extract the structural information from the visual scene. Such a distance can be exploited in classifiers that are based on the estimation of distances. Most of the semi-supervised learning approaches rely on the cluster assumption, which assumes that examples associated to the same cluster, or the same group of clusters, will share the same label. The techniques rely also on the manifold assumption. It considers that examples that are close to each other will have the same label. The label prediction of an example x will depend on both the labelled and unlabelled examples that are very close to x. Depending on the data, these two assumptions can be difficult to satisfy. The goal of this paper is to combine graph-based semi-supervised learning and the SSIM index in order to build the graph that is used for label propagation. A key issue is to determine the ideal size of the neighbourhood and to what extent different image deformations may provide a bridge across images of the same class.

2 Methods

The SSIM index is based on the computation of three terms (the luminance, the contrast, and the structural term) [Wang et al., 2004]. For the graph-based SSL method, we have considered label propagation. This method requires only the size of the neighbourhood (k) to create the adjacency matrix. A graph $g = (V, E)$ is defined by the nodes $V = \{1, \ldots, n\}$, which represent all the n examples of a training database $X = \{x_1, \ldots, x_n\}$, and edges E, which represent the similarities between examples. The similarities are typically represented by a weight matrix $W \in \mathbb{R}^{n \times n}$. A cell $W(i, j)$ corresponds to the similarity between the example x_i and x_j, i.e., the edge (i, j) in E. If x_i and x_j are close to each other (they belong to the same neighbourhood), then $W(i, j)$ has a non-zero value. In this study, $W(i, j)$ is estimated with SSIM. For the evaluation of the method, we have also considered the mean square error (MSE) as a distance between two images. Furthermore, $W(i, j)$

 ISBN 978-0-9934207-1-9

is set to 1 if an image belongs to the k^{th} closest neighbours, otherwise it is set to 0. In a multiclass problem, the label propagation algorithm is used for each class (one vs. all), and then the results are combined to determine the class of each example [Bengio et al., 2006]. We have used the TID2013 database (available at http://ponomarenko.info/tid2013.htm). TID2013 contains 25 colour images, and each image is deformed into 120 images. We consider here all the deformed images (3000). For the evaluation, we consider one image per class, i.e. one of the 120 deformed image for each image template.

3 Results

The matrices representing the distances between each image using SSIM and MSE are depicted in Fig. 1. For SSIM, each value is between 0 and 1. Values close to 1 represent a high similarity. For MSE, values (in the order of x10^4) close to 0 represent a high similarity between two images. The evolution of the accuracy in relation to the size of the neighbourhood (k) for the creation of the adjacency matrix highlight the importance of this parameter. With a large neighbourhood, all the deformed examples of an image are clustered together and the labels can be easily propagated. With $k = 110$, all the labels are perfectly propagated, i.e., with 25 labelled images, it is possible to label the remaining 2975 images with a perfect accuracy. However, when k is small (e.g. 20), the performance is significantly lower as there is no bridge allowing the labels to be propagated from one type of deformations to another.

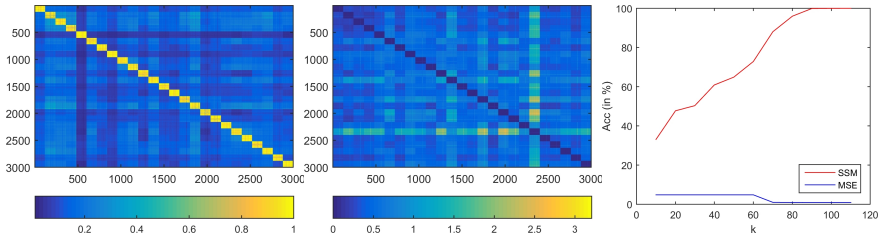


Figure 1: Matrices representing the distances between each image (**left**: SSIM **middle**: MSE). Accuracy (in %) in relation to the size of the neighbourhood (k) in the graph-based SSL method (**right**).

4 Conclusion

IQA techniques such as SSIM do not require any training for estimating the difference between images, therefore suitable for large databases that can contain sets of images that are very similar, and where it is not required to cluster images based on high level features (e.g. semantic content). In this study, we have considered the TID2013 database and evaluated to what extent graph based semi-supervised learning could be used to retrieve the label of all the deformed images based only on a randomly selected image of each class. Further work will include the analysis of other distances and the evaluation of the type of deformations that can be managed.

References

[Bengio et al., 2006] Bengio, Y., Dellalleau, O., and Roux, N. L. (2006). Label propagation and quadratic criterion. In O. Chapelle, B. S. and Zien, A., editors, *Semi-supervised learning*, pages 35–58. MIT Press.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error measurement to structural similarity. *IEEE trans. Image Processing*, 13(1).

Image quality assessment through brain signal analysis

H. Cecotti and B. Gardiner

Intelligent Systems Research Centre, Ulster University, Londonderry, UK.

Abstract

The presence of noise in images has a key impact on the difficulty of visual target detection tasks. In this study, we propose to explore the potential of brain signal analysis for discriminating the level of noise in a sequence of images. The data were recorded using a magnetoencephalography from four healthy individuals during a rapid serial visual presentation task in four conditions corresponding to four different levels of noise in the images. The results indicate a clear link between behavioural performance, single-trial detection, and noise level.

Keywords: Noise estimation, Brain-Machine Interface, Magnetoencephalography.

1 Introduction

Brain-Machine Interface (BMI) systems have been mainly used as a new means of communication for severely disabled people, and for rehabilitation [Millán et al., 2010]. BMIs based on the detection of event-related potentials (ERPs) typically require subjects to pay attention to a specific sequence of stimuli in order to produce a robust and detectable neural response. In a Rapid Serial Visual Presentation (RSVP), a rapid sequence of images are presented sequentially to subjects in the same location on a screen. The stream of images contains different types of visual stimuli, which can be classified as targets or non-targets [Pohlmeyer et al., 2011]. In this paper, we propose to investigate the use of brain signal analysis with magnetoencephalography for estimating the effect of the noise in images during a target detection task, and its relationship with behavioural performance.

2 Methods

Four healthy volunteer subjects (s1-s4) participated in the study. Each participant provided written informed consent, reported normal or corrected-to-normal vision, and no history of neurological problems. The experimental protocol was reviewed by the Faculty Ethics Filter Committee of Ulster University, and was in accordance with the Helsinki Declaration of 1975, as revised in 2000. Participants to the experiment had to perform a rapid serial visual presentation task (speed at 2 Hz) where different graylevel images of people (men and women) were presented on the screen. The task was to press a button each time an image of a woman was presented on the screen. Four sessions were recorded. Each session corresponded to a different level of noise in the background of the images (L1: low noise, to L4: high noise). The noise was generated through a uniformly distributed random numbers (see Fig. 1). The data was recorded with an Elekta Neuromag 306-channel MEG system at the Intelligent Systems Research Centre (ISRC), Ulster University, Londonderry, UK. The MEG signal was recorded with a sampling rate of 1 kHz using 204 planar gradiometers and 102 magnetometers, based on thin-film technology. Five head position indicator (HPI) coils were placed on the head to determine how close the head is to the sensors that are collecting the signal. The analysis of the brain response evoked by the presentation of a target (i.e., the image of a woman) was obtained by the area under the ROC curve of the single-trial classification of the different images (i.e., men vs. women). The signal was first bandpassed

 ISBN 978-0-9934207-1-9

Table 1: Behavioural performance and Area under the curve (AUC) for single-trial performance for the four conditions (L1, L2, L3 and L4). Each couple represents the hit rate and the precision of the response.

	Behavioural				Single-trial			
	L1	L2	L3	L4	L1	L2	L3	L4
s1	98.33/96.72	100.0/95.24	100.0/98.36	95.00/95.00	0.964 ± 0.062	0.983 ± 0.029	0.984 ± 0.025	0.910 ± 0.090
s2	98.33/98.33	96.67/95.08	81.36/87.27	45.00/87.10	0.941 ± 0.090	0.981 ± 0.019	0.888 ± 0.077	0.793 ± 0.099
s3	100.0/64.368	90.00/81.818	78.33/92.157	68.33/95.35	0.991 ± 0.017	0.912 ± 0.067	0.943 ± 0.084	0.827 ± 0.074
s4	82.76/34.783	68.52/28.682	50.85/28.302	30.91/20.99	0.856 ± 0.077	0.814 ± 0.138	0.660 ± 0.160	0.592 ± 0.156
Mean	94.86/73.55	88.80/75.21	77.63/76.52	59.81/74.61	0.938	0.922	0.869	0.781
SD	8.10/30.21	14.14/31.65	20.27/32.47	28.08/35.95	0.059	0.080	0.145	0.135

between 0.1 and 10.40 Hz, then downsampled to 31.25 Hz. After spatial filtering, the sets of features from each evoked response, considering a time segment of 800 ms post stimulus, were classified with a stepwise linear discriminant analysis classifier.

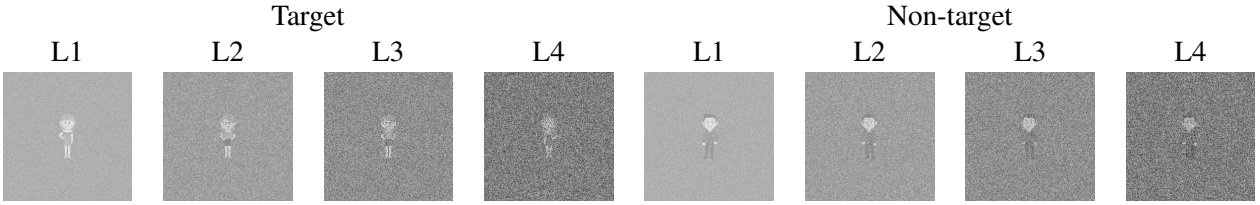


Figure 1: Examples of images for the different levels of noise.

3 Results

The results corresponding to the behavioural response and the AUC are given in Table 1. The mean hit rate decreases in relation to the increase of the difficulty of the task, from 94.86% to 59.81%, however the precision seems to not be changed by the level of noise. The mean AUC across subjects decreases in relation to the noise level in the condition, from 0.938 with the condition L1 that has the less noise, to 0.781 for the condition L4 that has the noisiest stimuli.

4 Conclusion

The level of noise in an image can have different impacts on the observers. Image quality assessment through brain signal analysis can provide a novel way to analyse the level of noise in images and describe to what extent it can impact the performance of target detection tasks. Future works will include the addition of more subjects to demonstrate the interest of the approach in relation to common image processing techniques.

References

[Millán et al., 2010] Millán, J. d. R., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., Neuper, C., Müller, K.-R., and Mattia, D. (2010). Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in Neuroscience*, 4(161):1–15.

[Pohlmeyer et al., 2011] Pohlmeyer, E. A., Wang, J., Jangraw, D. C., Lou, B., Chang, S., and Sajda, P. (2011). Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases. *J. Neural Eng.*, 8:036025.

 ISBN 978-0-9934207-1-9

An observation regarding multiple correct decryption keys in optical image encryption

Lingfei Zhang, Thomas J. Naughton

Department of Computer Science, Maynooth University–National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

Abstract

The strength of optical image encryption techniques, such as double random phase encoding, has been routinely measured using statistical and heuristic attacks. The most common attack scenario is a known plaintext attack in which an attacker tries to infer the Fourier plane encryption phase key from a 'plaintext' input and 'ciphertext' encrypted image pair. For intensity-encoded inputs, it is well known that there are multiple correct keys in the keyspace. This was previously considered an advantage to an attacker because finding any one of the correct keys was sufficient. However, we propose that this is the reason for the well-documented performance limitations of heuristic attacks. For a small sample of inputs, we have observed that a class of heuristic attack performs suboptimally when its keyspace contains multiple equivalent solutions. We propose that a solution to the problem of underperforming heuristic attacks might be achieved by transforming the keyspace so that it contains only one correct solution. We illustrate this with an example of an idealised keyspace with only one correct key. As a result, in our limited tests, the performance of the heuristic attack is significantly improved.

Keywords: Optical information processing, Optical image processing, Optical image encryption

1 Introduction

Complex-valued images encoded in cross-sections of laser beams can have their phase values modified by common pixellated devices such as LCD and LCoS panels, and can undergo Fourier transformation using simple lenses or free-space propagation. This is the basis for image encryption using optics. Optical image encryption offers the promise of high speed parallel encryption of image data (with frame rates as fast as the employed optoelectronic displays and cameras can support), and it has been widely studied in the field of optical image processing. Since the classic double random phase encoding (DRPE) algorithm was proposed by Réfrégier and Javidi [Refregier and Javidi, 1995], there has been significant interest in this field [Javidi et al., 1997] [Goudail et al., 1998, Unnikrishnan et al., 1998, Unnikrishnan et al., 2000, Nomura and Javidi, 2000] [Situ and Zhang, 2004, Cheng et al., 2008, Situ et al., 2008, Alfalou and Mansour, 2009].

Security analyses of the system have shown it to be vulnerable to both chosen-plaintext attacks [Frauel et al., 2007] and chosen-ciphertext attacks [Carnicer et al., 2005], in which an attacker needs to be able to acquire access to the system or to induce an authenticated user to encrypt or decrypt some specific plaintexts and ciphertexts. As a specific example, if an impulse function can be induced to be encrypted or decrypted, the exact keys in decryption can be readily obtained. In response, impulse-free architectures have been proposed to prevent this type of attack [Kumar et al., 2009b, Kumar et al., 2009a]. More sophisticated and practical attacks on DRPE have followed, for cases where the attacker has intercepted one or more plaintext-ciphertext pairs, but does not have access to the optical architecture. A early proposed known-plaintext attack introduced the use of a phase retrieval algorithm to estimate approximate decryption keys [Peng et al., 2006]. A variety of phase retrieval algorithms followed for known-pair and ciphertext-only

 ISBN 978-0-9934207-1-9

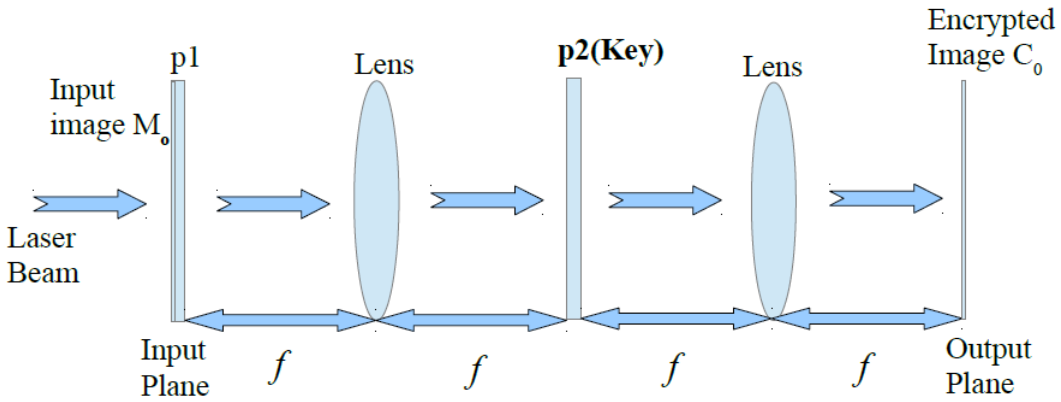


Figure 1: The most studied optical image encryption algorithm: symmetric DRPE, showing the location of the two phase-encoded image masks (p_1 and p_2) that constitute the encryption key (and the decryption key). Image mask p_1 is located immediately after the input image (which effects a pointwise multiplication between the input and p_1). Image mask p_2 is located in the Fourier plane, where it will be pointwise multiplied by the product of the input and p_1. A second Fourier transform returns the encrypted image to the space domain.

cases [Situ et al., 2007, Liu et al., 2015]. Frauel et al. [Frauel et al., 2007] demonstrated that by using only two known-plaintext pairs, the exact decryption keys can be found by solving a linear system of equations. To resist a selection of previously successful attacks, conventional cryptographic modes have been introduced into optical encryption [Naughton et al., 2008]. Cheng et al. [Cheng et al., 2008] presented a security enhanced DRPE encryption scheme that uses a complex-valued mask rather than a phase mask, with specially restricted amplitude values. A known-plaintext attack has yet to be found for this arrangement.

The exponential growth of decryption keyspace ensures that it is too computationally expensive to perform an exhaustive search. Analysis of the keyspace [Monaghan et al., 2007, Monaghan et al., 2008] [Nakano et al., 2013] has justified heuristic search strategies. Partial searches of the keyspace, such as simulated annealing [Gopinathan et al., 2006, Liu et al., 2009], have allowed an attacker to efficiently find approximations to the decryption keys. However, to date, no statistical attack has convincingly demonstrated that the approximated decryption keys found through a known-plaintext attack will reliably decrypt other images encrypted with the same encryption key.

2 Double random phase encoding

DRPE (see illustration in Fig. 1) is a symmetric encryption system, which means the encryption and decryption keys are simple functions of one another. The implementation of the encryption procedure is that the input image f is sequentially multiplied by two statistically independent random masks, each followed by a Fourier transform, equivalent to

$$\Psi(x,y) = \mathscr{F}\{f(x,y)p_1(x,y)\} * p_2(x,y), \tag{1}$$

where $\mathscr{F}$ means Fourier transformation, $*$ means convolution, and p_1 and p_2 (jointly, the encryption key) are two randomly generated phase masks with dimensions equal to f. Because the second Fourier transform does not affect the security of the system, we can choose to omit it from Eq. (1). For approximate decryption, the normalized root mean squared (NRMS) error is used to quantify the amount of error in the imperfect decryption output. It is worthwhile to note some generally agreed error thresholds. For example, an NRMS error if 0 means perfect decryption. In general, the NRMS error <= 0.1 is accepted as acceptable decryption. The NRMS error

 ISBN 978-0-9934207-1-9

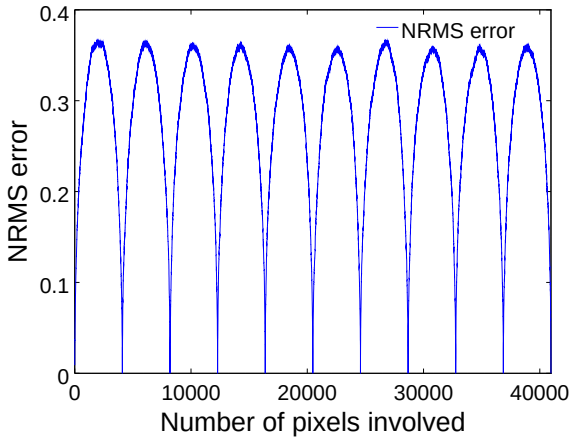


Figure 2: A systematic search of a subset of the keyspace to find multiple different but equivalent keys (see explanation in body of paper) that achieve perfect intensity-based decryption. A 64×64 pixel phase mask used to encrypt an image is modified by sequentially incrementing each pixel by $2\pi/Q$, and then repeating the procedure, until $64 \times 64 \times Q$ separate phase masks are generated. Each one is used to decrypt, and the resulting NRMS error plotted. The figure shows an example for $Q = 10$, where exactly Q phase masks yield an NRMS error of 0. Approximately 1700 phase masks yield an NRMS error less than 0.1 in this subset of the keyspace.

> 0.4 denotes an unsuccessful decryption result. NRMS is defined as

$$\mathbf{NRMS} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} |I_{\mathrm{d}}(i,j) - I(i,j)|^2}{\sum_{i=1}^{N}\sum_{j=1}^{M} |I(i,j)|^2}} \tag{2}$$

where I_{d} is the intensity of the decryption output. For real-valued inputs, the intensity of the decryption image I_{d} can be found without the use of the input plane mask p_1. In this context, the Fourier domain mask p_2 can be regarded as the only key in the system.

The keyspace of the system is dependent on the dimensions (M, N) of p_2 and the set of quantization levels Q, by $\mathrm{K} = |Q|^{MN}$. Q is the set of different phase levels possible for each mask pixel, and for intensity-encoded inputs $|Q|$ is the number of different but equivalent keys that achieve perfect decryption [Monaghan et al., 2007, Liu et al., 2009], as shown in Fig. 2. Given the original mask $p_2(x, y)$, each of the $|Q|$ masks $(p_2(x, y) + a) \bmod 2\pi$ (where $a \in Q \subset [0, 2\pi)$ is a constant phase value to be added to each pixel) is a mask that will achieve perfect decryption.

3 Simulated annealing

The simulated annealing (SA) algorithm is a strategy of global optimization that has been applied in many fields, and which also is appropriate for testing the practical strength of optical encryption (through simulated attacks) because the keyspace is so large.

This heuristic attack has been performed with two different classes of input: binary A_{b} and greyscale A_{gs} images with 64×64 pixels. Example plaintext inputs are shown in Fig. 3(a) and (b). Each plaintext input was encrypted with a uniformly randomly-generated key k with 8 bits (Q = 256) to generate a ciphertext image, and the plaintext-ciphertext pair was used to approximate k by SA. The SA algorithm was run 20 times, with independent uniformly randomly-generated starting values for k, and the running times are shown in Fig. 4. A Dell Optiplex 780 desktop PC was used for the simulations, with an Intel® Core™2 Duo E7500 CPU and 4 GB of RAM, running Ubuntu 14.04 Linux. Optical encryption was simulated with the Python programming language, using the scikit-image image processing library [van der Walt et al., 2014]. For the binary input

 ISBN 978-0-9934207-1-9

Figure 3: Original and decrypted 64×64 pixel images. (a) and (b): the plaintext part of the first two plaintext-ciphertext pairs, binary A_{b} and greyscale A_{g}, respectively, (c) and (d): the plaintext part of the second two plaintext-ciphertext pairs, binary B_{b} and greyscale B_{g}, respectively, (e) the result of decrypting the encrypted version of B_{b} using the key found with A_{b} yielding NRMS error of 0.370, (f) the result of decrypting the encrypted version of B_{g} using the key found with A_{g} yielding NRMS error of 0.844.

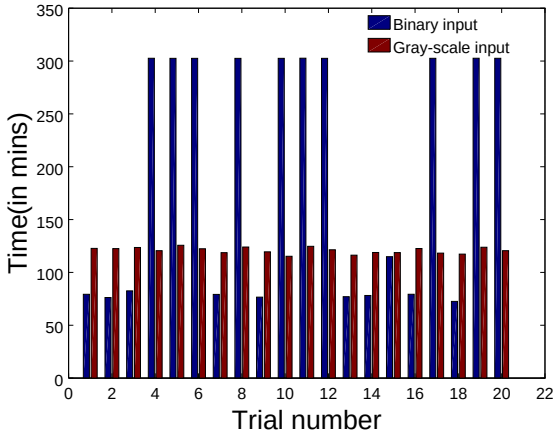


Figure 4: Running time for the known-plaintext SA algorithm on a modest desktop PC. A time limit of 300 minutes was imposed on SA to find keys yielding a NRMS error of 0.1. In cases where the time limit was reached, the keys yielded NRMS errors in range $[0.1, 0.2]$.

 ISBN 978-0-9934207-1-9

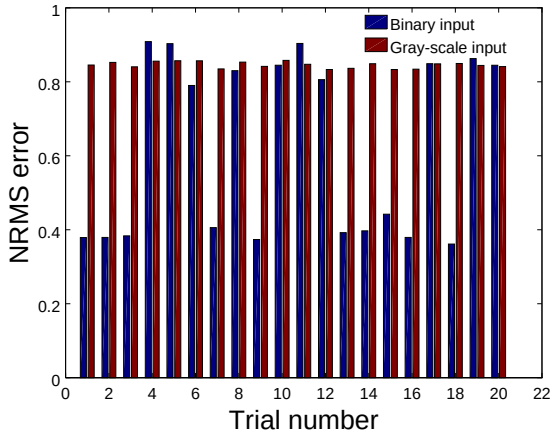


Figure 5: NRMS errors when decrypting the encrypted versions of B_b and B_g, using the approximated decryptions keys from plaintext-ciphertext pairs associated with A_b and A_g, respectively. Each trial number corresponds to the trial numbers in Fig. 4.

image shown, half of the 20 trials successfully found a key that yielded an NRMS error of 0.1, with those successful 10 trials taking an average time of 82 minutes (standard deviation of 12). This was typical for all binary images tested. For the greyscale image shown, all 20 trials successfully found keys that yielded NRMS errors of 0.1, taking an average time of 121 minutes (standard deviation of 3), also typical for all greyscale images tested.

Each of the 20 binary (greyscale, respectively) keys referred to in Fig. 4 is then used to decrypt a second binary image B_b (greyscale image B_g, respectively) that was encrypted with the same encryption key. Examples of these second images are shown in Figs. 3(c) and (d). The resulting decryption errors are shown in Fig. 5. For binary images, where SA did find an approximate key for A_b that yielded an NRMS error of 0.1, the decryption of an encrypted version of B_b with that key consistently yielded an error of approximately 0.4 (see Fig. 3(e)). This is considerably higher than would be expected for a key that yielded an NRMS error of 0.1 during the SA stage. With greyscale images, the NRMS errors for B_g is even worse, with all decrypted images yielding an NRMS error of 0.8 (see Fig. 3(f)).

We then artificially constrained the keyspace (as explained in [Zhang and Naughton, 2016]) so that SA could find only one of the possible correct keys shown in Fig. 2. The same input images A_b, A_g, B_b and B_g (see Fig. 6(a)-(d)) were reused to test this modified-keyspace version of the heuristic attack. Of the 40 trials performed, the average time taken to obtain keys yielding an NRMS error of 0.1 was 4.7 minutes and 5.7 minutes for the binary and greyscale classes, respectively. In addition, when these keys were used to decrypt the encrypted versions of the second images B_b and B_g, all yielded NRMS errors were in the range $[0.1, 0.15]$ (examples shown in Fig. 6(e) and (f)).

4 Conclusion

We have observed from a small sample of images that the well-known statistical technique, SA, for attacking intensity-based DRPE has the interesting behaviour that the encryption key k approximated using one plaintext-ciphertext pair may not be suitable to faithfully decrypt other plaintext images that were encrypted with the same key. While this was known for binary images, we have observed that the effect is even more pronounced for a small sample of greyscale images.

In this paper, we have observed with a limited sample set, that by artificially transforming the keyspace to have one unique solution, SA can more quickly find a key, and that this key is consistently more reliable in decrypting other plaintext images that were encrypted with the same key. We do not have a convincing explanation for our observation. Also, the observation will have to be repeatable for a wide range of images, in order to convince the community of its usefulness. However, if it is we believe that this experiment elucidates the

Figure 6: The experiment from Fig. 3 is repeated, this time for an artificial keyspace with only one solution, using the same images (a)-(d). However, in (e) and (f), the decryption of the encrypted versions of images B_{b} and B_{g} yield much lower NRMS errors of 0.09 and 0.11, respectively.

reason for the poor performance previously documented for SA. As such, it will be important to find a realistic technique to transform an intensity-based DRPE's multi-solution keyspace to a single-solution keyspace.

Acknowledgements. This publication has emanated from research conducted with the financial support of an Irish Research Council (IRC) Postgraduate Scholarship and of Science Foundation Ireland (SFI) under grant no. 13/CDA/2224.

References

[Alfalou and Mansour, 2009] Alfalou, A. and Mansour, A. (2009). Double random phase encryption scheme to multiplex and simultaneous encode multiple images. *Appl. Opt.*, 48(31):5933–5947.

[Carnicer et al., 2005] Carnicer, A., Montes-Usategui, M., Arcos, S., and Juvells, I. (2005). Vulnerability to chosen-cyphertext attacks of optical encryption schemes based on double random phase keys. *Opt. Lett.*, 30(13):1644–1646.

[Cheng et al., 2008] Cheng, X. C., Cai, L. Z., Wang, Y. R., Meng, X. F., Zhang, H., Xu, X. F., Shen, X. X., and Dong, G. Y. (2008). Security enhancement of double-random phase encryption by amplitude modulation. *Opt. Lett.*, 33(14):1575–1577.

[Frauel et al., 2007] Frauel, Y., Castro, A., Naughton, T. J., and Javidi, B. (2007). Resistance of the double random phase encryption against various attacks. *Opt. Express*, 15(16):10253–10265.

[Gopinathan et al., 2006] Gopinathan, U., Monaghan, D. S., Naughton, T. J., and Sheridan, J. T. (2006). A known-plaintext heuristic attack on the fourier plane encryption algorithm. *Opt. Express*, 14(8):3181–3186.

[Goudail et al., 1998] Goudail, F., Bollaro, F., Javidi, B., and Réfrégier, P. (1998). Influence of a perturbation in a double phase-encoding system. *J. Opt. Soc. Am. A*, 15(10):2629–2638.

[Javidi et al., 1997] Javidi, B., Zhang, G., and Li, J. (1997). Encrypted optical memory using double-random phase encoding. *Appl. Opt.*, 36(5):1054–1058.

[Kumar et al., 2009a] Kumar, P., Joseph, J., and Singh, K. (2009a). Impulse attack-free four random phase mask encryption based on a 4-f optical system. *Appl. Opt.*, 48(12):2356–2363.

[Kumar et al., 2009b] Kumar, P., Kumar, A., Joseph, J., and Singh, K. (2009b). Impulse attack free double-random-phase encryption scheme with randomized lens-phase functions. *Opt. Lett.*, 34(3):331–333.

[Liu et al., 2009] Liu, W., Yang, G., and Xie, H. (2009). A hybrid heuristic algorithm to improve known-plaintext attack on fourier plane encryption. *Opt. Express*, 17(16):13928–13938.

[Liu et al., 2015] Liu, X., Wu, J., He, W., Liao, M., Zhang, C., and Peng, X. (2015). Vulnerability to ciphertext-only attack of optical encryption scheme based on double random phase encoding. *Opt. Express*, 23(15):18955–18968.

[Monaghan et al., 2007] Monaghan, D. S., Gopinathan, U., Naughton, T. J., and Sheridan, J. T. (2007). Key-space analysis of double random phase encryption technique. *Appl. Opt.*, 46(26):6641–6647.

[Monaghan et al., 2008] Monaghan, D. S., Situ, G., Gopinathan, U., Naughton, T. J., and Sheridan, J. T. (2008). Role of phase key in the double random phase encoding technique: an error analysis. *Appl. Opt.*, 47(21):3808–3816.

[Nakano et al., 2013] Nakano, K., Takeda, M., Suzuki, H., and Yamaguchi, M. (2013). Evaluations of phase-only double random phase encoding based on key-space analysis. *Appl. Opt.*, 52(6):1276–1283.

[Naughton et al., 2008] Naughton, T. J., Hennelly, B. M., and Dowling, T. (2008). Introducing secure modes of operation for optical encryption. *J. Opt. Soc. Am. A*, 25(10):2608–2617.

[Nomura and Javidi, 2000] Nomura, T. and Javidi, B. (2000). Optical encryption using a joint transform correlator architecture. *Optical Engineering*, 39(8):2031–2035.

[Peng et al., 2006] Peng, X., Zhang, P., Wei, H., and Yu, B. (2006). Known-plaintext attack on optical encryption based on double random phase keys. *Opt. Lett.*, 31(8):1044–1046.

[Refregier and Javidi, 1995] Refregier, P. and Javidi, B. (1995). Optical image encryption based on input plane and fourier planerandom encoding. *Opt. Lett.*, 20(7):767–769.

[Situ et al., 2007] Situ, G., Gopinathan, U., Monaghan, D. S., and Sheridan, J. T. (2007). Cryptanalysis of optical security systems with significant output images. *Appl. Opt.*, 46(22):5257–5262.

[Situ et al., 2008] Situ, G., Monaghan, D. S., Naughton, T. J., Sheridan, J. T., Pedrini, G., and Osten, W. (2008). Collision in double random phase encoding. *Optics Communications*, 281(20):5122 – 5125.

[Situ and Zhang, 2004] Situ, G. and Zhang, J. (2004). Double random-phase encoding in the fresnel domain. *Opt. Lett.*, 29(14):1584–1586.

[Unnikrishnan et al., 1998] Unnikrishnan, G., Joseph, J., and Singh, K. (1998). Optical encryption system that uses phase conjugation in a photorefractive crystal. *Appl. Opt.*, 37(35):8181–8186.

[Unnikrishnan et al., 2000] Unnikrishnan, G., Joseph, J., and Singh, K. (2000). Optical encryption by double-random phase encoding in the fractional fourier domain. *Opt. Lett.*, 25(12):887–889.

[van der Walt et al., 2014] van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453.

[Zhang and Naughton, 2016] Zhang, L. and Naughton, T. J. (2016). Multiple correct decryption keys an advantage in optical image encryption. *Opt. Express*, 24. Submitted.

 ISBN 978-0-9934207-1-9

Classifying HER2 breast cancer cell samples using deep learning

Tomi Pitkäaho, Taina M. Lehtimäki, John McDonald and Thomas J. Naughton

Department of Computer Science, Maynooth University–National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

Abstract

Human epidermal growth factor receptor (HER2) is a gene that in 25% to 30% of malignant breast cancer cases leads to HER2 protein over-expression. This in turn leads to uncontrolled breast cell growth. In order to decide on the appropriate treatment, patients with this condition undergo HER2 testing. The test result is analysed visually by an expert or group of experts to decide on the stage of the condition. Unfortunately, opinions of experts, and therefore the appropriate treatment, vary significantly. Automated image analysis tools have been proposed to alleviate this problem. In this paper, we present our approach to cancer cell classification using deep convolutional neural networks.

Keywords: Deep learning, Convolutional neural network, Image processing, Biomedical imaging

1 Introduction

Human epidermal growth factor receptor (HER2), is a growth factor receptor gene that produces receptors [Coussens et al., 1985] on breast cells called HER2 proteins. In benign cases HER2 receptors control normal growth of breast cells, and control cell mitosis and repair. In 25 to 30 percent of breast cancers, the HER2 gene does not work correctly which leads to HER2 amplification and HER2 protein overexpression [Slamon et al., 1987, Slamon et al., 2001]. With this condition, growth and mitosis of breast cells is uncontrolled. A test called the immunohistochemical (IHC) test can be used to monitor HER2 protein overexpression. If HER2 overexpression is confirmed, drugs such as trastuzumab, lapatinib and pertuzumab can be used in the treatment. In this test, HER2 staining is performed on fresh or frozen breast cancer tissue, that had been removed during a biopsy, and analysed visually by an expert or group of experts. Their subjective, albeit expert, assessment constitutes the state-of-the-art in HER2 analysis [Rakha et al., 2014]. It has been reported that differences in HER2 scores for patients are due partially to differing expert assessments [Dowsett et al., 2007]. As a result, there is a demand for more consistent HER2 analysis.

In the spring of 2016, the University of Warwick publicly announced a HER2 Scoring Contest [HER,] in which the aim was to accelerate the development of state-of-the-art algorithms for automated scoring of HER2 images. The University of Warwick provided training and test data for the participating groups. The training data consisted of 53 whole-slide images that were HER2 analysed by an expert or a group of clinical experts. The ground truth analysis result comprised a HER2 score (a classification into one of {0+, 1+, 2+, 3+} representing increasing amounts of HER2 protein expression) and an estimation of percentage of cells that have a complete membrane staining. Each HER2 whole-slide image also contained an example class 3+ reference tissue sample to allow one to calibrate for the intensity of the staining. In addition to the HER2 slides, the contest organisers included supplementary data in the form of HE stained whole-slides that an expert typically uses to determine the presence of tumorigenic cells and the spatial location of the HER2 stained sample.

The test data consisted of 28 unlabelled whole-slide images of HE and HER2 stained samples. Contest participants were required, for each sample in the test set, to HER2 score (classify) the sample and estimate amount of cells with complete membrane staining. This paper summarises our entry in this contest.

 ISBN 978-0-9934207-1-9

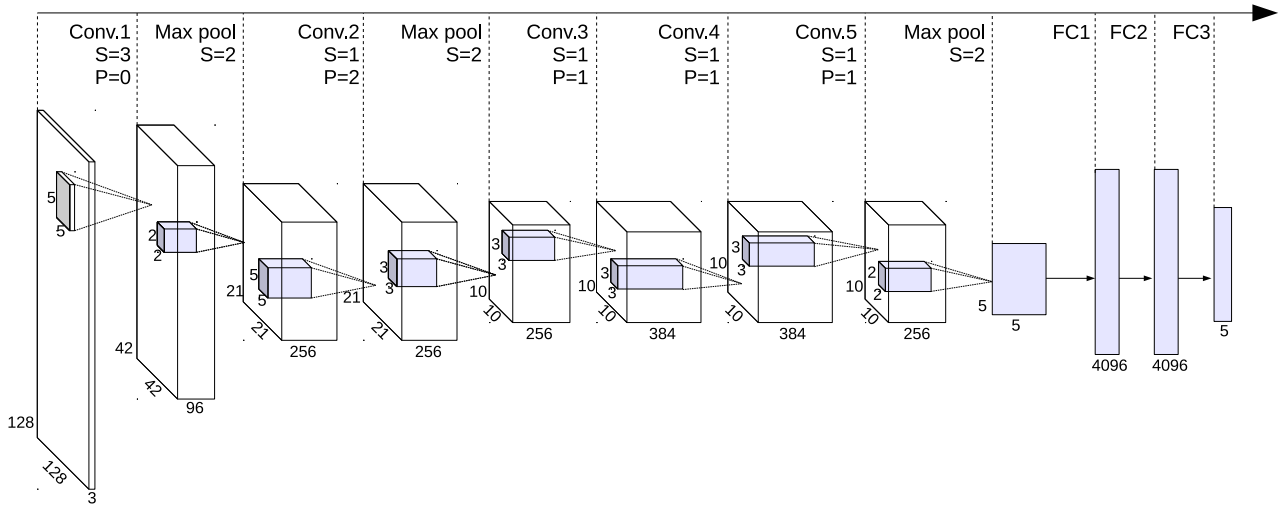


Figure 1: Network architecture. Conv = convolution, S = amount of stride, P = amount of zero padding, FC = fully connected.

2 Deep learning

Deep learning is a computational model that enables multiple levels of abstraction and can be used efficiently in various applications [Le Cun et al., 2015]. In practice each neural network with more than one hidden layer can be considered as a deep learning architecture. Deep convolutional neural networks that are one form of deep learning have been used successfully in various different visual object recognition and object detection applications [Le Cun et al., 1990, Krizhevsky et al., 2012, He et al., 2015, Szegedy et al., 2015].

We chose a convolutional neural network approach and the architecture of the network resembled strongly the AlexNet architecture that won the Large Scale Visual Recognition Challenge 2012 [Krizhevsky et al., 2012]. Our neural network implementation contained the same number of layers as the original AlexNet configuration (Fig. 1). As our inputs to the network were 128× 128 pixel images, layer specific parameters (kernel size, stride, and padding) were adjusted accordingly (see Table 1). The amount of outputs from the layers was kept the same than in the original AlexNet architecture.

The image classification dataset was created, and neural network training was executed, using Nvidia's Deep Learning GPU Training System (DIGIT) platform with an Nvidia GTX690 graphics card. The classification was performed using Caffe [Jia et al., 2014] into which the trained model was imported.

2.1 Training

As the training data contained only a single label for the whole slide, we synthesised training data by selecting carefully regions of the 53 training HER2 images at a low resolution (level 6 resolution) that were considered to contain the most representative samples from each class. The HE stained slide was used to determine the correct spatial location of the sample on the HER stained slide. In addition to the four HER classes (HER2 scores from 0+ to 3+), one more class was chosen for the background. This background class was considered to be a region with texture but without the clear appearance of nuclei (no blueish or brownish colour). In total, 119 regions were selected, with region sizes varying from 221184 to 59006976 pixels.

The training data set was extended by adding regions from the 28 test HER2 images. No classifications were released by the organisers for this test data. We followed the same procedure as that outlined above: we picked small regions that we felt we could classify reliably and we determined the classification ourselves. This was within the rules of the competition, but can lead to overfitting. In total, 32 regions were selected with sizes varying from 585728 to 7008256 pixels.

The coordinates of each region in the low (level 6) resolution image were mapped to the full resolution

 ISBN 978-0-9934207-1-9

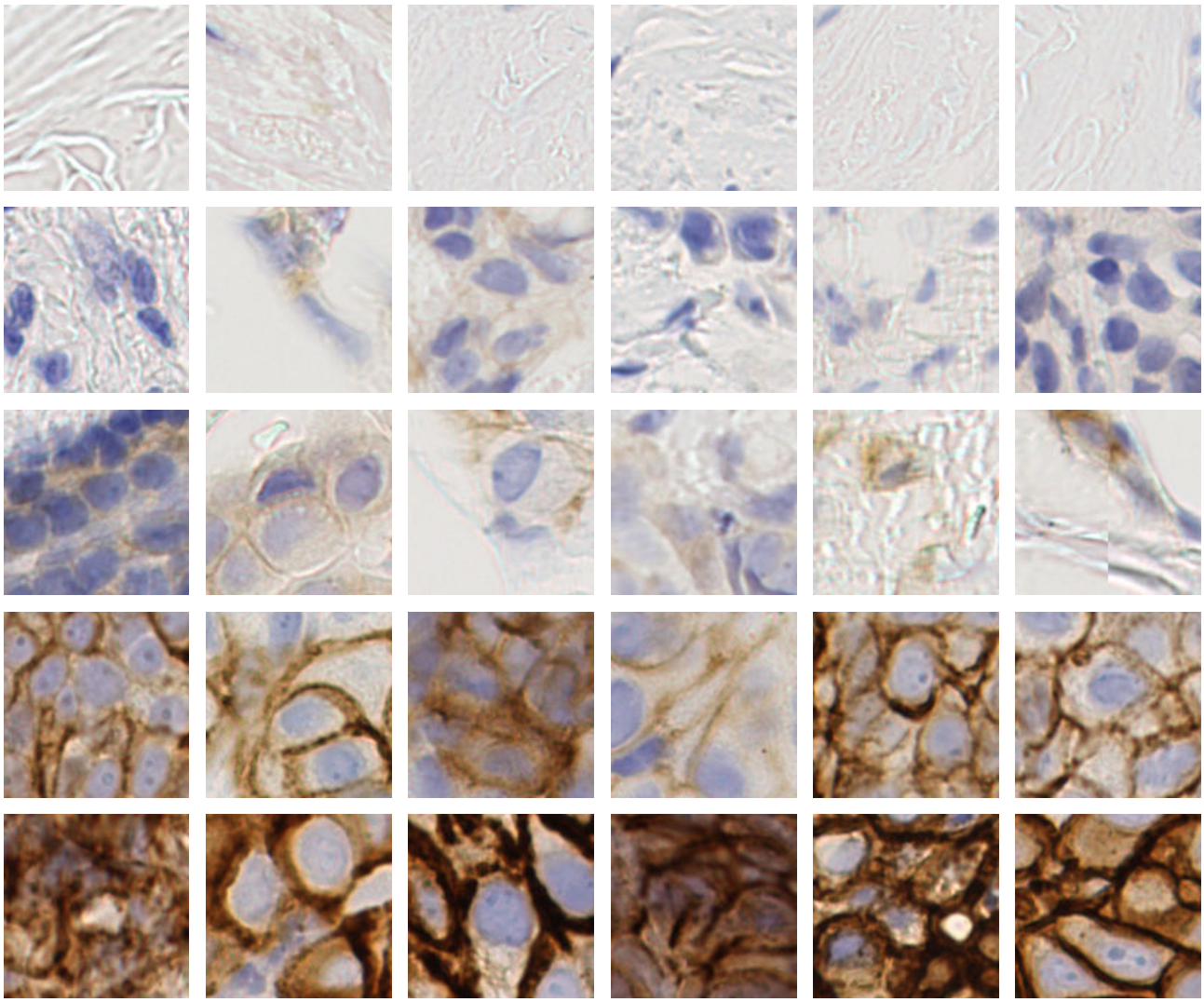


Figure 2: Training data examples. Starting from the top row classes are: background, 0+, 1+, 2+, 3+. Each block contains 128×128 pixels.

 ISBN 978-0-9934207-1-9

Table 1: Network layer parameters, size, and number of weights. Each of the convolution layers and first two fully-connected layers use rectified linear activation. The first two convolution layers are followed by response-normalisation layers. The last fully connected layer's output is fed into a 5-way softmax operation that produces a distribution over the five class labels (explained in Sect. 2.1). The total number of weights in the network is approx. 45 million.

Type	**Kernel size/Stride/Pad**	**Output size (height × width × # of outputs)**	**# of weights**
Convolution 1	5/3/0	42×42×96	7200
Max pool 1	2/2/0	21×21×96	
Convolution 2	5/1/2	21×21×256	307200
Max pool 2	3/2/0	10×10×256	
Convolution 3	3/1/1	10×10×384	884736
Convolution 4	3/1/1	10×10×384	663552
Convolution 5	3/1/1	10×10×256	442368
Max pool 3	2/2/0	5×5×256	
Fully connected 1		1×1×4096	26214400
Fully connected 2		1×1×4096	16777216
Fully connected 3		1×1×5	20480

image (level 0) and each region was divided into 128×128 pixel blocks. The data was augmented by rotating each block three times (90 degree rotations). Each resulting block was augmented through horizontal mirroring. After the data augmentation the training data was eight times larger than the original. In total, 319032 images were used in the training stage, which were divided between actual training data (75%, 239275) and validation data (25%, 79757). Example images used in the training are shown in Fig. 2.

The base learning rate was set to 0.001, and the learning rate was dropped every 1139667 iterations by a factor of 10 ($\gamma = 0.1$). During the training a mean pixel value was subtracted from each training image pixel. Loss and accuracy curves are shown in Fig. 3. The learned filters from the first convolution layer are shown in Fig. 4.

2.2 Classification

For the test data, regions that were common between the HE and HER2 slide were selected manually for classification. The manual selection was realised by defining a rectangle around the sample in a low resolution (level 6 resolution) representation of the sample slide. The chosen coordinates of the region were mapped to the full resolution image (level 0 resolution) and the region was divided to 128×128 pixel blocks. Adaptive thresholding was applied to each block with a block size of 10, and an offset of 10 pixels, to produce a binary image of size $n \times n$ pixels. If the total number of 1s in the binary image was smaller than a predefined threshold, $\tau = 0.9n^2$, the block was classified with the trained neural network model, otherwise the block was considered not to contain any texture and therefore did not require classification.

Each 128×128 pixel block was classified into one of the five classes. A preliminary classification example is shown in Fig. 5 in which the five classes are colour coded for visualisation purposes. The actual HER2 scoring for a whole slide was determined using the classified blocks as follows:

- Score 3, if amount of blocks with class 3 ≥ 10 % of total blocks
- Score 2, if amount of blocks with class 2 ≥ 10 %, or with class 3 ≥ 1 % and <10 %, of total blocks
- Score 1, if amount of blocks with class 1 ≥ 10 % of total blocks
- Score 0, otherwise.

 ISBN 978-0-9934207-1-9

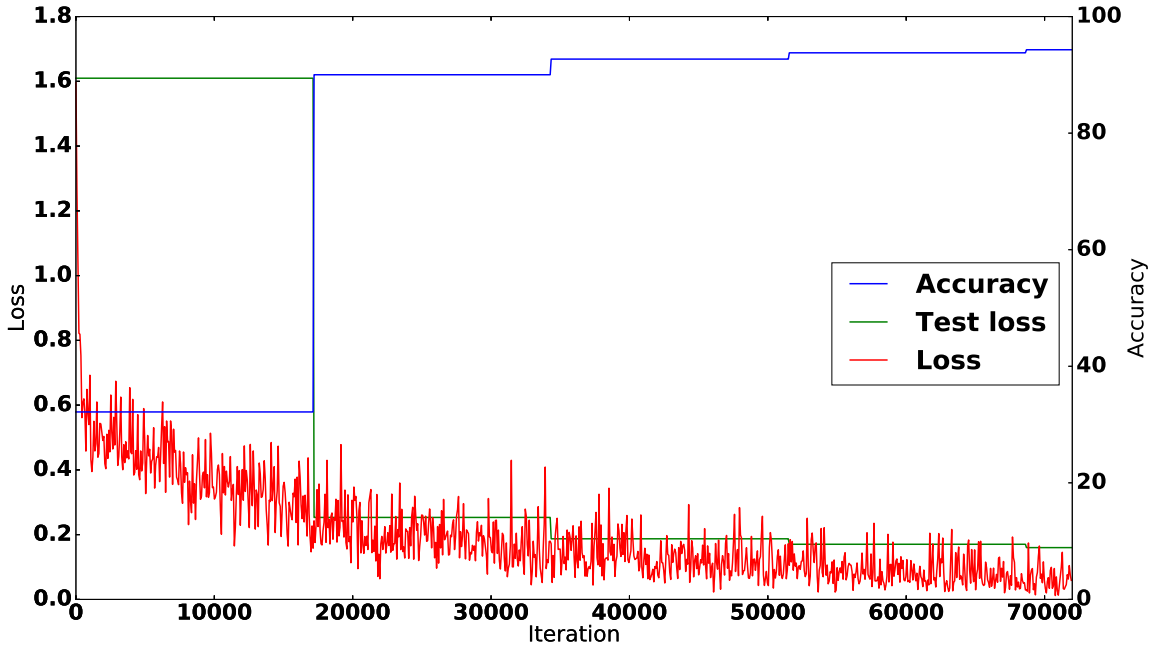


Figure 3: Loss, test accuracy, and test loss of the first 72000 iterations. At the end of learning the loss was 0.0012, test accuracy 97.7%, and test loss 0.081.

The neural network returned a confidence value for each 128×128 pixel block. The confidence value for a slide was calculated by averaging the confidence values for each 128×128 pixel block classified with the highest probability class. The scoring results of our convolutional neural network on the test data were submitted to the contest organisers on 21st June 2016. The results from the competition will be announced in July 2016 [Rajpoot and Qaiser, 2016].

3 Conclusion

We have shown that a deep convolutional neural network can be used to classify HER2 whole-slide images with a loss of 0.0012, a test accuracy of 97.7%, and a test loss of 0.081. It is highly probable that results could be improved if the training data was labelled by an expert at a more fine-grained level than on a whole-slide basis (as is the available ground truth data currently). Also, the documented variation between clinical experts means that our ground truth still has some element of subjectivity, and thus we are asking our neural network to emulate a human clinician rather than learning a completely objective classification task.

A more sophisticated network architecture should be employed that that detailed in this paper. One problem with the architecture in this paper is the significant data reduction in the lower layers of the network. The large kernel and stride numbers in these layers were necessary due to graphics card memory restrictions which can be overcome by using more modern graphics processing units.

Acknowledgements

This publication has emanated from research conducted with the financial support of an Irish Research Council (IRC) Postgraduate Scholarship and of Science Foundation Ireland (SFI) under grant no. 13/CDA/2224.

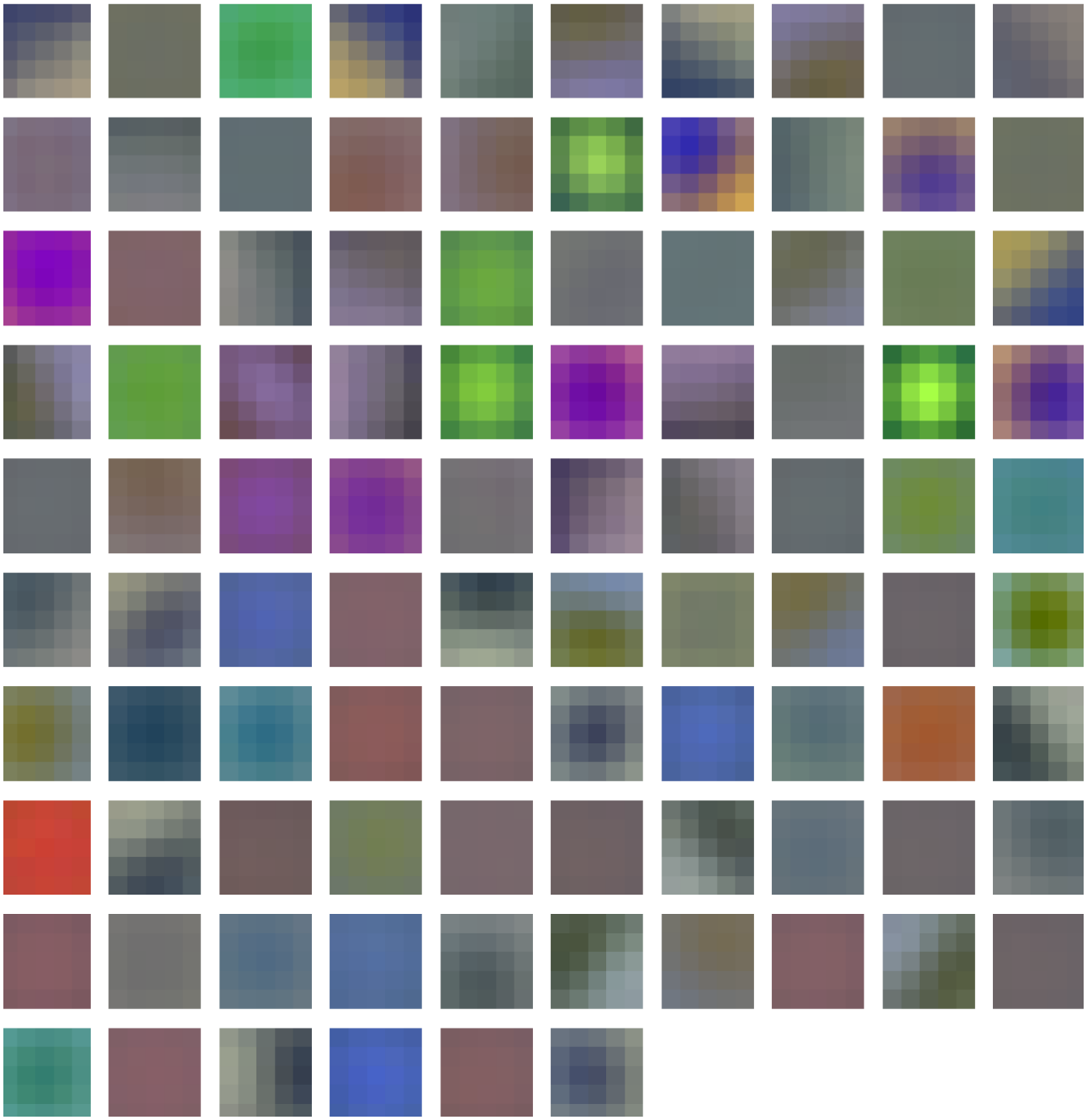


Figure 4: The 95 learned 5×5 filters from the first convolution layer.

 ISBN 978-0-9934207-1-9

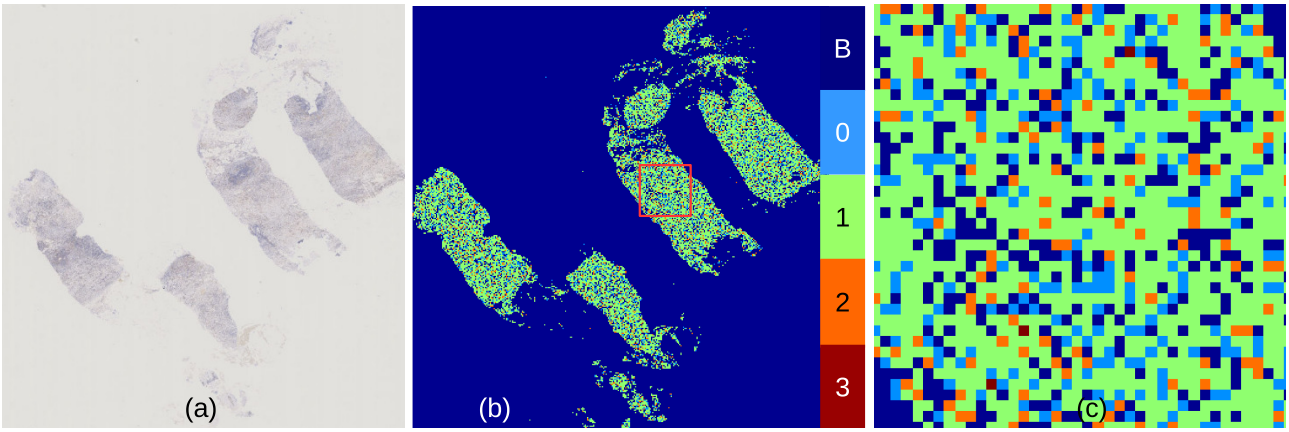


Figure 5: Preliminary classification result from the convolutional neural network: (a) the region of interest of a HER2 slide (training data slide no. 23), (b) classification of blocks (colour coded for visualisation purposes into 'B' for background, and '0' through '3' corresponding to the standard four HER2 classes) within that region of interest, (c) zoomed-in region of (b) that is marked with the red rectangle. Each pixel in (c) corresponds to a classified 128×128 pixel nonoverlapping block from the full-resolution slide. The HER2 score for the whole slide was determined to be 1.

References

[HER,] HER2 contest. `http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/research/bic/her2contest`. Accessed: 2016-06-20.

[Coussens et al., 1985] Coussens, L., Yang-Feng, T., Liao, Y., Chen, E., Gray, A., McGrath, J., Seeburg, P., Libermann, T., Schlessinger, J., Francke, U., and et, a. (1985). Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science*, 230(4730):1132–1139.

[Dowsett et al., 2007] Dowsett, M., Hanna, W. M., Kockx, M., Penault-Llorca, F., Ruschoff, J., Gutjahr, T., Habben, K., and van de Vijver, M. J. (2007). Standardization of HER2 testing: results of an international proficiency-testing ring study. *Mod Pathol*, 20(5):584–591.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

[Jia et al., 2014] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105.

[Le Cun et al., 2015] Le Cun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Le Cun et al., 1990] Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404.

[Rajpoot and Qaiser, 2016] Rajpoot, N. M. and Qaiser, T. (2016). Results from the University of Warwick HER2 scoring contest. To appear.

[Rakha et al., 2014] Rakha, E. A., Pinder, S. E., Bartlett, J. M. S., Ibrahim, M., Starczynski, J., Carder, P. J., Provenzano, E., Hanby, A., Hales, S., Lee, A. H. S., and Ellis, I. O. (2014). Updated UK recommendations for HER2 assessment in breast cancer. *Journal of Clinical Pathology*.

[Slamon et al., 1987] Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER2-2/neu oncogene. *Science*, 235(4785):177–82.

[Slamon et al., 2001] Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *The New England Journal of Medicine*, 344(11):783–792.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

 ISBN 978-0-9934207-1-9

Imaging and tracking MDCK cell vesicles using digital holographic microscopy

Tomi Pitkäaho,[1] Aki Manninen,[2] and Thomas J. Naughton[1]

[1] *Department of Computer Science, Maynooth University–National University of Ireland Maynooth, Maynooth, County Kildare, Ireland*

[2] *Biocenter Oulu, University of Oulu, P.O.Box 5000, FI-90014 University of Oulu, Finland*

Abstract

Vesicles are inner organelles within cells that perform many different functions. Confocal fluorescent microscopy is the state-of-the-art in imaging these structures. Due to the transparent nature of cells and their organelles, fluorescing markers are required. Other downsides of confocal fluorescent microscopy are high light intensities, relatively long shutter times, and the fact that as a scanning technique the imaging step is not instantaneous. Vesicle tracking is an image analysis problem for which versatile methods have been proposed. In this paper we propose to overcome the disadvantages of confocal fluorescent microscopy by using digital holographic microscopy together with image analysis to track vesicles of Madin Darby canine kidney (MDCK) cells. Multiple-depth amplitude reconstructions of a single hologram are used as the basis for tracking. Simultaneous tracking of independent vesicles' movement in three dimensions is demonstrated.

Keywords: Digital holographic microscopy, Cell imaging, Image processing, Object tracking

1 Introduction

Vesicle trafficking is a biological process whereby sub-cellular particles transport materials within cells and to/from cells. The healthy behaviour of eukaryotic cells is dependent on the faithful transport of these materials to, and fusion with, the appropriate cellular membrane [Donovan and Bretscher, 2015]. Vesicles are critical to a cell's internal organization and are involved in a variety of functions such as metabolism and enzyme storage, and can have fatal consequences when they malfunction. Automated tracking of vesicles is of great interest in cancer research, and cellular biology in general, [Yang et al., 2003, Li et al., 2004, Ku et al., 2007, Ku et al., 2009, Kalaidzidis, 2009, Chenouard et al., 2014, Kusumi et al., 2014]. However, because of their transparent nature, fluorescent tagging is used which can affect functionality, and the high light intensities of 3D microscopes such as confocal microscopes impede their regular use over long timescales. The inability to accurately identify, and track in three-dimensions, thousands of single particles, non-invasively in a label-free manner, via high-throughput microscopy has impeded dynamic studies of vesicle trafficking.

Digital holographic microscopy (DHM) is a label-free, single-shot technique that is well suited for imaging three dimensional objects [Cuche et al., 1999]. DHM has been studied extensively in particle tracking and various different methods for tracking particles have been proposed [Memmolo et al., 2015]. We propose to use multiple-depth amplitude reconstructions in the tracking as the amplitude reconstruction of a Madin Darby canine kidney (MDCK) cell reveals in-focus vesicle positions accurately and efficiently.

 ISBN 978-0-9934207-1-9

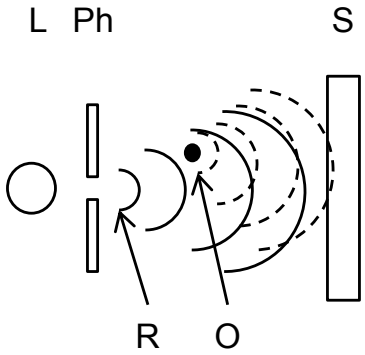


Figure 1: Principle of holography. The light from the light source (L) that emerges from the pinhole (Ph) forms the reference wave (R). A part of the reference wave is perturbed by the object, forming the object wave (O). The interference between the reference and object waves that is recorded by the sensor (S) is a hologram.

2 State-of-the-art vesicle tracking

The state-of-the-art imaging technology for tracking intracellular organelles such as vesicles is fluorescence microscopy [Kalaidzidis, 2007, Chenouard et al., 2014]. The reason for preferring fluorescence microscopy is the fact that individual proteins or other structures within cells are not clearly visible in brightfield or phase-contrast microscopy and therefore require fluorescent labelling and imaging [Meijering et al., 2012]. Introducing fluorescent labels enables imaging of the transparent cells together with their intracellular organelles. Different labelling markers that express specific organelles are used. Markers such as green fluorescent protein (GFP) and its derivatives can be introduced genetically into the cells providing known spectral characteristics for a specific protein [Kalaidzidis, 2009]. Fluorescence microscopy allows one to track particles in four dimensions (three spatial dimensions plus time) [Genovesio et al., 2006]. In fluorescence microscopy, the tracking is applied to fluorescing proteins and not directly to the vesicles themselves.

Manual tracking has been the most common way to follow the dynamics of vesicles [Kalaidzidis, 2009]. Since manually following multiple particles such as vesicles is time-consuming and error prone, tracking has evolved into an image processing and analysis problem with various methods proposed to tackle this challenging application. Generally, image analysis tracking algorithms are applied to captured fluorescence microscopy image stacks obtained through confocal microscopy, which are often captured in time-lapse mode. Typically, tracking is accomplished in two steps: particle detection (segmentation) followed by temporal particle linking. Various studies have been published in the literature, comparing tracking approaches employing template matching, watershed transformation, and deformable models [Wu et al., 2008] as well as more sophisticated machine vision approaches [Smal et al., 2010].

In this paper, we apply an existing and reported tracking method, namely template matching, to DHM. The novelty of this paper is in applying DHM to track the actual vesicles themselves and not attached fluorescing proteins as with the state-of-the-art. One advantage is biological relevance: the vesicles are more likely to behave naturally as it is a non-invasive approach (no genetic marking/manipulation) and the low light levels do not induce phototoxicity (compared to confocal microscopy). There is also the potential for more accurate tracking results as the 3D scene is sensed with a single camera frame (no artefacts associated with using a scanning technology to sense a dynamic 3D scene) and we can use shape information to distinguish individual vesicles during tracking. To our knowledge this approach has not been reported before.

3 Digital holographic microscopy

Holography enables the capture of both the amplitude and phase of an optical wavefront reflected from or transmitted through a real-world 3D object [Gabor, 1948]. (In contrast, a conventional camera or a digital camera

 ISBN 978-0-9934207-1-9

records only intensity, which is the square of the complex-valued optical wavefront entering its aperture). This allows one to image simultaneously multiple objects at different depths in the volume within the camera's field of view. In digital holography, we use a digital camera instead of traditional glass holographic plate. The digital camera performs a two-dimensional sampling of the magnified hologram, which is the interference between the light scattered from the object (the object wave) and the light passing straight through the volume (the reference wave), as illustrated in Fig. 1. Mathematically this interference can be represented as

$$H(x,y) = |R|^2 + |O|^2 + R^* O + R O^*, \tag{1}$$

where R^* and O^* denote the complex conjugates of the reference wave R and the object wave O, respectively. The magnification can be effected by adding a microscope objective after the sample volume.

The visualisation of the recorded object is achieved by illuminating the hologram with a replica of the reference wave. Goodman [Goodman, 1967] showed how this re-illumination can be replaced by digital processing that simulates the propagation of light, and referred to as numerical reconstruction. In principal, the intensity reconstruction can be realized at any depth z with the Fresnel approximation [Goodman, 2005] as

$$U(x,y;z) = \left| \frac{-i}{\lambda z} \exp(\mathrm{i}kz)\, H(x,y) \otimes \exp\left[\mathrm{i}\pi \frac{x^2+y^2}{\lambda z} \right] \right|^2, \tag{2}$$

where λ is the wavelength of the light, $\otimes$ denotes a convolution operation, and $k = 2\pi/\lambda$. This allows refocussing at arbitrary depths through the reconstruction volume, off-line after the hologram has been captured, something that is not possible with an image captured using a regular camera. From the complex valued reconstruction, the amplitude component is defined as

$$A(x,y;z) = \sqrt{\mathrm{Re}[U(x,y;z)]^2 + \mathrm{Im}[U(x,y;z)]^2}, \tag{3}$$

the intensity component is defined as $I(x,y;z) = [A(x,y;z)]^2$, and phase component is defined as

$$\phi(x,y;z) = \arctan\left\{ \frac{\mathrm{Im}[U(x,y;z)]}{\mathrm{Re}[U(x,y;z)]} \right\}. \tag{4}$$

4 Holographic vesicle tracking

MDCK cells can be considered as phase objects that in general do not admit a significant signal in an amplitude reconstruction, however vesicles in the amplitude reconstruction can be distinguished extremely well. Amplitude reconstructions from digital holograms can therefore be used to track vesicles and their position both axially and laterally.

Our procedure is as follows. After capturing a time-lapse sequence of living MDCK cells, a hologram from which the tracking will start is chosen. By searching for the local minimum along the longitudinal axis (z-axis) of the amplitude reconstruction stack one can identify an in-focus vesicle as shown in Fig. 2. For this, the chosen hologram is reconstructed through a volume using predefined reconstruction depths. The depth map $D(x,y)$ is formed as

$$D(x,y) = \begin{cases} \arg\min_z [V(x,y;z)], & \text{if} \min[V(x,y;z)] \geq \tau \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $V(x,y;z)$ is the amplitude reconstruction volume in which each amplitude reconstruction is multiplied by a binary mask, and τ is a manually adjusted threshold. The mask is obtained for each depth by an adaptive thresholding of each amplitude reconstruction. The threshold value to obtain the mask is a weighted sum (Gaussian window) of the $n \times n$ neighbourhood followed by subtraction of an offset o.

$D(x,y)$ is binarised to $M(x,y)$ by setting all the values greater than 0.001 (normalised) to 1, and running a morphological image opening operation on the result. A threshold of 0.001 was found experimentally. Unique labels are given to each independent region r in $M(x,y)$. The size in pixels of each region is calculated; if the

 ISBN 978-0-9934207-1-9

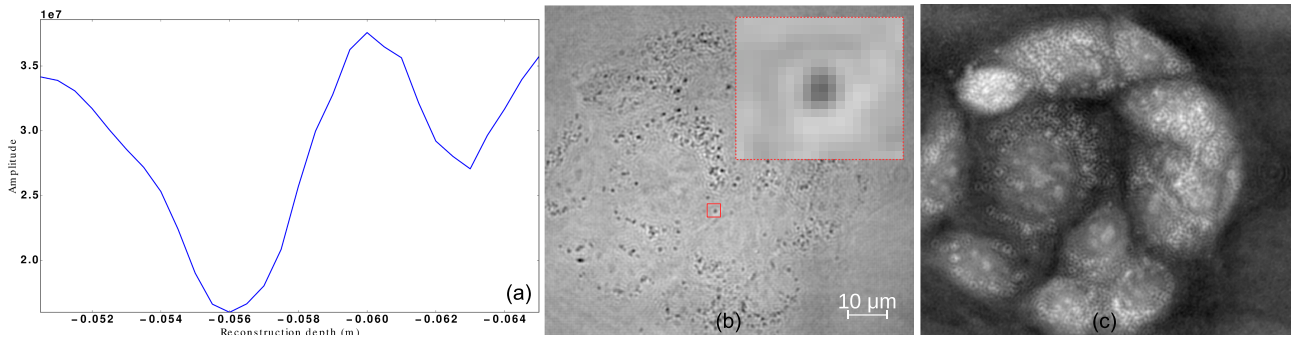


Figure 2: Focus metric. (a) amplitude values of a single pixel of a vesicle at different reconstruction depths. (b) amplitude reconstruction from a MDCK cell hologram at −5.6 cm from the hologram plane where some of the vesicles are in focus. The inset in (b) shows an in-focus vesicle. (c) unwrapped phase reconstruction at −5.6 cm from the hologram plane.

size is below a threshold, the region is used to obtain a mask M'_r for that vesicle. The median value of D within the extent of the mask for each vesicle, is the starting depth for the tracking of that vesicle.

Tracking can be realized by using template matching in which the template T is an amplitude reconstruction $M'_r \cap V$. The template with size $w \times h$ is convolved over a larger region C_z with the same centre as T. The matching is performed on the next frame of the sequence for each reconstructed depth z as

$$R(x,y;z) = \frac{\sum_{x',y'}[T'(x',y') \cdot C'(x+x',y+y';z)]}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} C'(x+x',y+y';z)^2}}, \tag{6}$$

where

$$T'(x',y') = T(x',y') - 1/(w' \cdot h') \cdot \sum_{x'',y''} T(x'',y''), \text{ and} \tag{7}$$

$$C'(x+x',y+y';z) = C(x+x',y+y';z) - 1/(w \cdot h) \cdot \sum_{x'',y''} C(x+x'',y+y'';z), \tag{8}$$

and where $x' = 0...w-1$, $y' = 0...h-1$, $w'' = w+\eta$, and $h'' = h+\eta$. The best match is extracted by finding the global maximum of $R(x,y;z)$ and the template is updated accordingly for the next tracking round. If the maximum coefficient value is below a threshold, the size of η is incremented and therefore the size of $C(x,y;z)$ is increased for the next round. This allows the vesicle to be lost in a small number of noisy frames, and picked up again.

5 Results

Figures 3 and 4 show an example tracking result with MDCK cells. MDCK cells were grown in a traditional growth medium and the time-lapse imaging was performed by using an off-axis Mach-Zehnder digital holographic microscope from Lyncée Tec, Lausanne, Switzerland (`www.lynceetec.com`). The imaging used a dry 40X microscope objective with a 0.7 numerical aperture (Leica HCX PL Fluotar). One 1024×1024 pixel hologram (using a digital camera with 6.45 µm pixel pitch) was captured every 20 seconds. Aperture apodization was applied to the holograms and the reconstruction volume was reconstructed with 0.5 mm steps from −65.0 mm to −50.5 mm. Adaptive Gaussian thresholding with a neighbourhood size n of 40 and a subtraction offset o of 40 was applied to the amplitude-normalised reconstructions. These parameter values were found experimentally to give the best results for our data. Of the other manually adjusted parameters, τ was set to 0.5 for these amplitude-normalised reconstructions, the offset of C was 10 pixels, the coefficient threshold was 0.95, and η was 2 pixels. Again, these values were found experimentally to work well for our data.

 ISBN 978-0-9934207-1-9

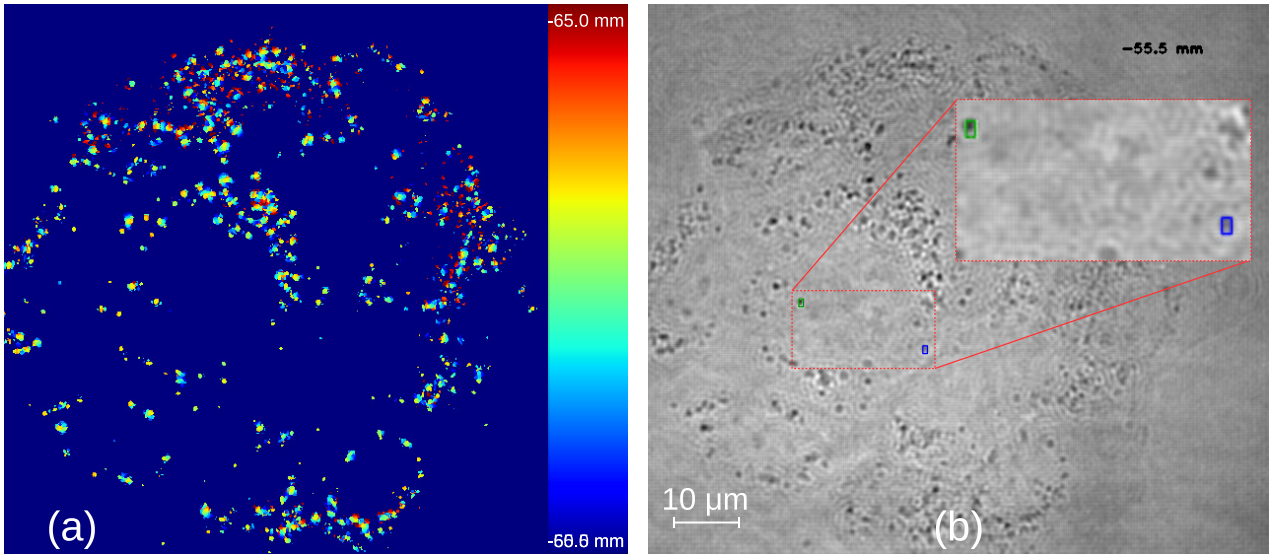


Figure 3: Example tracking result. (a) depth map of the starting frame, and (b) last frame of the sequence with blue and green rectangles showing the positions of two example identified objects.

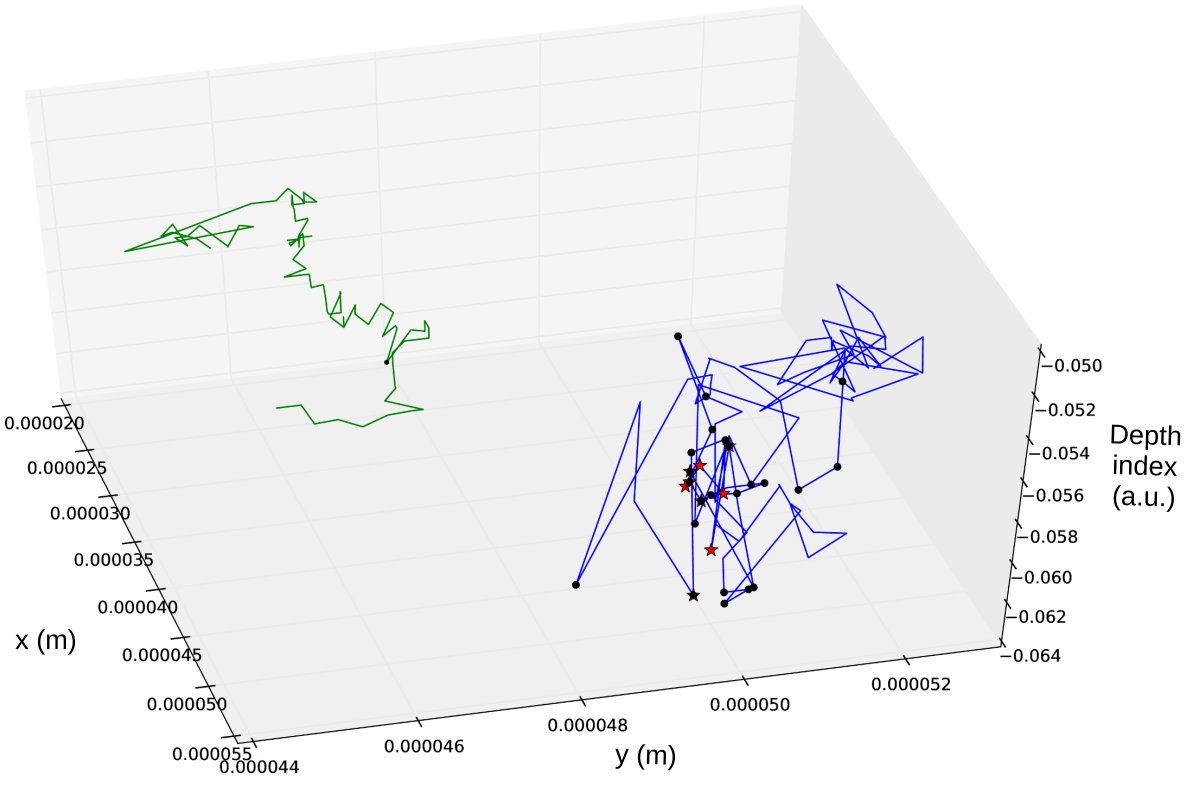


Figure 4: Trajectory of the tracked vesicles highlighted in Fig. 3. Each vesicle was tracked for 99 frames, each taken at 20 s intervals. One vesicle was tracked very accurately compared to the ground-truth data, and the other (in a more noisy part of the reconstruction volume) was tracked less so. Black spheres show frames where the 3D coordinate determined by the algorithm differs from manual tracking by five or more depth indices, red stars show frames where the difference is 10 pixels (1.5 µm) or more in either the x- or y-directions, and black stars indicate frames with both the aforementioned lateral and longitudinal differences. Most of the differences in depth indices are caused by the fact that multiple vesicles are at close proximity to each other at some frames.

 ISBN 978-0-9934207-1-9

Since manual tracking has been the most common way to follow the dynamics of vesicles [Kalaidzidis, 2009], and we would like our algorithm to be as accurate as a human operator (while having the advantage of significantly higher throughput), we obtained ground truth data by manually tracking two of the vesicles in our data. For this, the centre of the in-focus vesicle was obtained manually for each amplitude-reconstruction stack with the assistance of the manual tracking plugin of ImageJ [Schneider et al., 2012]. The in-focus plane was chosen where the amplitude of the vesicle was observed visually to be at a minimum. Comparing the manual tracking and the algorithm (see Fig. 4), over 198 frames, the mean absolute difference in the x-direction was 2.2 pixels (0.33 μm), in the y-direction was 2.7 pixels (0.41 μm), and in depth was 2.5 slice indices.

6 Conclusion

In this paper, it was proposed that vesicle tracking could be realised efficiently using template matching with amplitude reconstructions from digital holograms. All of the advantages of digital holographic microscopy over state-of-the-art microscopy techniques (such as no scanning employed, no harmful fluorescent tagging needed, and no harmful levels of light intensity) apply. The method was theoretically described and experimental results were demonstrated. The next steps involve comparison with a state-of-the-art confocal microscopy approach to determine if tracking the vesicles themselves (rather than attached fluorescing proteins) yields more accurate results.

Acknowledgements

This publication has emanated from research conducted with the financial support of an Irish Research Council Postgraduate Scholarship, Science Foundation Ireland (SFI) under grant no. 13/CDA/2224, and Kerttu Saalasti Foundation.

References

[Chenouard et al., 2014] Chenouard, N., Smal, I., de Chaumont, F., Maska, M., Sbalzarini, I. F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M., Cohen, A. R., Godinez, W. J., Rohr, K., Kalaidzidis, Y., Liang, L., Duncan, J., Shen, H., Xu, Y., Magnusson, K. E. G., Jaldén, J., Blau, H. M., Paul-Gilloteaux, P., Roudot, P., Kervrann, C., Waharte, F., Tinevez, J.-Y., Shorte, S. L., Willemse, J., Celler, K., van Wezel, G. P., Dan, H.-W., Tsai, Y.-S., de Solórzano, C. O., Olivo-Marin, J.-C., and Meijering, E. (2014). Objective comparison of particle tracking methods. *Nat. Methods*, 11(3):281–289.

[Cuche et al., 1999] Cuche, E., Bevilacqua, F., and Depeursinge, C. (1999). Digital holography for quantitative phase-contrast imaging. *Optics Letters*, 24:291–293.

[Donovan and Bretscher, 2015] Donovan, K. W. and Bretscher, A. (2015). Tracking individual secretory vesicles during exocytosis reveals an ordered and regulated process. *JCB*, 210:181–189.

[Gabor, 1948] Gabor, D. (1948). A new microscopic principle. *Nature*, 191:777–778.

[Genovesio et al., 2006] Genovesio, A., Liedl, T., Emiliani, V., Parak, W. J., and C, C.-M. M. O.-M. J. (2006). Multiple particle tracking in 3-d+t microscopy: method and application to the tracking of endocytosed quantum dots. *IEEE Transactions on Image Processing*, 15(5):1062–1070.

[Goodman, 1967] Goodman, J. W. (1967). Digital image formation from electronically detected holograms. *Applied Physics Letters*, 11:77–79.

[Goodman, 2005] Goodman, J. W. (2005). *Introduction to Fourier optics*. Roberts and Company Publishers.

 ISBN 978-0-9934207-1-9

[Kalaidzidis, 2007] Kalaidzidis, Y. (2007). Intracellular objects tracking. *European Journal of Cell Biology*, 86(9):569–578.

[Kalaidzidis, 2009] Kalaidzidis, Y. (2009). Multiple objects tracking in fluorescence microscopy. *J. Math. Biol.*, 58:57–80.

[Ku et al., 2007] Ku, T. C., Huang, Y. N., Huang, C. C., Yang, D. M., Kao, L. S., Chiu, T. Y., Hsieh, C. F., Wu, P. Y., Tsai, Y. S., and Lin, C. C. (2007). An automated tracking system to measure the dynamic properties of vesicles in living cells. *Microsc. Res Tech.*, 70:119–134.

[Ku et al., 2009] Ku, T. C., Kao, L. S., Lin, C. C., and Tsai, Y. S. (2009). Morphological filter improve the efficiency of automated tracking of secretory vesicles with various dynamic properties. *Microsc. Res Tech.*, 72:639–649.

[Kusumi et al., 2014] Kusumi, A., Tsunoyama, T. A., Hirosawa, K. M., Kasai, R. S., and Fujiwara, T. K. (2014). Tracking single molecules at work in living cells. *Applied Physics Letters*, 10:524–532.

[Li et al., 2004] Li, C. H., Bai, L., Li, D. D., Xia, S., and Xu, T. (2004). Dynamic tracking and mobility analysis of single GLUT4 storage vesicle in live 3T3-L1 cells. *Cell Research*, 14:480–486.

[Meijering et al., 2012] Meijering, E., Dzyubachyk, O., and Smal, I. (2012). Methods for cell and particle tracking. In Conn, P. M., editor, *Imaging and Spectroscopic Analysis of Living Cells*, chapter 9, pages 183–200. Elsevier.

[Memmolo et al., 2015] Memmolo, P., Miccio, L., Paturzo, M., Di Caprio, G., Coppola, G., Netti, P. A., and Ferraro, P. (2015). Recent advances in holographic 3D particle tracking. *Advances in Optics and Photonics*, 7:713–755.

[Schneider et al., 2012] Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). Nih image to imagej: 25 years of image analysis. *Nature Methods*, 9:671–675.

[Smal et al., 2010] Smal, I., Loog, M., Niessen, W., and Meijering, E. (2010). Quantitative comparison of spot detection methods in fluorescence microscopy. *IEEE Transactions on Medical Imaging*, 29:282–301.

[Wu et al., 2008] Wu, Q., Merchant, F. A., and Castleman, K. R. (2008). *Microscope Image Processing*. Academic Press.

[Yang et al., 2003] Yang, D. M., Huang, C. C., Lin, H. Y., Tsai, D. P., Kao, L. S., Chi, C. W., and Lin, C. C. (2003). Tracking of secretory vesicles of PC12 cells by total internal reflection fluorescence microscopy. *J. Microsc.*, 209:223–227.

 ISBN 978-0-9934207-1-9

IRISH MACHINE VISION & IMAGE PROCESSING

Conference proceedings 2016

25 - 26 August 2016

National University of Ireland, Galway

Ireland

Published by the Irish Pattern Recognition & Classification Society (web: iprcs.org)

ISBN 978-0-9934207-1-9