LITERATURE SURVEY REPORT ON

Real Time Emotion Detection from Face

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

In

Computer Science And Engineering

By

Sachin Jacob

Reg no - 14004104

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®

ANGAMALY-683577, ERNAKULAM (DIST)

Affiliated to

MAHATMA GANDHI UNIVERSITY

Kottayam-686560

November 2017

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®

Mookkannor(P.O), Angamaly-683577

CERTIFICATE

This is to certify that Literature Survey report for the project titled **Real Time Emotion Detection from Face** is a bonafide work carried out by **Sachin Jacob(Reg no - 14004104)** in partial fulfilment for the award of Bachelor of Technology in Computer Science and Engineering from Mahatma Gandhi University, Kottayam, Kerala during the academic year 2017-2018.

Staff In-charge **Head of the Department**

Place:

Date:

ABSTRACT

The human face plays a prodigious role for automatic recognition of emotion in the field of identification of human emotion and the interaction between human and computer for some real application like driver state surveillance, personalized learning, health monitoring etc. Most reported facial emotion recognition systems, however, are not fully considered subject-independent dynamic features, so they are not robust enough for real life recognition tasks with subject (human face) variation, head movement and illumination change. For human-computer interaction facial expression makes a platform for non-verbal communication. The emotions are effectively changeable happenings that are evoked as a result of impelling force. So in real life application, detection of emotion is very challenging task. Facial expression recognition system requires to overcome the human face having multiple variability such as color, orientation, expression, posture and texture so on.A literature survey is done to investigate the various frameworks available for emotion detection from face.

ACKNOWLEDGMENT

Apart from the efforts put in by us, the success of this project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this stage of the project:

Mr.Paul Mundadan, chairman, FISAT Governing Body, who provided us with the vital facilities required by the project right from planning to completion of 1st Stage.

Dr. George Issac, Principal, FISAT for the amenities he provided, which helped us in the fulfillment of our project's initial Stage.

Dr. Prasad J.C, HOD(CSE Dept), FISAT who always guided us and rendered his help in this phase of our project.

Mr.Pankaj Kumar G, ffor his constant encouragement and enthusiastic supervision and for guiding us with patience in 1st of the project .Without his help and inspiration, this would not have been materialized.

Mrs. Divya John ,Mrs Resmi R ,Mrs. Preethi N P and Mr. Paul P Mathai and for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the 1st stage of the project The faculty of the CSE Dept., FISAT and Lab Instructors for providing us with the necessary Lab facilities and helping us throughout this project.

Our families who inspired, encouraged and fully supported us in every trial that came our way. Also, we thank them for giving us not just financial, but moral and spiritual support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In recent years, the increase of generally available computational power has allowed us to train very deep neural network models. From that, researchers have proposed various different models to solve practical problems in our life. Many previous hard tasks, especially, computer vision tasks and natural language processing, are able to be solved simply by using deep neural networks and it makes human life easier and more convenient.

Facial expression recognition is a key task with many applications, for instance, smile detector in commercial digital cameras or interactive advertisements. Robots can also benefit from automated facial expression recognition. If robots can predict human emotions, they can react upon this and have appropriate behaviors. Emotion recognition task, however, is such a difficult problem that the human performance on this task is only about 65% accuracy.

An effective learning approach that can improve the discriminative power of DenseNet, which has won the CVPR 2017 best paper award recently. Particularly, we impose an regularizer on deep features produced by the pre-last layer of DenseNet to reduce the within-class scatter of the deep features and then further facilitate the work of softmax classifier. We demonstrate the efficiency of the proposed approach in emotion recognition task, and, particularly, on the well-known FERC-2013 dataset provided by Kaggle.

Detection of face in an image or video is the fundamental step in any recognition system. It is difficult for computer to find the face in light of the fact that the number of features in an image is extremely high. P.Viola and Jones introduces different classifiers which groups the number of features into different classifiers. Using these classifiers, the different facial features are detected which can be used for further processing. But this method is limited to only frontal view of the face. Therefore in order to detect the facial features of the persons face under different poses, face landmark annotation is required. For annotation of facial landmarks, the popular method Active Shape Model (ASM) or open source software

dlib can be used. Emotion detection is based on different expressions of face and these expressions are generated by variations in facial features.The face recognition is the basic part in modern authentication/identification applications; the accuracy of this system should be high for better results. Fisherface algorithm presents high accurate approach for face recognition; it performs two classes of analyses to achieve recognition i.e. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) respectively.

Chapter 2

Related Works

2.1 Facial Action Coding System

Classical approaches for facial expression recognition are often based on Facial Action Coding System (FACS) [4], which involves identifying various facial muscles causing changes in facial appearance. It includes a list of Action Units (AUs). A model based on an approach called the Active Appearance Model . Given input image, preprocessing steps are performed to create over 500 facial landmarks. From these landmarks, the authors perform PCA algorithm and derive Action Units (AUs). Finally, they classify facial expressions using a single layered neural network.

In recent years, deep learning has proved its superior power in many fields, especially, in computer vision. Along with Recurrent Neural Networks (RNNs), deep CNNs are thought to be the brightest stars in the deep network family. The first well-known CNN models can be mentioned are LeNet , and AlexNet - the winner of ImageNet ILSVRC challenge in 2012. Some latest CNNs such as VGG , Inception , ResNet and DenseNet tend to be deeper and deeper. Meanwhile, some other CNN architectures like WideResNet or ResNeXt tend to be wider.

All of these CNNs have demonstrated their impressive performances in one of the most prestigious competitions in computer vision - the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC). A new loss function, which is called center loss, to improve the intra-class compactness of deep features in face recognition. Besides designing more effective architectures of neural networks, making the deep features more discriminative is a promising approach to increase the power of neural networks.

A multi-class SVM loss instead of usual cross-entropy loss, and exploit augmentation techniques to create more training data. A VGG-like architecture called BKNet to emotion recognition problem and achieve better accuracy than previous methods. More recently, proposes a multi-task learning model to jointly learn 3 tasks: smile detection, gender classification, emotion recognition and achieves pretty good results. To the best of our knowledge, this is the state-of-the-art on the FERC2013 dataset so far.

2.2 Face recognition using Fisherface algorithm and elastic graph matching

It is a face recognition technique that effectively combines Elastic Graph Matching (EGM) and the Fisherface algorithm. EGM as one of the dynamic link architectures uses not only faceshape but also the gray information of image, and the Fisherface algorithm as a class-specific method is robust about variations such as lighting direction and facial expression[4].

In the proposed face recognition adopting the above two methods, the linear projection per node of an image graph reduces the dimensionality of labeled graph vector and provides a feature space to be used effectively for the classification. In comparison with the conventional method, the proposed approach could obtain satisfactory results from the perspectives of recognition rates and speeds. In particular, could get maximum recognition rate of 99.3% by the leaving-one-out method for experiments with the Yale face databases. As Information Age develops, the security of information is becoming more and more important and access to a reliable personal identification is becoming increasingly essential. Because conventional methods of identification based on possessiori of ID card or exclusive knowledge like a social security number or a password are not altogether reliable, biometrics that make out ones identity and authentication can be used.

Especially, in the perspective of ease of use and accuracy, face recognition has an advantage compared with other biometrics. In general face recognition procedure, the most important thing is which feature vector is used. In early 1990s, face recognition by using Karhunen-Loeve (K-L) projection was proposed. And several methods such as Fisherface and Elastic Graph Matching (EGM) the systemre researched.

Principal Component Analysis (PCA) and Fisherface using K-L projection are used to reduce the dimensionality of the feature vector and classify the feature space. But these methods have a defect that recognition rate decreases rapidly as the transition of a face region happens. In the case of EGM, that problem can be solved by Global Move and its recognition rate is higher than the above methods also. But compared with methods using K-L projection, its recognition speed is so slow that the recognition procedure is impossible at real time.

2.3 Emotion Recognition Using Principal Component Analysis

Emotion recognition plays vital role in Human Computer Interface. It focuses on facial expression to identify seven universal human emotions such as, happy, disgust, neutral, anger, sad, surprise and fear. This is carried out by trying to extract unique facial expression features among emotions using Principal Component Analysis with Singular Value Decomposition and Euclidean Distance Classifier. Using public database Japanese Female Facial Expression (JAFFE) recognition is obtained nearly 78.57% recognition rate and Accuracy of various expressions using Principal Component Analysis alone and Principal

Component Analysis with Singular Value Decomposition is compared.

Over the past decades, human-computer interaction together with computer vision has been an important field in computer study. The direct communication between the computer and human beings is matter of concerned. Much research has been conducted on improving and developing the interaction between human and the computer. One of the significant factors that contributed to increasing and developing the interaction between the computer and humans is studying the computers ability to distinguish facial expressions for human. With emotion recognition systems, the computer will be able to assess the human expressions depending on their effective state in the same way that humans senses do.

The intelligent computers will be able to judge, analyze and give response to person behaviors, expressions and moods. The emotion recognition system applied in different areas of life such as security and surveillance, they can predict the offender or criminal's behavior by analyzing the images of their faces that are captured by the control-camcorder. Furthermore, the emotion recognition system has been used in communication to make the answer machine more interactive with people. The answer machine has become more intelligent by analyzing the client's voice and dealing with the responses according to their emotions. Moreover, it is powerful in signed language recognition system that deals with the deaf and dumb people.

The facial expression recognition system has a considerable impact on the game and besides its use to increase the efficiency of robots for specific military tasks, medical robots, and manufacturing servicing . Generally, the intelligent computer with facial expression recognition system has been used to improve our daily lives.

2.4 An accurate active shape model for facial feature extraction

The Active Shape Model (ASM) has been used successfully to extract the facial features of a face image under frontal view. However, its performance degrades when the face concerned is under perspective variations.A modified shape model is proposed to make the model represent a face more flexibly, under different orientations. The model of the eyes, nose and mouth, and the face contour are separated. An energy function is defined that links up these two representations of a human face.

Three models are employed to represent the important facial features under different poses. A Genetic Algorithm (GA) is applied to search for the best representation of face images. Experiments show a better face representation under different perspective variations and facial expressions than the conventional ASM can.

Facial feature extraction is an important process in facial image analysis, which can be applied to face recognition and verification, animation, face image compression, clc. Important facial features include the face contour, eyes, nose. month. etc. Many approaches have been proposed for the extraction of these facial features. Snake s and deform able templates have been used to represent the face contour,

eyes and mouth in a face image . A global searching approach, Active Shape Model (ASM), which can adapt to any pride fined shapes more electrical and accurately, such as medical images, face images, hand gestures, etc. Modeling faces under different poses is a challenging problem, since the appearance of the racial features will significantly differ.

ASM is constructed based on a linear combination of a set of 2D face appearances. which arc usually frontal view images. Consequently, if the input face is not of frontal view, the model cannot work properly. Therefore, in the approach, model the face contour and tile facial features separately. For the facial features, three models are used to represent the features when the face is frontal view, turned left, and turned right, respectively. Totract the facial features in an image, it needs to fit the defined face model to a face image. To search for the best match between the model and the face image optimally. the Genetic Algorithm (GA) is used in the algorithm.

2.5 An Efficient Method to Face and Emotion Detection

Face detection and emotion selection is the one of the current topic in the security field which provides solution to various challenges. Beside traditional challenges in captured facial images under uncontrolled settings such as varying poses, different lighting and expressions for face recognition and different sound frequencies for emotion recognition. For the any face and emotion detection system database is the most important part for the comparison of the face features and sound Mel frequency components. For database creation features of the face are calculated and these features are store in the database.

This database is then use for the evaluation of the face and emotion by using different algorithms.The system implements an efficient method to create face and emotion feature database and then this will be used for face and emotion recognition of the person. For detecting face from the input image the system uses Viola-Jones face detection algorithm and to evaluate the face and emotion detection KNN classifier is used.

2.6 A Robust Method for Face Recognition and Face Emotion Detection System using Support Vector Machines

This research presents framework for real time face recognition and face emotion detection system based on facial features and their actions.The key elements of Face are considered for prediction of face emotions and the user.The variations in each facial feature are used to determine the different emotions of face. Machine learning algorithms are used for recognition and classification of different classes of face emotions by training of different set of images.

In this context, by implementing herein algorithms would contribute in several areas of identification, psychological researches and many real world problems. The algorithm is implemented using Open Source Computer Vision (OpenCV) and Machine learning with python.Face emotion recognition uses support

vector machine for finding the different emotions of face and also for classifying them. PCA is used to extract the facial features and to reduce the image dimensions. Face is a two dimensional image, for face analysis it is preferred to use two dimensional vector space.

2.7 One Millisecond Face Alignment with an Ensemble of Regression Tree

It addresses the problem of Face Alignment for a single image. it show how an ensemble of regression trees can be used to estimate the face's landmark positions directly from a sparse subset of pixel intensities, achieving super-realtime performance with high quality predictions[20]. It presents a general framework based on gradient boosting for learning an ensemble of regression trees that optimizes the sum of square error loss and naturally handles missing or partially labelled data. It shows how using appropriate priors exploiting the structure of image data helps with efficient feature selection.

Different regularization strategies and its importance to combat overfitting are also investigated. In addition,it analyses the effect of the quantity of training data on the accuracy of the predictions and explore the effect of data augmentation using synthesized data. The system presents a new algorithm that performs face alignment in milliseconds and achieves accuracy superior or comparable to state-of-the-art methods on standard datasets.

The speed gains over previous methods is a consequence of identifying the essential components of prior face alignment algorithms and then incorporating them in a streamlined formulation into a cascade of high capacity regression functions learnt via gradient boosting.As others have , that face alignment can be solved with a cascade of regression functions. In this case each regression function in the cascade efficiently estimates the shape from an initial estimate and the intensities of a sparse set of pixels indexed relative to this initial estimate. The work builds on the large amount of research over the last decade that has resulted in significant progress for face alignment. In particular, it incorporates into learnt regression functions two key elements that are present in several of the success.

2.8 Computer vision for detection of body expressions of children

This article is the result of an investigation to improve the communication with a case study that suffers Cerebral Palsy through the use of Computer Vision. At present, there is a 15% of the world's population that suffers some form of disability which prevents them from complying with the common activities of a normal person in a social environment. The technology can help to facilitate the implementation of processes that support people with special needs to improve their lifestyle; in this project the main objective is to improve the communication with the patient in order to facilitate the patient care.

For conducting the investigation it was necessary the development of a prototype that detects body expressions using the OpenCV library with the Python programming language. The results are promising

because the computer vision system is able to detect with high accuracy the following body patterns: headache 77%, happiness 75%, hunger 82%, fear 88% and recreation 77%. Finally, when a body pattern is detected it is communicated to the patient's caregiver through a mobile application.

Cerebral Palsy (CP) is the main cause of disability in children which prevents the development of movement and posture of the child who suffers it. The motor disorder of cerebral palsy is frequently accompanied in sensory, cognitive, communication and perceptual disorders of conduct, or by epilepsy. The CP is a common global problem so there are some families who have one member with special disabilities, having an incidence of 2 to 2.5 cases per 1,000 live births.

Several of the children born under this condition do not communicate easily due to several factors, what comes to hinder direct understanding of the needs of the child; resulting in stress for both the family and the child with CP. Since there are several types of cerebral palsy, it can affect in different ways on each person who possesses it; CP is not as a disease but as a set of syndromes, so it is advisable to investigate a specific case to get ideas and get the necessary conclusions for the present project.

2.9 Facial Emotion Recognition Based on Visual Information

Facial Emotion Recognition (FER) is an important topic in the fields of computer vision and artificial intelligence owing to its significant academic and commercial potential. Although FER can be conducted using multiple sensors, this review focuses on studies that exclusively use facial images, because visual expressions are one of the main information channels in interpersonal communication. It provides a brief review of researches in the field of FER conducted over the past decades.

First, conventional FER approaches are described along with a summary of the representative categories of FER systems and their main algorithms. Deep-learning-based FER approaches using deep networks enabling end-too-end learning are then presented. This review also focuses on an up-to-date hybrid deep-learning approach combining a Convolutional Neural Network (CNN) for the spatial features of an individual frame and Long Short-Term Memory (LSTM) for temporal features of consecutive frames.

A brief review of publicly available evaluation metrics is given, and a comparison with benchmark results, which are a standard for a quantitative comparison of FER researches, is described. This review can serve as a brief guidebook to newcomers in the field of FER, providing basic knowledge and a general understanding of the latest state-of-the-art studies, as well as to experienced researchers looking for productive directions for future work.

2.10 Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information

The interaction between human beings and computers will be more natural if computers are able to perceive and respond to human non-verbal communication such as emotions. Although several approaches

have been proposed to recognize human emotions based on facial expressions or speech, relatively limited work has been done to fuse these two, and other, modalities to improve the accuracy and robustness of the emotion recognition system[10].

It analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also discusses two approaches used to fuse these two modalities: decision level and feature level integration. Using a database recorded from an actress, four emotions were classified: sadness, anger, happiness, and neutral state. By the use of markers on her face, detailed facial motions were captured with motion capture, in conjunction with simultaneous speech recordings.

The results reveal that the system based on facial expression gave better performance than the system based on just acoustic information for the emotions considered. Results also show the complementarily of the two modalities and that when these two modalities are fused, the performance and the robustness of the emotion recognition system improve measurably.

2.11 Automatic face emotion recognition and classification using Genetic Algorithm

Automatic face emotion recognition is still challenging emerging problem with many applications such an automatic surveillance, robot motion, video indexing and retrieval and monitoring systems. Emotion recognition and classification depends upon gesture, pose, facial expression, speech and behavioral reactions, etc.An automatic emotion recognition and classification method is based on Genetic Algorithm and on neural network.

This system consists of 3 steps which automatically detect the face emotion image: First, pre-processing such as adjusting contrast, colour segmentation, filtering, and edge detection is applied on the input image. Secondly, features are extracted with projection profile method due to high speed which has taken as processed input image. Finally, in third stage to compute optimized parameters of eyes and lip through the GA, then emotions (neutral, happy, sad, dislike, angry, surprise and fear) is classified using artificial neural network. The proposed system is tested on a face emotion image. The obtained results show that better performance of genetic algorithm along with neural network.

2.12 Face Detection using Color Segmentation Energy Thresholding

Color segmentation is an effective process to separate skin from its background. The color segmentation process will be followed by energy thresholding. Face detection has been a fascinating problem for image processing researchers during the last decade because of many important applications such as video face recognition at airports and security check-points, digital image archiving, etc. The system attempts to

detect faces in a digital image using various techniques such as skin color segmentation, morphological processing, template matching, it determined that the more complex classifiers did not work as well as expected due to the lack of large databases for training. Reasonable results were obtained with color segmentation, template matching at multiple scales, and clustering of correlation peaks.

Here it tries to replicate on a computer that which human beings are able to do effortlessly every moment of their lives, detect the presence or absence of faces in their field of vision. The model will take three different color spaces into consideration namely HSV,RGB and YCbCr. Assuming that a person framed in any random photograph is not an attendee at the gathering or get-together, it can be assumed that the face is not white, green, red, or any unnatural color of that nature.

While different ethnic groups have different levels of melanin and pigmentation, the range of colors that human facial skin takes on is clearly a subspace of the total color space. With the assumption of a typical photographic scenario, it would be clearly wise to take advantage of face-color correlations to limit face search to areas of an input image that have at least the correct color components. The color segmentation process will be followed by energy thresholding.

Thresholding is the operation of converting a grayscale image into a binary image. Thresholding is a widely applied preprocessing step for image segmentation. Often the burden of segmentation is on the threshold operation, so that a properly thresholded image leads to better segmentation. There are mainly two types of thresholding techniques available: global and local. In the global thresholding technique a grayscale image is converted into a binary image based on an image intensity value called global threshold.

All pixels having values greater than the global threshold values are marked as 1 and the remaining pixels are marked as 0. In local thresholding technique, typically a threshold surface is constructed that is a function on the image domain.The system develops a model for face detection based on color segmentation. The color segmentation process will be followed by energy thresholding. The model tries to take advantage of face color correlation. The model will take three different color spaces into consideration namely HSV,RGB and YCbCr.

2.13 Human Face Detection in Color Images Using Eye Mouth Triangular approach

Human face detection has received substantial attention from researchers in biometrics, computer vision, pattern recognition, and cognitive psychology communities because of the increased attention being devoted to security, man-machine communication, content-based image retrieval, and image/video coding. Face detection is a challenging computer vision problem because of lighting conditions, a high degree of variability in size, shape, background, color, etc.

To build fully automated systems, Gaussian model and efficient face detection algorithms are required.The system uses a feature based algorithm for detecting human faces color images with multiple faces that is sufficiently generic and is also easily extensible to cope with more demanding variations of the imaging conditions.The algorithm first corrects the color bias by a lighting compensation technique.It overcomes the difficulty of detecting the low-luma and high-luma skin tones by applying a nonlinear transformation to the YCbCr color space with lighting compensation technique.It first detects the skin region from color images using Gaussian Model and then finding face candidates from grouping skin region using filtering.It constructed eye, mouth boundary maps to verify each face candidate. To detect eye it uses Daugman's method detect mouth using HSB color transformation. After that the system constructed eye and mouth maps in order to work on the triangle relationship between them.The face detector has been applied to several test images, and satisfactory results have been obtained.

Human face detection is a computer technology that determines the locations and sizes of human faces in arbitrary (digital) images. It detects facial features and ignores anything else, such as buildings, trees and bodies. Face detection can be regarded as a specific case of object class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class. Examples include upper torsos, pedestrians, and cars. Face detection can be regarded as a more general case of face localization. In face localization, the task is to find the locations and sizes of a known number of faces (usually one). In face detection, one does not have this additional information.

Early face detection algorithms focused on the detection of frontal human faces, whereas newer algorithms attempt to solve the more general and difficult problem of multi-view face detection [1]. That is, the detection of faces that are either rotated along the axis from the face to the observer (in-plane rotation), or rotated along the vertical or left-right axis (out-of-plane rotation), or both. The newer algorithms take into account variations in the image or video by factors such as face appearance, lighting, and pose. Face detection plays a very important role in human computer interaction field.

2.14 Discriminative Deep Feature Learning for Facial Emotion Recognition

2.14.1 Introduction

Facial expressions convey non-verbal information between humans in face-to-face interactions. Automatic facial expression recognition, which plays a vital role in human-machine interfaces, has attracted increasing attention from researchers since the early nineties. Classical machine learning approaches often require a complex feature extraction process and produce poor results.Apply recent advances in deep learning to propose effective deep Convolutional Neural Networks (CNNs) that can accurately interpret semantic information available in faces in an automated manner without hand-designing of features descriptors.It also applies different loss functions and training tricks in order to learn CNNs with a strong

classification power.

The experimental results show that the networks outperform state-of-the-art methods on the well-known FERC-2013 dataset provided on the Kaggle facial expression recognition competition. In comparison to the winning model of this competition, the number of parameters in the networks intensively decreases, that accelerates the overall performance speed and makes the proposed networks well suitable for real-time systems.

2.14.2 Dataset

FERC-2013 dataset is provided on the Kaggle facial expression competition. The dataset consists of 35,887 gray images of 48 x 48 resolution. Kaggle has divided into 28,709 training images, 3589 public test images and 3589 private test images. Each image contains a human face that is not posed (in the wild). Each image is labeled by one of seven emotions: angry, disgust, fear, happy, sad, surprise and neutral.

2.14.3 Data augmentation

Due to small amount of samples in the dataset, it uses data augmentation techniques to generate more new data for the training phase. These techniques help us to reduce overfitting and, hence, to train a more robust network. The system uses 3 main methods for data augmentation as follows: - Randomly crop: Add margins to each image in the datasets and then crop a random area of that image with the same size as the original image;

- Randomly flip an image from left to right;
- Randomly rotate an image by a random angle from -15 to 15.

The space around the rotated image is then filled with black color. In practice, find that applying augmentation techniques greatly improves the performance of the network.

2.14.4 Dense Convulutional network

Dense Convolutional Networks (DenseNet) [8] is thought to be one of the most efficient CNN architectures so far. In order to enhance the information flow between layers, DenseNet uses direct connections from any layer to all subsequent layers.

Dense blocks: Each dense block contains some convolutional layers. The input of a convolutional layer is the combination of all feature-maps of all preceding layers in that dense block.It demonstrates the architecture of a dense block. Each convolutional layer has a composite function that includes three consecutive operations: batch normalization , followed by a Rectified Linear Unit (ReLU) and a 3 x 3 convolution. Each convolutional layer in a dense block has k kernels for convolution operator. The parameter k is called growth rate. The growth rate k regulates how much new information each layer contributes to the global state.

Although each layer only produces k output feature-maps, it typically has many more inputs. Thus, to reduce the computational complexity, DenseNet uses a bottleneck layer before the composite function. Therefore, each convolutional layer consists of some operations: batch normalization, ReLU, 1x1 convolution, batch normalization, ReLU, 3 x 3 convolution. Each 1 x 1 convolution produces 4k feature maps. DenseNet with bottleneck layer is called DenseNet-B.It illustrates the operations in each convolutional layer.

Transition blocks: An important part of convolutional networks is down-sampling layers that decreases the size of feature-maps. To make down-sampling layer in DenseNet, the authors employ transition block after each dense block. A typical transition block consists of a batch normalization layer and a 1 x 1 convolutional layer followed by a 2 x 2 average pooling layer. Nevertheless, in the last transition block before softmax layer, global average pooling is used instead of 2 x 2 average pooling layer.

To improve model compactness, one can reduce the number of feature-maps in transition blocks. The 1 x 1 convolution layer produces Îÿp feature-maps, where p is number of featuremaps produced by the preceding dense block .It will obtain a compressed structure. DenseNet with compressed structure is denoted by DenseNet-C. DenseNet with both bottleneck layers and compressed structure is called DenseNet-BC.

2.14.5 Center Loss

Center loss aims to simultaneously learn a center for deep features of each class and penalize the distances between the deep features and their corresponding class centers. Suppose d denotes the deep feature of i-th sample, which can be achieved as the output of the pre-last layer, which is the global average pooling layer in DenseNet as shown in the Fig 2.1

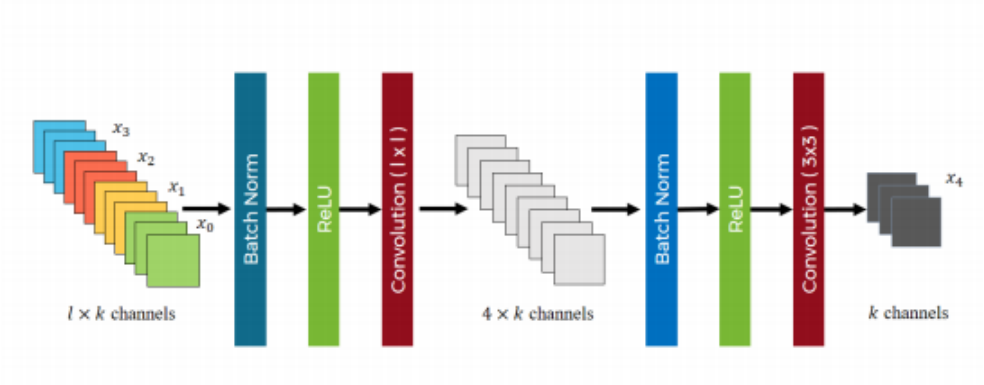


Figure 2.1. Dense block architecture

The center loss effectively describes the intraclass compactness. Minimizing the center loss can enhance

the discriminability of deep features, and then further improve the accuracy of softmax classfier. Ideally, the class centers should be updated in each iteration with respect to the entire training set. Nevertheless, the update in such a way is inefficient and even impractical.

To deal with this problem, the authors in [21] propose to perform the update based on mini-batch of size m samples as shown in the Fig 2.2. In each iteration, the centers are computed by averaging the features of the corresponding classes and, thus, some of the centers may keep unchanged.

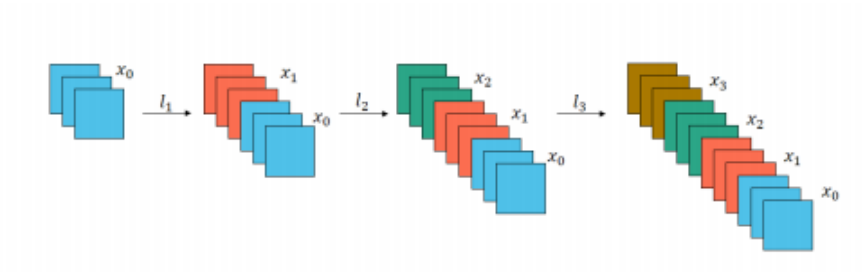


Figure 2.2. Composite function with bottleneck layer

2.14.6 Discriminative Deep Feature Learning with DenseNet

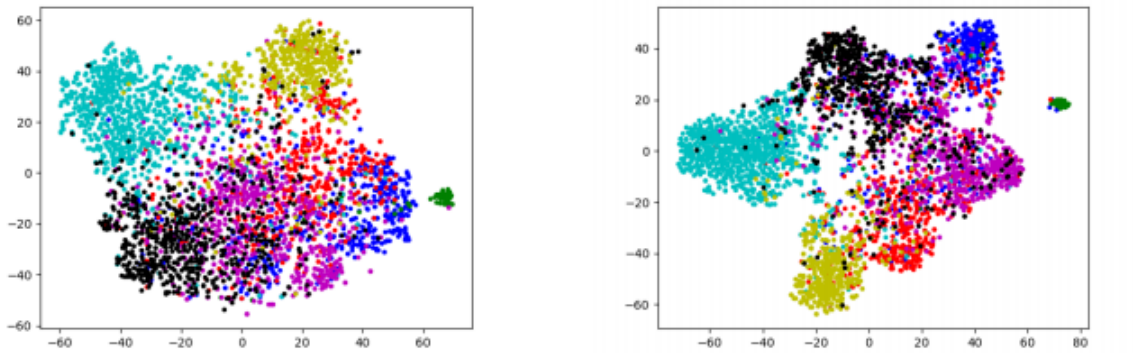
A DenseNet with 3 dense blocks to deal with emotion recognition task. The depth of the DenseNet must be 3L+4, where L is the number of convolutional layers in each dense block. In the experiments, set L = 12. Use two different values of growth rate, k = 12 (the network in this case is called DenseNet as usual) and k = 32 (the network in this case is called WideDenseNet).It also uses bottleneck layers and compressed structure in the models. Table I illustrates the proposed DenseNet and DenseNet-BC with k = 12 for emotion recognition task. In order to reduce the intra-class variation, it impose the center loss on the deep features obtained after the prelast layer. The center loss plays a key role in enhancing the discrimination capacity of the network and, hence, improving the accuracy of classification task.

In the experiments, it also try turning off the impact of center loss as shown in Fig 2.3. With the loss function, the models are trainable and can be directly optimized by the standard SGD algorithm. Moreover to avoid large perturbations caused by few mislabeled samples, a scalar parameter Îś is used to control the learning rate of the center update.

2.14.7 Implementation details

the experiments are conducted by using Python with Tensorflow framework on a workspace computer with the following specifications: Intel Xeon E5-2650 v2 Eight-Core Processor 2.6GHz 8.0GT/s 20MB, Ubuntu Operating System 14.04 64 bit, 32GB RAM and GPU Tesla K20c. Preparing dataset: Since the images in FERC-2013 dataset are cropped already, the preprocessing step is quite simple. Firstly, it compute the mean image over the training set.

Each pixel in the mean image is computed from the average of all corresponding pixels (i.e. with the same coordinates) across all training images. For each training image, it then subtract from each pixel its mean value, and then set the standard deviation of each pixel over all training images to 1. Training phase: the model is trained end-to-end by using SGD algorithm with Nesterov momentum 0.9.It set the batch size equal to 128.It initialize all weights using variance scaling initializer. The L2 weight decay is 10-4 for DenseNet and 10-5 for WideDenseNet.



(a) Without center loss

(b) With center loss

Figure 2.3. The impact of center loss on the scatter of deep features from the private test set.

In the experiments which use softmax loss combined with center loss. Moreover, it train softmax loss and center loss with different learning rates. With softmax loss, it use learning rate 0.1 and then decreased by 10 times whenever the training loss stops improving. Meanwhile, a constant learning rate 10-3 is used for training center loss.

Testing phase: In the testing phase, it evaluate the model in two different test sets as mentioned above: public test set and private test set. Furthermore, it also try to ensemble different checkpoints obtained during the training phases to increase the robustness and the power of the learned classifier.It simply keep five last checkpoints corresponding to five last training epochs for inference.

2.14.8 Face Detection Method

The detection of the face region is a based technology for feature extraction related to the face. In this study, the ed, Green, Blue (RGB) and YCbCr color models were used for face region detection. Face detection consists of a pre-processing module that uses skin color segmentation, two basic morphological operations , blob detection , and the maximum morphological gradient combination . The face detection architecture is shown in Figure 2.5

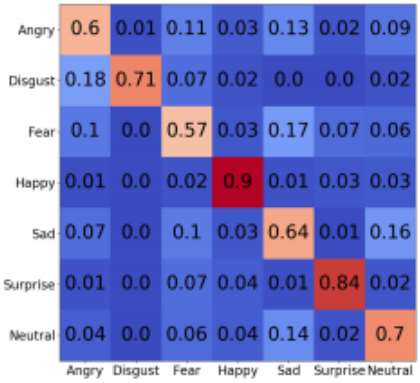


Figure 2.4. Confusion matrix on private test set

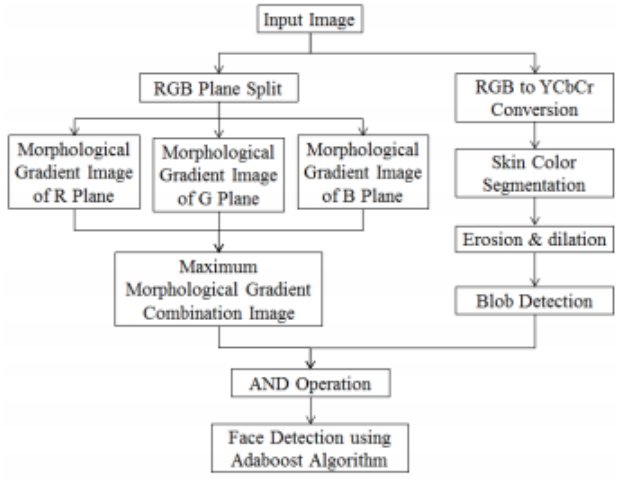


Figure 2.5. Face detection architecture

To reduce the effects of small background objects in the binary image, two basic morphological operations (erosion and dilation) are performed. Non-skin color objects that are larger than the 3 x 3 structuring element remain. To remove large objects while retaining the face region, it labeled each blob.

An obvious gradient value cannot be acquired by using only a gray image because of the equalizing effects of RGB to gray conversion. Thus, it devise the maximum morphological gradient values in the split R, G, and B planes and combine them into a single image. This facilitates clearer gradient values than those of a gray image. Finally, it can obtain an AND image between MMGC and the resulting image of the skin color segmentation and blob detection. The face detection from the AND image, which includes the clear gradient information and the non-skin color subtraction, has higher performance than that from the original image.

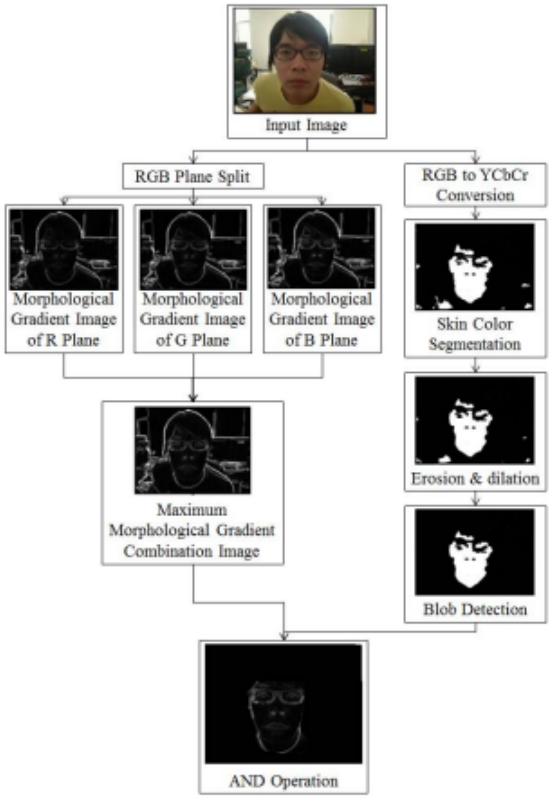


Figure 2.6. Results of pre-processing procedure of the face detection

For the face detection, it collected the circle area images as positive samples of the face from the results of preprocessing. The samples were collected under various illumination conditions. The number of collected positive samples is 2,240, and 4,500 images from the results of preprocessing are collected as the negative samples for the training process. The face cascade classifier is a 13-stage cascade that is 20 x 10 in size. The face region is detected through Harr-likebased Adaboost algorithm , the classifier generated as described above.Figure 2.6 shows the results of the pre-processing procedure of face detection.

2.14.9 Facial Components Detection Method

Figure 2.7 shows the facial component detection method used in this study. First, the face region is detected from a single image captured by a camera. The searching region of the facial components is limited to using the geometric information from the detected face region. Within a limited region, each facial component is detected.

2.14.10 Emotion Detection

Face emotion detection is used to predict the emotion state of the person based on their face expressions. Here input images are classified into two types,

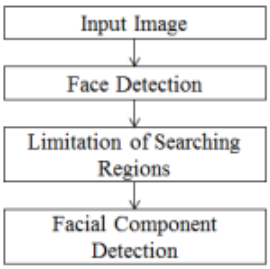


Figure 2.7. Facial components detection architecture

- Training images
- Testing images

Training images are used for training of classifier. Testing images are used to verify the algorithm by predicting the different emotions of the face. Expression analysis is the major part of the emotion detection, the schematic of expression analysis for classifying different emotions PCA is applied to training images to reduce the dimensionality. Because training images are more compared to testing and if the dimension is high then the time taken for processing will be more. Support vector machine classification is done for classifying different emotions namely, Happy, Sad, Angry, Fear, Disgust and Surprise.

Facial features such as eyes, nose, lips and face contour are considered as the action units of face and are responsible for creation of expressions on face, are extracted using open source software called dlib. SVM classifier compares the features of training data and testing data to predict any emotion of the face. Here facial features are considered as the key points which are used for training and testing. Support vector machine is the supervised learning method of machine learning. Machine learning algorithms are advantageous over other algorithms, because of less error rate and faster results. LinearSVC which is also called as MultiSVM is used for classification.

2.15 Facial expression recognition using deep convolutional neural networks

2.15.1 Introduction

Ever since computers were invented, people have wanted to build Artificially Intelligent (AI) systems that are mentally and/or physically equivalent to humans. In the past decades, the increase of generally available computational power provided a helping hand for developing fast learning machines, whereas the Internet supplied an enormous amount of data for training. Among a lot of advanced machine learning techniques that have been developed so far, deep learning is widely considered as one of the

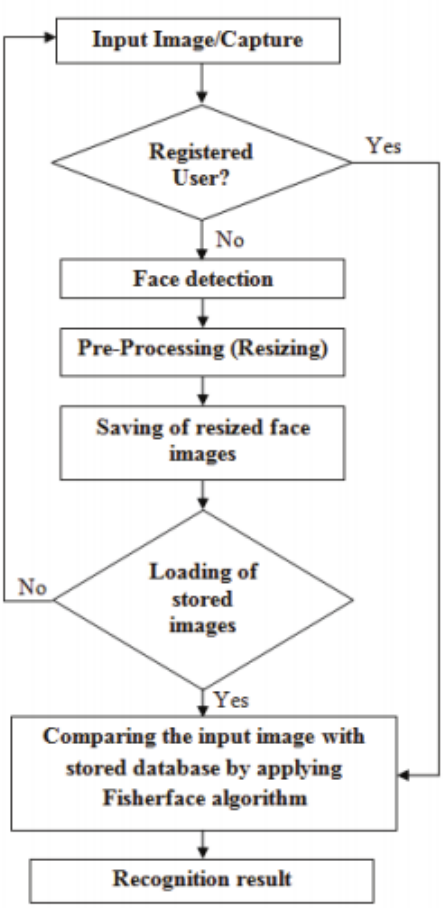


Figure 2.8. Workflow of Emotion detection system

most promising techniques to make AI machines approaching human-level intelligence. Facial expression recognition is the process of identifying human emotion based on facial expressions.

Humans are naturally capable of recognizing emotions. In fact, children, which are only 36 hours old, can interpret some very basic emotions from faces. In older humans, this ability is considered one of the most important social skills. There is a universality in facial expressions of humans in expressing certain emotions. Human develop similar muscular movements belonging to a certain mental state, despite their place of birth, race, education.

Therefore, if properly being modelled, this universality can be a convenient feature in human-machine interaction: a well trained system can understand emotions, independent of who the subject is.Automated facial expression recognition has numerous practical applications such as psychological analysis, medical diagnosis, forensics (lie-detection), studying effectiveness of advertisement and so on.

The ability to read facial expressions and then recognize human emotions provides a new dimension to human-machine interactions, for instance, smile detector in commercial digital cameras or interactive ad-

vertisements. Robots can also benefit from automated facial expression recognition. If robots can predict human emotions, they can react upon this and have appropriate behaviors. Deep learning technique and propose effective architectures of Convolutional Neural Networks to solve the problem of facial expression recognition.

The system also apply different loss functions associated with supervised learning and several training tricks in order to learn CNNs with a strong discriminative power. The system shows that Multiclass SVM loss works better than cross-entropy loss (combining with softmax function) in facial expression recognition. Besides, the evaluation of the test sets using multiple crops with different scales and rotations can yield an accuracy boost compared to single crop evaluation as shown in Fig 2.8 .

To the best of knowledge, this is the state of the art on the FERC-2013 dataset so far. The experiments show that the method achieves better accuracy than their results on both two test sets.

2.15.2 Data preprocessing:

The preprocessing step is quite simple: firstly normalizing data per image, and then normalizing data per pixel.

1)Normalizing data per image: firstly, the system subtracts from each image the mean value over that image and then set the standard deviation of the image to 3.125.

2) Normalizing data per pixel: firstly, compute the mean image over the training set. Each pixel in the mean image is computed from the average of all corresponding pixels (i.e. with the same coordinates) across all training images. For each training image,then subtract from each pixel its mean value, and then set the standard deviation of each pixel over all training images to 1.

Figure 2.9. Some samples in the FERC-2013 dataset

2.15.3 Data augmentation

Due to the small amount of training data,apply data augmentation techniques to increase the amount of training samples in order to avoid overfitting and improve recognition accuracy. For each image,perform the following successive transforms:

Detection Components	Proposed Algorithm		Adaboost Algorithm	
	Detection Rate(%)	Detection Time(ms)	Detection Rate(%)	Detection Time(ms)
Eye	84.3	74	68.7	194
Nose	88.4	41	67.4	206
Lip	71.6	119	60.2	247
Average	81.4	53.6	65.4	215.6

Table 2.1. The result of performance evaluation

Mirror the image with a probability of 0.5

1) Rotate the image with a random angle from -45 to 45 (in degrees).

2)Rescale the image, whose original size in the FERC-2013 dataset is 48 x 48, to a random size in the range from 42 x 42 to 54 x 54.

3) Take a random crop of size 42 x 42 from the rescaled image.

2.15.4 Training

In the training phase,minimize the loss function by using mini-batch gradient descent with momentum and the back-propagation algorithm) . The batch size is 256, the momentum is 0.9. To avoid overfitting,apply the dropout technique to the fully connected layers (except for the output one) with a dropout probability of 0.5. During training phase, use a strategy that decrease the learning rate 10 times if the training loss stops improving. The experiments show that the learning rate is often decreased about 5 times, and the training phase is often finished after about 1400 epochs.

For each iteration, the next 256 images are taken from the training data. After performing data augmentation, will have a batch of 256 augmented images, each of which has size of 42 x 42. These images are then fed to the network for training. After each epoch, the training data is randomly shuffled.

2.15.5 Testing

Divide the original dataset into a training set and a validation set. During the training phase, monitor the validation accuracy and save the model state with the highest accuracy on the validation set. The trained model is then applied to the test sets to estimate the final accuracy. The accuracy on the training set and validation set during training process.

In the test phase, having a trained network and an input image of size 48x48, use multiple crops in two ways to predict the true class label:

1) Method Eval1: Take 10 crops of size 42x42 of the image (5 crops and their mirrors).

2) Method Eval2: First rescale the image to various sizes 48 x 48, 50 x 50, 52 x 52, and 54 x 54. For each size, rotate the image with different angles -45, -33.75, -22.5, -11.25, 0, 11.25, 22.5, 33.75, 45 (in degrees). For each angle, take 18 crops of size 42 x 42 of the image (9 crops and their mirrors).

In the test phase, having a trained network and an input image of size 48x48, use multiple crops in two ways to predict the true class label.In both methods Eval1 and Eval2,for each crop, compute scores over the classes(in case of L2 multi-class SVM loss) or estimate the class probability distribution (in case of cross-entropy loss) as shown in Table 2.1.The final resulting vector is computed by averaging over the results of all crops. The class label associated with the highest value in the final vector is selected as the predicted label.

Chapter 3

Scope of the work

Psychological studies identified six types of facial expressions which are generally recognized. These are: fear, disgust, anger, happiness, sadness and surprise. When a person enters into one of these states, the face will change appearance significantly. In some cases, the eyes may become partially or fully closed, making the iris less visible. In other cases, such as when the person is surprised, the eyes might become much wider, producing a larger shining glint. All these changes of the eye appearance and of the other facial features (lowering eyebrows for example) will make precise eye localization more difficult, compared to the case of neutral state of the face.

This work gives a comparative study of three solutions to this non-trivial problem of finding the eyes on faces expressing emotions. Our approach is outlined next. It is based on machine learning techniques i.e. training a set of classifiers so that they can distinguish between eye and non- eye regions on the face. Firstly, the classifiers will be trained with a set of positive and negative samples, followed by an evaluation of their performance. Several iterations will be made in order to find the best combination of classifier parameters and training environment. The algorithm is robust and works under various illumination conditions. The sole problems that might occur are in cases when very strong shadows appear on the eye region. In these extreme situations, the shape features of the eyes can be significantly modified, thus leading to errors.

The feature vectors suffer from high dimensionality, which can cause over-fitting during classification. One approach to reducing the dimension of the feature vectors is to apply principal component analysis. Principal Component Analysis (PCA) is a statistical technique used for dimension reduction and recognition and is widely used for facial feature extraction and recognition. PCA is known as Eigen Space Projection which is based on linearly Projection the image space to a low dimension feature space that is known as Eigen space. Many PCA-based face-recognition systems have also been developed in the last decade [19]. In this paper PCA is used along with neural network for more efficient results. When using the PCA, there is no need to calculate the facial features like lips, cheeks, etc. But rather, the whole face

is considered as the principal component for facial expression recognition. In this paper first Eigenfaces are calculated for each of the different expressed image. After calculating the Eigenfaces of each image, the eigenvector will be calculated. With these Eigenvectors, a threshold value will be calculated for each of the facial expression.

Chapter 4

Conclusion

Face recognition which is implemented in real-time helps to recognize the human faces can be used for person identification and authentication purposes. Various Methods for implementing the same is studied. Most reported facial emotion recognition systems, however, are not fully considered subject independent dynamic features, so they are not robust enough for real life recognition tasks with subject (human face) variation, head movement and illumination change. The accuracy of both face recognition and emotion detection can be increased by increasing the number of images during training.From this survey it has been understood that extracting the feature of the training images is the most challenging task. So Proposed system should have a effective mechanism for feature extraction. By creating a model using conventional neural network it would be much more easier to implement the project.

References

[1] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza.Challenges in respresentation learning: Facial expression recognition challenge, 2013.

[2] A. Botev, G. Lever, and D. Barber. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In Neural Networks (IJCNN), 2017 International Joint Conference on, pages 1899-1903. IEEE, 2017.

[3] T. F. Cootes, C. J. Taylor, et al. Statistical models of appearance for computer vision, 2004.

[4] P. Ekman and E. L. Rosenberg. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026- 1034, 2015

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630-645. Springer, 2016.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448-456, 2015.

[10] R. T. Ionescu, M. Popescu, and C. Grozea. Local learning to improve bag of visual words model for facial expression recognition. In Workshop on challenges in representation learning, ICML, 2013.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097-1105, 2012.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, 1998.

[13] L. v. d. Maaten and G. Hinton. Visualizing data using tsne. Journal of machine learning research, 9(Nov):2579- 2605, 2008.

[14] Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10), pages 807-814, 2010.

[15] D. Sang, L. Tran Bao Cuong, and V. Van Thieu. Multitask learning for smile detection, emotion recognition and gender classification. In Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT 2017), pages 340-347, 2017.

[16] D. V. Sang, N. V. Dat, and D. P. Thuan. Facial expression recognition using deep convolutional neural networks. In The 9th International Conference on Knowledge and Systems Engineering (KSE 2017), pages 144-149, 2017.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1-9, 2015.

[19] Y. Tang. Deep learning using support vector machines. CoRR, abs/1306.0239, 2, 2013.

[20] H. Van Kuilenburg, M. Wiering, and M. Den Uyl. A model based method for automatic facial expression recognition. In Proceedings of the 16th European Conference on Machine Learning (ECML 05), pages 194-205. Springer, 2005.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In European Conference on Computer Vision, pages 499-515. Springer, 2016.

[22] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. arXiv preprint arxiv:1611.05431, 2016.

[23] S. Zagoruyko and N. Komodakis. Wide residual networks. arxiv preprint arxiv:1605.07146, 2016.