# VIDEO AND IMAGE BASED FACE ANALYSIS WITH EXTREME LEARNING MACHINES

by

Furkan Gürpınar

B.S., Mechanical Engineering, Boğaziçi University, 2012

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computational Science and Engineering

Boğaziçi University

2017

VIDEO AND IMAGE BASED FACE ANALYSIS WITH EXTREME LEARNING MACHINES

APPROVED BY:

| | |
|---|---|
| Assoc. Prof. Albert Ali Salah<br>(Thesis Supervisor) | ..... |
| Assist. Prof. Heysem Kaya<br>(Thesis Co-supervisor) | ..... |
| Prof. Lale Akarun | ..... |
| Prof. Fikret Gürgen | ..... |
| Assoc. Prof. Alkım Almıla Akdağ Salah | ..... |

DATE OF APPROVAL: 14.08.2017

# ACKNOWLEDGEMENTS

With my deepest gratitude, I would like to thank to my thesis supervisor Albert Ali Salah and my thesis co-supervisor Heysem Kaya for their friendship, guidance and support during the course of my thesis.

I would like to thank Lale Akarun, Fikret Gürgen and Alkım Almıla Akdağ Salah for participating in my thesis committee and their valuable comments.

I would like to thank the members of Media Laboratory and Perceptual Intelligence Laboratory Alp Kındıroğlu, Cihan Camgöz, Barış Evrim Demiröz, Gül Varol, Sadaf Afshar, İlhan Adıyaman and Nihan Karslıoğlu for providing a nice working environment.

I would like to express my gratitude to my friends and family for their support throughout my life.

This thesis is supported by Boğaziçi University projects BAP 16A01P4 and BAP 12A01P3, by the Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 114E481, and by the BAGEP Award of the Science Academy.

# ABSTRACT

## VIDEO AND IMAGE BASED FACE ANALYSIS WITH EXTREME LEARNING MACHINES

Automatic analysis of human behavior has been a difficult problem due to noise, environmental differences and lack of annotation. While lab-controlled data provides an easier learning experiment, "in the wild" datasets require systems complex enough to fit to unseen data, at the same time, deal with the problem of overlearning. In this thesis, we propose a fast and robust multimodal system that analyzes humans from facial images, videos and voice. We extract dense appearance descriptors as well as Deep Convolutional Neural Network (DCNN) features from the faces and we train kernel Extreme Learning Machine (ELM) classifiers, which are then combined by various fusion schemes. We apply our pipeline to a number of affective and biometric challenges and we show that ELM provides fast and accurate learning compared to traditional learning methods. We also show that multimodal fusion and DCNN fine-tuning improves the accuracy in almost all tasks. Our method has ranked $2^{nd}$ in the Emotion Recognition in the Wild (EmotiW) challenge and $1^{st}$ in the second round of ChaLearn Apparent Personality Analysis from First Impressions (FI) challenge as well as the ChaLearn Job Candidate Screening (JCS) challenge. Our results show that using extreme learning machine, efficient learning can be performed in terms of both time and computational complexity while preserving high performance.

# ÖZET

## UÇTA ÖĞRENME MAKİNELERİYLE VİDEO VE İMGE TABANLI YÜZ ANALİZİ

İnsan davranışının otomatik analizi gürültü, çevresel farklılıklar ve etiketlenmiş veri yetersizliği gibi sebeplerden dolayı zor bir problemdir. Laboratuvar ortamında toplanmış veri bu analizi kolaylaştırırken, kontrolsüz ortamda toplanmış verideki büyük varyansı modelleyebilmek için karmaşık bir model gerekmektedir, bu da aşırı öğrenme problemiyle başa çıkmayı gerektirir. Bu tezde çokkipli bir sistemle insan sesi, yüz imgeleri ve videolarından hızlı ve gürbüz bir analiz sistemi önerilmektedir. İnsan yüzlerinden görünüm ve derin öğrenme öznitelikleri çıkararak çekirdek uçta öğrenme makinesi sınıflandırıcılarını birleştirmekteyiz. Önerilen yöntem bir dizi insan analizi yarışmasında değerlendirilmiştir. Deneylerimiz sonucunda uçta öğrenme makinelerinin geleneksel alternatiflere göre hızlı ve isabetli bir yöntem olduğunu göstermekteyiz. Ayrıca, derin öğrenme parametrelerinin ilgili görev için güncellenmesi ve çokkipli birleştirmenin tahmin isabetini neredeyse bütün görevler için yükselttiğini gözlemlemekteyiz. Önerdiğimiz sistem Emotion Recognition in the Wild (EmotiW) yarışmasında 2. sırada, ChaLearn Apparent Personality Analysis from First Impressions (FI) ve ChaLearn Job Candidate Screening (JCS) yarışmalarında 1. sırada yer almıştır. Sonuçlar göstermektedir ki uçta öğrenme makineleri zaman ve hesaplamasal karmaşıklık anlamında maliyetsiz ama aynı zamanda isabetli öğrenme yapılmasına elverişlidir.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF SYMBOLS

| | |
|---|---|
| $C$ | Regularization parameter of ELM |
| $D$ | Dimensionality |
| $h$ | Number of hidden neurons in ELM |
| $H$ | Hidden layer output matrix of ELM |
| $H^{\dagger}$ | Moore-Penrose Generalized Inverse of $H$ |
| $I$ | Identity matrix |
| $K$ | Number of cluster centers |
| $L$ | Empirical loss |
| $N$ | Number of instances |
| $\mathbb{R}$ | The set of real numbers |
| $\mathbb{R}_+$ | The set of positive real numbers |
| $T$ | Label matrix |
| $y_i$ | True label of instance $i$ |
| $\hat{y}_i$ | Predicted label of instance $i$ |
| $w$ | Parameter matrix |
| | |
| $\beta$ | Output weight matrix of ELM |
| $\epsilon$ | Normal score |
| $\eta$ | Learning rate |
| $\mu$ | Mean |
| $\sigma$ | Variance |
| $\Omega$ | Kernel matrix |

# LIST OF ACRONYMS/ABBREVIATIONS

| | |
|---|---|
| 2D | Two Dimensional |
| 3D | Three Dimensional |
| AFEW | Acted Facial Expressions in the Wild Database |
| BOF | Bag of Features |
| DCNN | Deep Convolutional Neural Network |
| DPM | Deformable Parts Model |
| ELM | Extreme Learning Machine |
| FER | Facial Expression Recognition 2013 Database |
| FI | First Impressions Database |
| GMM | Gaussian Mixture Model |
| FUN | Functional statistics |
| HOG | Histogram of Oriented Gradients |
| KDEF | Karolinska Directed Emotional Faces Dataset |
| LAP | Looking at People |
| LBP | Local Binary Patterns |
| LGBP | Local Gabor Binary Patterns |
| LLD | Low Level Descriptor |
| LPQ | Local Phase Quantisation |
| MAE | Mean Absolute Error |
| PCA | Principal Component Analysis |
| PLS | Partial Least Squares |
| RBF | Radial Basis Function |
| SIFT | Scale-invariant Feature Transform |
| SLFN | Single-hidden-layer Feed-forward Network |
| SDM | Supervised Descent Method |
| SVM | Support Vector Machine |
| TOP | Three Orthogonal Planes |

# 1. INTRODUCTION

## 1.1. Motivation

Machines that are able to recognize biometric, affective and social traits of people can be useful in many applications, such as computer assisted tutoring systems, forensics, business intelligence and social robotics. Especially in human-computer interaction, there are many applications that require multimodal processing of human behavioral data.

Traditionally, observing humans and analyzing their social signals automatically has been a difficult problem. Both audio and video, two of the most frequently used modalities (and the ones that we focus in this thesis) are plagued by differences in acquisition conditions, sensor and environment noise, and most importantly, from the lack of adequate annotation in the data used to train the automatic systems.

Multimodal analysis requires machine learning systems that are complex enough to deal with the great variance shown in the data, as well as powerful enough to catch subtle cues (e.g. the wrinkles around the eye to indicate that a smile is real [5]). However, complex models with a lot of parameters can easily over-learn the data sets with which they are trained [6], and fail to generalize to new samples. These systems should be carefully crafted to prevent overlearning.

The research question of this thesis is how to build computer vision systems that learn to analyze human social signals. We focus on the human face, analyzing its appearance and dynamics, and augment it with voice whenever necessary. We tackle emotional expression recognition, age estimation, and apparent personality trait recognition in this thesis, and propose a general processing pipeline that can be easily adapted to similar problems. We use well-known feature extraction methods and optimize them in terms of speed and accuracy. Moreover, we investigate fusion strategies with deep learning to further increase the estimation accuracy.

## 1.2. Contributions of the thesis

In this thesis, we propose the use of Extreme Learning Machine (ELM) classifiers for fusion of image, audio and video features. We use state of the art visual and audio features for our models. A set of popularly used visual features is investigated for image and video representation, including Deep Convolutional Neural Network (DCNN) based features. The OpenSMILE [7] tool is used for extracting a strong set of audio features. Depending on the target domain, feature-, decision- or multi-level fusion strategies have been contrasted. We have obtained state of the art results with the proposed pipeline.

When data are collected under natural conditions, they pose much greater challenges compared with data collected from controlled laboratory conditions. Such data are called "in the wild," and real applications typically require systems that can operate under such conditions.

Particularly for handling in the wild face and video analysis efficiently, we proposed in this thesis a "transfer learning" approach. In the type of transfer learning we use, a system that is trained for a certain purpose is adapted to a new task. This is especially useful when there are complex models pre-trained with great amounts of data, which is one of the reasons why deep learning became so popular in recent years. Our method involves fine tuning of a pre-trained DCNN model for face recognition on an emotional face dataset. Since the network already extracts good intermediate features, transfer learning enables us to quickly adapt it for other face analysis tasks.

The outcomes of the thesis have been published in a journal paper [8], a book chapter [9], and international conference proceedings [10–14].

The works of image and video based emotion recognition with particular focus on in the wild estimation are collected in Chapter 4 and published in [8, 10]. Age and gender estimation can be considered as soft biometrics [15]. Their usage in this area is explored in [11]. The multimodal method applied for personality trait estimation and corresponding results are published in [12, 13] and collected under Chapter 5.

Furthermore, the systems produced in the thesis have been measured against others in international competitions. We obtained the $2^{nd}$ place in the Emotion Recognition in the Wild ICMI 2015 Challenge [10], the $6^{th}$ place in ChaLearn LAP Apparent Age Estimation CVPR 2016 Challenge [11], the first place in ChaLearn LAP Apparent Personality Trait Estimation from First Impressions ICPR 2016 Challenge [13] and lastly, using the same base set of features as in [13], we obtained the first place in two stages of the Explainable Automatic Job Candidate Screening CVPR 2017 Challenge [14].

### 1.3. Organization of the thesis

This thesis is structured as follows. In Chapter 2, we provide a brief summary of the methodology involved in face analysis applications. This includes the methods used in face detection, feature extraction, video analysis and model learning. Chapter 3 provides an overview of the general methodology we used in various applications that constitute this thesis. The main idea behind our approach is to use state of the art feature extraction methods from different modalities and allowing multiple methods to extract information from the same modality if it is able to complement existing feature sets. We propose efficient and simple decision level fusion to bring together different types of information. This way, we end up with systems of considerable complexity and powerful capabilities but escape the trap of overlearning.

In Chapters 4, 5, and 6, we investigate three problems with this general pipeline, namely, emotion recognition, personality trait recognition, and age estimation, respectively. To keep the exposition as self-contained as possible, we provide literature surveys, detailed explanations of the methodology employed and the experimental results in the respective chapters. The last chapter of the thesis provides an overview and our conclusions.

# 2. FACE ANALYSIS

This chapter provides the background work on facial analysis and does not contain original contributions.

A visual face analysis pipeline often consists of multiple processing steps such as face detection and alignment (either with or without landmark detection), feature extraction, model learning/validation and estimation.

Feature extraction for images is often done by traditional appearance descriptors in combination with DCNN features as explained in Section 2.2. For temporal signals like audio and video, temporal feature extraction and encoding methods are often used, some of which are explained in Sections 2.3 and 2.4.

Model learning is often done by training classifiers and regressors as explained in Section 2.5. DCNNs, however, can directly give the estimations related to the task, eliminating the need for an additional learning algorithm. In this work, we use DCNN as a feature extraction method and combine it with other features to feed into discriminative learners. In the following sections, we explain the background of the methodology used in this thesis.

## 2.1. Face Detection

A visual face analysis pipeline usually starts with face registration, which consists of face detection [16, 17] and sometimes followed by landmark localization [18–21]. Landmark detection is often necessary for proper alignment of faces, especially if the output of the face detection is not very reliable.

### 2.1.1. Challenges

For face detection and landmarking, there has been a number of challenges that served as a benchmark for evaluation of such algorithms. 300 faces in-the-wild (300-W) challenge [22–24] provides ground-truth annotations for 68 landmark points per face, and the training set consists of a combination of AFW [21] and the training sets of LFPW [25] and HELEN [26] datasets.

### 2.1.2. Viola & Jones face detector

The Viola & Jones object detection algorithm [16] has been the state-of-the-art face detection method for many years, and it's still very commonly used due to its desirable accuracy and high detection speed. The method proposes to use a boosted cascade of simple (weak) classifiers. The classifiers are constructed using the Haar features, which involve sums of the pixels in certain rectangular regions of the image. The cascaded classifiers are then applied to a sliding window over the whole image in order to detect the face-like regions of the image. The Viola & Jones object detector is desirable due to its capability of generalization (to other object detection tasks), its testing speed and accuracy. However, the algorithm also has some disadvantages such as it only works for frontal images, hence very sensitive to pose and lighting conditions.

### 2.1.3. Zhu & Ramanan face detector

The face detection algorithm proposed by Zhu and Ramanan uses a part-based model for detection as well as estimation of pose and landmark locations [21]. The authors propose a mixture of tree-structured models, which they show to be easily optimizable and robust in terms of capturing elastic deformations of the faces. The authors also created a manually annotated faces in-the-wild (AFW) dataset in order to further evaluate their results and made the dataset publicly available.

### 2.1.4. Deformable Parts Model-based face detector

The work by Mathias *et al.* [17] proposes a deformable part-based model (DPM) for robust detection of faces. Similarly to Zhu & Ramanan's method, the authors propose to represent faces as a combination of deformable parts and implement the learning as a boosted cascade of weak classifiers. The authors first implement a DPM-based detector to serve as a baseline, however, they surprisingly show that building a detector with enough data and a proper training procedure, a traditional part-based model can reach top performance. The authors attribute the success of their algorithm to adequate amount of training data, as well as the appropriate use of non-maximum suppression.

The authors further improve their work with a detector called HeadHunter, which makes use of rigid templates. While the performance improvement is minimal, they show that with enough training data to cover a diversity of poses, rigid templates work as good as part based models.

### 2.1.5. IntraFace

IntraFace is an implementation of the Supervised Descent Method (SDM) applied to facial landmark localization [18]. The algorithm proposes to train a regressor that aims to find the correct amount (and direction) of updates to an initial set of landmarks, using the features from local windows around them. Concatenation of the features of each local window is then multiplied by this learned matrix (that is called a descent direction) to obtain the update vector. The update vector is then added to the initial set of landmarks and the same process is repeated until convergence, which is usually reported to take 4 or 5 steps.

## 2.2. Image Features

Feature extraction often means processing an input (image, video or audio) with mathematical functions, often called *descriptors* that convert the input to a fixed-size

numeric vector, which will then be fed as an input to the model learning phase. For inputs such as images of different sizes and videos of different durations, the feature vector size can change. To overcome this issue, the feature set modeling techniques discussed in Sections 2.3.2 and 2.3.3 are used. The descriptors also need to have certain properties in order to be useful in discriminative learning algorithms. Ideally, a feature vector should be invariant to factors such as rotations, translations, scale, illumination conditions and occlusions, at the same time, they need to be informative/responsive to factors that are related to the task at hand. Most of the methods described below overcome these issues, only leaving occlusions as an open problem. The feature extraction methods used in this thesis are explained in detail in the following sections.

### 2.2.1. Local Binary Patterns

Local Binary Patterns (LBP) is an appearance descriptor, which encodes a $3 \times 3$ pixel neighborhood by representing each pixel with a binary number depending on its relation to the pixel in the center [27]. Excluding the central pixel itself, there are 8 such numbers, hence a window is described by an 8-bit number. After processing all the windows, an image is represented by the 256-bin histogram of these numbers. Moreover, it's found that using only the 58 possible uniform patterns gives almost the same explanatory power, hence the LBP descriptor is often represented as a 58 (or 59, counting the non-uniform patterns) dimensional histogram. Since the descriptor is dependent on the binary relations between pixel intensities, LBP is known to be invariant to illumination. An example LBP image under different lighting conditions is provided in Figure 2.1.

### 2.2.2. Gabor features and LGBP

The LBP descriptor is often used in combination with a Gabor feature extraction method beforehand, resulting in the Local Gabor Binary Patterns (LGBP) descriptor [28].

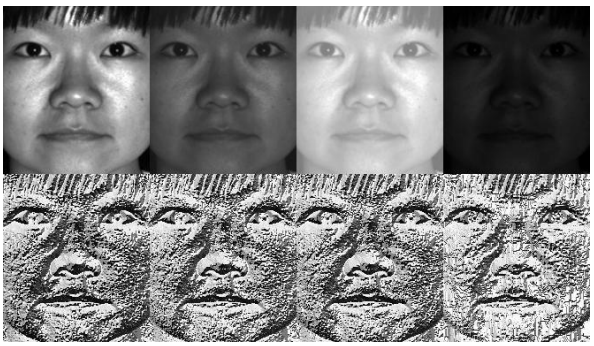


Figure 2.1. Example LBP images under different illumination conditions [1].

The Gabor features are the result of processing the local windows with Gabor filters (or Gabor kernels). A 2D complex Gabor filter is the convolution of a 2D sinusoid with phase $P$, spatial frequencies $u_0$ and $v_0$ with a 2D Gaussian kernel (envelope) with amplitude $A$, orientation $\theta$,and spatial scales $a$ and $b$. In line with [28], for simplicity we take $a = b = \sigma, u_0 = v_0 = \phi$ and $A = 1$ to obtain

$$G(x, y) \exp(-\pi\sigma^2((x - x_0)_r^2 + (y - y_0)_r^2)) \exp(j(2\pi\phi(x + y) + P)) \tag{2.1}$$

where the subscript $_r$ stands for a clockwise rotation operation around reference point $(x_0, y_0)$ such that

$$\begin{aligned} (x - x_0)_r &= (x - x_0)cos\theta + (y - y_0)sin\theta \\ (y - y_0)_r &= -(x - x_0)sin\theta + (y - y_0)cos\theta \end{aligned} \tag{2.2}$$

For Gabor feature extraction, we use an open source script [29] and implement our own LGBP routine that, with 6 orientations and 3 scales, results in 18 Gabor responses per input image. Further processing the responses with the 58-dimensional LBP operator, we obtain a $18 \times 58 = 1044$ dimensional descriptor per local portion of the image.

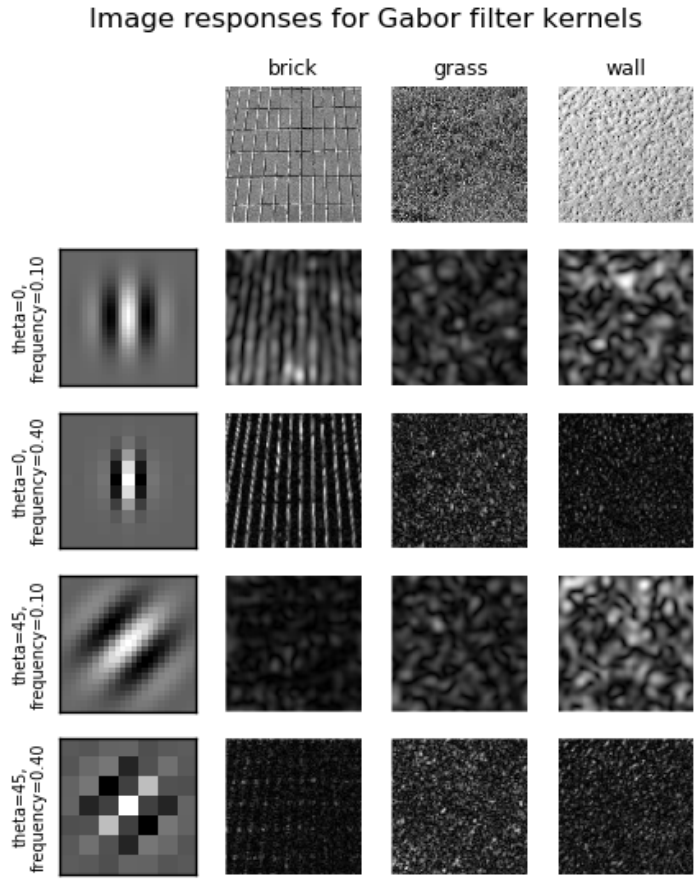



Figure 2.2. Example responses of Gabor kernels [2].
Example images are in the first row and kernels are in the first column.

### 2.2.3. Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) [3] is another popularly used histogram-based image descriptor, which is originally proposed for pedestrian detection. HOG descriptor counts the frequencies of gradients in discretized directions over local regions of the image. In this thesis, the HOG variant in [30] is used, which gives a 31-dimensional feature vector per local window.

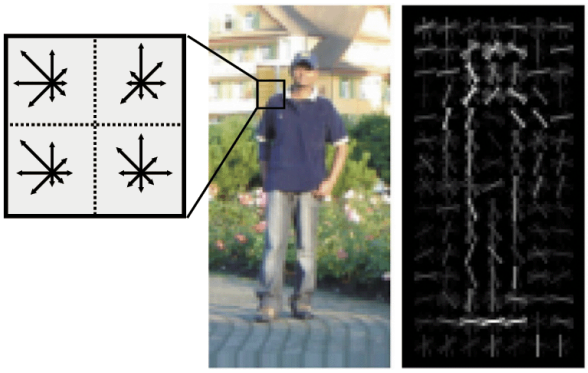


Figure 2.3. Example HOG descriptors [3].

### 2.2.4. Scale-invariant Feature Transform

Scale-invariant Feature Transform (SIFT) is a two-stage feature extraction method in which the keypoints in an image are detected in the first phase, and the interest points are described with the statistics of gradient directions in windows around those keypoints [31]. In this thesis, assuming the images are already spatially aligned, we use the Dense-SIFT variant, which replaces the keypoint detection stage with a regular grid over the image. The description part results in a 128-dimensional feature vector per local window of the image.

### 2.2.5. Local Phase Quantisation

Local Phase Quantisation (LPQ) is a popularly used image descriptor that is known to be resistant to blurring [32]. The descriptor summarizes the phase information obtained using the short-term Fourier transform (STFT) computed over a local window of the image. Using four frequency points and recording the real and imaginary parts of the Fourier coefficients, an 8-bit number per local region is obtained, which is then used to summarize the whole image with a 256-bin histogram.

### 2.2.6. Deep CNN Features

Deep convolutional neural networks (DCNN) are typically implemented as feed-forward neural networks which takes an input $x$ and processes it with a series of activation functions:

$$f(x) = f_L(...f_2(f_1(x; w_1); w_2)...), w_L) \tag{2.3}$$

where $f_l$ denotes the activation function at layer $l$ and $w_l$ is the corresponding parameter set. The organization of layers, *i.e.* determining the number and dimensions of convolution filters, is typically done manually. However, the values of the parameters, *i.e.* $w_l$ are learned from the data in a supervised manner.

Low-level layers of DCNNs typically consist of convolution and pooling layers. The convolution operation in two-dimensional domain is defined as follows:

$$z[m, n] = f[m, n] * g[m, n] = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f[u, v]g[m - u, n - v] \tag{2.4}$$

For practicality, deep learning libraries often implement cross-correlation instead of convolution, which is almost identical to convolution, only with the kernel $g$ flipped:

$$z[m, n] = f[m, n] * g[m, n] = \sum_{i} \sum_{j} f[m + i, n + j]g[m, n] \tag{2.5}$$

Another important aspect of DCNNs is pooling, which is the operation of combining nearby feature values by the applying a suitable operator, therefore reducing the dimensionality of the feature map for the next layer. Common choices for the operator include max-pooling (using the max operator) or sum-pooling (using summation), or stochastic pooling (choosing a value randomly with a probability that depends on the

activation value). For example, max-pooling is defined as follows:

$$z_{ijk} = max\{z_{i'j'k} : i \leq i' < i + p, j \leq j' < j + p\} \tag{2.6}$$

Training a DCNN, or fine-tuning a pre-trained network is usually done with the backpropagation algorithm. The objective function that needs to be optimized under supervision involves a loss function $\ell(z, \hat{z})$ that quantizes the penalty for predicting $\hat{z}$ where the true label was $z$. Averaging over all instances, the empirical loss function is defined as follows:

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(z_i, f(x_i; w)) \tag{2.7}$$

The loss can be related to modeling the distribution of the data in a generative application, however, in this thesis, DCNNs are used to learn discriminative tasks such as regression and classification. During the fine-tuning of DCNNs as multi-class classifiers, we use the softmax loss, or the cross-entropy loss, which is defined as follows:

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \tag{2.8}$$

In order to minimize the objective function $L$, a gradient descent based method is used, which is defined by starting from an initial set of parameters and compute the update that is along the fastest descent direction of $L$:

$$w^{t+1} = w^t - \eta_t \frac{\partial f}{\partial w}(w^t) \tag{2.9}$$

where $\eta \in \mathbb{R}_+$ is the learning rate.

Pre-trained and fine-tuned versions of a variety of deep convolutional neural networks (DCNN) are used to extract features. An example feature extraction from DCNNs are shown in Figure 2.4. Fine tuning of these DCNNs are explained in further detail in Section 2.5.4.

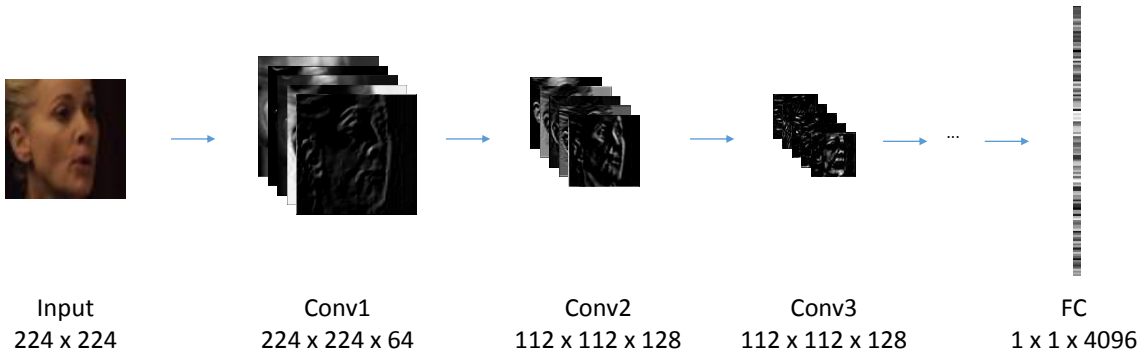



Figure 2.4. Feature extraction from the VGG-Face network.

DCNNs encode low-level information such as basic shapes and lines in the lower layers and contain higher level information that is related to the task they are trained for in the final layers. The dimensionality of the feature vector decreases with the layer height. In order to reduce computational requirements, we extract the features from the final, fully-connected layers of the networks, which usually results in a feature dimensionality of 1024 to 4096.

### 2.2.7. Geometric Features

In addition to appearance descriptors, we also encode the shape information with a set of geometric features extracted from the landmarks. The set of landmarks given by IntraFace, alongside with their indices, are displayed in Figure 2.5.

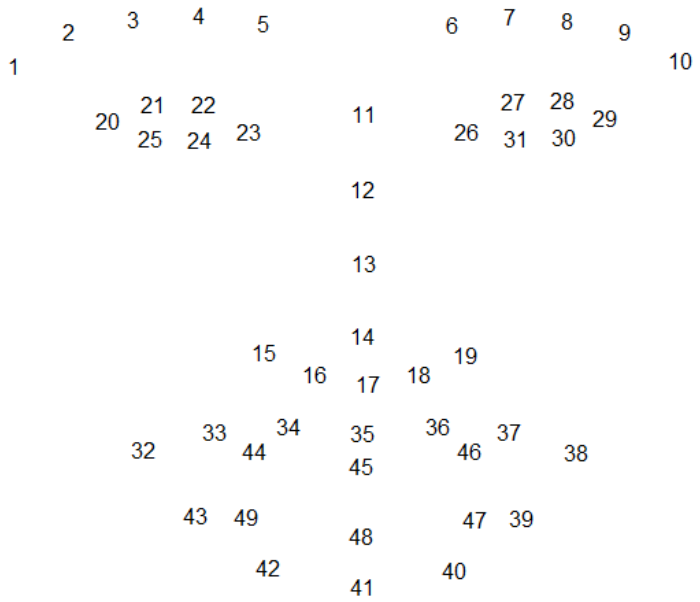



Figure 2.5. Landmarks given by IntraFace.

Using this set of 49 landmarks per face, we extract a number of geometric features that mostly involve the normalized lengths, ratios, angles, and curvatures related to the fiducial points around the mouth and the eyes. The geometric features used in this thesis are displayed in Table 2.1.

## 2.3. Video Features

Video analysis differs from image analysis in the sense that it requires an additional description method over the image-based feature extraction methods. In order to tackle this problem, we employ two classes of approaches, the first one being features from three orthogonal planes (see Section 2.3.1), and the other methods include various encoding methods (see Sections 2.3.2 and 2.3.3).

### 2.3.1. Features from Three Orthogonal Planes

In order to describe a 3D data such as a video, traditional appearance descriptors (such as LBP, LGBP HOG, SIFT and LPQ) are extracted from a video in a volumetric way, resulting in descriptors called LBP-TOP, LGBP-TOP and LPQ-TOP. An example

Table 2.1. Hand-crafted geometric features.

LR: Averaged features from left and right parts of the face.

Features 18 through 23 are motivated by [33].

| Feature # | Explanation | Landmarks Involved | Feature Type |
|---|---|---|---|
| 1 | Eye aspect ratio (LR) | [20:25], [26:31] | Distance |
| 2 | Mouth aspect ratio | 32, 35, 38, 41 | Distance |
| 3 | Upper lip angles (LR) | 32, 35, 38 | Angle |
| 4 | Nose tip - mouth corner angles (LR) | 17, 32, 38 | Angle |
| 5 | Lower lip angles (LR) | [32, 42] , [38, 40] | Angle |
| 6 | Eyebrow slope (LR) | [1, 5] , [6, 10] | Angle |
| 7,8 | Lower eye angles (LR) | [20, 23, 24, 25], [26, 29, 30, 31] | Angle |
| 9 | Mouth corner - mouth bottom angles | 32, 38, 41 | Angle |
| 10 | Upper mouth angles (LR) | [32, 34], [36, 38] | Angle |
| 11 | Curvature of lower-outer lips (LR) | [32, 43, 42], [38, 39, 40] | Curvature |
| 12 | Curvature of lower-inner lips (LR) | [32, 42, 41], [38, 40, 41] | Curvature |
| 13 | Bottom lip curvature | [32, 38, 41] | Curvature |
| 14 | Mouth opening / mouth width | 45, 48, 32, 38 | Distance |
| 15 | Mouth up/low | 35, 41, 45 | Distance |
| 16 | Eye - middle eyebrow distance (LR) | [3, 20, 23], [8, 26, 29] | Distance |
| 17 | Eye - inner eyebrow distance (LR) | [5, 20, 23], [6, 26, 29] | Distance |
| 18 | Inner eye - eyebrow center (LR) | [3, 23], [8, 26] | Distance |
| 19 | Inner eye - mouth top distance | 23, 26, 35 | Distance |
| 20 | Mouth width | 32, 38 | Distance |
| 21 | Mouth height | 35, 41 | Distance |
| 22 | Upper mouth height | 32, 38, 35 | Distance |
| 23 | Lower mouth height | 32, 38, 41 | Distance |

LBP-TOP feature extraction is illustrated in Figure 2.6.

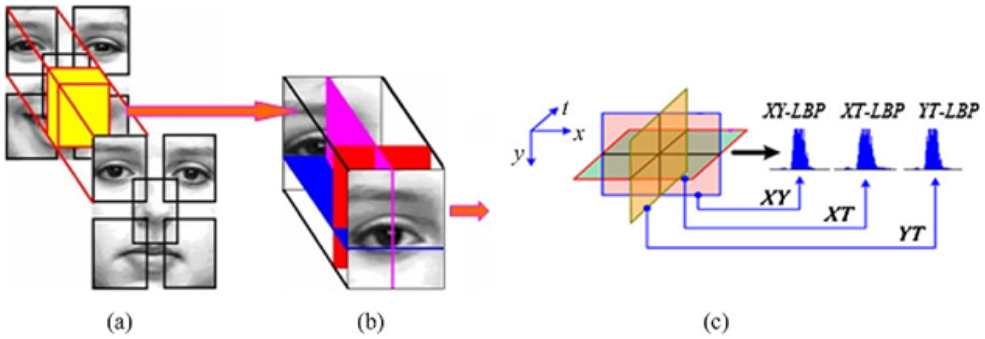



Figure 2.6. Example LBP-TOP feature extraction.

### 2.3.2. Bag of Features

Bag of features (BOF) (or bag of visual words) model is an unsupervised feature encoding method that is initially proposed for text-based document retrieval [34]. This technique is also widely used in computer vision [35], replacing the term "word" with "visual word", or simply "feature".

The idea behind BOF is a two-step encoding of an input (or its features) of any sample size into a histogram-based representation of fixed dimensionality. In the first step, a number ($K$) of important/discriminative words are determined from the training set. This is often done by k-means clustering. This set of words, *i.e.* the $K$ cluster centers form a codebook that is called as the "vocabulary". In the second step, all the words, *i.e.* local instances are queried against the vocabulary and each word is assigned to the nearest cluster center. By counting the number of occurrences of each cluster center, a $K$-dimensional histogram representation is obtained, which is then used as a feature vector to the subsequent learning processes.

The resulting feature vectors in BOF are often normalized via $\ell_1$ or $\ell_2$ normalization, which are achieved by dividing every element of a vector by the sum of the absolute values of its elements, and dividing every element to the sum of the squares of its elements, respectively. The formal definitions of $\ell_1$ and $\ell_2$ normalization are

provided in the following equations.

$$||z||_1 = \sum_{i=1}^{N} |z_i| \tag{2.10}$$

$$||z||_2 = \sqrt{\sum_{i=1}^{N} |z_i|^2} \tag{2.11}$$

In computer vision, words are usually learned from the set of local appearance descriptors such as LBP, HOG and SIFT, each modality resulting in a different codebook. Then all the local instances are encoded with the corresponding codebook to obtain the bag of visual words representation. Such a feature processing is useful in terms of enabling the comparison of inputs of different dimensionality, as well as providing a spatially-invariant feature representation.

### 2.3.3. Fisher Vectors

Fisher Vector (FV) encoding is proposed recently by Perronnin and Dance [36]. The method is very similar to bag of features encoding, with two differences. In the vocabulary calculation phase, instead of k-means clustering, Gaussian mixture models are used. This enables to capture additional statistics like mean and variance of the visual word distribution. The second difference is in the encoding phase. The BOF model only includes 0-th order information, *i.e.* frequencies. However, FV encodes both first and second order statistics, although it can also encode 0-th order statistics. In this case, it degenerates to BOF. But using a markedly smaller number of clusters compared to those typically used in BOF, the contribution of 0-th order statistics are shown to be insignificant in FV encoding.

The power of FV encoding comes from the fact that it combines generative and discriminative approaches. Given a generative visual codebook, FV calculates the difference that needs to be applied to the vocabulary in order to fit the data better [37].

In order both to decrease the size of the resulting feature and to obtain decorrelated features (hence achieve diagonal covariance matrices), one can apply principal components analysis (PCA) prior to building the vocabulary in BOF or the background probability model (GMM) in FV.

### 2.3.4. Functional statistics

Inspired by acoustic feature extraction techniques used in paralinguistic challenges [38, 39], we encode each low-level descriptor (LLD) with a number of functional statistics (FUN) that include mean, standard deviation, slope, offset and curvature. Using these five functionals, the input frame-wise feature dimensionality $D$ is converted to a $5D$-dimensional video descriptor. The statistics used in this thesis are explained in detail in the following paragraphs.

We calculate the arithmetic mean $\mu$ of each frame-wise LLD over the number of frames $N$ as follows:

$$\mu = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \tag{2.12}$$

The standard deviation $\sigma$ is defined as the square root of the variance:

$$\sigma = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x(n) - \mu)^2} \tag{2.13}$$

We fit a first-order polynomial to the data that is parameterized as follows:

$$y = ax + b \tag{2.14}$$

We use parameters $a$ (slope) and $b$ (offset) to describe the LLDs' dynamics over time. Similarly, we fit a second-order polynomial to the data as follows:

$$y = ax^2 + bx + c \tag{2.15}$$

In the second-order polynomial fit, we only use the parameter $a$ in order to describe the temporal curvature of the data points.

## 2.4. Audio Features

For the video processing pipeline, we include audio-based features that are inspired by the baseline systems provided by computational paralinguistics challenges [38–41].

The acoustic features include LLDs such as energy, spectral and cepstral coefficients, voicing-related features, logarithmic harmonic-to-noise ratio, spectral harmonicity, and psychoacoustic spectral sharpness. These LLDs are often post-processed with functional statistics on sliding windows in order to describe the features temporally, as described in 2.3.4. In order to extract these audio-related features, we use the open-source OpenSMILE toolkit [7]. On audio modality of video corpora, we analyzed several standard configurations with the OpenSMILE toolkit and found the version used in INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) the most effective [39]. This feature set contains 6 373 suprasegmental features, where 65 low-level descriptors as well as their first order temporal derivatives (see Table 2.2) are summarized by 54 statistical functionals (see Table 2.3). Note that some functionals are not applied to all LLD contours.

Table 2.2. 65 low-level descriptors used in INTERSPEECH 2013 ComParE Challenge [39].

| |
|---|
| 4 energy related LLDs |
| Sum of auditory spectrum (loudness) |
| Sum of RASTA-style filtered auditory spectrum |
| Root Mean Square Energy |
| Zero-Crossing Rate |
| 55 Spectral LLDs |
| RASTA-style auditory spectrum, bands 1-26 (0–8 kHz) |
| Mel-Frequency Cepstral Coefficients (MFCC) 1–14 |
| Spectral energy 250–650 Hz, 1 k–4 kHz |
| Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90 |
| Spectral Flux, Centroid, Entropy |
| Skewness, Kurtosis, Variance, Slope |
| Psychoacoustic Sharpness, Harmonicity |
| 6 voicing related LLDs |
| $F_0$ by SHS + Viterbi smoothing |
| Probability of voicing |
| Logarithmic Harmonics to Noise Ratio (HNR) |
| Jitter (local, delta), Shimmer (local) |

Table 2.3. Applied Functionals. [1]: Arithmetic mean of LLD / positive Δ LLD, [2]: Only applied to voice related LLD, [3]: Not applied to voice related LLD except $F_0$, [4]: Only applied to $F_0$.

| Functionals applied to LLD / Δ LLD |
|---|
| quartiles 1–3, 3 inter-quartile ranges |
| 1 % percentile (≈min), 99 % percentile (≈ max) |
| position of min / max, percentile range 1 – 99% |
| arithmetic mean[1], root quadratic mean |
| contour centroid, flatness |
| standard deviation, skewness, kurtosis |
| rel. duration LLD is above / below 25 / 50 / 75 / 90 % range |
| rel. duration LLD is rising / falling |
| rel. duration LLD has positive / negative curvature[2] |
| gain of linear prediction (LP), LP Coefficients 1–5 |
| mean, max, min, std. dev. of segment length[3] |
| Functionals applied to LLD only |
| mean / std.dev. of peak distances |
| mean value of peaks |
| mean value of peaks – arithmetic mean |
| mean / std.dev. of rising / falling slopes |
| mean / std.dev. of inter maxima distances |
| amplitude mean of maxima / minima |
| amplitude range of maxima |
| linear regression slope, offset, quadratic error |
| quadratic regression a, b, offset, quadratic error |
| percentage of non-zero frames [4] |

## 2.5. Model Learning

In this section, we provide explanations of various model learning algorithms, *i.e.* classifiers and regressors that are used in the recognition parts of various applications in this thesis.

### 2.5.1. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm first proposed by Cortes and Vapnik in 1995 [42]. The algorithm is a kernel-based discriminative learner that aims to solve binary classification problems, but the algorithm can also be modified to learn regression.

SVM aims to find a discriminant function that maximizes the margin between two classes. This is done by mapping the training feature vectors into a higher dimensional space and compute a separating hyperplane in the new space. An SVM classifier is represented by "support vectors", which are training data points that are closer to the separating hyperplane.

### 2.5.2. Extreme Learning Machine

Extreme Learning Machine (ELM) is a supervised classification and regression method proposed by Huang *et al.*, which is a very fast and robust learning algorithm compared to conventional popular algorithms [43].

In this section, we present the learning strategy of ELM and various applications derived from it. The proposed algorithm aims to map the input features to the labels using a single-hidden-layer feed-forward network (SLFN) architecture. The main difference from traditional artificial neural networks is that the hidden layer need not be tuned, but it's initialized randomly in terms of weights and biases [44]. The SLFN architecture of ELM is shown in Figure 2.7.

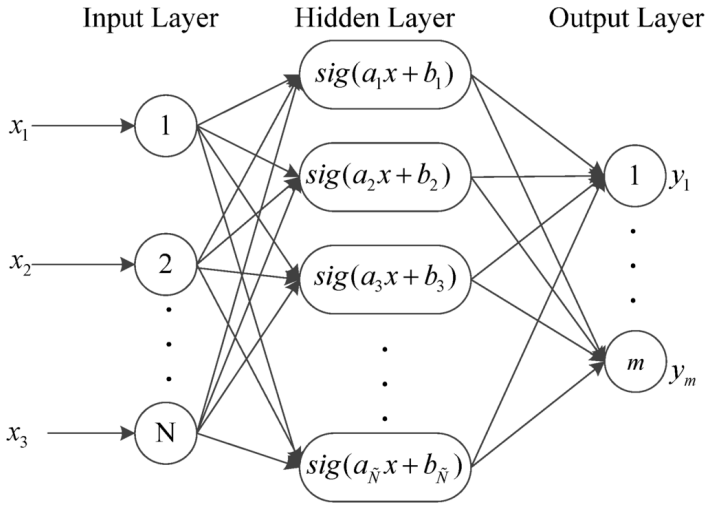



Figure 2.7. Single-hidden-layer feed-forward architecture of ELM [4].

The hidden layer output matrix $H \in \mathbb{R}^{N \times h}$, where $N$ and $h$ denote the number of instances and the hidden neurons, respectively, is what connects the input (features) to the output (labels). In the basic version of ELM, $H$ is generated in an unsupervised, even random manner. The actual learning takes place in the second layer between $H$ and the label matrix $T \in \mathbb{R}^{N \times L}$, where $L$ is the number of classes. $T$ is a vector of continuous annotations in case of regression. In the case of $L$-class classification, $T$ is represented in one vs. all coding:

$$T_{t,l} = \begin{cases} +1 & \text{if } y^t = l \\ -1 & \text{if } y^t \neq l \end{cases} \tag{2.16}$$

The second level weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $H\beta = T$. The output weights can be learned via:

$$\beta = H^{\dagger}T \tag{2.17}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse [45] that gives the minimum $L_2$ norm solution to $||H\beta - T||$, simultaneously minimizing the norm of $||\beta||$. To increase the robustness and generalization capability, the optimization problem of ELM is reformulated using a regularization coefficient on the residual error $||H\beta - T||$. The learning

rule of this alternative ELM is related to Least Square SVMs (LSSVM) via the following output weight learning formulation, where $I$ is the $N \times N$ identity matrix and $C$, which is used to regularize the linear kernel $HH^T$, corresponds to the complexity parameter of LSSVM [46]:

$$\beta = H^T(\frac{I}{C} + HH^T)^{-1}T \tag{2.18}$$

This formulation is further simplified by noting that the hidden layer matrix need not be generated explicitly given a kernel $\Omega$, which can be seen identical to Kernel Regularized Least Squares [43, 47]:

$$\beta = (\frac{I}{C} + \Omega)^{-1}T \tag{2.19}$$

The kernel $\Omega(x, y)$ can be calculated in various ways. In this thesis, we try three popular alternatives, which are the linear kernel:

$$\Omega(x, y) = x^T y \tag{2.20}$$

the radial basis function (RBF) kernel:

$$\Omega(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{2.21}$$

and the polynomial kernel:

$$\Omega(x, y) = (x^T y + c)^d \tag{2.22}$$

Since the RBF and polynomial kernels have additional parameters that need to be optimized, we often use the linear kernel in this thesis for simplicity and to avoid the risk of overlearning.

### 2.5.3. Partial Least Squares

PLS is a method to learn a regression model between two variables $X \in \mathbb{R}^{N\times d}$ and $Y \in \mathbb{R}^{N\times p}$. The model is trained by decomposing the matrices as $X = U_x V_x + r_x$, $Y = U_y V_y + r_y$, where $U$ represents the latent factors, $V$ represents the loadings and $r$ is the residuals. The projection weights $W_x, W_y$ of this decomposition are found such that they maximize the covariance of corresponding columns of $U_x = XW_x$ and $U_y = YW_y$. In this thesis, PLS is used as a one-versus-all classification method where $X$ stands for the feature (or kernel) matrix and $Y$ is the binary label vector. After the projections are learned, the class with the biggest regression score is taken as the predicted class. This model depends on only one parameter, which is the number of latent factors. This parameter is tuned via cross-validation.

### 2.5.4. Transfer Learning from Deep Neural Networks

Convolutional Neural Network (CNN), first proposed by LeCun *et al.* [48] is usually the main data structure in a deep learning system. Typically, a CNN consists of convolution and pooling layers in the lower levels and fully connected layers followed by a modeling layer at the end, which produces the targeted estimations. During the training phase, back propagation is used to update the parameters of the network.

With the recent advancement of computational resources, Deep CNNs (DCNN) have become a reality and achieved state-of-the-art results in almost all machine learning tasks such as face recognition [49], age estimation [50–53] and emotion recognition [8, 54].

The performance of a DCNN depends on a number of factors such as input normalization, convolution filters, pooling operators, regularization and the choice of objective function.

Another advantage of DCNN over other traditional feature extraction and model learning techniques is that DCNNs can be used to replace any or both steps of the

analysis pipeline. A DCNN can be trained from scratch to optimize the task at hand, a DCNN that is pre-trained on a similar task can be fine-tuned in order to optimize the result, and it can be used only as a feature extraction method as showed in [11]. Comparison of these various uses of DCNNs is provided in [55].

# 3. GENERAL METHODOLOGY

In this thesis, we tackle the problem of video and image based face analysis for a number of applications including emotion recognition, apparent age estimation and apparent personality trait estimation.

For most of the applications, the pipeline of our system is very similar. We start by a face detection/alignment step using the methods described in Section 2.1. For image-based challenges, we often use the DPM-based face detector because of its accuracy and no need for landmark detection and an extra alignment step. For video-based challenges where using DPM is infeasible in terms of time, we use the Viola & Jones face detector of IntraFace which results in a much faster but slightly less accurate face detection. We provide the detection statistics of the two detectors in Tables 6.1 and 5.1. This shows that the Viola & Jones face detector catches 95.6% of the faces whereas the DPM-based face detector yields 97.5% hit rate.

For feature extraction from images, we use the methods described in Sections 2.2. Dense appearance descriptors (such as LBP, HOG and SIFT) are often used with the IntraFace alignment where we crop the face strictly from outer landmarks. This approach ensures the visual descriptors do not contain noisy background information. DCNN features need to be extracted from a wider bounding box of the face, therefore we add 40% interocular distance as extra padding in our alignment.

For video-based challenges, we process (or extract) the appearance descriptors using the methods described in 2.3. We also include the features related to speech using audio feature extraction techniques explained in 2.4. The functional statistics encoding of audio and video features ensures that the resulting feature vectors are of the same size for each video. For appearance descriptors from three orthogonal planes, we achieve this by dividing the video into two non-overlapping temporal windows and concatenating their feature vectors.

Before feeding the feature vectors into our modeling algorithm, we often normalize the data using a combination of min-max normalization to $[-1, 1]$ and $\ell_2$ normalization. Normalized feature vectors are then transformed into a kernel representation using techniques described in 2.5.2. ELM computes the output weights that relate the kernel to the label matrix. We later apply these weights to the test kernel. The regularization parameter (and the kernel parameters in case of RBF and polynomial kernels) are optimized by a grid search where the performance with each parameter set is estimated using k-fold subject-independent cross validation.

Depending on the application, the architecture of the learning algorithm can change. In the case of emotion recognition (see Chapter 4), we use kernel ELM as a multi-class classifier. For apparent personality trait estimation (see Chapter 5), we learn a separate regression model for each personality trait. For age estimation (see Chapter 6), we first train multiple one-versus-all classifiers to assign each sample into a number of overlapping age groups, we then use the estimations of local kernel ELM regressors that are trained within the relevant group.

We visualize our general workflow in Figure 3.1.

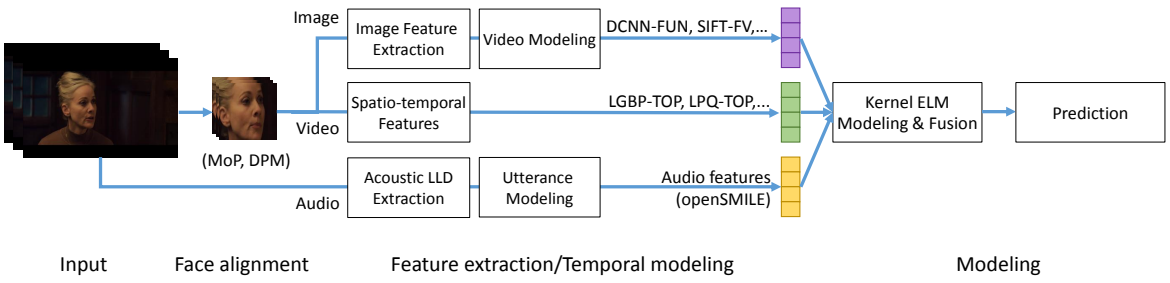



Figure 3.1. Overview of the proposed face analysis system.

# 4. VIDEO BASED EMOTION RECOGNITION

Emotion recognition from facial images and videos typically fall into two categories; multi-class classification between the 6 basic emotion classes and continuous estimation in the arousal-valence space. Facial emotion analysis often involves face detection, registration, feature extraction and emotion modeling.

In this chapter, we provide a brief literature review on studies related to emotion recognition, followed by a description of the methodology we propose for this task.

## 4.1. Related Work

In the last decade, studies in multimodal emotion recognition gained momentum and reached maturity on lab-controlled data. Currently, there is a shift towards challenging "in the wild" conditions that include background noise, large idiosyncratic variance and sensor-related differences. Since affective data collection in natural conditions is hard, costly and even impossible in some cases (due to ethical issues), sharing the limited data is of crucial importance to push forward the state-of-the-art. Moreover, fixed-protocol challenges in this field provide a unique opportunity to compare a variety of approaches under the same conditions.

Among the series of events carried out in this vein, the Emotion Recognition in the Wild (EmotiW) challenge provides out of laboratory data -Acted Facial Expressions in the Wild (AFEW)-, collected from videos that mimic real life [56–60]. In 2016, in addition to the video-based challenge, the EmotiW series introduced an image-based group happiness intensity level estimation challenge [56, 61].

## 4.2. Emotion Recognition in the Wild (EmotiW) Challenge

EmotiW challenge dataset, also known as Acted Facial Expressions in the Wild (AFEW) dataset, consists of short videos extracted from movie scenes, hence serves

as an in-the-wild facial expression dataset [60]. The dataset is divided into training, validation and test partitions and test set labels are not provided. The performance evaluation of the challenge is the overall classification accuracy on the testing dataset. The number of given and aligned videos are provided in Table 4.1.

Table 4.1. Summary of the AFEW-5 dataset.
Instance distribution over partitions, alignment and modality types. GA: Given alignment, IF: IntraFace-based alignment.

| # | Train | Val | Test |
|---|---|---|---|
| Clips | 711 | 383 | 539 |
| GA | 698 | 369 | 539 |
| IF | 663 | 340 | 455 |
| Audio | 707 | 383 | 539 |

In the EmotiW challenge, the performance measure is the overall multi-class classification accuracy. For readability purposes, we report this score as a percentage. The challenge organizers provide a baseline system which achieves 39.13% classification accuracy on the validation set.

## 4.3. Proposed Approach

In this thesis, we propose a multimodal emotion estimation method that uses audio, appearance and DCNN features and their weighted score-level fusion with kernel ELM. Our proposed method is visualized in Figure 4.1.

For dense appearance descriptors (LBP, LGBP, HOG, SIFT and LPQ), we use the alignment given by the challenge organizers, which crops the face from a tight bounding box. Since DCNNs require a wider alignment, we process input videos with the DPM-based face detector and enhance the bounding box by 40%. We register each detected face as a $224 \times 224$ image in order to fit the input dimensionality of the DCNN. Figure 4.2 shows examples of the given alignment and the DPM-based alignment.

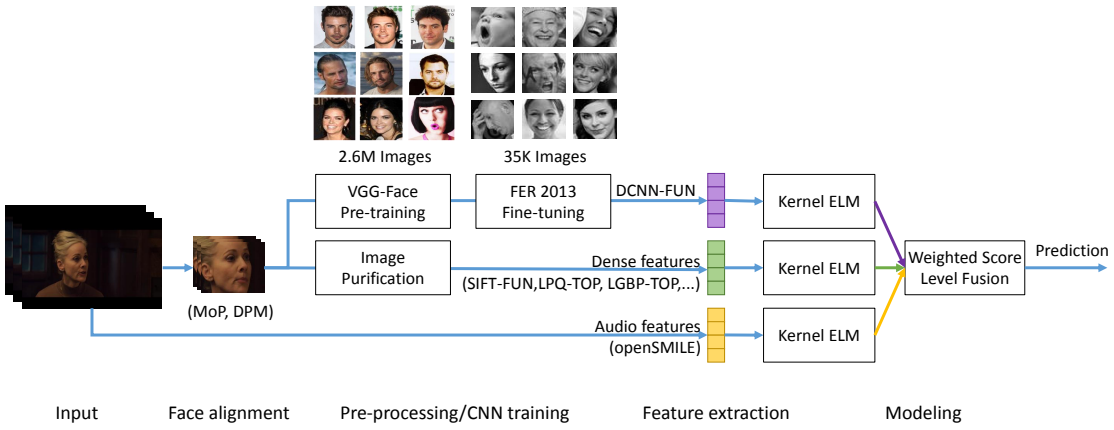



Figure 4.1. Overview of the proposed emotion recognition system.


Figure 4.2. Example face alignment in AFEW-5 dataset.
The first row shows the alignment given by challenge organizers, whereas the second row shows our DPM-based alignment.

We extract deep features from the VGG-Face network that is originally trained for face recognition [49]. We also fine-tune the network with the Facial Expression Recognition (FER) 2013 dataset that contains 35,000 images labeled with the seven basic emotion classes [62]. For this, we remove the final layer of VGG Face and replace it with a seven dimensional fully connected layer. In the following paragraphs, we describe the fine tuning process in detail and present the effect of this fine tuning on classification accuracy in Table 4.4.

For the fine-tuning of the VGG-Face network for the emotion recognition task, we investigated various options in our preliminary analysis. We found that combining weight decay and dropout for regularization gives the best results on the FER validation

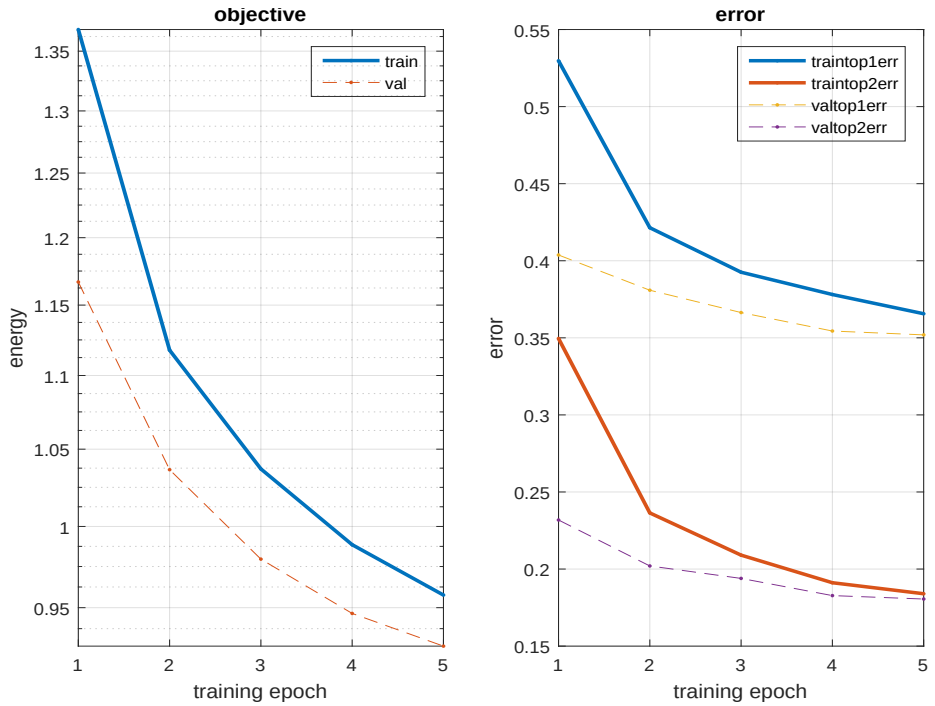
set. We carried out a multi-stage fine tuning. In the first stage, we fine tune on the FER public test set and run weight updates for five epochs. In the second stage, we update the upper layers (higher than layer 27) using the FER private test set and update for another ten epochs. We finally use the AFEW validation set for the third fine tuning stage.

In Figure 4.3, we show the training curves of two stage fine-tuning of the network with FER dataset, where we set the learning rate and weight decay to 0.0005, momentum to 0.9, and dropout probability to 0.8. We observe that the validation set error is lower than that of the training set, which suggests that overlearning is not an issue here. Note that these curves represent error on FER training and validation partitions, and not AFEW.

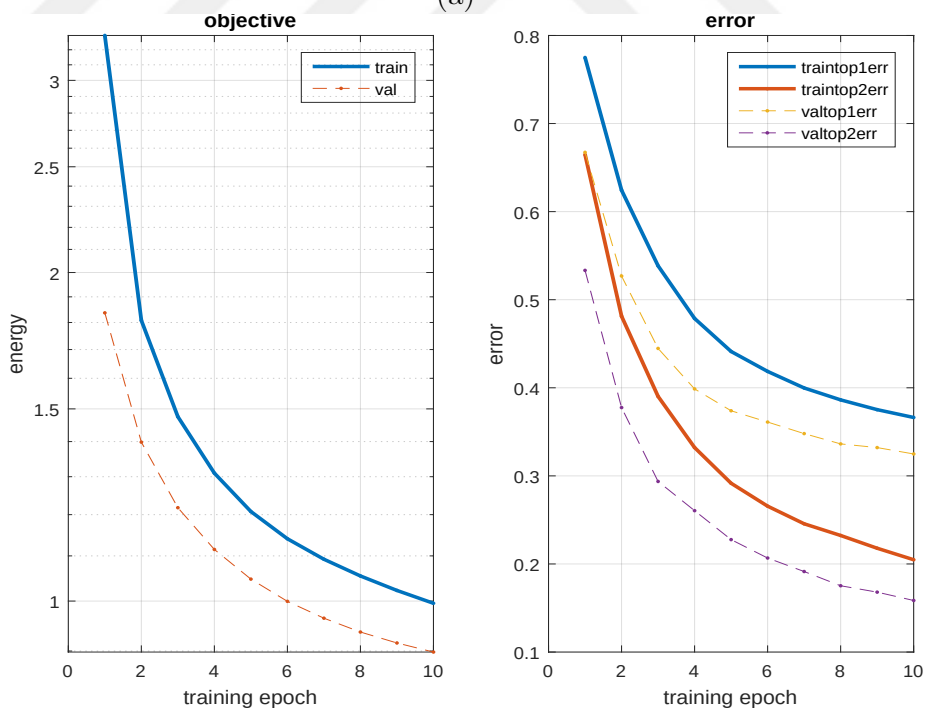
For model learning with kernel ELM, we try two different kernels, namely the linear kernel and the RBF kernel. For the linear kernel $K(x, y) = x^T y$, the only parameter of the classifier is the regularization coefficient $C$ (see Equation 2.19). We optimize this parameter with a 5-fold cross validation on the development set, where we try values from the set $2^{[-6,-4,...,4]}$ and choose the one that yields the best classification accuracy. For the RBF kernel, we use a two-dimensional grid search in order to also optimize the parameter $\sigma$ in Equation 2.21.

## 4.4. Experimental Results

For the EmotiW challenge, we provide the overall classification accuracies with various visual descriptors, learning algorithms and kernel types in Table 4.2. From this table, we see that the audio features yield a performance that is below the challenge baseline, and the DCNN features provide a performance similar to conventional appearance descriptors. This is due to the fact that the alignment does not contain the extra padding around the face that the DCNN is accustomed to. We increase the explanatory power of DCNN features with our DPM-based alignment.




(a)




(b)

Figure 4.3. Fine-tuning VGG-Face with dropout on FER-2013.
(a) public (b) private test set. The left column shows the softmax loss whereas the right column shows the top-1 and top-2 classification errors.

Note that since our DPM alignment contains faces cropped with a big padding which includes unnecessary information in dense appearance descriptors, we initially use the alignment given by the challenge organizers (see Figure 4.2) for extracting visual features.

Table 4.2. Performance of audio-visual features in AFEW-5 validation set.

| | Lin | | RBF | |
|---|---|---|---|---|
| Feature | ELM | PLS | ELM | PLS |
| Audio | *36.59* | *36.29* | 35.51 | 34.73 |
| Visual Features with Given Alignment | | | | |
| DCNN-FUN | 43.40 | 41.78 | *44.47* | 43.13 |
| HOG-FUN | 39.02 | *44.99* | 41.46 | 42.01 |
| SIFT-FUN | 40.92 | *43.63* | 42.28 | 41.19 |
| SIFT-FV | 40.11 | *41.73* | 40.38 | 40.38 |
| LBP-TOP | 41.19 | 41.46 | *42.01* | 40.65 |
| LGBP-TOP | 43.63 | 43.90 | *44.44* | 44.17 |
| LPQ-TOP | 40.65 | 41.19 | 40.65 | *42.01* |

We also provide the pairwise fusion performances of the features extracted from the given alignment in Table 4.3. From here, we see that the best performance is obtained via combining the two convolution-based features, namely DCNN-FUN and LGBP-TOP.

Table 4.3. Best AFEW-5 validation set results of pairwise feature fusion/simple weighted score fusion on given aligned images.

| | HOG-FUN | LBP-TOP | LGBP-TOP | SIFT-FV |
|---|---|---|---|---|
| DCNN-FUN | 43.40/46.61 | 45.01/46.34 | *47.71*/46.88 | 44.74/45.26 |
| HOG-FUN | - | 45.82/47.70 | 44.47/47.15 | 43.67/*48.24* |
| LBP-TOP | | - | 42.86/44.44 | 44.74/45.26 |
| LGBP-TOP | | | - | 43.67/45.80 |

In order to improve the performance of our DCNN classifiers, we use our DPM-based alignment that crops the face with a wider bounding box. We show this increase in performance alongside with the effect of fine-tuning the network in Table 4.4. For comparison, we do the same fine tuning on the VGG-M-2048 network, which is pre-trained for object recognition [63]. This shows that fine tuning a network that is pre-trained on a face-related task performs better in emotion recognition.

Table 4.4. Comparison of AFEW-5 validation set accuracies (%) of DCNN features over DCNN model, fine tuning and classifier/kernel alternatives.

| | | ELM | | PLS | |
|---|---|---|---|---|---|
| DCNN Model | Version | Linear | RBF | Linear | RBF |
| VGG-Face [49] | Original | 39.89 | 42.29 | 41.76 | 44.15 |
| | Fine-tuned | 48.94 | 48.67 | 49.20 | *50.80* |
| VGG-M-2048 [63] | Original | 36.68 | 36.15 | 35.88 | 35.09 |
| | Fine-tuned | 41.42 | 42.74 | 37.47 | 39.58 |

Finally, we try various combinations of the feature extraction and alignment schemes and we provide the classification accuracy in validation and test sets of the AFEW-5 dataset in Table 4.5. In the first part, we show the best performance without DCNN features that had the $2^{nd}$ place in the challenge [60]. In the second part, we show the improvement in classification performance with the introduction of DCNN features. We obtain the best performance with a fusion of audio features and the best three appearance descriptors, which results in state-of-the-art (54.55%) classification accuracy on the test set (see Table 4.6).

Table 4.5. Validation and test set accuracies in AFEW-5.
WF: weighted score level fusion, FF: feature level fusion, $GA$: given alignment, $DPM$: DPM based alignment, $IF$: IntraFace alignment [18].

| System | Val | Test |
|---|---|---|
| WF(Audio, LBP-TOP$_{GA}$, LGBP-TOP$_{GA}$) | 50.14% | 50.28% |
| WF(Audio, LGBP-TOP$_{GA}$, LBP-TOP$_{GA}$, FF(SIFT-FUN$_{GA}$, GEO-FV$_{IF}$), LPQ-TOP$_{GA}$) | 52.30% | 49.17% |
| *WF (Audio, LGBP-TOP$_{GA}$, HOG-FUN$_{GA}$, SIFT-FV$_{GA}$, LBP-TOP$_{GA}$, LPQ-TOP$_{GA}$)* | *52.30%* | *53.62%* |
| DCNN-FUN$_{GA}$ | 44.47% | 42.86% |
| WF(Audio, DCNN-FUN$_{GA}$, SIFT-FV$_{GA}$ LBP-TOP$_{GA}$, LGBP-TOP$_{GA}$, HOG-FUN$_{GA}$) | 53.70% | 51.76% |
| DCNN-FUN$_{DPM}$ | 51.60% | 51.39% |
| *WF(Audio, DCNN-FUN$_{DPM}$, LGBP-TOP$_{GA}$, HOG-FUN$_{GA}$)* | *57.02%* | *54.55%* |

Table 4.6. Comparison of our approach with the top three systems of the EmotiW 2015 Challenge.

| Work | Val | Test |
|---|---|---|
| Ours | 57.02 | *54.55* |
| Yao et al. [64] | 49.09 | 53.80 |
| Kaya et al. [10] | 52.30 | 53.62 |
| Kahou et al. [65] | - | 52.88 |
| Baseline [60] | 36.08 | 39.33 |

# 5. VIDEO BASED PERSONALITY TRAIT ESTIMATION

It is not possible to judge the personality of a person by a mere glimpse of the face, but people attribute apparent personality traits for a face they newly encounter, in a stereotypical way, and with remarkable consistency [66].

It is not surprising that emotional expressions influence the attribution of personality traits. It is more likely for a smiling person to be perceived as more trustworthy and friendly. Todorov et al. convincingly argued that rapid, unreflective trait inferences from faces can influence consequential decisions [67]. This is why people do not typically use frowning or angry pictures in their resumés. Also, the context of the image can affect the perception of the face. In our proposed approach, we estimate emotional facial expressions, as well as cues from the context of the face to predict first impressions.

Before describing the followed approach, we provide a brief literature review on automatic personality trait recognition.

## 5.1. Related Work

In the past, various approaches have been used for recognizing apparent personality traits from different modalities such as audio [68, 69], text [70–72] and visual information [73, 74]. As in other recognition problems, multimodal systems are also investigated to improve robustness of prediction [75–78]. These works often aim to estimate the "Big Five" personality traits that are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN).

A number of scientific events have been dedicated to the subjects related to personality trait analysis, including the INTERSPEECH 2012 Speaker Trait Challenge [38], Audio/Visual Emotion Recognition Challenges (AVEC) [79–81], ACM ICMI 2014 Mapping Personality Traits (MAPTRAITS) Challenge [82], and the Emotion

Recognition in The Wild (EmotiW) Challenge series [57, 58, 60]. These challenges provided benchmarks, enabling objective comparisons of state-of-the-art systems. Our contribution to the current ChaLearn Looking at People (LAP) Challenge on First Impression Recognition follows the same idea.

## 5.2. ChaLearn First Impressions Challenge

The ChaLearn Looking at People - First Impressions (FI) dataset consists of 10,000 clips that are short (5 to 10 seconds) excerpts from 5,563 different YouTube videos, which are labeled by multiple human annotators for five personality traits [83]. Each clip is labeled for the Big Five personality traits that are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). Basic statistics of the dataset partitions alongside with the face detection rates are provided in Table 5.1.

The face poses are more or less uniform in terms of pose, but the resolution, lighting and background conditions are not controlled, providing an in-the-wild dataset. The performance evaluation in this dataset is 1-MAE, which is formulated as follows:

$$1 - \sum_{i}^{N} \frac{|\hat{y}_i - y_i|}{N} \tag{5.1}$$

where $N$ is the number of samples, $\hat{y}$ is the predicted label and $y$ is the true label ($0 \leq y \leq 1$). This score is then averaged over five tasks. This means the final score varies between 0 (worst case) and 1 (best case).

Table 5.1. Summary of ChaLearn First Impressions dataset.

| | Train | Val | Test |
|---|---|---|---|
| #Clips | 6,000 | 2,000 | 2,000 |
| #YouTube videos | 2,624 | 1,484 | 1,455 |
| #Given frames | 2.56M | 0.86M | 0.86M |
| #Detected frames | 2.45M | 0.82M | 0.82M |

## 5.3. ChaLearn Job Candidate Screening Challenge

Following apparent personality trait analysis, we employ the output of our model in ChaLearn Job Candidate Screening (JCS) Challenge. The challenge dataset consists of the same videos and personality trait assessments, with an additional variable that represents the average vote for each subject whether they would be called to a job interview or not. The quantitative phase of the challenge is to predict this interview variable with least mean absolute error. In the qualitative phase, the task is to produce explanations for the interview decision. The performance score in this stage is assessed by a committee according to several criteria such as clarity, explainability, soundness, model interpretability and creativity.

## 5.4. Proposed Approach

For apparent personality trait estimation from facial videos, we propose a multimodal fusion method that uses audio, appearance and deep learning features and their late fusion with kernel ELM. We illustrate our proposed method in Figure 5.1.

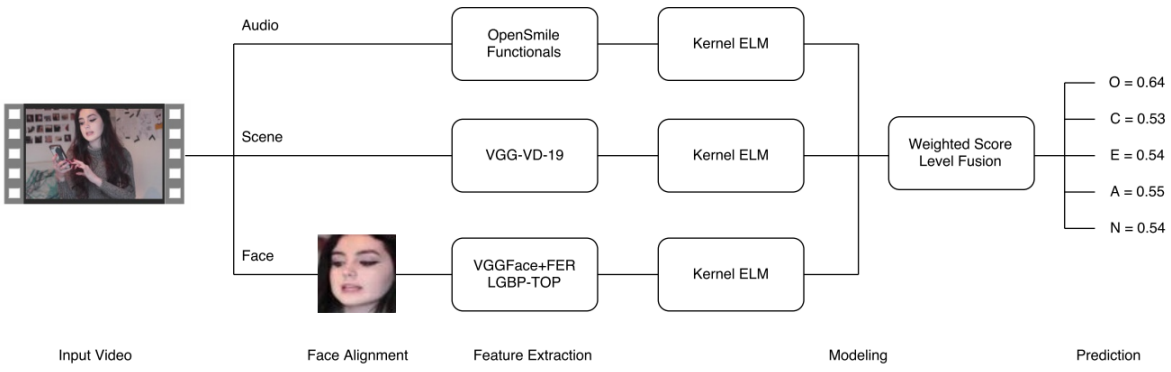



Figure 5.1. Overview of the proposed apparent personality estimation system.

For face detection and alignment, we use the IntraFace library due to its speed compared to the DPM-based face detector. The number of given and detected frames are summarized in Table 5.1.

We extract audio and face-based features, as well as scene features that are extracted from the first frame of each input video. For the scene features, we use the VGG VD-19 network that is trained for object recognition [84]. Features from these three modalities are then used to train kernel ELM regressors for each personality trait, which we combine using weighted score level fusion in order to get the final estimation. The parameters of the kernel ELM regressors are optimized via 5-fold subject-independent cross-validation on the training set.

For the DCNN feature extraction from faces, we use both the original VGG Face network as well as our fine tuned version for emotion recognition, as described in Section 4.3, and we show that the fine tuned network performs better in apparent personality trait recognition as well.

For the quantitative part of the JCS challenge, we followed a slightly different fusion scheme on the same set of features. As illustrated in Figure 5.2, the predictions of the ELM models are stacked to a random forest, which is an ensemble of decision trees grown with a random subset of instances (sampled with replacement) and a random subset of features. The randomness in both features and samples allow diversity of the base learners and help avoid over-fitting [85].

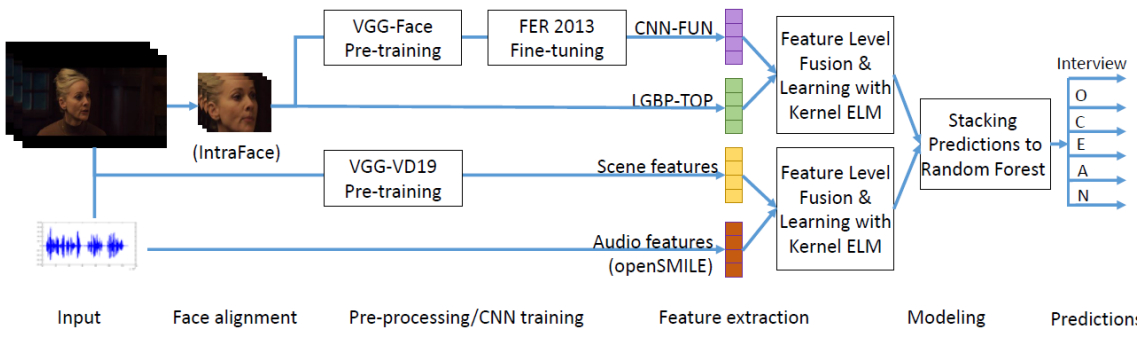



Figure 5.2. Overview of the proposed trait prediction and JCS system.

For the qualitative stage, the personality trait predictions from the quantitative model are binarized by comparing with their training set mean value (0/1 corresponding to low/high scores). The five binarized personality scores are then mapped to the binarized ground truth interview variable using a decision tree classifier.

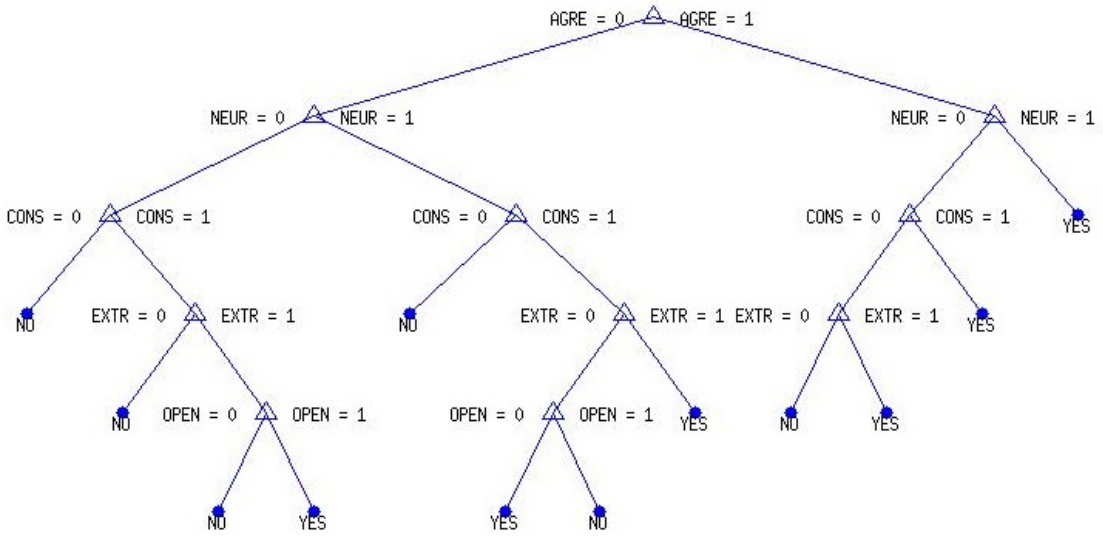



Figure 5.3. Illustration of the trained decision tree for job interview invitation.

The decision tree trained on the predicted Big Five personality dimensions gives a classification accuracy of 94.2% for binarized interview variable. The illustration of the decision tree (DT) is given in Figure 5.3. On the overall, the model is intuitive in that the higher scores of traits generally increase the chance of interview invitation. As can be seen from the figure, the DT ranks the relevance of the predicted Big Five traits from highest (Agreeableness) to lowest (Openness to Experience) with respect to information gain between corresponding trait and the interview variable. The second most important trait for job interview invitation is Neuroticism, which is followed by Conscientiousness and Extraversion. The high/low scores of these top four traits are correlated with target variable and are observed to be consistent throughout the DT. If the Openness score is high, then having a high score in any of the Neuroticism, Conscientiousness or Extraversion variables suffices for an invitation. Chances of invitation decrease if Agreeability is low: only three out of eight leaf nodes are "YES" in this branch. In two of these cases, one has to have high scores in three out of four remaining traits.

For verbal explanations, we converted the DT structure into a compact set of "if-then" rules in the form mentioned earlier. The metadata provided by the organizers do not contain gender annotations, which could have been useful in explanatory sentences.

For this purpose, we have manually annotated 4 000 development set (training + validation) videos using the first face-detected frames, then trained a gender prediction model based on the audio and video features used in the apparent personality trait recognition. The ELM based gender predictors gave 97.6% and 98.9% validation set accuracies using audio (openSMILE) and video (CNN-FUN) features, respectively. We fused the scores of audio and video models with equal weight and obtained a validation set accuracy of 99.3%, which is close to perfect. We then used all annotated data for training with the optimized hyper-parameters and cast predictions on the remaining 6 000 (validation + test set) instances.

The verbal explanations are finally accompanied with the aligned image from the first face-detected frame and the bar graphs of corresponding mean normalized scores. When we analyzed the results, we observed that individually processed clips cut from different places of a single input video have very similar scores, and the exactly same reasons for invitation decision, showing the consistency of the proposed approach. Figure 5.4 illustrates automatically generated verbal and visual explanations for this stage.

## 5.5. Experimental Results

We first examine the performance of each functional individually and choose their best fusion in Table 5.2. As can be seen, the mean is the most informative functional, and the best feature-level fusion consists of mean, standard deviation and offset.

After examining the functionals, we measure the performance with traditional visual descriptors as well as audio, scene and DCNN features, and we report the results in Table 5.3. The last four rows show the performance of various fusion schemes. We show that the best performance is obtained via weighted score-level fusion of the convolution-based features for all tasks except agreeableness. This task is better explained by including the audio features, hence we use this fusion scheme for the agreeableness task.

Table 5.2. Performance of functional statistics in ChaLearn FI dataset.

| Feature | Mean | Extr. | Agre. | Cons. | Neur. | Open. |
|---|---|---|---|---|---|---|
| Mean | 0.900 | 0.906 | 0.902 | 0.897 | 0.894 | 0.902 |
| Std | 0.883 | 0.891 | 0.881 | 0.876 | 0.880 | 0.886 |
| Curvature | 0.880 | 0.876 | 0.891 | 0.874 | 0.876 | 0.882 |
| Slope | 0.880 | 0.876 | 0.892 | 0.874 | 0.876 | 0.882 |
| Offset | 0.899 | 0.904 | 0.901 | 0.895 | 0.893 | 0.901 |
| Fusion of all 5 | 0.902 | 0.908 | 0.903 | 0.898 | *0.898* | 0.904 |
| Mean + Std + Offset | *0.902* | *0.909* | *0.903* | *0.899* | 0.897 | *0.904* |

The results in Table 5.3 indicate that even though individual audio-visual feature types perform high and very similar, there is still room for improvement using multi-modal fusion. We finally combine the training and validation sets and retrain a model with the optimized parameters on the validation set. Using the selective fusion approach (System 14 in Table 5.3), we submit a single set of predictions for the challenge test videos. The test set ranking of top competitors are shown in Table 5.4. We see that our system ranks the first and that the mean performance obtained in validation and test sets are very similar, which collectively show high generalization ability of our system.

In Table 5.5, we present the results obtained in the test set of the quantitative stage of the JCS challenge. Our system achieves the top estimation accuracy in all six tasks. The test set scores of the winners for the qualitative stage are shown in Table 5.6. Our team ranks the first in terms of the overall mean score.

In Figure 5.4, we show example outputs of the qualitative stage of the challenge, combined with the personality trait estimations normalized by their respective training set mean values.

Table 5.3. Regression performance of various systems in FI validation set. FF: Feature-level fusion, WF: Weighted score-level fusion, SF: Selective fusion.

| System | Feature | Dim | Mean | Extr. | Agre. | Cons. | Neur. | Open. |
|---|---|---|---|---|---|---|---|---|
| 1 | GEO | 115 | 0.892 | 0.896 | 0.896 | 0.883 | 0.888 | 0.896 |
| 2 | LPQ-TOP | 12288 | 0.901 | 0.904 | 0.901 | 0.898 | 0.899 | 0.903 |
| 3 | LBP-TOP | 5568 | 0.900 | 0.903 | 0.900 | 0.895 | 0.897 | 0.902 |
| 4 | LGBP-TOP | 100224 | 0.912 | 0.915 | 0.913 | 0.910 | 0.910 | 0.911 |
| 5 | VGG Face+FER | 20480 | 0.910 | 0.915 | 0.911 | 0.906 | 0.907 | 0.910 |
| 6 | VGG VD-19 | 4096 | 0.899 | 0.895 | 0.906 | 0.899 | 0.893 | 0.900 |
| 7 | Audio-IS09 [40] | 384 | 0.894 | 0.893 | 0.898 | 0.886 | 0.892 | 0.898 |
| 8 | Audio-IS10 [86] | 1582 | 0.895 | 0.895 | 0.899 | 0.889 | 0.895 | 0.899 |
| 9 | Audio-IS12 [38] | 6125 | 0.895 | 0.895 | 0.900 | 0.889 | 0.895 | 0.899 |
| 10 | Audio-IS13 [39] | 6373 | 0.896 | 0.894 | 0.900 | 0.890 | 0.895 | 0.899 |
| 11 | FF(5,6) | 24576 | 0.911 | 0.914 | 0.913 | 0.913 | 0.906 | 0.910 |
| 12 | WF(4:10) | - | *0.914* | 0.918 | *0.914* | 0.912 | *0.912* | 0.913 |
| 13 | WF(FF(4,5),6) | - | *0.914* | *0.919* | 0.913 | *0.914* | *0.912* | *0.914* |
| 14 | SF(12,13) | - | *0.915* | *0.919* | *0.914* | *0.914* | *0.912* | *0.914* |

Table 5.4. Regression performance in FI test set.

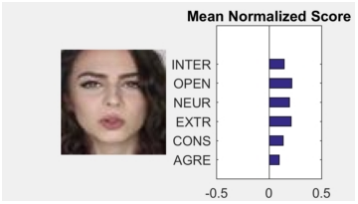
| Rank | Team | Accuracy |
|---|---|---|
| 1 | Ours [13] | *0.913* |
| 2 | Arulkumar *et al.* [87] | 0.912 |
| 3 | Rai | 0.910 |
| 4 | Aydın *et al.* [88] | 0.898 |

Table 5.5. JCS Challenge quantitative stage test set results.

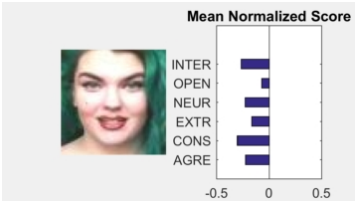
| Participant | INTER | AGRE | CONS | EXTR | NEUR | OPEN | MEAN TRAITS |
|---|---|---|---|---|---|---|---|
| Ours | *0.9209* | *0.9137* | *0.9198* | *0.9213* | *0.9146* | *0.9170* | *0.9173* |
| Baseline | 0.9162 | 0.9112 | 0.9152 | 0.9112 | 0.9104 | 0.9111 | 0.9118 |
| First Runner Up | 0.9157 | 0.9103 | 0.9138 | 0.9155 | 0.9083 | 0.9101 | 0.9116 |
| Second Runner Up | 0.9019 | 0.9032 | 0.8949 | 0.9027 | 0.9011 | 0.9047 | 0.9013 |

Table 5.6. JCS Challenge qualitative stage test set results.

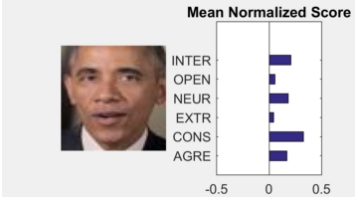
| Participant | Our Team | First Runner Up |
|---|---|---|
| Clarity | 4.31±0.54 | 3.33±1.43 |
| Explainability | 3.58±0.64 | 3.23±0.87 |
| Soundness | 3.40±0.66 | 3.43±0.92 |
| Interpretability | 3.83±0.69 | 2.40±1.02 |
| Creativity | 2.67±0.75 | 3.40±0.8 |
| Mean Score | *3.56* | 3.16 |



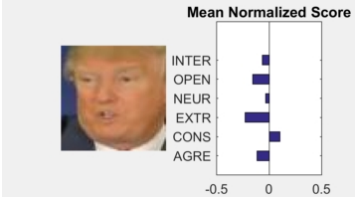

This lady is invited for an interview due to her high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.




This lady is not invited due to her low apparent agreeableness, neuroticism, conscientiousness, extraversion and openness scores. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.




This gentleman is invited for an interview due to his high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.




This gentleman is not invited due to his low apparent agreeableness, neuroticism, extraversion and openness scores. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.

Figure 5.4. Sample verbal and visual explanations from qualitative stage of JCS.

# 6. FACIAL IMAGE BASED AGE ESTIMATION

Automated age estimation from facial images is one of the most difficult challenges in face analysis. It can be very favorable in a number of real life applications such as age-based authorization systems, demographic data mining, business intelligence and video surveillance systems. The difficulty of this task originates from many factors such as the lack of enough labeled samples to model the aging patterns of subjects, as well as uncontrolled conditions in data collection such as illumination, pose, occlusions and other environmental variables. The aging process is also known to be very subject-dependent, *i.e.* subjects might differ in terms of aging patterns, resulting in high variations within the samples from the same age.

## 6.1. Related Work

One of the earliest works involving age estimation from face images is conducted in the early 2000s by Lanitis *et al.* [89,90]. After the emergence of large age databases such as MORPH [91], FRGC [92] and FG-NET [93], the interest on this subject has significantly grown. In the following paragraphs, we provide a brief literature review on studies.

A variety of feature extraction methods were applied for the task of modeling the aging pattern from facial images. For example, Active Shape Models and Active Appearance Models have been employed as features for age estimation [90, 94–96].

Histogram-based local appearance features have been a very popular choice for age estimation. These features include the Local Binary Patterns (LBP) descriptor, which encodes a local patch of the image based on the binary relations of the center pixel with its neighbors, and is widely used in age estimation [97–99]. Similarly, the Histogram of Oriented Gradients (HOG) descriptor has shown to be informative for age modeling [98, 100].

The LBP descriptor is modified by processing the input image with Gabor filters, resulting in the Local Gabor Binary Patterns (LGBP) descriptor, which has shown to be informative in age estimation [99].

Gabor filters are also employed in the calculation of Bio-Inspired Features (BIF) [101], which is consistently used for age estimation in recent years [15, 98]. BIF feature processes an image using a multi-layer feed-forward model where the first layer convolves the image with a set of Gabor filters from multiple orientations and scales, and the resulting vector is downsized with a pooling step, usually with STD or MAX operators. A simplified version of this model is used in [102], where the authors choose the number of bands and orientations manually.

Convolutional Neural Networks (CNN) has been successfully applied to the age estimation problem in a number of applications. Almost all the top-ranking participants of ChaLearn Looking at People 2015 - Apparent Age Estimation challenge used deep learning and achieved very good results [50]. Rothe *et al.* [51] won the LAP-2015 challenge by using a deep network that was trained for image classification, and fine-tuning it for the task of apparent age estimation by collecting a custom dataset of 524,230 images automatically from the Internet. Liu *et al.* [52] and Zhu *et al.* [53] also proposed to fine-tune deep networks for age estimation with augmented data, and achieved very good results.

For modeling in age estimation problems, many different algorithms have been employed. For example, Support Vector Machine Regression (SVR) is a commonly used algorithm for this task [15, 95, 97, 98, 102]. Other learning algorithms that have been utilized in the task of age estimation include Neural Networks [90, 103], Random Forests (RF) [53, 104], projection based learners such as Partial Least Squares regression and Canonical Correlation Analysis, which are often used in combination with kernel and regularization techniques [105–107]. Ranking based methods are also commonly used for age estimation [95, 97, 108, 109]. Extreme Learning Machine (ELM) is used for classification into four non-overlapping age groups and shown to yield good classification performance in [99].

## 6.2. ChaLearn Apparent Age Estimation Challenge

ChaLearn Looking at People 2016 - Apparent Age Estimation challenge dataset [110] includes 7,591 facial images. The images are labeled by multiple human annotators collectively, and the resulting mean and standard deviation for each image is provided with the dataset. The performance evaluation includes both the mean and standard deviation, in order to better compare the algorithm with human performance. The performance score of the challenge is the normal score, which is formulated as follows:

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

This score is then averaged over all instances in the testing set. Therefore, the challenge score can vary between 0 (best case) and 1 (worst case).

We also measure the mean absolute error (MAE) of the regressors for better comparison with the literature. MAE is the average absolute deviation of each estimation from its ground truth value, which is formulated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{6.2}$$

## 6.3. Proposed Approach

For age estimation from facial images, we propose a deep feature extraction method combined with a two-stage kernel ELM regressor. We first classify a test instance into overlapping age groups while allowing multiple group assignments, we then average the estimations of the local regressors of each group the instance is assigned to. For instances with no group assignment, we use the global regression model as a backup. Our proposed method is illustrated in Figure 6.1.

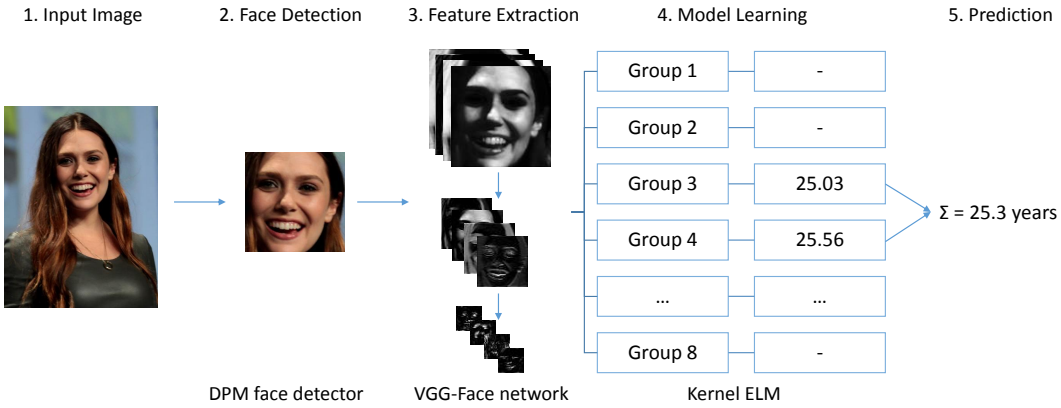



Figure 6.1. Overview of the proposed age estimation system.

For face registration, we use the DPM-based face detector due to its accuracy. The DPM face detector gives the coordinates of the bounding box (if any face is detected), as well as the detection score. Inspired by [51], we run the face detector on rotated version of the original image between -45° and 45° in 5° increments, in order to deal with in-plane rotations. We then take the output with the maximum face score. DPM's face detection statistics on this dataset is provided in Table 6.1.

Table 6.1. Face alignment summary in LAP-2016 dataset.

| # | Train | Val | Test |
|---|---|---|---|
| Given | 4113 | 1500 | 1978 |
| Detected | 4016 | 1462 | 1920 |

We extract deep features from the VGG-Face network that is originally trained for face recognition [49]. We then feed these features to our binary kernel ELM classifiers in order to assign each instance to a number of overlapping age groups. Inside each group, we train a kernel ELM regressor and we average the decisions of local regressors in order to get the final estimation.

In each modeling algorithm, we optimize the relevant parameters via a 5-fold cross validation in the training set and we report the accuracies on the validation set. We show the classification statistics of each binary classifier as well as the regression performance of each model in Table 6.5.

## 6.4. Experimental Results

In this section we present the results of our classification and regression experiments with various feature extraction and normalization options.

In order to choose the most suitable deep feature extraction method, we try various layers of VGG-Face and we provide the performance of the global regression method in Table 6.2.

Table 6.2. Comparison of different layers of VGG-Face in age estimation.

| Layer | Num. features | $\epsilon_{val}$ | $\text{MAE}_{val}$ |
|---|---|---|---|
| 32 | 25088 | 0.4284 | 4.68 |
| 33 | 4096 | *0.4021* | *4.35* |
| 35 | 4096 | 0.4150 | 4.48 |
| 37 | 2622 | 0.4066 | 4.38 |

We test the performance of three different kernel types with features extracted from the $33^{rd}$ layer of VGG-Face. Table 6.3 shows that the best performance is obtained by using the RBF kernel.

Table 6.3. Age estimation performance with different kernel types.

| Linear | | RBF | | Poly | |
|---|---|---|---|---|---|
| $\epsilon$ | MAE | $\epsilon$ | MAE | $\epsilon$ | MAE |
| 0.49 | 5.26 | *0.42* | *4.35* | 0.45 | 4.48 |

We examine the effect of feature normalization options prior to computing the kernel matrix on the global regression model. We show the regression performance with various normalization methods, namely min-max normalization to $[-1, 1]$, z-normalization, power normalization and sigmoid function in Table 6.4. We try features extracted from two different layers of the VGG-Face network, and see that $\ell_2$ normalization results in the smallest normal error.

Table 6.4. Age estimation performance with different normalization options.

| | Layer 33 | | Layer 37 | |
|---|---|---|---|---|
| Norm. Type | $\epsilon$ | MAE | $\epsilon$ | MAE |
| No norm. | 0.4487 | 4.91 | 0.4403 | 4.79 |
| $\ell_2$ | *0.4021* | 4.35 | 0.4066 | 4.38 |
| Pow. + $\ell_2$ | 0.4028 | *4.32* | 0.4079 | 4.44 |
| Sig. + $\ell_2$ | 0.4152 | 4.49 | 0.4137 | 4.46 |
| MM + $\ell_2$ | 0.4355 | 4.77 | 0.4301 | 4.63 |
| Z + $\ell_2$ | 0.4102 | 4.48 | 0.4036 | 4.33 |
| MM + Sig. + $\ell_2$ | 0.4861 | 5.46 | 0.4652 | 5.12 |
| Z + Sig. + $\ell_2$ | 0.4220 | 4.59 | 0.4164 | 4.51 |
| MM + Pow. + $\ell_2$ | 0.4565 | 5.01 | 0.4438 | 4.88 |
| Z + Pow. + $\ell_2$ | 0.4083 | 4.43 | 0.4078 | 4.37 |

In Table 6.5, we summarize the classification accuracy and recall for the 8 overlapping age groups we used. The $9^{th}$ row is the performance of the backup system, and the final row is the performance of the overall system on the validation set of LAP-2016 dataset.

The proposed system is applied on the test set using a single submission option, and ranked $6^{th}$ in the challenge (see Table 6.6). We observe that our test set error is even smaller than that of the validation set. Note that the top performing systems employed a direct application of DCNNs to the image based recognition problem. Although very effective, this way was not efficient and feasible considering our limited computational resources.

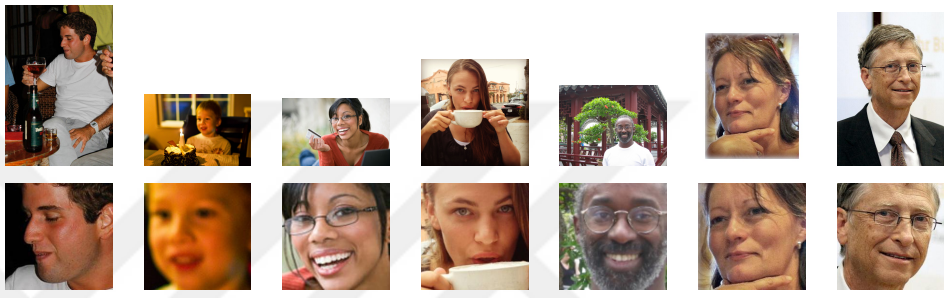
We show some of the good age estimation examples from the LAP development set in Figure 6.2, and some examples where our apparent age estimation system fails due to wrong modeling or alignment in Figure 6.3.

Table 6.5. Age estimation performance in LAP-2016 validation set.
Classification accuracy, recall and regression performance on different age groups.

| Group | $N_{tr}$ | $N_{val}$ | Acc. | Rec. | $\epsilon$ | MAE |
|---|---|---|---|---|---|---|
| 0-15 | 860 | 152 | 0.96 | 0.78 | 0.45 | 2.46 |
| 10-25 | 2366 | 436 | 0.84 | 0.65 | 0.31 | 2.90 |
| 15-30 | 3686 | 662 | 0.84 | 0.83 | 0.31 | 3.19 |
| 20-35 | 4072 | 705 | 0.81 | 0.86 | 0.33 | 3.52 |
| 30-40 | 1764 | 311 | 0.81 | 0.35 | 0.34 | 3.82 |
| 35-50 | 1568 | 288 | 0.85 | 0.45 | 0.34 | 4.26 |
| 45-60 | 976 | 184 | 0.91 | 0.48 | 0.28 | 3.87 |
| 55-$\infty$ | 554 | 106 | 0.96 | 0.57 | 0.28 | 4.36 |
| 0-$\infty$ | 8032 | 1462 | - | - | 0.40 | 4.35 |
| Overall | 8032 | 1462 | - | - | 0.33 | 3.85 |

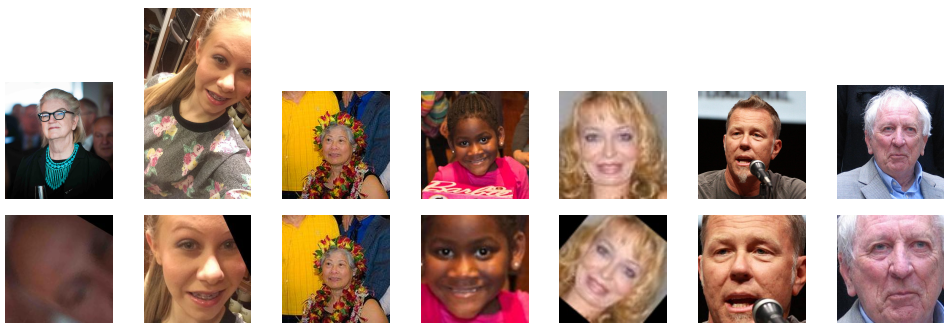
Table 6.6. Age estimation performance in LAP-2016 test set.

| Position | Team | $\epsilon$ |
|---|---|---|
| 1 | OrangeLabs | 0.2411 |
| 2 | palm_seu | 0.3214 |
| 3 | cmp+ETH | 0.3361 |
| 4 | WYU_CVL | 0.3405 |
| 5 | ITU_SiMiT | 0.3668 |
| 6 | Ours | 0.3740 |
| 7 | MIPAL_SNU | 0.4569 |
| 8 | DeepAge | 0.4573 |



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Input image | | | | | | | |
| Aligned face | | | | | | | |
| Apparent age | 23 | 4 | 28 | 25 | 45 | 46 | 60 |
| Predicted age | 23.22 | 4.35 | 27.71 | 25.61 | 43.94 | 46.57 | 57.83 |

Figure 6.2. Examples from the LAP validation set with good estimations.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Input image | | | | | | | |
| Aligned face | | | | | | | |
| Apparent age | 61 | 17 | 74 | 6 | 53 | 43 | 73 |
| Predicted age | 22.91 | 21.85 | 6.13 | 20.19 | 44.83 | 53.86 | 65.31 |

Figure 6.3. Examples from the LAP validation set with bad estimations.

# 7. FACIAL IMAGE BASED AFFECTIVE AND BIOMETRIC APPLICATIONS

We test our face analysis framework on a variety of other face-related tasks such as action unit detection, gender and pose and identity estimation. In the following sections, we provide the methodology and experimental results obtained from these tasks.

## 7.1. Karolinska Directed Emotional Faces Dataset

The Karolinska Directed Emotional Faces (KDEF) Dataset consists of 4 900 images that contain facial expressions in different viewing conditions [111]. There are 70 subjects displaying the seven basic emotional expressions under five different poses. We use this database to test our method's performance under three tasks, namely emotion, pose and identity recognition.

Here, we use the outputs of the two steps of our fine-tuning method described in 4.3, as well as the original VGG Face network for comparison. We show the performance boost effect of fine-tuning clearly, however, we also show that there is not much difference between the first and second folds of fine-tuning.

In order to optimize the regularization parameter $C$ of the linear kernel ELM classifier, we apply 10-fold subject-independent cross validation and report the average accuracy of all folds.

In Table 7.1, we show the pose classification accuracy, *i.e.* the binary classification accuracy for each one of the five poses that exist in the database. We show that the fully profile poses are the hardest to detect, and the highest performance is on the frontal pose since most of the images that were used for the training of the DCNNs were also mostly frontal.

Table 7.1. Pose classification accuracy (%) on KDEF dataset.
FR: Fully right, HR: Half right, F: Frontal, HL: Half left, FL: Fully left.

| Feature | FR | HR | F | HL | FL |
|---|---|---|---|---|---|
| VGG Face | 62.65 | 75.59 | 77.55 | 75.08 | 64.35 |
| VGG FER (1) | 65.20 | *80.39* | 83.06 | *81.61* | 63.43 |
| VGG FER (2) | 64.89 | 80.18 | *83.16* | 81.00 | 64.25 |
| SIFT | *71.22* | 77.12 | 78.98 | 75.59 | *70.48* |
| HOG | 58.47 | 66.39 | 64.08 | 65.17 | 58.63 |
| LBP | 52.04 | 63.33 | 64.59 | 60.06 | 53.93 |
| LGBP | 57.24 | 68.34 | 73.98 | 66.29 | 59.45 |

In Figure 7.1, we report the emotion classification accuracy under two different scenarios. In the first case, we train a pose-independent emotion classifier, that is using all the samples as the training set during cross validation. In the second scenario, we first train a five-class classifier for pose classification. We then train pose-based emotion classifiers using ground truth pose labels of the training instances. For testing, we first classify each instance into one of the five pose classes, then we use the corresponding pose-based model's estimation. We observe that the pose-based classifier increases the performance in all cases, however, the amount of increase is minimal with DCNN based features while being marked with other dense features (e.g. SIFT, HOG, LBP). Considering the finding that the DCNN features provide the best results both with and without pose-based hierarchical classification, we can conclude that DCNN features are inherently robust to pose variations for emotion recognition.

Finally, we test our system on the identity recognition task with matched and mismatched conditions. We conduct three experiments, in the first one we use all the images of one photo session (half of the images) for training. In the other two, we use only neutral and only frontal images as training. As shown on Table 7.2, the original VGG Face outperforms all other features since it was originally trained for identity recognition. Fine tuning on FER degrades the identity recognition performance as it biases the model towards emotional feature extraction disregarding the identity of

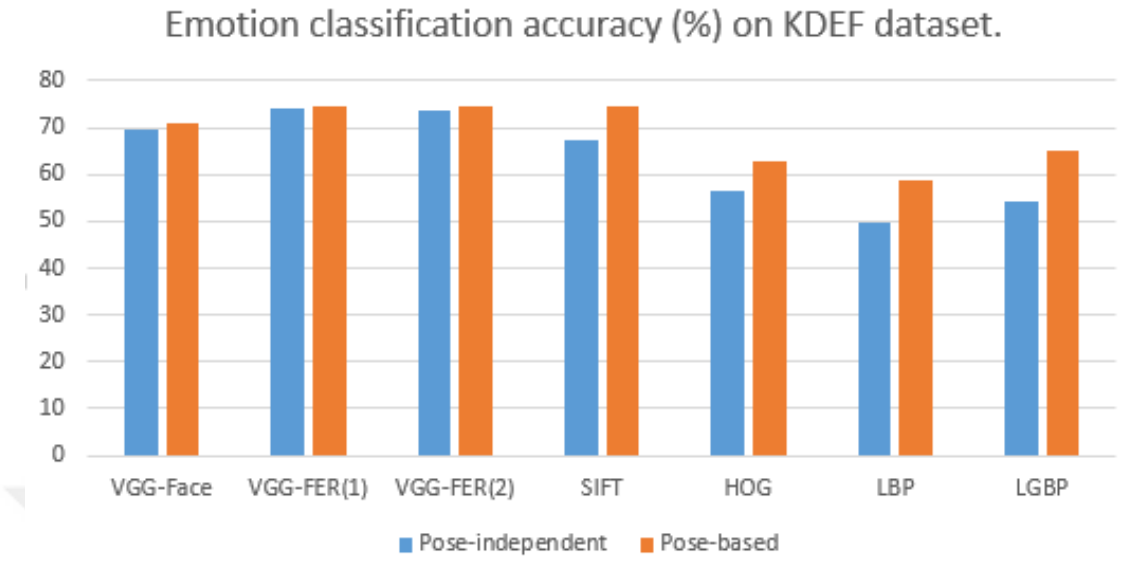



Figure 7.1. Emotion classification accuracy (%) on KDEF dataset.

the face. Moreover, we observe that mismatched emotion conditions do not cause a high performance reduction as long as pose variations are contained in the training set (with VGG-Face feature, the performance drop is around 1.25%). However, when only frontal images are used in training, mismatched pose conditions cause a dramatic performance degradation (from gold accuracy level of 98.80% to 53.87% with VGG-Face feature). The results indicate that pose mismatch has much higher impact on performance compared to emotion mismatch in identity recognition.

## 7.2. EmotioNet Dataset

The EmotioNet v1.0 dataset [112] contains 975,000 images of facial expressions in the wild that are provided with their web links. For 25,000 images in the dataset, manual annotations for 11 action units are given.

In order to test the action unit recognition performance of our system, we train 11 different one-versus-all classifiers with linear kernel ELM and we optimize the regularization parameter with a 5-fold cross validation. We report the average detection accuracy per action unit index in Table 7.3. The results indicate that (1) on the over-

Table 7.2. Identity recognition accuracy (%) on KDEF dataset.

| | Training Condition | | |
|---|---|---|---|
| Feature | All poses, All emotions | All poses, Only neutral | Frontal, All emotions |
| VGG Face | *99.80* | *98.54* | *53.87* |
| VGG FER (1) | 99.78 | 94.03 | 45.96 |
| VGG FER (2) | 99.78 | 93.22 | 44.80 |
| SIFT | 99.69 | 93.33 | 25.57 |
| HOG | 99.49 | 94.99 | 22.60 |
| LBP | 99.63 | 95.44 | 29.34 |
| LGBP | 99.59 | 96.59 | 25.15 |

all DCNN based features outperform conventional densely extracted features and (2) as hypothesized, our fine tuned VGG FER based features provide the highest overall performance.

Table 7.3. Action Unit detection performance on EmotioNet v1.0 dataset.

| Feature | 1 | 2 | 4 | 5 | 6 | 9 | 12 | 17 | 20 | 25 | 26 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GEO | 0.69 | 0.67 | 0.70 | 0.69 | 0.82 | 0.77 | 0.82 | 0.68 | 0.82 | 0.74 | 0.64 | 0.73 |
| LBP | 0.73 | 0.73 | 0.71 | 0.67 | 0.86 | 0.79 | 0.86 | 0.69 | 0.77 | 0.74 | 0.66 | 0.75 |
| HOG | 0.74 | 0.74 | 0.68 | 0.68 | 0.85 | 0.77 | 0.84 | 0.64 | 0.77 | 0.75 | 0.66 | 0.74 |
| AlexNet | 0.76 | 0.74 | 0.77 | 0.75 | 0.88 | 0.84 | 0.88 | 0.74 | 0.83 | 0.82 | 0.72 | 0.79 |
| VGG Face | 0.79 | 0.79 | 0.81 | 0.77 | 0.90 | 0.87 | 0.90 | 0.76 | 0.89 | 0.82 | 0.69 | 0.82 |
| VGG FER | 0.81 | 0.77 | 0.84 | 0.79 | 0.91 | 0.90 | 0.91 | 0.78 | 0.93 | 0.84 | 0.75 | *0.84* |
| # Instances | 642 | 324 | 1416 | 380 | 2609 | 222 | 4301 | 210 | 64 | 6019 | 1001 | - |

# 8. DISCUSSION & CONCLUSIONS

In this thesis, we constructed a framework that can be applied to a number of face analysis tasks including emotion, age and personality trait estimation. We applied various feature encoding methods for video analysis and employed early and late fusion schemes for multimodal learning.

Our experiments involved analysis of various model parameters and methods, as well as comparative investigations of DCNNs with well-known traditional feature extraction methods. Each building block of our pipeline is examined in detail with explanations of the results.

Depending on the target task and available modalities/features and annotations, we proposed a special variant of the base signal processing and machine learning pipeline. Each tailored pipeline provided state-of-the-art results, which are verified on a variety of challenge corpora and against competitive opponents.

For the 7-class emotion recognition task in AFEW-5 dataset, we report 54.55% overall classification accuracy on the test set. Using only DCNN features extracted after proposed alignment and fine tuning processes, a test set accuracy over 51% is obtained, which is the highest single-modality, single-feature performance obtained on this challenge corpus to date.

In the age estimation task, we achieve a mean absolute error of 3.85 years, which is at the level of human performance. In the apparent personality trait estimation from first impressions challenge, we achieve 91.5% estimation accuracy, leading to the best results in the challenge. The same set of potent features are used in job candidate screening competition with a different fusion scheme, also giving the top results in both qualitative and quantitative stages of ChaLearn CVPR 2017 Challenge.

The contribution of using ELMs in these challenge settings are twofold. The first is fast model learning, which enabled us to test a large set of alternative hypothesis including various features and fusion schemes with limited time and moderate computational resources. The second is high generalization power, which was practically proved with similar scores obtained in the challenge validation and test sets, as well as top results.

## 8.1. Remarks

We see that, for almost all the tasks, DCNN feature based models outperform their traditional equivalents. We also show that with an appropriate fusion scheme, two (or more) approaches can be combined to yield an even better performance. However, it should also be noted that the fusion of traditional appearance descriptors with DCNN features does not improve the performance markedly compared to deep features alone. This shows that DCNNs are successful in describing the appearance information that is encoded by traditional descriptors as well.

Our methodology allows any type of feature to be incorporated in the system. However, we realize that there is an important difference between conventional appearance descriptors and DCNN features in terms of alignment. Appearance descriptors such as LBP, HOG and SIFT require a strict face alignment, which does not contain background information. However, due to the convolution and spatial pooling operations in DCNNs, the input images need to be cropped from outside the face in order to work better, as a strict alignment causes DCNNs to lose information. Therefore, a single alignment does not work well for both DCNN and conventional descriptors. Moreover, as the results on a both acted (e.g KDEF) and in-the-wild (e.g. EmotiW) corpora indicate, DCNN features are more resilient to pose variations compared to the conventional features. This can also be attributed to their training conditions with loose alignment.

For the age estimation task, although the DCNN features and kernel ELM learners are powerful, we show that learning a global regression model from all data is not very

effective as opposed to learning local regression models within each age group. Almost all the local regressors yield smaller error than the global model, and the two-step age estimation scheme drops the mean absolute error from 4.35 to 3.85 years.

For the video-based tasks, we show that audio features are not as informative as visual features, however, combining the two modalities with an appropriate fusion scheme increases the overall performance in every task.

Transfer learning using deep networks, although more demanding in terms of both time and computational resources, is shown to be more accurate. We also show the importance of pre-trained model selection in transfer learning, as a generic object classifier, even fine-tuned for the emotion recognition task, yields less accuracy compared to a model that is pre-trained on a face-related task.

Using features from an identity recognition network tends to yield better performance in biometric tasks such as age estimation, compared to affective tasks such as emotion and personality trait estimation.

We conclude that fine-tuning a DCNN that is originally trained with face images yields better performance compared to traditional descriptors, but it is also very computationally demanding compared to hand-crafted descriptors. There are many factors that affect the performance of a DCNN that need to be selected carefully both during training, fine-tuning and feature extraction parts of a visual analysis pipeline.

In this thesis, due to lack of sufficient time and computational resources, we only used the original VGG Face network and its fine-tuned version for emotion recognition, for all tasks. A possible improvement would be training or fine-tuning DCNNs in all related tasks. This would also enable the fusion of features or estimations from multiple domain-specific networks.

Apart from vague cues of emotion and high level of background noise / occlusions, a problem with the AFEW dataset is the fact that labeling is often ambiguous /

inconsistent. We attribute the low classification accuracies on this dataset to this fact. In order to overcome this, a future work could be allowing an annotation where each clip can be assigned to multiple emotion classes. This could result in a more accurate emotion assessment. The trend in facial emotional expression analysis is also in this direction, and databases with overlapping and compound expression labels are being collected [113].

In the age estimation task, we show that a combination of local experts outperforms a global model. This approach could be applied to other domains such as dividing the data according to age, gender and other demographics. This would require more effort for annotation, but it could improve the overall accuracy of our learning algorithm.

We conclude with the remark that ELM classifier can be improved by ensemble methods such as multimodal feature and score level fusion, as well as the addition of DCNN models. Fine-tuning the deep face networks on specific tasks in order to employ them directly in the estimation stage can also be explored.

# REFERENCES


1. OpenCV 2.4.13.3 Documentation, "Face Recognition with OpenCV", 2014, `http://docs.opencv.org/2.4/modules/contrib/doc/facerec/facerec_tutorial.html`, accessed at August 2017.

2. Scikit-image development team, "Gabor filter banks for texture classification", 2013, `http://scikit-image.org/docs/dev/auto_examples/features_detection/plot_gabor.html`, accessed at August 2017.

3. Dalal, N. and B. Triggs, "Histograms of oriented gradients for human detection", *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893, IEEE, 2005.

4. Bai, Y., M. S. Wong, W.-Z. Shi, L.-X. Wu and K. Qin, "Advancing of land surface temperature retrieval using extreme learning machine and spatio-temporal adaptive data fusion algorithm", *Remote Sensing*, Vol. 7, No. 4, pp. 4424–4441, 2015.

5. Dibeklioğlu, H., A. A. Salah and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles", *European Conference on Computer Vision*, pp. 525–538, Springer, 2012.

6. Alpaydin, E., *Introduction to Machine Learning*, The MIT Press, 2nd edn., 2010.

7. Eyben, F., M. Wöllmer and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor", *Proceedings of the International Conference on Multimedia*, pp. 1459–1462, ACM, 2010.

8. Kaya, H., F. Gürpınar and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion", *Image and Vision Computing*, 2017.

9. Dibeklioğlu, H., A. A. Salah and F. Gürpınar, "Measurement of Facial Dynamics for Soft Biometrics", Q. Ji, T. B. Moeslund, G. Hua and K. Nasrollahi (Editors), *Face and Facial Expression Recognition from Real World Videos*, pp. 69–84, Springer, Cham, 2015.

10. Kaya, H., F. Gürpınar, S. Afshar and A. A. Salah, "Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 459–466, ACM, 2015.

11. Gürpınar, F., H. Kaya, H. Dibeklioğlu and A. A. Salah, "Kernel ELM and CNN Based Facial Age Estimation", *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 80–86, Las Vegas, Nevada, USA, June 2016.

12. Gürpınar, F., H. Kaya and A. A. Salah, "Combining deep facial and ambient features for first impression estimation", *Computer Vision–ECCV 2016 Workshops*, pp. 372–385, Springer, 2016.

13. Gürpınar, F., H. Kaya and A. A. Salah, "Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation", *23rd International Conference on Pattern Recognition*, pp. 43–48, 2016.

14. Kaya, H., F. Gürpınar and A. A. Salah, "Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs", *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9, Honolulu, Hawaii, USA, 2017.

15. Han, H., C. Otto and A. K. Jain, "Age estimation from face images: Human vs. machine performance", *International Conference on Biometrics*, pp. 1–8, IEEE, 2013.

16. Viola, P. and M. J. Jones, "Robust real-time face detection", *International journal of computer vision*, Vol. 57, No. 2, pp. 137–154, 2004.

17. Mathias, M., R. Benenson, M. Pedersoli and L. Van Gool, "Face detection without bells and whistles", *Computer Vision–ECCV 2014*, pp. 720–735, 2014.

18. Xiong, X. and F. De la Torre, "Supervised Descent Method and Its Application to Face Alignment", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 532–539, 2013.

19. Xiong, X. and F. De la Torre, "Global supervised descent method", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2664–2673, 2015.

20. Ren, S., X. Cao, Y. Wei and J. Sun, "Face alignment at 3000 fps via regressing local binary features", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

21. Zhu, X. and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2879–2886, IEEE, 2012.

22. Sagonas, C., G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge", *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

23. Sagonas, C., G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "A semi-automatic methodology for facial landmark annotation", *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 896–903, 2013.

24. Sagonas, C., E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, C. Sagonas, I. Marras, I. Kasampalidis, I. Pitas, K. Lyroudia *et al.*, "300 faces in-the-wild challenge: Database and results", *Image and Vision Computing*, Vol. 4, 2015.

25. Belhumeur, P. N., D. W. Jacobs, D. J. Kriegman and N. Kumar, "Localizing

parts of faces using a consensus of exemplars", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp. 2930–2940, 2013.

26. Le, V., J. Brandt, Z. Lin, L. Bourdev and T. S. Huang, "Interactive facial feature localization", *European Conference on Computer Vision*, pp. 679–692, Springer, 2012.

27. Ojala, T., M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, pp. 971–987, 2002.

28. Almaev, T. R. and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition", *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356–361, IEEE, 2013.

29. Haghighat, M., S. Zonouz and M. Abdel-Mottaleb, "Identification using encrypted biometrics", *Computer Analysis of Images and Patterns*, pp. 440–448, Springer, 2013.

30. Felzenszwalb, P. F., R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1627–1645, 2010.

31. Lowe, D. G., "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, 2004.

32. Ojansivu, V. and J. Heikkilä, "Blur insensitive texture classification using local phase quantization", *Image and signal processing*, pp. 236–243, Springer, 2008.

33. Saeed, A., A. Al-Hamadi, R. Niese and M. Elzobi, "Effective geometric features

for human emotion recognition", *IEEE 11th International Conference on Signal Processing (ICSP)*, Vol. 1, pp. 623–627, 2012.

34. Joachims, T., "Text categorization with support vector machines: Learning with many relevant features", *Machine learning: ECML-98*, pp. 137–142, 1998.

35. Csurka, G., C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints", *Workshop on statistical learning in computer vision, ECCV*, 1-22, pp. 1–2, Prague, 2004.

36. Perronnin, F. and C. Dance, "Fisher kernels on visual vocabularies for image categorization", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

37. Jégou, H., M. Douze, C. Schmid and P. Pérez, "Aggregating local descriptors into a compact image representation", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, IEEE, 2010.

38. Schuller, B., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The INTERSPEECH 2012 Speaker Trait Challenge", *INTERSPEECH*, pp. 254–257, 2012.

39. Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism", *Proc. INTERSPEECH*, pp. 148–152, ISCA, Lyon, France, August 2013.

40. Schuller, B., S. Steidl and A. Batliner, "The Interspeech 2009 Emotion Challenge", *Proc. INTERSPEECH*, pp. 312–315, ISCA, Brighton, UK, September 2009.

41. Schuller, B., S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge:

Cognitive & Physical Load", *Proceedings of INTERSPEECH*, ISCA, Singapore, Singapore, September 2014.

42. Cortes, C. and V. Vapnik, "Support-vector networks", *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

43. Huang, G.-B., H. Zhou, X. Ding and R. Zhang, "Extreme learning machine for regression and multiclass classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 42, No. 2, pp. 513–529, 2012.

44. Huang, G.-B., Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: theory and applications", *Neurocomputing*, Vol. 70, No. 1, pp. 489–501, 2006.

45. Rao, C. R. and S. K. Mitra, *Generalized inverse of matrices and its applications*, Vol. 7, Wiley New York, 1971.

46. Suykens, J. A. and J. Vandewalle, "Least squares support vector machine classifiers", *Neural processing letters*, Vol. 9, No. 3, pp. 293–300, 1999.

47. Rifkin, R., G. Yeo and T. Poggio, "Regularized least-squares classification", *NATO Science Series Sub Series III Computer and Systems Sciences*, Vol. 190, pp. 131–154, 2003.

48. LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.

49. Parkhi, O. M., A. Vedaldi and A. Zisserman, "Deep Face Recognition", *British Machine Vision Conference*, 2015.

50. Escalera, S., J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. Escalante, D. Misevic, U. Steiner and I. Guyon, "ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results", *Proceedings of the IEEE*

*International Conference on Computer Vision Workshops*, pp. 1–9, 2015.

51. Rothe, R., R. Timofte and L. Gool, "DEX: Deep EXpectation of apparent age from a single image", *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 10–15, 2015.

52. Liu, X., S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan and X. Chen, "AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation", *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 16–24, 2015.

53. Zhu, Y., Y. Li, G. Mu and G. Guo, "A Study on Apparent Age Estimation", *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 25–31, 2015.

54. Kim, B.-K., H. Lee, J. Roh and S.-Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427–434, ACM, 2015.

55. Li, Z. and D. Hoiem, "Learning without forgetting", *European Conference on Computer Vision*, pp. 614–629, Springer, 2016.

56. Dhall, A., R. Goecke, J. Joshi and T. Gedeon, "The Fourth Emotion Recognition in the Wild Challenge 2016: Baseline, Data and Protocols", *Proceedings of of the 18th ACM Intl. Conf. on Multimodal Interaction (ICMI 2016)*, ACM, 2015.

57. Dhall, A., R. Goecke, J. Joshi, K. Sikka and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol", *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 461–466, ACM, 2014.

58. Dhall, A., R. Goecke, J. Joshi, M. Wagner and T. Gedeon, "Emotion recognition in the wild challenge 2013", *Proceedings of the 15th ACM on International*

*conference on multimodal interaction*, pp. 509–516, ACM, 2013.

59. Dhall, A., R. Goecke, S. Lucey and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies", *IEEE MultiMedia*, Vol. 19, No. 3, pp. 34–41, July 2012.

60. Dhall, A., O. Ramana Murthy, R. Goecke, J. Joshi and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 423–426, ACM, 2015.

61. Dhall, A., R. Goecke and T. Gedeon, "Automatic Group Happiness Intensity Analysis", *IEEE Transactions on Affective Computing*, Vol. 6, No. 1, pp. 13–26, 2015.

62. Goodfellow, I. J., D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests", *International Conference on Neural Information Processing*, pp. 117–124, Springer, 2013.

63. Chatfield, K., K. Simonyan, A. Vedaldi and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets", *British Machine Vision Conference*, 2014.

64. Yao, A., J. Shao, N. Ma and Y. Chen, "Capturing AU-aware facial features and their latent relations for emotion recognition in the wild", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 451–458, ACM, 2015.

65. Ebrahimi Kahou, S., V. Michalski, K. Konda, R. Memisevic and C. Pal, "Recurrent neural networks for emotion recognition in video", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 467–474, ACM, 2015.

66. Cuddy, A. J., S. T. Fiske and P. Glick, "Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map", *Advances in experimental social psychology*, Vol. 40, pp. 61–149, 2008.

67. Todorov, A., A. N. Mandisodza, A. Goren and C. C. Hall, "Inferences of competence from faces predict election outcomes", *Science*, Vol. 308, No. 5728, pp. 1623–1626, 2005.

68. Valente, F., S. Kim and P. Motlicek, "Annotation and Recognition of Personality Traits in Spoken Conversations from the AMI Meetings Corpus.", *INTERSPEECH*, pp. 1183–1186, 2012.

69. Madzlan, N., J. Han, F. Bonin and N. Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs", *Speech Prosody, Dublin, Ireland*, pp. 91–94, 2014.

70. Alam, F., E. A. Stepanov and G. Riccardi, "Personality traits recognition on social network-facebook", *WCPR (ICWSM-13), Cambridge, MA, USA*, 2013.

71. Nowson, S. and A. J. Gill, "Look! Who's Talking?: Projection of Extraversion Across Different Social Contexts", *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 23–26, ACM, 2014.

72. Gievska, S. and K. Koroveshovski, "The impact of affective verbal content on predicting personality impressions in YouTube videos", *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 19–22, ACM, 2014.

73. Fernando, T. *et al.*, "Persons' Personality Traits Recognition using Machine Learning Algorithms and Image Processing Techniques", *Advances in Computer Science: an International Journal*, Vol. 5, No. 1, pp. 40–44, 2016.

74. Qin, R., W. Gao, H. Xu and Z. Hu, "Modern Physiognomy: An Investigation

on Predicting Personality Traits and Intelligence from the Human Face", *CoRR*, Vol. abs/1604.07499, 2016, `http://arxiv.org/abs/1604.07499`.

75. Sarkar, C., S. Bhatia, A. Agarwal and J. Li, "Feature analysis for computational personality recognition using youtube personality data set", *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 11–14, ACM, 2014.

76. Alam, F. and G. Riccardi, "Predicting personality traits using multimodal information", *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 15–18, ACM, 2014.

77. Farnadi, G., S. Sushmita, G. Sitaraman, N. Ton, M. De Cock and S. Davalos, "A multivariate regression approach to personality impression recognition of vloggers", *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 1–6, ACM, 2014.

78. Sidorov, M., S. Ultes and A. Schmitt, "Automatic Recognition of Personality Traits: A Multimodal Approach", *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge*, pp. 11–15, ACM, 2014.

79. Schuller, B., M. Valstar, F. Eyben, G. McKeown, R. Cowie and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge", *International Conference on Affective Computing and Intelligent Interaction*, pp. 415–424, Springer, 2011.

80. Schuller, B., M. Valster, F. Eyben, R. Cowie and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge", *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 449–456, ACM, 2012.

81. Valstar, M., B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie and M. Pantic, "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge", *Proceedings of the 3rd ACM international*

*workshop on Audio/visual emotion challenge*, pp. 3–10, ACM, 2013.

82. Celiktutan, O., F. Eyben, E. Sariyanidi, H. Gunes and B. Schuller, "MAPTRAITS 2014: The first audio/visual mapping personality traits challenge", *Proceedings of the 2014 ICMI Workshop on Mapping Personality Traits Challenge*, pp. 3–9, ACM, 2014.

83. Lopez, V. P., B. Chen, A. Places, M. Oliu, C. Corneanu, X. Baro, H. J. Escalante, I. Guyon and S. Escalera, "ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results", *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings*, 2016.

84. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *CoRR*, Vol. abs/1409.1556, 2014.

85. Breiman, L., "Random forests", *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

86. Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller and S. S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge", *Proc. INTERSPEECH*, pp. 2794–2797, 2010.

87. Subramaniam, A., V. Patel, A. Mishra, P. Balasubramanian and A. Mittal, "Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features", *CoRR*, Vol. abs/1610.10048, 2016.

88. Aydin, B., A. A. Kindiroglu, O. Aran and L. Akarun, "Automatic personality prediction from audiovisual data using random forest regression", *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 37–42, IEEE, 2016.

89. Lanitis, A., C. J. Taylor and T. F. Cootes, "Toward automatic simulation of aging effects on face images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 4, pp. 442–455, 2002.

90. Lanitis, A., C. Draganova and C. Christodoulou, "Comparing different classifiers for automatic age estimation", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 34, No. 1, pp. 621–628, 2004.

91. Ricanek Jr, K. and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression", *7th International Conference on Automatic Face and Gesture Recognition*, pp. 341–345, IEEE, 2006.

92. Phillips, P. J., P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, "Overview of the face recognition grand challenge", *IEEE computer society conference on Computer vision and pattern recognition*, Vol. 1, pp. 947–954, IEEE, 2005.

93. Lanitis, A., "The FG-NET Aging Database", `http://sting.cycollege.ac.cy/~alanitis/fgnetaging/index.htm`, 2002, accessed at March 2016.

94. Geng, X., Z.-H. Zhou and K. Smith-Miles, "Automatic age estimation based on facial aging patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 12, pp. 2234–2240, 2007.

95. Chang, K.-Y., C.-S. Chen and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 585–592, IEEE, 2011.

96. Luu, K., K. Ricanek Jr, T. D. Bui and C. Y. Suen, "Age estimation using active appearance models and support vector machine regression", *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pp. 1–5, IEEE, 2009.

97. Weng, R., J. Lu, G. Yang and Y.-P. Tan, "Multi-feature ordinal ranking for facial age estimation", *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, IEEE, 2013.

98. Liu, K.-H., S. Yan and C.-C. J. Kuo, "Age estimation via grouping and decision fusion", *Information Forensics and Security, IEEE Transactions on*, Vol. 10, No. 11, pp. 2408–2423, 2015.

99. Sai, P.-K., J.-G. Wang and E.-K. Teoh, "Facial age range estimation with extreme learning machines", *Neurocomputing*, Vol. 149, pp. 364–372, 2015.

100. Fernández, C., I. Huerta and A. Prati, "A comparative evaluation of regression learning algorithms for facial age estimation", *Face and Facial Expression Recognition from Real World Videos*, pp. 133–144, Springer, 2015.

101. Riesenhuber, M. and T. Poggio, "Hierarchical models of object recognition in cortex", *Nature neuroscience*, Vol. 2, No. 11, pp. 1019–1025, 1999.

102. Guo, G., G. Mu, Y. Fu and T. S. Huang, "Human age estimation using bio-inspired features", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 112–119, IEEE, 2009.

103. Geng, X., C. Yin and Z.-H. Zhou, "Facial age estimation by learning from label distributions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 10, pp. 2401–2412, 2013.

104. Montillo, A. and H. Ling, "Age regression from faces using random forests", *16th IEEE International Conference on Image Processing*, pp. 2465–2468, IEEE, 2009.

105. Guo, G. and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 657–664, IEEE, 2011.

106. Guo, G. and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS", *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, IEEE, 2013.

107. Guo, G. and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database", *Image and Vision Computing*, Vol. 32, No. 10, pp. 761–770, 2014.

108. Yang, P., L. Zhong and D. Metaxas, "Ranking model for facial age estimation", *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3404–3407, IEEE, 2010.

109. Li, S., S. Shan and X. Chen, "Relative forest for attribute prediction", *Computer Vision–ACCV 2012*, pp. 316–327, Springer, 2012.

110. Escalera, S., M. Torres, B. Marténez, X. Baró, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri and M. Valstar, "ChaLearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge 2016", *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

111. Lundquist, D., A. Flykt and A. Öhman, "The Karolinska directed emotional faces", *Department of Neurosciences, Karolinska Hospital, Stockholm, Sweden*, 1998.

112. Fabian Benitez-Quiroz, C., R. Srinivasan and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5562–5570, 2016.

113. Du, S., Y. Tao and A. M. Martinez, "Compound facial expressions of emotion", *Proceedings of the National Academy of Sciences*, Vol. 111, No. 15, pp. E1454–E1462, 2014.