

AMERICAN UNIVERSITY OF BEIRUT

FACIAL EXPRESSION RECOGNITION FROM
IMAGES FOR VARIOUS HEAD POSES

by

CHADI HANNA TRAD

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
July 2015

AMERICAN UNIVERSITY OF BEIRUT

FACIAL EXPRESSION RECOGNITION FROM IMAGES FOR VARIOUS HEAD POSES

by
CHADI HANNA TRAD

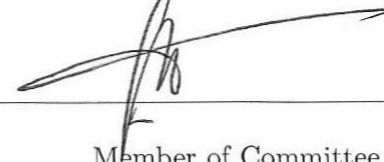
Approved by:

Dr. Hazem Hajj, Associate Professor
Electrical and Computer Engineering



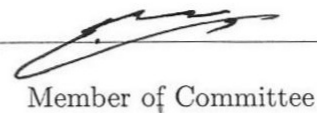
Advisor

Dr. Fadi Karamah, Associate Professor
Electrical and Computer Engineering



Member of Committee

Dr. Wassim El-Hajj, Associate Professor
Computer Science



Member of Committee

Dr. Daniel Asmar, Associate Professor
Mechanical Engineering



Member of Committee

Date of thesis defense: July 2, 2015

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: _____
Last First Middle

Master's Thesis Master's Project Doctoral Dissertation

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, **three years after the date of submitting my thesis, dissertation, or project**, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

Signature

Date

Acknowledgments

My deepest recognition goes to Professor Hazem Hajj who encouraged me through all the phases of this thesis while sharing his knowledge and providing me with helpful aid in the analysis of data. This work would not have been possible without his patience. I would like to thank the committee members for their valuable feedback during the course of this thesis.

I also would like to thank the American University of Beirut and the Department of Electrical and Computer Engineering for offering a great research environment.

Finally, I am very grateful for the support of my parents who have been there with me from the start. I also thank a number of friends who made this work possible. Particularly, many thanks to the following people for having a positive impact on my life in the last three years: Lynn, Nay, Manuel and Joelle.

An Abstract of the Thesis of

CHADI HANNA TRAD for Master of Engineering
Major: Electrical and Computer Engineering

Title: Facial expression recognition from images for various head poses

In the process of facial expression detection from image or video modalities, the variation of head poses with respect to the camera causes a challenging problem for any robust recognition. Several studies have been conducted on the effect of the pose on the recognition rate. The prevalent methodology to solve this problem consists of transforming the facial features back to a frontal pose before inferring the facial expression. Some work has further considered splitting the face into multiple parts then performing a simple maximum combination of the classifications. In this work, we propose a new approach for splitting and fusing the facial features in cases with head yaw rotations. The approach consists of splitting the face into left and right features. Then, two methods are proposed to classify the facial expression. In the first method, we detect facial Action Units (AUs) in the left and right parts then combine the results using a logical OR operation. In the second method, we propose an optimized fusion of the facial expressions. The outcome of the optimized method is a set of weights to combine the classifications from each side of the face at different yaw angles. The weights are determined dynamically based on the yaw angle of the head through a polynomial regression. Experiments were conducted on the two methods using a custom-made database and a set of benchmark 3D facial images. The results showed a 7.1% improvement for our proposed split-and-fuse method over full facial features approach. Furthermore, the optimized fusion method showed superiority in comparison to max-based fusion.

Contents

Acknowledgments	v
Abstract	vi
1 Introduction	1
2 Background	3
2.1 Taxonomy of Facial Expressions	3
2.2 Facial Expression Recognition Approach	4
2.3 Pose Normalization Approaches	5
3 Related Work	6
3.1 AU detection	6
3.2 Emotion Recognition	7
4 Method 1: Angle Independent Action Unit Detection Method	8
4.1 Preliminary Concepts	8
4.2 Proposed Method	9
4.2.1 Feature Extraction from Left and Right Face Regions	9
4.2.2 Action Unit Detection Model	11
4.3 Experiments and Evaluation	11
4.3.1 Benchmarking for Action Unit Detection	12
4.3.2 Benchmarking for Emotion Detection	14
4.3.3 Various Pose Evaluation for Brow Raisers	14
4.4 Summary of Observations for Method 1	16
5 Method 2: Angle Dependent Emotion Detection Method	19
5.1 Proposed Method	19
5.1.1 Overview	19
5.1.2 Ground Truth Data	21
5.1.3 Feature Extraction and Classification	23
5.1.4 Fusion Model	24

5.2	Experiments and Results	27
5.2.1	Experiment Setup	28
5.2.2	Method Evaluation	28
5.2.3	Fusion Approaches	31
5.2.4	Baseline Classification	32
5.3	Summary of Observations for Method 2	32
6	Conclusion	34
	References	35

List of Figures

2.1	AUs activated for the emotion happy	4
4.1	AU detection algorithm	10
4.2	Left/Right face model	12
4.3	Samples in an images sequence from the recorded database. The subject is first asked to rotate their face, then to perform the AU. The third image is annotated as the neutral face, and the fifth is annotated as the apex frame.	17
5.1	Proposed classification system	20
5.2	Proposed fusion approach for one emotion	20
5.3	Steps to generate 2D points for each specific pose.	21
5.4	Approach to train the classifiers	22
5.5	Approach to learn the fusion parameters	24
5.6	Estimating a smoothed posterior probability distribution	24
5.7	Fusion weights and approximations for the happy emotion	29
5.8	Examples of generated facial points with 15 degrees increments	29
5.9	Learning fusion parameters	29
5.10	Applying fusion parameters	30
5.11	Accuracies of the left, right and fusion classifiers, calculated per angle	32

List of Tables

4.1	Size of the facial landmark sets and their corresponding features . .	13
4.2	Comparison between the work in this paper (DRC) and the one of Valstar et al (TMP).	15
4.3	Confusion matrix for emotion classification using the method by Valstar et al. (TMP)	16
4.4	Confusion matrix for emotion classification using the features from the left face	16
4.5	Evaluating algorithm on the recorded database for 3 head orientation intensities and two AUs	17
5.1	Confusion matrix comparison between the proposed algorithm and the baseline approach, for 4 head orientations of 0, 15, 30 and 45 degrees	31

Chapter 1

Introduction

Facial Expression Recognition (FER) has attracted research works for the past two decades. Broadly, FER consists of automatically detecting emotions or muscle changes in the face, typically from a camera. While significant progress has been made towards automated FER, the problem of head pose remains a key challenge. In this thesis, we present a background on this particular area of research and our contributions. The importance of FER lies in the number of applications enabled by this area of research, in fields such as Human-Computer-Interaction (HCI), crime prevention and even marketing. We begin this chapter by discussing the relevance of FER to these three fields.

A modern user-interface should interpret the facial movements as well as the emotional state of the user [1]. One of the well-known systems that employ FER-based user-interfaces is Intel's Assistive Context Aware Toolkit (ACAT). By only detecting the movement of a single cheek muscle, ACAT enables Prof. Stephen Hawking to interface with his computer and the world via voice synthesis and text prediction [2]. Other types of emotion-based HCI systems have been proposed for entertainment purposes. For instance, Durnaika et al [3] proposed to add an emotion recognition functionality to Sony's social AIBO robot. These additions allowed the AIBO to detect human emotions and respond to them. Such applications represent a sample of FER field's contributions to the HCI field. Additionally, new applications in crime prevention can now be enabled by automated FER. Due to the long hours of operator monitoring a camera feed, operators are considered to be the "weakest link" in surveillance systems according to Bullington [4]. Therefore, by offloading the burden of human information processing off the operators, Bullington suggested that automated FER can make surveillance more effective. Furthermore, new contributions to the marketing field have been made with the use of FER technology. RealEyes, a UK start-up, has recently partnered with more than 20 major companies to assess the effectiveness of their current and future video commercials. By tracking the emotions of a number of subjects during

a commercial, RealEyes claims to evaluate the user engagement while watching a video and predict the impact of the video.

Despite the progress of these applications, the variation in head pose remains one of the major setbacks for the accuracy of automated FER. The majority of the works available solves the problem of head pose by *pose normalization* which consists of transforming the face to a frontal view first or by choosing features that are not highly affected by the pose. It has been reported that the pose has an impact on the recognition rate and that the best rate is achieved at non frontal poses [5]. To improve on the accuracy obtained for non frontal poses, we propose two approaches that consist of splitting the face into two regions, and then combining the decisions from both regions.

For head rotations around the vertical axis, also called yaw rotations, we split the face into left and right regions, classify the expression in each region separately, and then fusing the decisions. The reasoning behind the fusion of decisions from two regions is that these parts provide *competitive* rather than *complementary* information. For instance, a smile can be detected independently using the left or right side of the face. In contrast, when regions provide complementary information, a fusion of decisions may suffer from loss of information. To illustrate this effect, an analogy can be made to the mythological story, "the blind men and the elephant". In an attempt to learn what the elephant is like, each person feels a different part of the elephant. The group of persons end up in complete disagreement, thinking that the elephant is either a hand fan (ear) or a pillar (leg), etc. Similarly, choosing regions that present *complementary* information can impede the performance of a decision fusion.

Therefore, we present in this thesis two methods to recognize facial expressions with head rotation: one is angle independent, and the other is angle dependent. In the first method, we propose a heuristic approach to detect Action Units (AUs), codes for muscle changes in the face, by applying classifiers on two regions of the face and merging the decisions. In the second method, we propose a fusion based approach to recognize the emotion by optimizing the detection with respect to the angle. The particulars of these approaches are further described in this thesis.

This thesis is planned as follows. Chapter 2 covers the backgrounds related to the thesis and the field of FER. Chapter 3 describes the related research work in this field. Chapter 4 provides the details for the first proposed method for AU detection from either sides of the face. Chapter 5 presents the second method based on optimized fusion for emotion recognition from both sides of the face. Chapter ?? summarizes our contributions and concludes the thesis.

Chapter 2

Background

Facial expression recognition consists of determining the affective state of an individual. A number of methods have been proposed in the last two decades to achieve this recognition. The methods vary in multiple aspects such as the use of coding for facial expressions, the facial features and the classification methods. This chapter provides a brief overview of the common methods used.

Another differentiating aspect is the pose normalization approach applied. Briefly, this chapter presents the similarities and differences between these expression detection methods, starting with the different taxonomies of facial expressions, followed by an overview of the methods applied.

2.1 Taxonomy of Facial Expressions

For facial expression detection, there are typically two types of approaches: One type is based on facial expression measurement and the other is based on facial muscle action detection. These types are also known as message-judgment and sign-judgment approaches, respectively [6]. The aim of message-judgment is to detect the affect underlying the facial expression, while the aim of sign-judgment is to purely describe the state of facial components such as their movements or shapes, leaving affect judgment to a higher level process.

Facial Action Coding System (FACS) is considered as a sign-judgment approach, while the emotion detection is a message-judgment approach. FACS, which was introduced by Ekman et al. [7], is the most used coding scheme that describes the muscular activity of a face. This coding scheme describes visually discernible facial movements in terms of Action Units (AUs). Ekman and Friesen first identified 44 AUs, which were associated with the contraction of facial muscles. They also provided rules for recognition of the onset (start), apex (peak) and offset (end) of the AUs. Additional AUs were later added in newer revisions. On

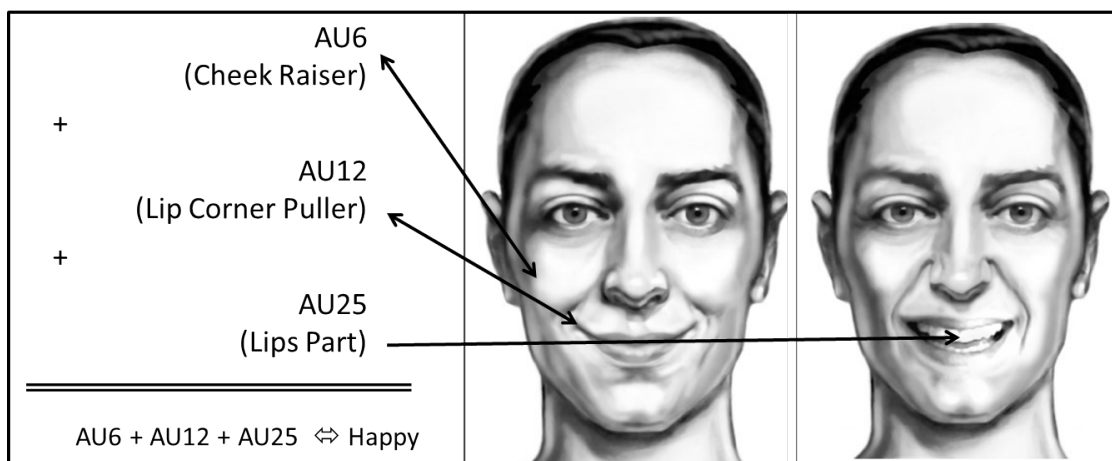


Figure 2.1: AUs activated for the emotion happy

the other hand, a similar coding scheme is also used to describe the underlying affective state. In his research, Ekman first discovers a set of emotions that are expressed similarly across all cultures, which he coined universal emotions (anger, disgust, fear, happiness, sadness and surprise). Each of Ekman’s emotions can be represented by a combination of AUs, as illustrated in Figure 2.1. Virtually every existing work and application on facial expressions studies the six basic emotions. Aside from being widely available in most databases, the main reason of their popularity is their stability over culture and age. Another notable emotion labelling scheme is Russell’s valence-arousal model [8]. While the use of such scheme is less common when describing facial expressions, the advantage provided by this model is that emotions are classified along two continuous dimensions instead of being labelled discretely. In this thesis, our first proposed method in Chapter 4 detects AUs while the second one in Chapter 5 detects universal emotions.

2.2 Facial Expression Recognition Approach

Generally, most facial expression recognition approaches apply similar procedures for detection. Starting with a set of images or a video, the first step consists of detecting the position of the face in the image. Facial detection has been extensively solved, and one of the prevalent algorithms used to solve this problem is the Viola-Jones algorithm [9]. The detected face has a certain position and orientation with respect to the camera, called head pose. The variation in head pose between one video and the other is problematic since it can negatively affect the accuracy of the system. Once the face is detected, head pose normalization

is typically applied. The types of normalizations are further explained in section 2.3. The normalization is then followed by feature extraction. The main idea behind feature extraction is to compute features that are relevant to the facial expressions. For instance, to detect if a person is smiling, the distance between the lip corners can be used as a feature. If this distance is high with respect to a reference, the person is likely to be smiling. Alternatively, the forming of folds and riddles around the mouth, specifically in the nasolabial region, can also indicate a smile. These two types of features are later referred to as geometric and appearance-based features. After computing these features, a machine learning algorithm is applied for the detection of the facial expression.

2.3 Pose Normalization Approaches

There are three types of common transformations for head pose normalization: Euclidean transformations, affine transformations, and model based transformations.

Euclidean transformation is simplest transformation, and consists of a translation, a rotation and a scaling. The method considers the inner eye corners, which are invariable with respect to the expressed affect. A rotation of the face is performed, aiming to align the eye corners with the horizontal plane. This rotation is then followed by the scaling of the face to match a reference interocular distance (the distance between the eyes). The rotation, which is a rotation within the camera plane, can be used to normalize the face when it has rotated within the same plane. Very small out-of-plane rotations might be tolerable when applying this transformation.

Affine transformations are more complex and consist of a linear combination of translation, rotation, scaling and shearing (non-uniform scaling). This kind of transformation requires three points of reference, and the tip of the nose is typically considered in addition to the inner eye corners. Affine transformations are generally used for small angles ranging between -30 and 30 degrees, as the reported accuracies drastically decrease for images outside of those bounds.

Finally, model-based transformations consist of mapping the pixel coordinates detected to a 3D model of the face and then projecting them back to a frontal view. Common frameworks that are used for this type of transformation are Active Shape Models (ASM) [10] or Mixture of Parts (MoPs) [11].

Chapter 3

Related Work

In this chapter, we discuss the work related for AU detection in section ?? and emotion detection in section 3.2.

3.1 AU detection

Virtually most of the AU methods reported have been based on near frontal views data [12] or on data with moderate head pose variation [13]. For instance, the work of Valstar et al. [12] investigated facial AU recognition from near-frontal views using geometrical features, and modeled the temporal phases of AUs. One of the weaknesses reported is that significant out-of-plane rotations affected the recognition accuracy. Other work such as the work of Tyan et al. [14] used a combination of geometrical and appearance based features. This method was reported to be robust for moderate face rotations, but no direct measure for the accuracy/angle dependency was reported.

To extract facial expressions in less-constrained environments, such as different head poses, Pantic and Patras [15] investigated facial AU recognition from profile views. Most methods used to recognize AUs are based on geometric features. In the work of Valstar et al. [12], the authors extracted the facial points using a tracking scheme based on particle filtering using factorized likelihoods (PFFL). Affine transformation was then performed on the obtained coordinates to reverse the effect of scaling and small head orientations. Geometrical features as well as temporal features were extracted from the image sequences. Finally, a combination of Gentle-Boost and Support Vector Machine (SVM) was used in the classification stage. The system was further extended to detect the temporal activation model (neutral, onset, apex and offset). The main disadvantages of their system in detecting AUs for pose variations can be summarized by the following: (1) the affine transformation cannot model out-of-plane rotations assigned with the head

pose, and (2) PFFL cannot handle facial point occlusions associated with head pose variation.

3.2 Emotion Recognition

Broadly, emotion detection can be achieved by using two types of features: geometric and appearance-based. Prominent works have utilized the following types of appearance based descriptors: Histogram of Gradient (HoG) (e.g [16]), Local Binary Patterns (LBP) (e.g [5]) and Scale Invariant Feature Transform (SIFT) (e.g [17]). On the other hand, a number of works have focused on geometric features. Features such as the displacement of important facial points in Hu [18], to more sophisticated ones including temporal features in Valstar [12] have been proposed. For a more elaborate survey on the types of features used, we refer to the survey in [19]. While most of the works improve on the feature extraction, dimensionality reduction [16] or classification approaches, a very limited number of algorithms have directly studied the effect of the pose (e.g. [18, 16, 20]). Also few works have attempted to split the face into regions and perform a decision fusion to study the effect of pose (e.g. [21, 22]).

The work of Tariq et al. [21] reported that splitting the face into multiple regions and applying a fusion on the decision level can improve the performance of the recognition. The fusion applied in their work was MAX fusion. MAX fusion consists of selecting the class that has the highest decision score from all the classifiers. However, we argue that this type of fusion does not model the effect of the pose on the performance of each region, and thus has room for improvement. In contrast, we propose a fusion approach that optimizes the fusion depending on the yaw angle.

Chapter 4

Method 1: Angle Independent Action Unit Detection Method

In this study, we propose a system that can recognize FACS AUs for various poses, which was published in [22]. Facial expression recognition has been an active research topic for many years, with Facial Action Coding Systems (FACS) being among the widely used methods. FACS is a well-established scheme in psychology to annotate facial muscle contractions and relaxations, also called Action Units (AUs). Previous works on FACS-based methods focused on frontal or near-frontal head poses. In this work, we propose a method to recognize expressions in side head poses. This method builds one classifier for each possible group of occlusions. Facial expression recognition of a side facial pose is then based on a boosting approach of the different classifiers. The method is first tested with frontal and near-frontal head poses, and the results are shown to be comparable to state of the art work for AU and emotion detection. The method is then tested with a small training set for various orientations and AUs, and shown to be accurate.

The rest of this chapter is planned as follows. In Section 4.1, we present some preliminary concepts. Section 4.2 presents the proposed method. Experimental results are shown in Section 4.3. This method is concluded in Section 4.4.

4.1 Preliminary Concepts

In this section, we describe the general parts in a geometrical based AU recognition system and introduce the methods used in our system. First, a facial tracker is employed to detect and track the facial points. One of the most used models for facial tracking is the Deformable Model Fitting (DMF). DMF is a classic problem formulation in which the shape of object deformations is modeled using the Point Distribution Model (PDM) founded by Taylor [10]. In this model, the facial points'

positions are calculated using the following equation:

$$\mathbf{x}_i = s\mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i\mathbf{q}) + t, \quad (4.1)$$

where \mathbf{x}_i denotes the location of the i^{th} landmark, s denotes a scale, \mathbf{R} a rotation matrix, $\bar{\mathbf{x}}_i$ the mean location of the i^{th} landmark, \mathbf{q} a set of non-rigid parameters, and Φ a submatrix of basis variations. The aim is to determine the landmarks positions x_i , by determining the set of shape parameters (shape, rotation, translation and non-rigid parameters). Particularly, the Regularized Deformable Model Fitting (RDMF) tracker [23] follows the DMF model. RDMF uses a logistic regressor function to determine the likelihood of a facial point position, given an input image. The values of the landmark positions are determined by minimizing the misalignment error according to the PDM model as well as maximizing the new position likelihood. The search space of an optimal solution is minimized using hill climbing methods. The advantage of this tracker is that it is robust to multiple occlusions since it leverages the relationship among the facial points in the PDM model. After extracting features from the obtained facial points' positions, a machine learning algorithm such as SVM can be used to classify an AU. However, if the feature dimension is greater than the training data, overfitting to the training data is rather probable. Many feature reduction techniques can be used at this stage. Gentle-Boost combines a weighted vote of weak classifiers in the final classification. In the next section, we propose how to combine RDMF with Gentle-Boost to detect the AUs for various orientations.

4.2 Proposed Method

In this section, we describe our proposed method illustrated in Fig. 4.1. The first step is to detect and track a set of facial point coordinates using the RDMF tracker proposed in [23]. These coordinates are then separated into two groups: left-face points and right-face points. The features extracted are the distances' variation among the points. Finally, we describe the model for detecting the activation status of each AU. In the following subsections, we first describe how the features are extracted for each face region, and then we explain the AU classification scheme. Finally, we describe a classification system for emotion detection.

4.2.1 Feature Extraction from Left and Right Face Regions

Each image sequence is first processed using the RDMF tracker [23] obtain facial points coordinates across all frames. We note the coordinates of these points as:

$$X = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)). \quad (4.2)$$

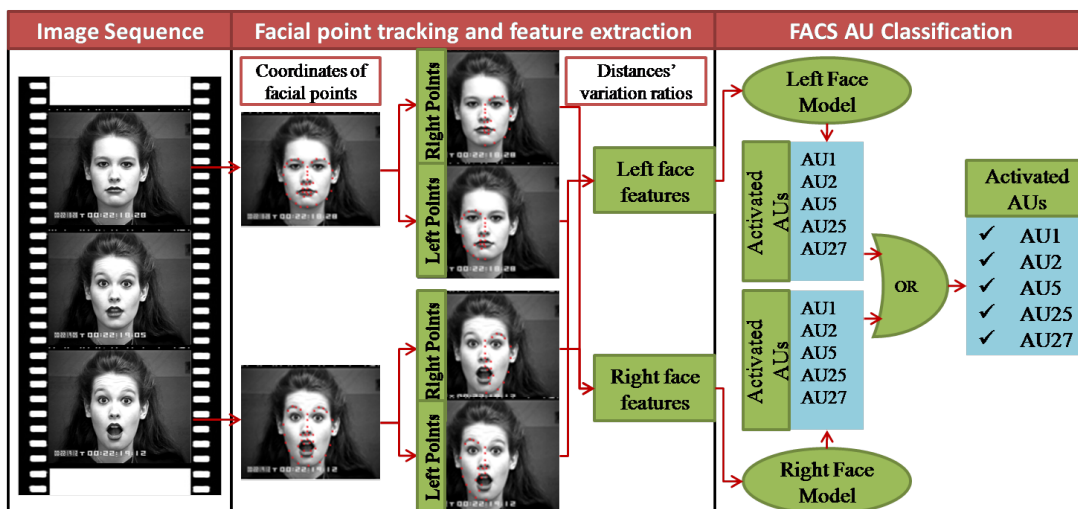


Figure 4.1: AU detection algorithm

In this work, the number of facial points is 66, which is based on the RDMF implementation. We note at this point that not all of these points can be extracted when an area from the face is occluded, such as the right or the left facial area. After detecting the facial points, a set of features are then extracted. Euclidean distances $d_{i,j}$ among the points are calculated, such that $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Then, the set of features are extracted from the specified facial area. The features are defined by the ratios:

$$R_{ij} = \frac{d_{ij}}{d_{ij}^r} \quad (4.3)$$

where d_{ij}^r is the distance between the facial points i and j in a reference frame: a frame where the facial expression is neutral. The choice of these features is suitable when the head poses in the neutral frame and the tracked frame are relatively close. Rather than employing all the points in the AU classification, we propose to train multiple classifiers for each facial area (in this case left or right). It follows that only pair of points within one part of the face are used as features in each classifier. For instance, consider a facial tracker that can track three disjoint sets of facial points: **A**, **B** and **C**, and assume that either one of the set of points **A** or **C** can be occluded at once (for example the left or right facial area). Rather than training a single classifier M_{ABC} on all the landmarks that belong to **A**, **B** or **C**, we propose to train two classifiers: M_{AB} and M_{BC} , that is for all possible combinations of landmarks being present/absent. For instance, in the testing phase, in the case where **A** (respectively **C**) is occluded, M_{BC} classifier should be used (respectively M_{AB}). The occlusion status of the facial area can be

directly extracted from the RDMF tracker by checking the point coordinates. On the other hand, if no landmarks from \mathbf{A} or \mathbf{B} are occluded, the decision should be weighted between $\mathbf{M}_{\mathbf{BC}}$ and $\mathbf{M}_{\mathbf{AC}}$. The same procedure applies for the training phase, where the classifiers can be trained only when their corresponding sets are not occluded. In our implementation, we consider three sets of landmarks: left-face-only set \mathbf{L} , right-face-only set \mathbf{R} and common points \mathbf{J} . In the case where no area is occluded, the final decision is based on a logical OR between $\mathbf{M}_{\mathbf{LJ}}$ and $\mathbf{M}_{\mathbf{RJ}}$. In the remaining part of the paper, the term “left face” (respectively “right face”) will refer to the points in \mathbf{L} and \mathbf{J} (respectively to the points in \mathbf{R} and \mathbf{J}). It is worth noting that the difference between this method and conventional ones is that multiple classifiers with various features are being used for each face region rather than using one classifier for the whole face.

4.2.2 Action Unit Detection Model

The training algorithm for each AU classifier is illustrated in Fig. 4.1 and Fig. 4.2. In the training phase, the neutral and apex frames are extracted from each video. The features from the left and right areas of the face are collected separately. One classifier is trained to classify the activation state of each AU (activated or not) and for each area of the face, using the features collected. The activation state of each AU should be available in the database or manually annotated by a FACS coder. We employ Gentle-Boost algorithm to avoid data overfitting on one hand, since the number of features is higher than the number of training data, and to make our work more comparable with other works in the literature. In the testing phase, if a facial area is occluded, the classifier of the other area will be used for classification. In the case where no facial area is occluded, the activation state of the AU is calculated by performing a logical OR on both left and right classifications. In fact, the FACS manual states that if an AU is activated in one part of the face, e.g. left eye brow raiser, the AU is annotated to be activated.

4.3 Experiments and Evaluation

In our system, we employed the author’s implementation of the RDMF facial tracker in our system [23]. The code executes in real time and its output ranges from 20 - 30 fps based on the processor and the compiler used. In order to evaluate our method, we perform three experiments. In the first one, we test our system on the Cohn-Kanade (CK) database [10] which contains 480 gray scale videos that were made public. The head orientation of subjects in the recorded videos is near-frontal. This database was collected for the purpose of facial expression recognition and this is currently the most used database in this research area. We

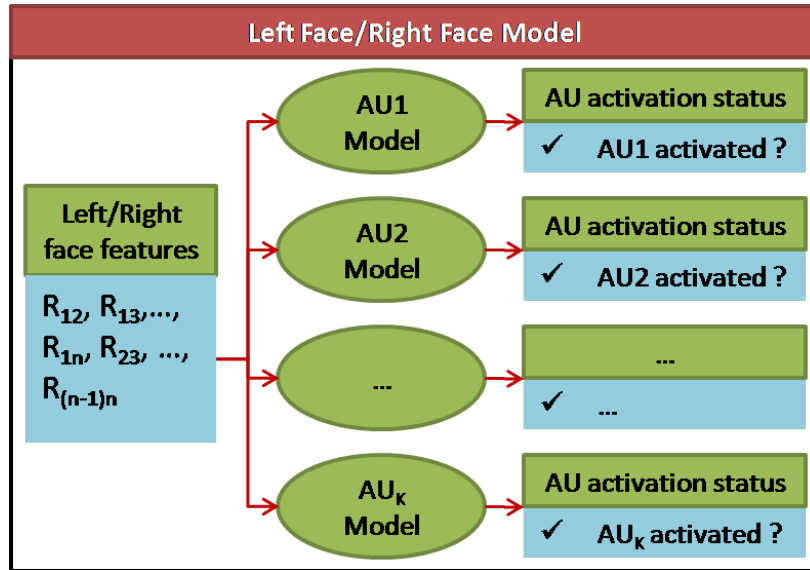


Figure 4.2: Left/Right face model

study our system on this database in order to validate our results by comparing them with a state-of-the-art geometrical approach for small head orientations. In this comparison, we use the results obtained in the frame-based experiments by Valstar et al. in [12]. In the second experiment, the Emotion detection system is tested against the one proposed also in [12]. Finally, we validate our system in the third experiment for various poses of the head, ranging from 0 degree to 90 degrees. For this case, multiple videos were recorded featuring different orientations

4.3.1 Benchmarking for Action Unit Detection

In the CK database, for each sequence of images, the facial landmarks were detected using the RDMF tracker. The coordinates of the 66 facial landmarks are tracked through each image sequence in the database. The left face area consists of 37 points while the right one contains 38. The numbers were determined experimentally and depend on the training of the facial tracker.

For each facial area, the ratios of the distances, described in section 3, were extracted from the neutral face and the apex frames. In the CK database, rather than manually labeling the apex frames for each AU in each video, which is time consuming, we considered the neutral and the apex frames to be the first and the last frames of the image sequence, respectively. We note that this assumption is fair for most AUs in this database, since most image sequences were recorded till the beginning of the apex stage for all AUs. However, we are aware of this assumption's limitation since it is not always valid, especially for certain AUs like

Table 4.1: Size of the facial landmark sets and their corresponding features

Size of landmark sets			Size of feature vector		
Full-Face	Left	Right	Full-Face	Left	Right
66	37	38	2145	666	703

AU45 (blink). As per features and class labels, the left face features and the right face features were extracted for each image as specified in the previous section. In summary, the size of the feature sets are presented in Table 4.1. Finally, we mention that we trained each of the Gentle-Boost classifiers for 10 rounds.

We conduct our experiment using the leave-one-subject-out strategy. In each fold, one subject is left out of the database, and all classifiers are trained on the remaining subjects and then tested on the subject that was left out. Binary confusion matrices are then summed together for all the experiments, i.e. for each subject in the database. Table 4.2 shows a comparison between the results obtained in our work, named Distance Ratio Classifier (DRC), and the one in the paper [12] by Valstar et al., illustrated as (TMP). In the third column, the number of positive examples for each AU is illustrated. All AUs that were previously studied, except for AU10, are also studied in this work. In our experiment, all 500 image sequences from the CK database were used, whereas the number of image sequences used in the TMP algorithm is 153. Four measures were calculated: accuracy, recall, precision and F1 measure. While the accuracy measure is a highly biased measure due to the unbalanced nature of the data, precision and recall are a better approximation of the data. The F1 measure combines the two latter measures by favoring them equally. The table is interpreted as follows. For each AU, compare the F1 column of the DRC and TMP algorithms. Precision and recall can be used for further investigation on the property of the classifier used. When needed, we refer to the accuracy of column. However, we note again that the latter measure is not very significant since the number of negative examples for each AU is much bigger than the ones with positive examples. Although, the results obtained in our method are highly optimistic, no direct conclusion on the superiority of our algorithm over the temporal based algorithm can be made since the selection of videos used is not the same.

As can be seen, AU1 (inner brow raisers) and AU24 (lip pressor) show very close results with superior measures for the DRC method. Additionally, our method is also superior for other AUs such as AU2, AU4, AU5, AU7 and AU9. We believe that the reason for this improvement is behind the DMF model used by the Facial Tracker. On the other hand, other AUs such as AU6 (cheek raiser) and AU12 (lip corner puller) show that the method proposed is not accurate. In fact, it can be observed that the landmarks of the lip corners are not tracked effectively using the

RDMF when performing AU12. Lastly, the poorest result achieved by DRC is for AU45 (blink). The lack of precision for this AU detector is mostly attributed to the preprocessing assumption that we made. In fact, most image sequences end after the offset of this AU has occurred, i.e. the final frame does not generally contain the apex for AU45. Any accurate result for this AU is due to the mere correlation between AUs in the database. We mention that we trained each of the Gentle-Boost classifiers for 10 rounds.

4.3.2 Benchmarking for Emotion Detection

In this section, the same features extracted previously are used in the emotion detection. We compare our method to the one proposed by Valstar et al. in [12] on the CK database. Note that the annotation for the emotions is provided in the database. The confusion matrix obtained in [12] is illustrated in Table 4.3. The results obtained using our method are illustrated in Table 4.4. We test our system by using the points from left face only. The classification accuracy in our method is very comparable to the one in [12]. The classification rate for the emotion anger and surprise is much better in our method. On the other hand, the sadness classification rate is lower in our case. We note that the database subsets used are not the same in our experiment and the one in [12]. Thus, we don't elaborate more on the comparison, and simply state that the results obtained in our method are comparable to the state of the art approach in [12].

4.3.3 Various Pose Evaluation for Brow Raisers

This section is from our work [22]. In this section, we test our method on various orientations. For this purpose, a training set was created featuring two subjects in 45 videos in total. The sequences were recorded for three discrete yaw orientations (horizontal rotations), namely no yaw, moderate yaw and extreme yaw, approximated by: 0 degree, 45 degrees and 90 degrees angles to the camera imager. All videos were taken using a DMC-F3 Panasonic camera at a resolution of 1280x720 pixel², a rate of 30 fps and a distance of 2 meters from the subject.

The subject was asked to stand in a frontal pose, and then to rotate his head by a specific angle until facing a marker on the wall and to perform an AU or a combination of AUs. Afterwards, we manually annotated the neutral frames and apex frames of each sequence. We note that only the neutral frame preceding the onset phase is considered. A sample recorded set of images is shown in Fig. 4.3. The subjects were asked to perform AU1 and AU2 (brow raisers). As a matter of fact, it is essential to assess the validity of any FACS system for the most common AUs (AU1 and AU2 consists about 20% of the CK database).

Table 4.2: Comparison between the work in this paper (DRC) and the one of Valstar et al (TMP).

AU	Meth.	Videos	Acc.	Recall	Prec.	F1
1	DRC	144	0.910	0.809	0.864	0.835
	TMP	68	0.918	0.808	0.844	0.826
2	DRC	97	0.964	0.871	0.946	0.907
	TMP	50	0.939	0.791	0.879	0.833
4	DRC	156	0.896	0.755	0.864	0.806
	TMP	54	0.870	0.604	0.658	0.630
5	DRC	78	0.926	0.708	0.761	0.734
	TMP	37	0.904	0.566	0.629	0.596
6	DRC	111	0.870	0.713	0.694	0.703
	TMP	39	0.930	0.789	0.811	0.800
7	DRC	108	0.862	0.685	0.679	0.682
	TMP	31	0.870	0.268	0.315	0.290
9	DRC	50	0.972	0.864	0.826	0.844
	TMP	30	0.928	0.676	0.497	0.573
12	DRC	113	0.904	0.780	0.780	0.780
	TMP	42	0.930	0.827	0.844	0.836
15	DRC	81	0.910	0.609	0.661	0.634
	TMP	19	0.969	0.500	0.283	0.361
20	DRC	70	0.924	0.638	0.772	0.698
	TMP	34	0.908	0.466	0.582	0.517
24	DRC	43	0.928	0.421	0.533	0.471
	TMP	17	0.935	0.395	0.497	0.440
25	DRC	303	0.888	0.917	0.905	0.911
	TMP	19	0.851	0.717	0.782	0.748
26	DRC	39	0.926	0.175	0.636	0.275
	TMP	27	0.902	0.336	0.380	0.357
27	DRC	77	0.972	0.919	0.895	0.907
	TMP	30	0.964	0.836	0.873	0.854
45	DRC	19	0.954	0.091	0.400	0.148
	TMP	23	0.943	0.584	0.408	0.480
DRC Avg.			0.920	0.664	0.748	0.689
TMP Avg.			0.917	0.611	0.619	0.609

Table 4.3: Confusion matrix for emotion classification using the method by Valstar et al. (TMP)

	An.	Di.	Fe.	H.	Sad.	Sur.	Rate
Ang.	2	3	2	0	9	1	0.118
Disg.	1	19	1	1	4	1	0.704
Fear	1	4	15	5	2	1	0.536
Hap.	1	0	3	33	0	1	0.868
Sad.	4	2	1	0	16	1	0.667
Sur.	0	1	1	1	0	34	0.919

Table 4.4: Confusion matrix for emotion classification using the features from the left face

	An.	Di.	Fe.	H.	Sad.	Sur.	N.	Rate
Ang.	19	3	0	2	4	0	1	0.655
Disg.	1	32	0	0	0	0	1	0.941
Fear	0	1	14	1	1	0	0	0.824
Hap.	0	0	1	60	0	0	0	0.984
Sad.	4	0	2	0	9	0	1	0.563
Sur.	0	0	8	0	0	63	0	0.887
Neu.	0	0	0	0	0	0	228	1

In the testing phase, only the tracked points from the neutral and apex frame were extracted from the video. A previously trained DRC classifier set from the CK database was used on the data set. Table 4.5 illustrates the evaluation of our method for the three yaw intensities, and for the two AUs. We note that the tracker failed to track some videos for extreme face orientations. These videos are excluded from the final statistics. The second column shows the intensity of the head orientation. The numbers of positive and negative examples are illustrated in the third and the fourth column. The same measures from experiment 1 are used in this experiment. Not surprisingly, the F1 increased when the orientation intensity was stronger. In fact, this is consistent with the work in [15] which concluded that profile views are better than frontal views for AU detection.

4.4 Summary of Observations for Method 1

Method 1 gives a working system that can detect AUs for various head poses without any prior training on these poses. The method has comparable results with the state-of-the-art geometrical algorithm in [12] for near-frontal head orientation.

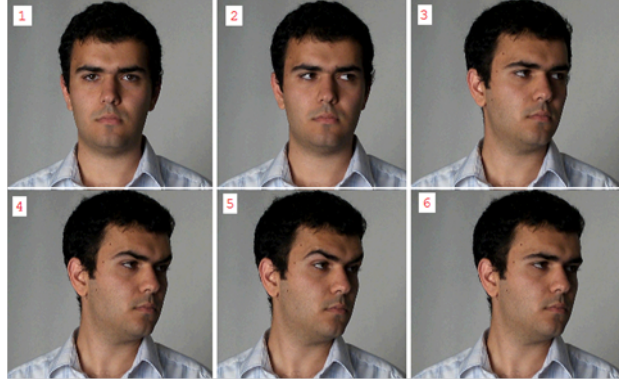


Figure 4.3: Samples in an images sequence from the recorded database. The subject is first asked to rotate their face, then to perform the AU. The third image is annotated as the neutral face, and the fifth is annotated as the apex frame.

Table 4.5: Evaluating algorithm on the recorded database for 3 head orientation intensities and two AUs

AU	Int.	P	N	Acc.	Recall	Prec.	F1
1	1	4	10	0.929	1.000	0.800	0.889
	2	14	22	0.861	0.786	0.846	0.815
	3	9	11	0.950	0.889	1.000	0.941
2	1	4	10	0.643	0.500	0.400	0.444
	2	7	29	0.889	0.571	0.800	0.667
	3	5	15	0.850	0.600	0.750	0.667

The model was able to generalize to a new database where AUs were still detectable at various pose orientations. One of the limitations of our proposed method is that the fusion (OR) considers that both left and right classifiers are equally reliable. In our next work in Chapter 5, we propose a fusion method that optimizes the fusion of left and right regions depending on the angle.

Chapter 5

Method 2: Angle Dependent Emotion Detection Method

In this chapter, we provide details for the method that can recognize basic emotions through an optimized fusion approach. The method is shown in Section 5.1. The performed experiments are shown in Section 5.2. Section 5.3 concludes the proposed proposed approach.

5.1 Proposed Method

In this section, we present our proposed method to detect emotions using a decision fusion. This section is divided as follows. In subsection 5.1.1, we describe an overview of the classification and fusion approach. Then, we discuss how to generate ground truth to train and test our algorithm in subsection 5.1.2. The third subsection describes the features and how the classifiers are trained. Finally, the process of training the fusion model parameters is explained.

5.1.1 Overview

We hereby describe the process of detecting an emotion given two sets of points from the left and right regions of the face. This process is illustrated in Figure 5.1 and Figure 5.2. Before describing the emotion detection process, we assume that the pose and a number of facial points are tracked using third-party software that applies algorithms such as Particle filtering with Factorized Likelihoods [24] or Kalman filtering schemes such as the method proposed in [25]. Alternatively, in the context of this work, we generate these tracked points using a 3D database. Also, we only limit the scope of this study to the variations in the yaw angle of the face.

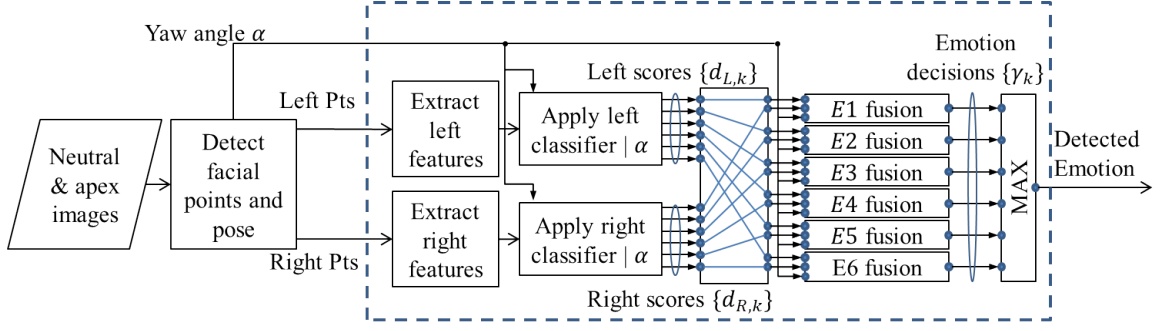


Figure 5.1: Proposed classification system

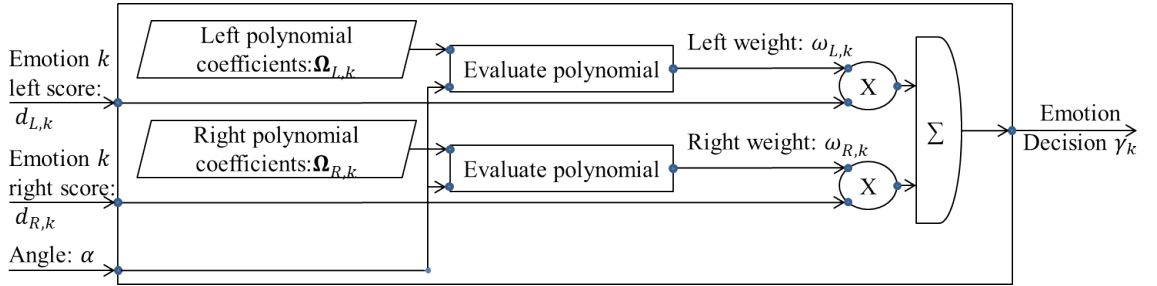


Figure 5.2: Proposed fusion approach for one emotion

The system starts by extracting features from the left and right regions independently, followed by applying a multi-class classifier that detects the emotion in each region. The features and classifiers used are further described in section 5.1.3. The output of the two classifiers is two sets of scores, $\{d_{L,k}\}$ and $\{d_{R,k}\}$, which represent the probability of having an emotion E_k , given the left and right facial features respectively. Then, for each emotion E_k , we perform a fusion on the two scores $d_{L,k}$ and $d_{R,k}$, depending on the yaw angle α . The proposed fusion, further illustrated in Figure 5.2, works as follows. Considering an emotion E_k , we propose to apply a linear combination of the left and right scores of that emotion. Additionally, the weights $w_{L,k}$ and $w_{R,k}$ of this combination are mainly determined by the yaw angle α through two polynomial approximations, using two stored vectors $\Omega_{L,k}$ and $\Omega_{R,k}$. The output of this fusion is a decision score γ_k for each emotion. Finally, the emotion with the highest fusion score is selected.

The rest of this section is divided as follows. We first describe how the training data is generated from a 3D database in section 5.1.2. Then, we discuss the features and classifiers in section 5.1.3. Finally, we show how the fusion parameters are trained in section 5.1.4.

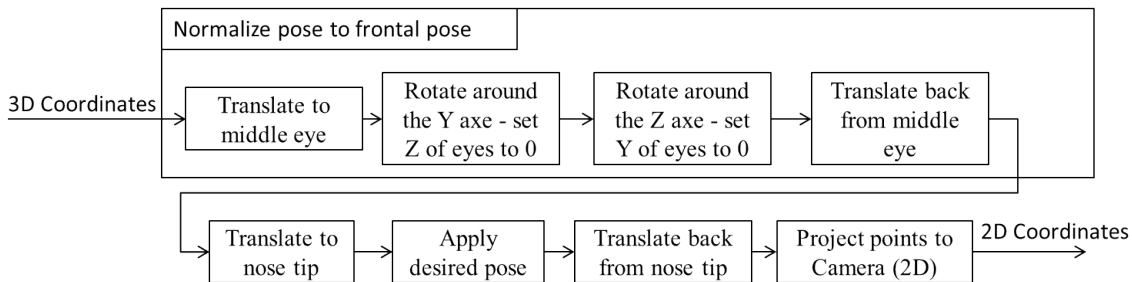


Figure 5.3: Steps to generate 2D points for each specific pose.

5.1.2 Ground Truth Data

Training the classifiers and fusion parameters requires training instances for multiple yaw angles. Therefore, we use a 3D database [26] and generate training instances for various yaw angles. The input of this procedure is a 3D frame from the database, containing 3D coordinates of specific facial points, a cloud of 3D points from the face, and a texture; the latter two are essential for *3D rendering* of an image. However, since only geometric features are needed in our proposed method, we only transform the 3D coordinates of the key facial points we want to use. The output of this procedure is the set of key facial points as seen by a 2D camera when the head performs a yaw rotation α . The generation procedure is described in this section and in Figure 5.3. At this point, we note that the transformations used to project 3D models onto a camera are very well established in computer graphics. However, the procedure explained hereby only shows how these transformations are applied to ensure further alignment in the training data.

Provided that the face of a subject is in a near frontal pose, i.e. a pose where the face of the subject is not directly facing the camera, we first transform the coordinates of the facial points to a frontal pose. This preliminary transformation is applied in order to ensure that all subsequent pose transformations provide consistent results. Two important characteristics of a frontal pose are that the coordinates of the eyes are vertically aligned, and that they have the same depth. We denote by pose normalization this described process. This normalization can be achieved by performing two rotations about the center of the eyes.

Formally, we assume a direct system of coordinates (X, Y, Z) centered at the camera where X is the horizontal axis, Y the vertical one, and Z the axis representing the depth. The suggested frontal pose normalization is equivalent to a translation to the center of the eyes, followed by two rotations about the Z and Y axis that set the Y and Z coordinates of the eyes to 0, respectively. Finally, a translation from the center of the eyes back to the camera system is performed. By using a *homogeneous* representation of the coordinates which is a representa-

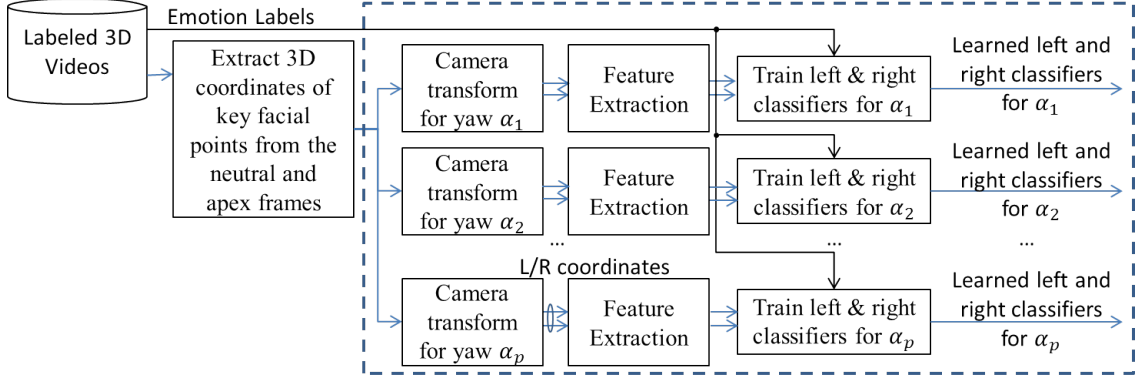


Figure 5.4: Approach to train the classifiers

tion in a 4 dimensional system of coordinates, all rotations and translations can be described by matrix multiplications. The normalization process can then be represented as

$$P_{\text{frontal}} = T_{\text{eyes}}^{-1} \times R_Z \times R_Y \times T_{\text{eyes}} \times P \quad (5.1)$$

where P is a $4 \times N$ matrix containing the homogeneous coordinates of the N points to transform, T_{eyes} is a 4×4 matrix representing the translation from the camera system to the center point between the eyes, R_Y is the 4×4 rotation matrix about the Y axis that sets the Z coordinates of the eyes to 0. Similarly, R_Z is the 4×4 rotation matrix about the Z axis that sets the Y coordinates of the eyes to 0. The values of these matrices can be computed by calculating the coordinates of the inner eye corners and the angle they form with each of the axis discussed above.

The second step of the process is to perform the desired head pose transformation. Ideally, a yaw rotation is a rotation of the head about a virtual axis passing through the center of the neck. However, the coordinates of the neck are not generally tracked in facial point trackers. Alternatively, we employ the nose tip coordinates and apply the transformation on a vertical axis passing through the nose tip. To apply the yaw rotation, a translation should be first performed from the camera to the nose tip coordinates. The inverse of this translation should also be performed after the rotation is applied. All of the previous steps are achieved through a sequence of transformation matrix multiplications.

$$P_{\text{pose}} = T_{\text{nose}}^{-1} \times R'_Y \times T_{\text{nose}} \times P_{\text{frontal}} \quad (5.2)$$

where R'_Y is the rotation matrix representing the yaw rotation around the axis Y . The transformed coordinates in P_{pose} represent the 3D coordinates as observed when the head performs the desired yaw angle. The final step is to project the

computed coordinates on the camera plane to obtain the 2D coordinates. Ideally, this can be achieved by applying a *perspective projection* if the *intrinsic parameters* (e.g. focal length) of the camera are available. However, in the case of the used database [26], we do not have access to these parameters. Thus, we apply a *weak perspective projection* to the camera, which simply considers scaled values of the X and Y components and ignores the Z component. This approximation is largely valid since the depth variations within the face are relatively small compared to the average depth of the face with respect to the camera.

At the end of this process, for each subject and emotion, we obtain a set of 2D facial point coordinates for the neutral and apex frame, and for several angles. This set of coordinates and labels are later used to train and test our classifiers.

5.1.3 Feature Extraction and Classification

After generating the facial points as discussed in section 5.1.2, we now discuss the geometrical features extraction as well as the emotion classification method. The facial points are first split to left and right coordinates by grouping the ids of the points provided by the database. Features are then extracted from each region to later decide on the classification scores for each emotion. The outline of the training approach is illustrated in Figure 5.4. After extracting the key 3D facial points, which are provided with the database, we transform the 3D coordinates for several yaw angles. Afterwards, we extract the geometrical features and train the left and right classifiers. The details of this process are further described in this section.

Similarly to the algorithm in [18] and our earlier study in [22], we extract a set of features representing the displacement of the facial points between the apex frame and the neutral frame of an image sequence, in the left and right regions independently. This is achieved by first calculating the distances between all pairs of points in the neutral frame, and the distances between all pairs of points in the apex frame. In this work, the final feature set for an image sequence is the set of ratios of those distances. Namely, for any pair of points $\{p, q\}$ in the left or right regions, the feature extracted is

$$f_{p,q} = \frac{d_{p,q}^{\text{apex}}}{d_{p,q}^{\text{neutral}}} \quad (5.3)$$

where $d_{p,q}^{\text{apex}}$ is the Euclidean distance between points p and q in the apex frame, and $d_{p,q}^{\text{neutral}}$ the one between the same points in the neutral frame.

Following the feature extraction process, a basic classifier can be applied. For similar features extracted from the entire face, Hu et al. [18] tested several classifiers and concluded that support-vector-machines (SVMs) performed best. Similarly, we apply the multi-class SVM algorithm implemented in LIBSVM [27].

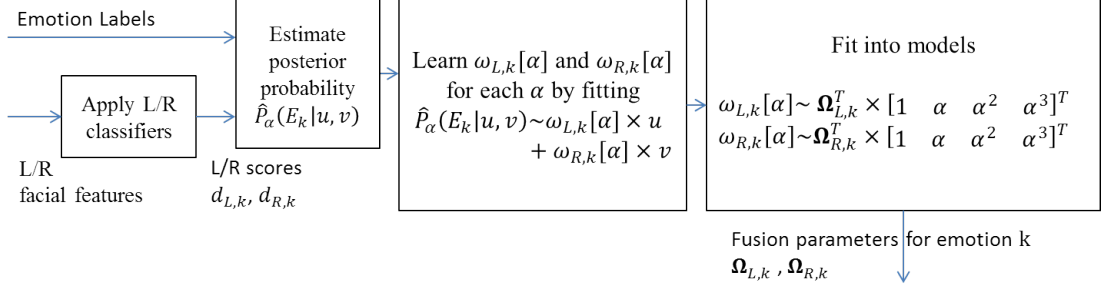


Figure 5.5: Approach to learn the fusion parameters

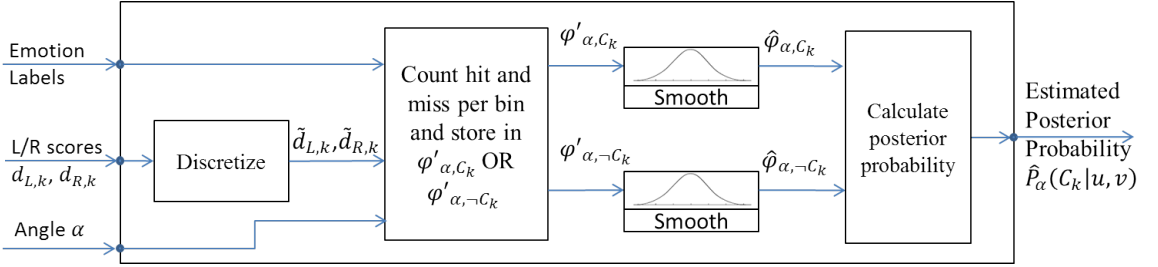


Figure 5.6: Estimating a smoothed posterior probability distribution

When applying these classifiers, and in order to perform an advanced fusion on the decision level, we perform a soft classification. In contrast to a hard classification where a testing instance strictly belongs to one single class at a time, a soft classification allows the instance to belong to all classes with different degrees of membership. The membership degree can be represented as a probability that the instance belongs to the specific class given the features observed. A common way to extract the probability estimates for multiple classes in SVM algorithms is the method proposed by Wu et al. [28] which is also implemented in LIBSVM. Therefore, whenever the trained SVM models are applied, we extract a vector of membership levels representing 6 scores per region. Each score represents the probability that the instance belongs to one of the six basic emotions. The fusion of these scores is further described in the next section.

5.1.4 Fusion Model

Typically, when a single classifier is applied and a vector of probabilities are obtained, we apply the Maximum A Posteriori (MAP) rule and choose the emotion with the highest *posterior probability*. On the other hand, when the face is split into two regions, two classifiers are used, and thus two probability vectors are extracted. Performing a convenient fusion of the two probability vectors is an

optimization problem, and we argue that this fusion should depend on the yaw angle. Therefore, we propose a fusion algorithm that optimizes the fusion of the left and right probability vectors, depending on the angle. This section contains our proposed method to train our MAP-based fusion. As previously described in the beginning of section 5.1 and in Figure 5.1, the fusion consists of a linear combination of the probabilities obtained in each region in the face, left or right.

Formally, a MAP fusion given the observed scores $d_{L,k}$ and $d_{R,k}$ can be formulated as follows

$$\hat{y} = \arg \max_{k \in \{1, \dots, c\}} P(E_k | d_{L,k}, d_{R,k}) \quad (5.4)$$

where \hat{y} is the estimated emotion, and c is the total number of emotion classes. Note that the sparsity of the observations provided by the left and right scores makes the fusion less generalizable for all angles. Therefore, to provide further generalization across angles, we perform the three following steps: (1) smooth the posterior probability distribution, (2) learn the weights of a linear combination to estimate the probability distribution, and (3) learn the polynomial parameters to estimate the latter weights given the angle. These three steps are further discussed in each of the following sections, and can be summarized by Figure 5.5 and Figure 5.6.

The rest of this section is organized as follows. In section 5.1.4, we provide a smoothing approach to estimate the posterior probability distribution. In section 5.1.4, we estimate the rest of the weights.

Posterior Probability Smoothing

In this part of our work, we propose a method to estimate a smoothed distribution of the posterior probability of a given emotion. The method, illustrated in Figure 5.6, consists of discretizing the left and right scores, $(d_{L,k}, d_{R,k})$, and computing 2D histograms for the pair of scores in two cases: when E_k is the ground truth emotion and when it is not. The obtained histograms are then smoothed. Subsequently, the histograms are used to calculate the likelihood and prior probabilities. Finally, the estimate of the posterior probability is calculated.

We begin by discretizing each input into N equal bins. Formally, we note $\tilde{d}_{i,k}$ the discretization of $d_{i,k}$ such that

$$\tilde{d}_{i,k} \stackrel{\text{def}}{=} \max \left\{ m \in G \mid m + \frac{q}{2} \leq d_{i,k} \right\} \quad (5.5)$$

with $G = \{\frac{q}{2}, q + \frac{q}{2}, 2q + \frac{q}{2}, \dots, (N-1)q + \frac{q}{2}\}$, and $i \in \{L, R\}$. In the experiments, we considered a number of bins $N = 10$, and $q = 0.1$. To calculate the two 2D histograms per emotion, when E_k is the ground truth and when it is not, we define the two histogram spaces as follows.

$$\varphi_{\alpha, E_k}(u, v) = \left| \left\{ \left(\tilde{d}_{L,k} \right) \mid u = \tilde{d}_{L,k}, v = \tilde{d}_{L,k} \right\} \right| \quad (5.6)$$

$$\varphi_{\alpha, \neg E_k}(u, v) = \left| \left\{ \left(\tilde{d}_{L,j} \right) \mid u = \tilde{d}_{L,j}, v = \tilde{d}_{L,j}, j \neq k \right\} \right| \quad (5.7)$$

where E_k indicates that the ground truth label is E_k , $\neg E_k$ indicates that it is not, and $|\cdot|$ indicates the cardinality of a set, i.e. the number of training instances that satisfy the conditions indicated. In order to estimate the posterior probability, Equation 5.6 is used in the estimation of the *likelihood probability* while both Equations 5.6 and 5.7 are used in the estimation of the *prior probability*. Therefore, before estimating the *posterior probability*, a Laplacian correction is due to avoid a division by 0.

$$\varphi'_{\alpha, E_k}(u, v) = \varphi_{\alpha, E_k}(u, v) + 1 \quad (5.8)$$

$$\varphi'_{\alpha, \neg E_k}(u, v) = \varphi_{\alpha, \neg E_k}(u, v) + 1 \quad (5.9)$$

Additionally, due to the sparsity of the histograms and to ensure a smoother model that can be generalized over all angles, we define $\hat{\varphi}$ a smoothed histogram derived from φ' . The smoothing process can be any standard smoothing operation on 2D images such as average weighting.

Therefore, for a given angle α , after computing the histograms, the likelihood probabilities can be estimated as follows.

$$\hat{P}_\alpha(u, v | E_k) = \frac{\hat{\varphi}_{\alpha, E_k}(u, v)}{\sum_{u,v} \hat{\varphi}_{\alpha, E_k}(u, v)} \quad (5.10)$$

$$\hat{P}_\alpha(u, v | \neg E_k) = \frac{\hat{\varphi}_{\alpha, \neg E_k}(u, v)}{\sum_{u,v} \hat{\varphi}_{\alpha, \neg E_k}(u, v)} \quad (5.11)$$

Consequently, the resulting total probability can be calculated

$$\hat{P}_\alpha(u, v) = \hat{P}_\alpha(u, v | E_k) \times P(E_k) + \hat{P}_\alpha(u, v | \neg E_k) \times P(\neg E_k) \quad (5.12)$$

Finally, the posterior probability estimate of having an emotion E_k given $d_{L,k}$, $d_{R,k}$ and α is:

$$P(E_k | d_{L,k}, d_{R,k}, \alpha) \approx \hat{P}_\alpha(E_k | \tilde{d}_{L,k}, \tilde{d}_{R,k}) \quad (5.13)$$

where

$$\hat{P}_\alpha(E_k | u, v) = \frac{\hat{P}_\alpha(u, v | E_k) \times P(E_k)}{\hat{P}_\alpha(u, v)} \quad (5.14)$$

Angle-Dependent Posterior Probability Regression

Using the smoothed estimated fusion probability $\hat{P}_\alpha(E_k|u, v)$ obtained, we estimate a multi-variable regression for each class given u , v and α . The objective of this training is to learn a fused decision for emotion k that fits the estimated posterior probability. This can be noted as

$$\hat{P}_\alpha(E_k|u, v) \approx \gamma_k(u, v, \alpha) \quad (5.15)$$

To fit this equation, two regressions are performed sequentially:

1. For each angle, a regression is performed for all pairs (u, v) to determine a linear combination of the left and right scores of an emotion E_k .
2. In the second regression, the model parameters are fitted for all angles α .

Namely, these regressions can be defined by the following two equations.

$$\gamma_k(u, v, \alpha) \approx \omega_{L,k}[\alpha] \times u + \omega_{R,k}[\alpha] \times v \quad (5.16)$$

and

$$\omega_{i,k}[\alpha] \approx \boldsymbol{\Omega}_{i,k}^T \times \mathbf{a} \quad (5.17)$$

where $\omega_{i,k}[\alpha]$ are the coefficients of the linear combination learned for a given yaw angle α , $\boldsymbol{\Omega}_{i,k}$ is a vector containing the polynomial coefficients used to estimate $\omega_{i,k}[\alpha]$, $\mathbf{a} = [1 \ \alpha^1 \ \alpha^2 \ \dots \ \alpha^m]^T$, and m is the polynomial degree. Alternatively, Equations 5.16 and 5.17 can be rewritten as:

$$(\omega_{L,k}[\alpha], \omega_{R,k}[\alpha]) = \arg \min_{(b_1, b_2)} \sum_{(u, v)} ((b_1 \times u + b_2 \times v) - \gamma_k(u, v, \alpha))^2 \quad (5.18)$$

and

$$\boldsymbol{\Omega}_{i,k} = \arg \min_{\boldsymbol{\Omega}_{i,k}} \sum_{\alpha} (\boldsymbol{\Omega}_{i,k}^T \times \mathbf{a} - \omega_{i,k}[\alpha])^2 \quad (5.19)$$

These least square minimizations are computed to generate the parameters $\boldsymbol{\Omega}_{L,k}$ and $\boldsymbol{\Omega}_{R,k}$ for each emotion.

5.2 Experiments and Results

In this section, we conduct a set of experiments to evaluate the performance of our algorithm. In Section 5.2.1, we evaluate our method by running a set of preliminary experiments. In Section 5.2.3, we compare various approaches to fuse the left and right classifications. Finally, Section 5.2.4 compares the fusion approach to a prior art method.

5.2.1 Experiment Setup

To test our algorithm, we employ the BU-4DFE [26] database. The BU-4DFE database features 101 subjects, with a balanced gender ratio and ethnic variety, where each subject performed 6 emotions: anger, disgust, fear, joy, sadness and surprise. The database consists of 606 videos along with 83 tracked facial points. 15 of these points that belong to the contour of the face were removed. Therefore, only 68 of the database points were considered.

For the purpose of employing our image-based method, we manually extract two frames from each video: one frame with a neutral expression and one with a full-blown (apex) expression. All the experiments reported were performed using 10-fold cross validation. In the training phase and for each instance, the tracked points are transformed to the specified pose as described in Fig. 5.3 and Section 5.1.2. The features were generated as described in Section 5.1.3. First, two multi-class linear SVM classifiers are trained to classify the 6 basic emotions, one for the left region of the face, and one for the right region. To classify the 6 basic emotions, from the left and right regions of the face independently as described in Section 5.1.3 and Figure 5.4. Next, the output of the two classifiers is then used to learn the fusion parameters as described in Section 5.1.4 and Figure 5.5. In the testing phase, the tracked points are synthesized in the same manner and both classifications and the fusion are applied as described in Figures 5.1 and 5.2. We note here that the same training data is used to train classifiers and fusion parameters.

5.2.2 Method Evaluation

In this section, we present some preliminary results to evaluate our method. Starting with the set of 3D coordinates, the coordinates were transformed to various poses, as shown in Figure 5.8. The angles considered in our experiments are between -45 and 45 degrees with increments of 5 degrees. Afterwards, the left and right coordinates were grouped together in two sets of points. Features were extracted from each group separately and a classifier was learned for all emotions per facial region.

After applying the learned classifiers, the remaining parameters were learned as described in Figure 5.9. Considering the emotion happy for illustration, the left and right scores for the happy emotion were extracted from all the training instances. For each angle, an estimate of the posterior probability was calculated as described in Equation 5.13. This estimate can be represented as a non decreasing function of the left and right weights between 0 and 1, as shown in the 3D plot. In the second step, we fitted the points from this plot to a linear summation of the weights as in Equation 5.16. For each angle, a left weight and a right weight,

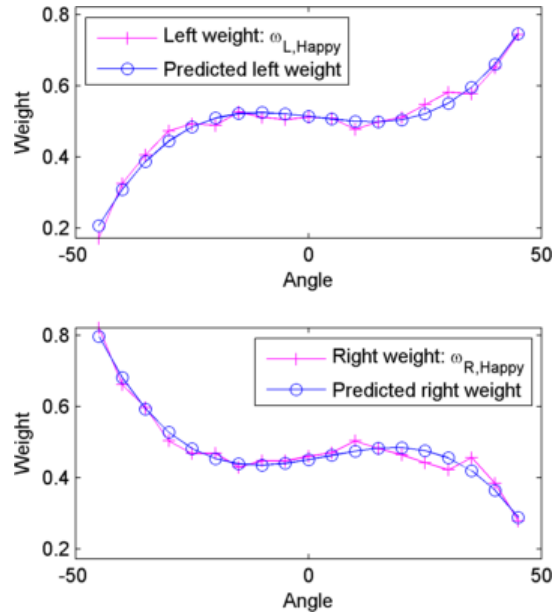


Figure 5.7: Fusion weights and approximations for the happy emotion

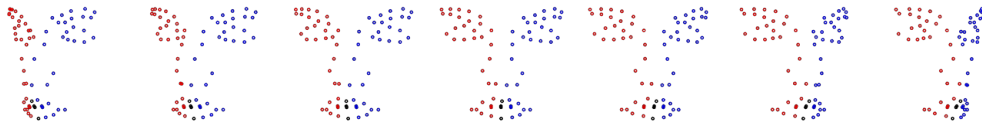


Figure 5.8: Examples of generated facial points with 15 degrees increments

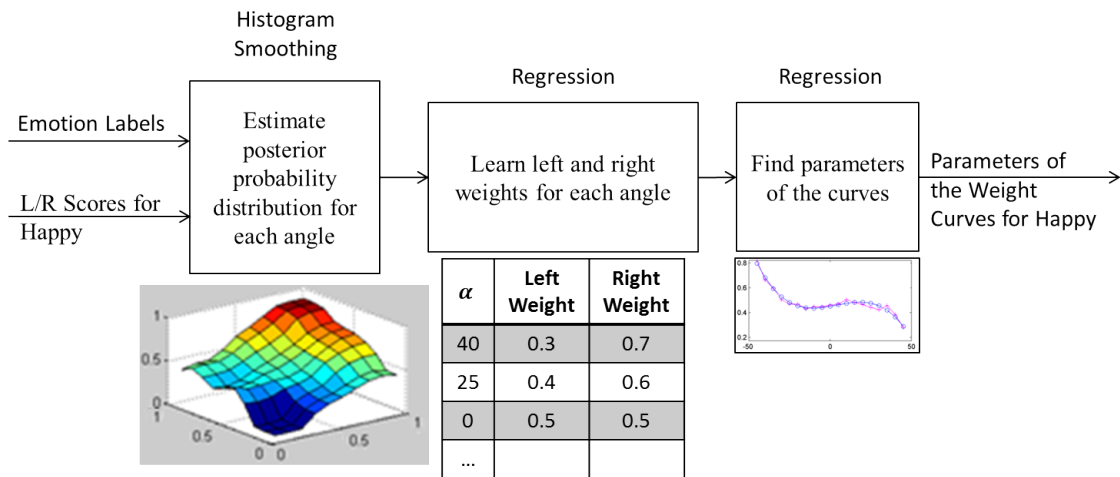


Figure 5.9: Learning fusion parameters

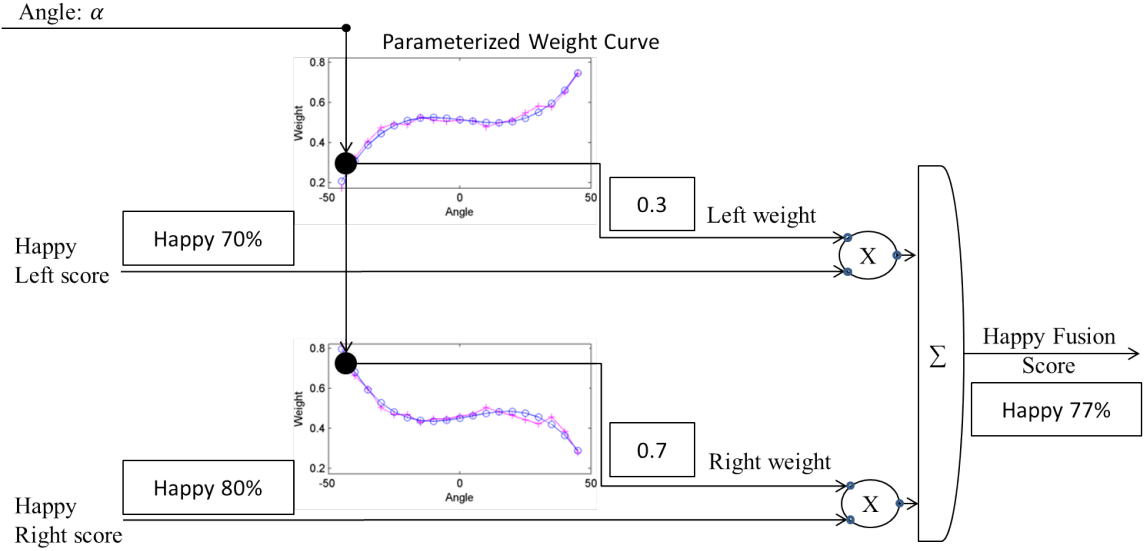


Figure 5.10: Applying fusion parameters

$\omega_{L,\text{Happiness}}[\alpha]$ and $\omega_{R,\text{Happiness}}[\alpha]$, were then obtained. In the third and final step, we fitted these weights through a polynomial regression where α is the variable, as in Equation 5.17. The set of coefficients $\Omega_{L,\text{Happiness}}$ and $\Omega_{R,\text{Happiness}}$ were then obtained. The same process was repeated for all six emotions in this database.

To evaluate the system, the learned classifiers and parameters were then applied to the cross validation test data. On the other hand, when recognizing the emotion of a test subject, the left and right classifiers were first applied on the face. The left and right scores for all emotions were then obtained. Considering the emotion happy, the fusion score can be illustrated in Figure 5.10. The angle α was applied to determine an estimate of the left and right weights $\omega_{L,\text{Happiness}}$ and $\omega_{R,\text{Happiness}}$. These weights were then multiplied with the left and right scores obtained from the classifiers, thus obtaining a fused score for the emotion happy. In the same manner, the fused scores for all emotions were calculated. Finally, the emotion with the highest fusion score was selected.

Figure 5.7 illustrates the parameters $\omega_{L,k}$ and $\omega_{R,k}$ obtained for the happy emotion. We applied a regression to the third degree polynomial ($m = 3$ in Section 5.1.4). A general ascending trend can be observed in the left weight curve, indicating that at the boundaries, a much higher weight should be applied to the region that is facing the camera. However, a local maximum in the same figure can also be observed around the angle -30 degrees. In fact, the reported accuracy of all emotions by the left and right classifiers seem to follow the same trend by looking at Figure 5.11, which suggests that the performance is increased at a slight yaw

	Baseline Approach						Fusion Approach						
	Anger	Disgust	Fear	Joy	Sad	Surprised	Anger	Disgust	Fear	Joy	Sad	Surprised	
Anger	72.52	9.41	0.00	0.99	11.63	5.45	Anger	75.00	9.16	0.00	0.99	14.6	0.25
Disgust	7.92	63.37	5.45	5.20	2.72	15.35	Disgust	10.15	76.49	4.21	4.21	2.97	1.98
Fear	0.99	13.61	40.10	14.60	10.89	19.80	Fear	0.99	14.11	45.54	13.12	13.86	12.38
Joy	0.25	3.96	2.48	79.70	2.72	10.89	Joy	0.00	4.95	1.98	86.88	5.45	0.74
Sad	13.12	1.73	7.18	0.00	75.25	2.72	Sad	12.62	1.24	4.46	0.00	80.69	0.99
Surprised	0.00	0.50	7.67	0.00	0.25	91.58	Surprised	0.00	0.99	10.40	0.00	0.50	88.12
Avg. RR	70.42						Avg. RR 75.45						

Table 5.1: Confusion matrix comparison between the proposed algorithm and the baseline approach, for 4 head orientations of 0, 15, 30 and 45 degrees

angle. This is consistent with previous works that reported a higher performance at non frontal angles. On the other hand, the same reasoning can be applied for the right weight curve.

The weights determined for the fusion vary smoothly with respect to the angle, in Figure 5.7. However, these weights do not generally add up to 1. This is attributed to the fact that the left and right weights are determined independently without any constraint.

Overall the proposed method is expected to work well under the following assumptions: (1) The orientation of the head with respect to the camera mainly consists of a yaw rotation and (2) the probability extracted from the classifiers can accurately fit the model described in Equation 5.16 and Equation 5.17, and (3) both regions of the face are equally reliable for a test subject in a frontal pose.

5.2.3 Fusion Approaches

The goal of this experiment is to evaluate the performance of the fusion based method and compare it to other fusion approaches. We first compare with the left and right detection approaches (consistently performing a hard decision from either the left or right region). The accuracy of the left and right classifiers are evaluated individually, and then compared with the obtained accuracy of the fusion method. The results are illustrated in Figure 5.11. For most of the angles, the recognition rate corresponding to the fusion is better than the rates for the left and right alone. At all times, it seems to do as good or better than the best standalone classifier of the left and right. In a second experiment, we compare our proposed fusion approach to a max-based fusion where the emotion with the highest score on both regions is selected. After conducting the experiment on angles between 0 and 45 with 15 degrees increment, the obtained accuracy for max-based fusion is 75.08% while the proposed fusion approach showed an accuracy of 75.45%, which shows a slight superiority of the proposed optimization.

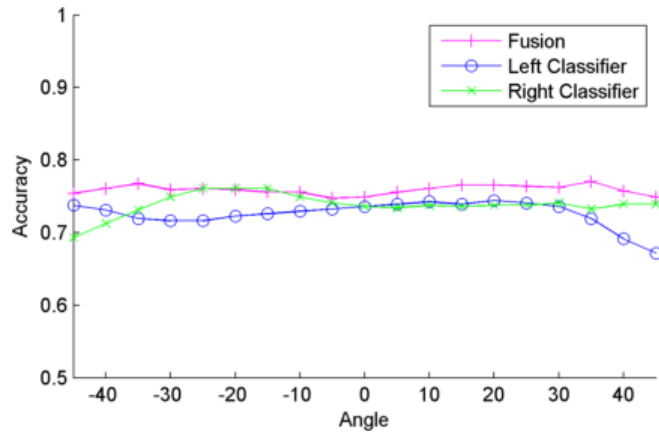


Figure 5.11: Accuracies of the left, right and fusion classifiers, calculated per angle

5.2.4 Baseline Classification

The objective of this experiment is to compare the proposed decision fusion against classifying the entire face at once. By considering the features from the entire face, this method becomes similar to the one proposed by Hu et al. [18]. The algorithm extracts the same type of features from the entire face, and applies a similar SVM linear classifier. The main differences are splitting the face and fusing the outputs of left and right emotion classifiers. We compare our method to the prior art approach. The angles are varied between 0 and 45 with 15 degrees increment. The reported results of the fusion approach and the baseline approach are presented in Table 5.1. For all emotions except for surprise which had a decrease in recognition rate, we have an increase in the individual recognition rates of our fusion method. An increase of 13% is achieved on the emotion disgust which is generally a difficult emotion to detect. Overall, the new proposed method accuracy showed a 7.1% increase in accuracy compared to the method proposed by Hu et al. [18].

5.3 Summary of Observations for Method 2

In this chapter, we proposed an approach to detect facial expressions for different head poses. Unlike most approaches which rely on evaluating features from the whole face, we split the face into two regions, calculate the features and classify the emotion in each region of the face independently. We then proposed an angle dependent fusion based method to combine the decisions of the classifiers. The outcome of this method is that the parameters of the decision fusion can be predicted given the angle using a simple regression. We show that the proposed decision fusion outperforms previous approaches. The main contribution of this

work is in the dynamic fusion approach. It should be noted that the classifiers used can be improved by using different features. Therefore, future improvements should consider other features and classifiers.

Chapter 6

Conclusion

In this thesis, two methods were proposed for FER. The first method is based on splitting the face to two regions and fusing the decisions using a logical OR operation. The second method optimizes the fusion weights by maximizing the posterior probability estimate with respect to the yaw angle. We showed that our first method compared well with state-of-the-art algorithms for frontal poses. Additionally, the method generalized for non-frontal head poses and different databases. In our second method, a 7.1% improvement in recognition rate was achieved over full facial features approach.

For future works, several improvement aspects can be considered. We note that the facial tracker plays an important role in the recognition of facial expressions. For instance, choosing a different tracker and applying it to our first method can improve the accuracy for certain AUs such as lip corner pullers. Additionally, the accuracy can further be improved by choosing more complicated features, geometrical or appearance based, such as SIFT-based descriptors. Aside from obtaining a higher classification performance overall, the posterior probability estimation can become more accurate, thus making the estimation of the left and right weights more accurate. Other important improvements can be made on the fusion model. The model can be extended to allow for other types of rotations (such as pitch rotations). Additionally, the regressions, which are applied to determine the left and right weights, can be improved by performing a fitting to non-linear models. One suggested model would be a Gaussian estimation. Consequently, the weights would be determined using expectation maximization. Finally, further experiments need to be conducted on bigger databases in order to compare our methods with other state-of-the-art approaches.

References

- [1] C. Breazeal and B. Scassellati, “Robots that imitate humans,” *Trends in cognitive sciences*, vol. 6, no. 11, pp. 481–487, 2002.
- [2] J. Medeiros, “Giving stephen hawking a voice,” *Wired.co.uk*, January 2015.
- [3] F. Dornaika and B. Raducanu, “Facial expression recognition for HCI applications,” in *Encyclopedia of Artificial Intelligence (3 Volumes)*, pp. 625–631, 2009.
- [4] J. Bullington, “‘affective’ computing and emotion recognition systems: the future of biometric surveillance?,” in *Proceedings of the 2nd annual conference on Information security curriculum development*, pp. 95–99, ACM, 2005.
- [5] S. Moore and R. Bowden, “Local binary patterns for multi-view facial expression recognition,” *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011.
- [6] Y. li Tian, T. Kanade, and J. F. Cohn, “Facial expression recognition,” in *Handbook of Face Recognition, 2nd Edition* (S. Z. Li and A. K. Jain, eds.), pp. 487–519, Springer, 2011.
- [7] P. Ekman and W. V. Friesen, “Facial action coding system,” 1977.
- [8] J. A. Russell, “A circumplex model of affect.,” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
- [9] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–511, IEEE, 2001.
- [10] T. Cootes and C. Taylor, “Active shape models—smart snakes,” in *Proc. British Machine Vision Conference*, vol. 266275, Citeseer, 1992.

- [11] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
- [12] M. Valstar and M. Pantic, “Fully automatic recognition of the temporal phases of facial actions,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 1, pp. 28–43, 2012.
- [13] Y. Tong, W. Liao, and Q. Ji, “Facial action unit recognition by exploiting their dynamic and semantic relationships,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, pp. 1683–1699, 2007.
- [14] Y. Tian, T. Kanade, and J. Cohn, “Recognizing action units for facial expression analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.
- [15] M. Pantic and I. Patras, “Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 2, pp. 433–449, 2006.
- [16] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, “Multi-view facial expression recognition,” in *FG*, pp. 1–6, 2008.
- [17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [18] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang, “A study of non-frontal-view facial expressions recognition,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, IEEE, 2008.
- [19] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, *et al.*, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.
- [20] O. Rudovic, I. Patras, and M. Pantic, “Coupled gaussian process regression for pose-invariant facial expression recognition,” in *Computer Vision–ECCV 2010*, pp. 350–363, Springer, 2010.
- [21] U. Tariq and T. S. Huang, “Features and fusion for expression recognition - a comparative analysis,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 146–152, IEEE, 2012.

- [22] C. Trad, H. Hajj, W. El-Hajj, and F. Al-Jamil, “Facial action unit and emotion recognition with head pose variations,” in *Advanced Data Mining and Applications*, pp. 383–394, Springer, 2012.
- [23] J. Saragih, S. Lucey, and J. Cohn, “Deformable model fitting by regularized landmark mean-shift,” *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.
- [24] I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for tracking facial features,” in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 97–102, IEEE, 2004.
- [25] H. Gu and Q. Ji, “Information extraction from image sequences of real-world facial expressions,” *Machine Vision and Applications*, vol. 16, no. 2, pp. 105–115, 2005.
- [26] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, “A high-resolution 3d dynamic facial expression database,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference On*, pp. 1–6, IEEE, 2008.
- [27] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [28] T.-F. Wu, C.-J. Lin, and R. C. Weng, “Probability estimates for multi-class classification by pairwise coupling,” *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.