2014-01-01

Multi-Dimensional Emotion Recognition From Geometry And Color Information

Geovany A. Ramirez
University of Texas at El Paso, garamirez@miners.utep.edu

MULTI-DIMENSIONAL EMOTION RECOGNITION FROM GEOMETRY AND COLOR INFORMATION

GEOVANY ABISAI RAMIREZ GARCIA

Department of Computer Science

APPROVED:

Olac Fuentes, Ph.D., Chair

Nigel Ward, Ph.D.

Stephen Crites, Ph.D.

Bess Sirmon-Taylor, Ph.D.,
Interim Dean of the Graduate School

to my

PARENTS

with love

MULTI-DIMENSIONAL EMOTION RECOGNITION FROM GEOMETRY AND COLOR INFORMATION

by

GEOVANY ABISAI RAMIREZ GARCIA

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2014

Abstract

Emotions play a fundamental role in everyday interactions among humans. Humans are adept at expressing themselves and interpreting others through a multi-modal, subtle and complex process using non-verbal cues including speech prosody, facial expression, eye gaze, body gestures, head motion, posture, and skin color changes. However, recognizing the affective state of humans is a difficult task for computers.

Automatic emotion recognition has focused on analysis of the six discrete basic emotions: happiness, sadness, surprise, fear, anger and disgust. However, humans express more complex and subtle affective states such as confusion, shame, pleasure, anxiety or depression. Therefore, a different representation based on a small number of continuous latent dimensions is more suitable for emotion recognition. In this dissertation I explored the problem of multi-dimensional emotion recognition. The first step for gathering visual features is face detection; I developed an approach for face detection based on a new set of Haar features. The approach also includes a method based on a genetic algorithm to reduce the training time. According to my experiments, these new set of Haar features can attain similar results to other methods while generating simpler classifiers with fewer Haar features.

I developed an emotion recognition approach based on a small set of high-level features instead of a large set of low-level features. First, I present a set of experiments with high-level features obtained from geometric information. These features include gaze direction, head tilt, smile level, and eyebrows interaction. The experimental results show an improvement over current approaches. Later, I used features from color information that measure changes in color for three face regions. My results show that facial skin color changes can be used to infer the emotional state of a person. Finally, I used a combination of features from geometric and color information that improved performance in emotion recognition.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Humans display affective behavior that is multi-modal, subtle and complex. People are adept at expressing themselves and interpreting others through the use of such non-verbal cues as speech prosody, facial expression, eye gaze, body gestures, head motion and posture, skin color changes and sweat. All of these modalities contain important affective information that can be used to infer the emotional state of a person (Zeng et al., 2009, Gunes and Pantic, 2010). Automated recognition and analysis of human emotions is an important part of the development of affect-sensitive artificial intelligent systems (Pantic and Rothkrantz, 2003).

Much work in automated emotion recognition (Zeng et al., 2009) has focused on analysis of the six discrete basic emotions (Ekman, 1992) (happiness, sadness, surprise, fear, anger and disgust). However in everyday interactions people exhibit non-basic and recognizable mental/affective states such as interest, boredom and confusion (Rozin and Cohen, 2003) and even more complex emotions including shame, pleasure, anxiety, and depression. Therefore, instead of using a single label (or multiple discrete labels from a small set), a different representation based on a small number of continuous latent dimensions is more suitable for emotion recognition, providing a continuous rather than a categorical view of emotions. Examples of such affective dimensions are valence (pleasant vs. unpleasant), activation (relaxed vs. aroused), power (sense of control), and expectancy (anticipation). Fontaine et al. (2007) argue that these four dimensions account for most of the distinctions among everyday emotion categories and hence form a good set to analyze.

Multiple modalities can be fused to improve the performance of emotion recognition, but the related complexity of fusing unsynchronized modalities has limited the number of

approaches working with multi-modal data. Thus, most work has concentrated on analyzing different modalities in isolation rather than looking for ways to fuse them. Additionally, there is a limited availability of suitably labeled multi-modal datasets. The optimal level at which the features should be fused is still an open research question (Zeng et al., 2009, Gunes and Pantic, 2010).

State of the art approaches for emotion recognition use an extensive set of visual or audio features, where each feature is a low-level representation of the appearance of the face or a low-level descriptor of the audio signal. The problem with this approach is that the feature space can be extremely large, for example, 5,900 dimensions of visual and 1,941 of audio features in the case of Schuller et al. (2011). This high dimensionality issue can be partially solved by performing dimensionality reduction or feature selection, but this usually does not improve the performance over the original set of features.

The goal of this dissertation is to explore the problem of automatic vision-based emotion recognition using geometric and color information to answer the following questions:

1. Could we improve the performance of automatic emotion recognition using a small set of high-level features from geometric and color instead of a large set of low-level features?

2. How can we model the temporal information included in the expression of emotions?

3. How does each high-level feature contribute to the performance of emotion recognition?

1.1 Models of Emotions

It is important to have a framework to model and describe emotions. As presented by Grandjean et al. (2008), there are three main models or approaches to describe emotions: (1) Basic or categorical model, (2) dimensional model, and (3) componential appraisal model.

1.1.1 Categorical Model

The categorical model is based on the interpretation of the theory presented by Darwin in his book "*The expression of the emotions in man and animals*" (Darwin, 1998). According to Tomkins, there are a limited number of emotions (Tomkins, 1962, 1963). The interpretation of Tomkins was supported by findings of Izard (1977) and Ekman (1992) concluding that there exist a fixed number of emotions that are universally recognized by humans. The basic emotions that are innate in humans and can be recognized universally are happiness, sadness, surprise, fear, anger, and disgust.

1.1.2 Dimensional Model

In the dimensional model, an emotion is a combination of affective states or feelings. Each affective state corresponds to a dimension. There are three basic dimensions: valence, arousal/activation and potency/power (Osgood et al., 1975). The valence dimension describes an emotion in the sense of how positive or negative it is. It can go from an unpleasant feeling to a pleasant feeling. The activation dimension describes the level of interest that a person shows. It can go from a lethargy state to a high level of excitement. The power dimension describes a feeling in terms of sense of control. It can go from a dominant to a submissive state. According to Russell (1977, 1980), it is possible to map the six basic emotions of the categorical model in the valence versus activation dimensional space as shown in Figure 1.1. Also, it is possible to add new labels to different locations in the activation-valence space that can better represent a person's emotional state.

1.1.3 Componential Appraisal Model

The componential appraisal model was introduced by Scherer (2001) based on the work of Arnold (1960) and Lazarus (1991). In this model, an emotion is the result of a dynamic process caused by the interaction between the internal state of a person and the outside world. An emotion, under this context, is the interaction of many different components

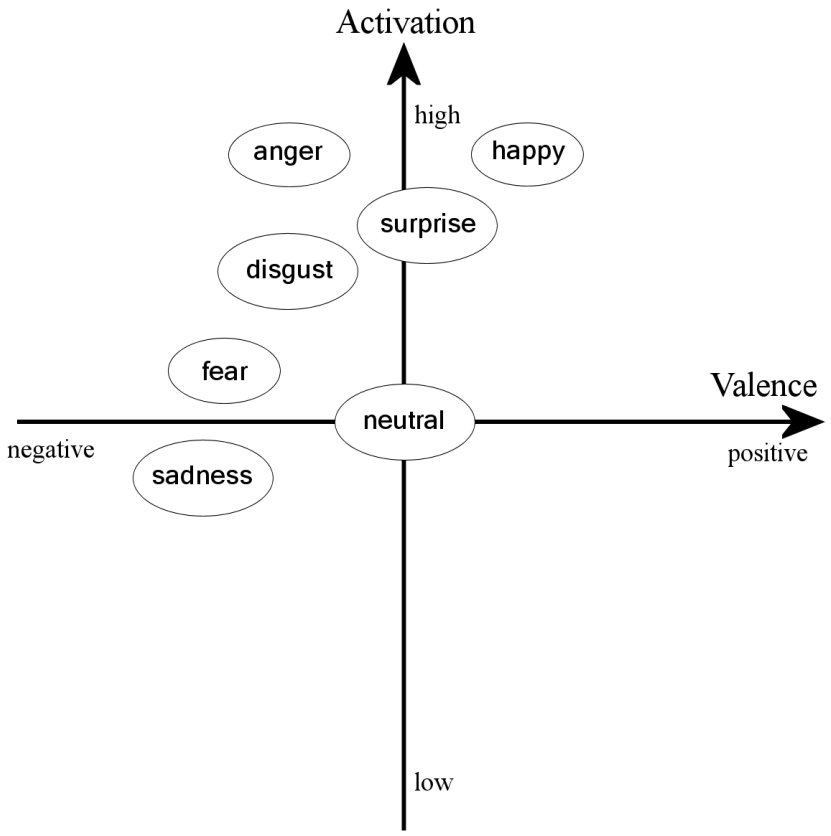


Figure 1.1: Emotion mapping of the six basic emotions in the categorical model in the activation vs. valence dimensional space.

such as cognition, motivation, physiological reaction, motor expression, and feeling that are changing continually. It involves the continuous appraisal and interpretation of the surrounding environment by the subject and the possible consequences that the subject could face. This model can be seen as an extension of the dimensional model that is not limited to a fixed number of dimensions. The componential appraisal model has a limited use in automatic emotion recognition due to the difficulties of measuring the complex and synchronized changes in all the components.

1.2 Automatic Emotion Recognition

For humans, emotion recognition is an effortless task. Humans are able to recognize emotions even in complex conditions and from a broad set of cues such as speech prosody, facial expressions, hand gestures, head movement, body postured, skin color changes and sweat. However, for computers, this is a difficult task. State of the art approaches are well below the human ability for this problem. In Figure 1.2 we can see some examples of the sensors used to gather affective signals from different modalities.

Automatic emotion recognition has attracted increasing attention not only in the computer vision and machine learning communities, but also in other areas such as psychology, cognitive science, neuroscience, behavioral science and linguistics. There are many potential applications (Picard, 1997), including human-computer interaction, realistic computer-generated humans, surveillance, evaluation of systems for customer service, monitoring the level of attention in education, measuring the level of fatigue and stress in working places, psychological therapy for treatment of posttraumatic stress disorder or autism spectrum disorder, and the study of human behavior.

The first attempts at automatic emotion recognition were performed on the six basic prototypical emotions with data that were obtained from posed, rather than spontaneous expressions. This creates successful approaches for emotion recognition under lab conditions, but yielded poor results under realistic and natural conditions. This created a limitation for practical applications because under natural conditions, humans show affective signals that are subtle and complex.

Researches have focused on analyzing different modalities in isolation (Zeng et al., 2009, Gunes and Pantic, 2010). This is partly due to the limited availability of suitably labeled multi-modal datasets and the difficulty of fusion itself. While there have been considerable advances in automatic emotion recognition from single modality approaches (Meng and Bianchi-Berthouze, 2011, Lajevardi and Wu, 2012, Scherer et al., 2013), there is still work to do by fusing modalities which is likely to improve the performance of emotion recognition

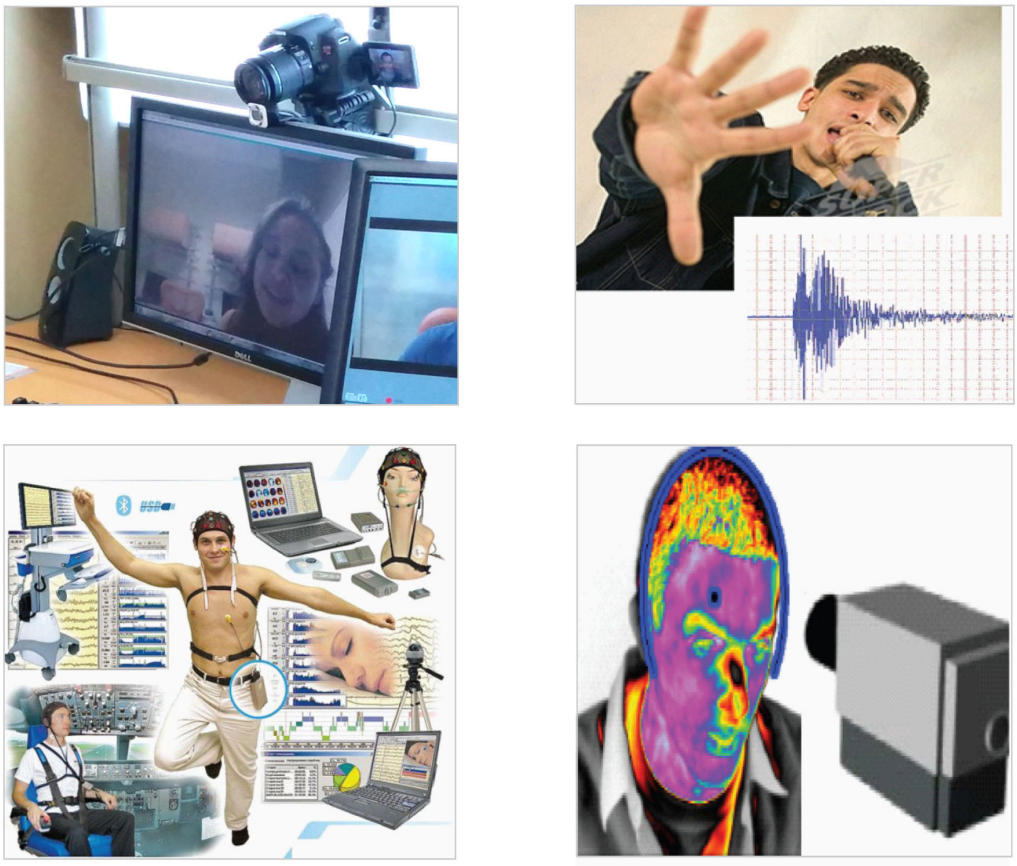
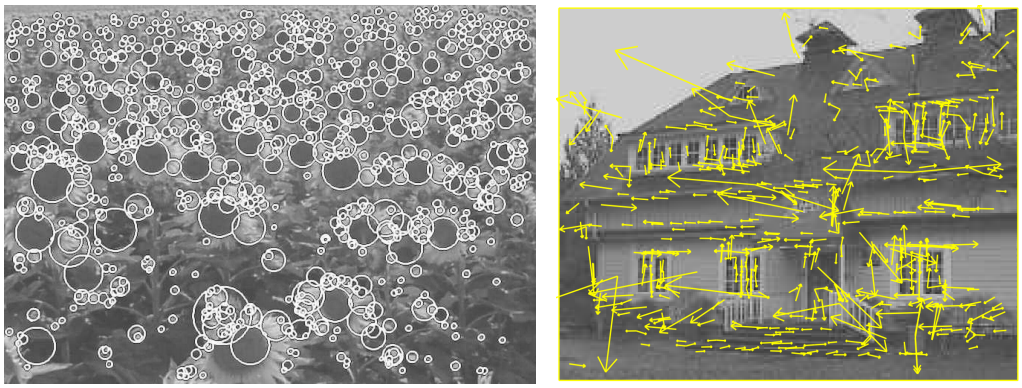


Figure 1.2: Some examples of sensors for gathering affective signal from different modalities. Images adapted from Gunes and Pantic (2010).

system.

1.2.1 High-Level Features

Object recognition has received a great deal of attention in the last few years. Visual emotion recognition can be seen as a sub-problem of object recognition where the face is the main object of attention. The goal in emotion recognition is to assign a label that represents the current emotional state of a person to a specific variation of the face appearance. In the general problem of object recognition, recent research contributions have focused mostly



(a) (b)

Figure 1.3: Example of Low-level features for object recognition. (a) Speeded-Up Robust Features (SURF) (Image extracted from Bay et al., 2008). (b) Scale-Invariant Feature Transform (SIFT) (Image extracted from Lowe, 2004).

on two main aspects of the problem: feature engineering and classifier design. Feature engineering consists of designing and selecting classes of features that can improve the performance of the classifiers that are based on them. Work on classifier design consists of adapting existing classification algorithms to the detection and recognition problems, or of designing special-purpose algorithms that are targeted to these problems.

The design of the features that can provide the best representation of the variability and characteristics of an object is still a open problem. Most of the work related to object recognition has been based on low-level features. Low-level features are obtained directly from the dataset and in most cases have little meaning for humans (Zhang and Chen, 2003). One of the problems of low-level features is the high density of features with minimal information. Visual descriptors such as Locally Binary Patters (LBP) (Ojala et al., 2002), Speeded-Up Robust Features (SURF) (Bay et al., 2008) or Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) have been used successfully for object recognition (see Figure 1.3). The problem with these approaches is that the resulting feature space is extremely large. This dimensionality issue can be partially solved by performing feature

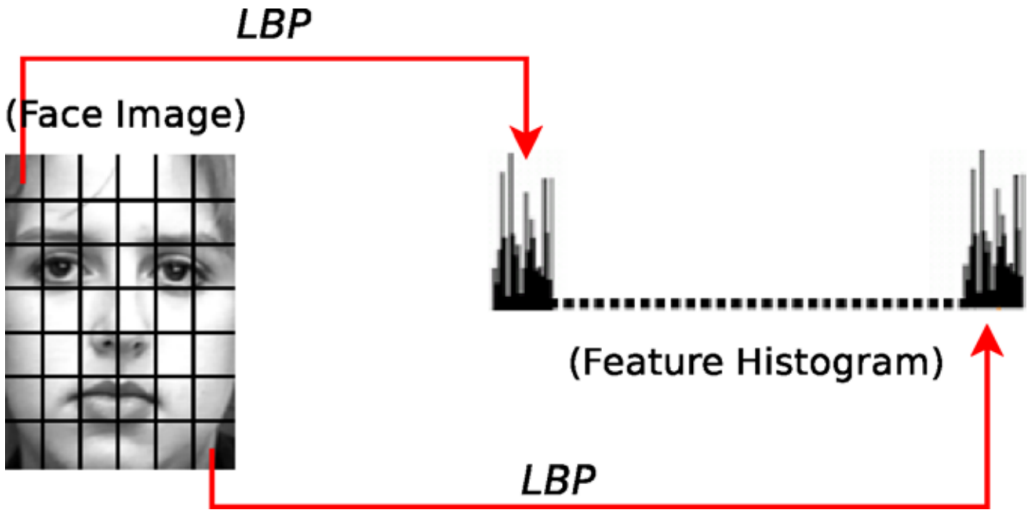


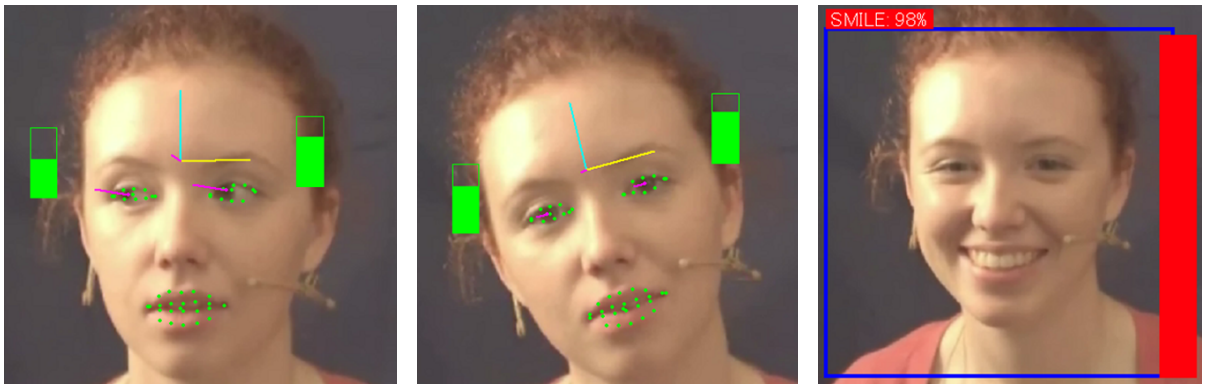
Figure 1.4: Locally Binary Patters (LBP) for facial expression. A LBP histogram is computed for each subregion of the face and concatenated in a vector. (Image extracted from Shan et al., 2009).

selection or by a statistical analysis, but in both cases this does not help the recognition problem. In Figure 1.4 we can see the feature extraction for a face image using LBP; the resulting feature vector contains 2,478 features.

Instead of using low-level features with low information content, we can use high-level features that can represent more information. High-level features or semantic features are features that can be evaluated in a similar way as humans do. Each high-level feature is measured in a similar way as humans perceive or recognize that feature. These high-level features include smile level, head tilt, gaze direction and skin color (see Figure 1.5). The main advantage is that few high-level features represent more information in a few values than many low-level features with several values (Zhang and Chen, 2003).

1.3 Contributions

The contributions of this dissertation are the following:



(a) (b) (c)

Figure 1.5: High-level features can be measured in a similar way as humans do. (a) Vertical and horizontal eye gaze. (b) Head tilt. (C) Smile intensity.

1. **Face Detection.** Face detection is an important first step for applications in several areas, including biometrics, human-computer interfaces, and surveillance. Face detection is a difficult task, due to factors such as varying size, orientation, poses, facial expression, occlusion, and lighting conditions. For vision-based emotion recognition, the face is the most important source of information and will be the first step for gathering visual features. I developed an approach for face detection with three contributions. The first contribution is the introduction of asymmetric Haar features. According to my experiments, this new set of Haar features can represent object appearance more accurately that previously used ones. The second contribution is a method based on a genetic algorithm to reduce the training time. This method allows to exploit the expressive advantage of asymmetric features in 0.02% of the time that would be required using the full feature set. The last contribution is the application of a skin color-segmentation scheme to reduce the search space (Ramirez and Fuentes, 2008).

2. **Emotion Recognition** State of the art systems for emotion recognition are based on a high density of low-level features. Each low-level feature can only contain a small

amount of information that results in a poor and redundant representation of affective information. For this dissertation, I developed an emotion recognition approach based on high-level features from geometric and color information. The first contribution is a study about different high-level features based on geometric information such as gaze direction, head tilt, smile level, and eyebrows motion. The results show that a small set of geometric high-level features can outperform previous methods based on low-level features. The second contribution is a novel set of high-level features based on skin color. These features were tested on a new dataset of human emotions carefully created to elicit spontaneous and natural emotions. The results show that skin color is a reliable feature for emotion recognition. The last contribution is the combination of geometric and color information for emotion recognition in a multi-dimensional space. According to our experimental results, the combination of geometric and color information can improve the recognition of the affective state in humans (Ramirez et al., 2011, 2014).

Chapter 2

Literature Review

2.1 Face Detection

Face detection is an important first step for applications in several areas, including human-computer interfaces, surveillance and facial component extraction. Face detection is also the first step for the extraction of visual features for emotion recognition. This literature review about face detection only covers works previous to my development of method presented in Chapter 3 in 2007.

Rowley et al. (1998a) developed a frontal face detection system that scanned every possible region and scale of an image using a window of 20×20 pixels. Each window is pre-processed to correct for varying lighting, then a retinally connected neural network is used to process the pixel intensity levels of each window to determine if it contains a face. In later work Rowley et al. (1998b) provided invariance to rotation perpendicular to the image plane by means of another neural network that determined the angle or rotation of a region; then the region was rotated by the computed angle and given to the original neural network for classification (see Figure 2.3).

Convolution neural networks, which are highly modular multi-layer feedforward neural networks that are invariant to certain transformations, were originally proposed by Bengio and Cun (1994) with the goal of performing handwritten character recognition. Later, they have been used for generic object recognition (Lecun et al., 1999) and they have also been shown to provide good results in face recognition (Lawrence et al., 1997, Garcia and Delakis, 2002).

Schneiderman and Kanade (2000) detected faces and cars from different points of view

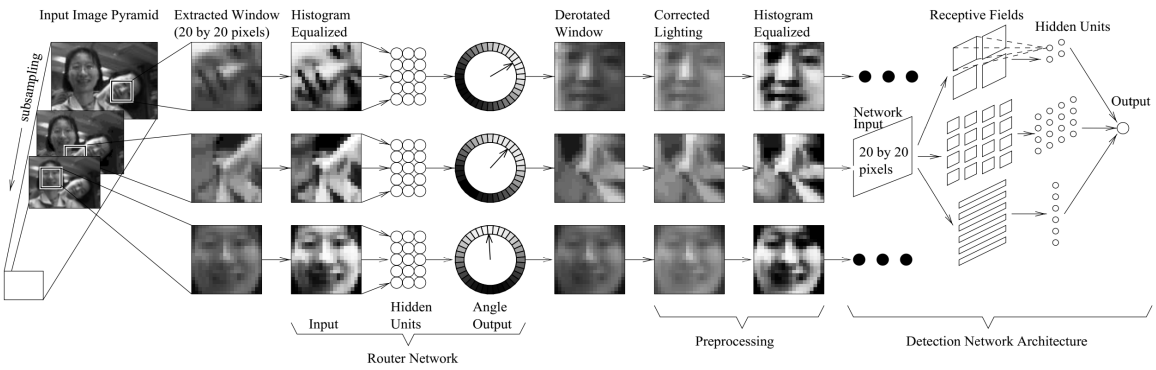
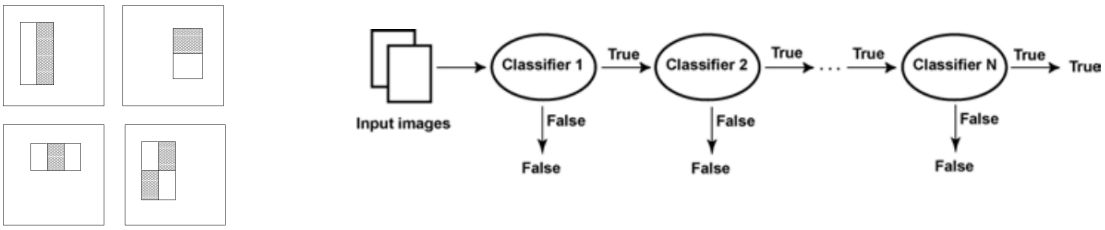


Figure 2.1: High level diagram of the face detection method based on artificial neural networks developed by Rowley et al. (1998b) (Image extracted from Rowley et al. (1998b)).

using specialized detectors. For faces, they used 3 specialized detectors in frontal, left profile and right profile, respectively. Each specialized detector is based on histograms that represent the wavelet coefficients and the position of the possible object, and then they used a statistical decision rule to eliminate false negatives.

Fröba and Ernst (2004) used the census transform as a feature for face detection. The census transform builds a bit-vector that functions as a descriptor of the 3-by-3 neighborhood of each pixel by comparing each pixel against the average intensity of the neighborhood. If the pixel has an intensity greater than the average, the bit value is one, otherwise it is zero. It is repeated for all the subregions of 3-by-3 pixels. Using these simple features and a cascade-style classifier, they obtained results that were comparable to the best systems presented to date.

Viola and Jones used Haar-like wavelets as features in their object detection systems (Viola and Jones, 2001, Tieu and Viola, 2004, Viola et al., 2005). A Haar feature is the difference between the sum of pixels in two or more adjacent regions (see Figure 2.2a). For classification, they used a modified version of the Adaboost algorithm, an ensemble method originally presented by Freund and Schapire (1996), that has proven to yield excellent results in several learning domains. Adaboost builds a sequence of classifiers based



(a) (b)

Figure 2.2: (a) Basic set of Haar features used by Viola et al. (2005). (b) Input images are processed by a cascade of classifiers, the first classifiers use few features that are evaluated quickly and the last classifiers use more features which results in a slow evaluation. Images extracted from Viola and Jones (2001) and Viola et al. (2005), respectively.

of a set of basic Haar features. The ith classifier in the sequence is biased to correct the misclassifications of classifiers $1, ..., i-1$. The final classification of Adaboost is given by the weighted average of the classification made by each of the individual classifiers (see Figure 2.2b). The fact that classifiers that appear later in the sequence attempt to correct the mistakes made by earlier ones results in accuracy levels that are superior to most other algorithms, particularly in situations where noise levels are moderate. The advantage of a cascade of classifiers is the efficiency for processing a large amount of possible face images. This is possible because the classifiers at the beginning of the cascade use fewer features than the classifiers at the end increasing significatively the classification speed.

Xiao et al. (2004) present a face detector that consists of three stages. The first stage is simple and is based on Haar features and a linear support vector machine classifier. The second stage is based on a cascade of classifiers, and each individual classifier is also based on Haar features. In the last stage they apply a color skin filter to eliminate false positives.

Li and Zhang (2004) present another approach to detect frontal and profile faces using Haar features and the FloatBoost algorithm. Their system can detect faces with ± 45 degree rotation perpendicular to the image plane using 10 detectors specialized for different face poses. In (Li et al., 2004), a set of specialized support vector machines was trained to detect

faces at specific angles in the view sphere, and, when an image needed to be classified, another support vector machine detected the pose and chose the appropriate specialized detector to use.

Wu et al. (2004) detect frontal and profile faces with arbitrary in-plane rotation and up to 90-degree out-of-plane rotation. They used Haar features and a look-up table to develop strong classifiers. To create a cascade of strong classifiers, they used Real AdaBoost, an extension to the conventional AdaBoost . They built a specialized detector for each of 60 different face poses. To simplify the training process, they took advantage of the fact that Haar features can be efficiently rotated by 90 degrees or reversed, thus they only needed to train 8 detectors, while the other 52 can be obtained by rotating or inverting the Haar features.

2.2 Emotion Recognition

In this section, different approaches to the problem of emotion recognition are reviewed and organized in four categories: audio-based, video-based using geometric information, video-based using color information, and combine audio and vision features. For an extended review of emotion recognition see recent surveys of dimensional and categorical affect recognition by Zeng et al. (2009), and by Gunes and Pantic (2010).

2.2.1 Audio-Based

Lee and Narayanan explore emotion recognition from spoken dialogs (Lee and Narayanan, 2005). They classify emotion as negative or non-negative based on three aspects of spoken language information: acoustic, lexical, and discourse. Their dataset consists of dialogs between a human user and a machine agent over the telephone. First, a classifier for each source of information is created independently. After that, the final decision is computed by fusing the output of the three classifiers using a different classifier. The experiments were performed using linear discriminant classifiers (LDC) and k-nearest neighborhood classifier

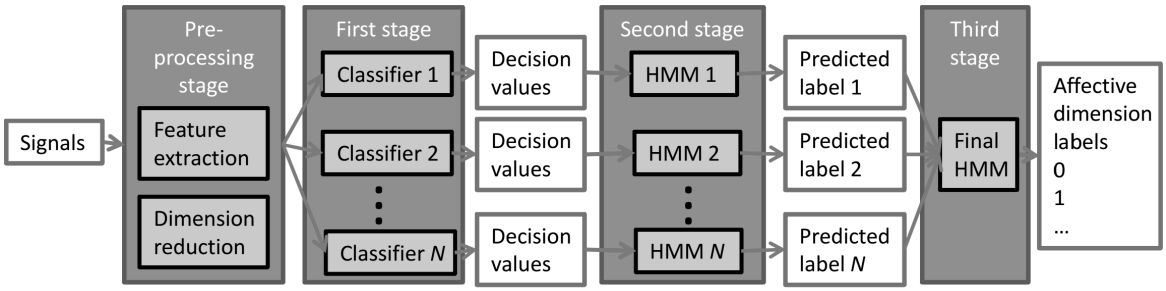


Figure 2.3: Multi-stage emotion recognition system for audio signals proposed by Meng and Bianchi-Berthouze (2011). The first stage uses KNN classifiers, second and third stages use HMM classifiers (Image extracted from Meng and Bianchi-Berthouze (2011)).

(KNN). Their results show a improvement of 39% in average fusing multiple sources in comparison to using a single source.

Wöllmer et al. (2008) use Conditional Random Fields (CRF) for discrete emotion recognition by quantising the continuous labels for valance and arousal based on a selection of acoustic features. In addition, they use Long Short-Term Memory Recurrent Neural Networks to perform regression analysis on these two dimensions. Both of these approaches demonstrate the benefits of including temporal information when approaching emotion recognition in dimensional space.

Meng and Bianchi-Berthouze (2011) propose a multi-stage emotion recognition system on audio signals. They exploit the strong temporal relationship of consecutive observations using Hidden Markov Models (HMMs). Their system is divided into three stages. The first one consists of a set of 13 classifiers based on KNN that do not use the temporal information. The second stage takes the output of the classifiers in the first stage as the input to the same number of HMM classifiers. The last stage fuses the output of the second stage to predict a label. To test their system they used the AVEC dataset (Schuller et al., 2011) to classify audio signals in four emotional dimensions with binary labels. They obtained an accuracy of 53.27% on average.

Arias et al. (2013) detect emotional modulation by modeling the shape of low level descriptors such as energy and F0 contours. Their method is based on generating neutral references from emotion neutral utterances from several speakers. After extracting the energy and F0 contour descriptors, they apply functional principal component analysis (functional PCA) to generate a set of orthogonal basis functions of principal components. The testing speech is aligned using dynamic time warping (DTW). The energy and F0 contour are extracted and projected onto the neutral reference model of principal components. The differences between the model and the testing speech are used to discriminate between neutral and emotional speech. Their approach outperformed previous results with the SEMAINE dataset (McKeown et al., 2010).

2.2.2 Vision-Based Using Geometric Information

Valstar et al. (2007) present a study of how the face, head, and shoulders are important to distinguish between a posed or spontaneous smile. The main analysis was based on tracking of different modalities. For head tracking they used a cylindrical head tracker with six degrees of freedom. For face tracking they defined a set of eight points of interest around the eyes and four points of interest around the mouth. Later they used Particle Filtering with Factorized Likelihoods (PFFL) to track the twelve points of interest. For tracking the shoulders they used two points of interest per shoulder and one stable point for the torso. For the shoulder tracker they used Auxiliary Particle Filtering (APF). They experimented with different levels for multimodal merging using 2 boosting algorithms: GentleSVM and GentleBoost. They created a set of posed smiles by asking subjects to smile. The set of spontaneous smiles was created by showing to subjects cartoons and nauseating videos. Their results show that fusing features from head pose, face, and shoulder increase the accuracy to discriminate between posed and spontaneous smiles. In addition, they found the head pose as the most relevant feature to associate with a spontaneous smile. Their method was able to recognize between posed and spontaneous smiles with 94.0% accuracy.

Figure 2.4: Tracked points used by Nicolaou et al. (2011) as visual features. (Image extracted from Nicolaou et al. (2011)).

Caridakis et al. (2008) explore the fusion of hand gestures and facial expressions for classification of affect in the dimensional space. They built a classifier based on an adaptive neural network to classify an emotional video segment as lying in one of the quadrants of activation and valance space.

Zhou et al. (2010) presented a method called Aligned Cluster Algorithm for automatic detection of facial events. Instead of using a predefined label scheme such as the Facial Action Coding System (FACS) (Ekman and Friesen, 1977), they used an unsupervised learning approach directly on the input video. Their set of features was divided into two parts, one corresponding to the face geometry and the other to the face appearance. To extract the face geometry features they used a face model to track the location of the face components. After that, they used the landmarks of the mouth and eyebrows to create appearance features using scale-invariant feature transform (SIFT) descriptors. Their results were comparable to methods based on FACS.

Nicolaou et al. (2011) propose the use of Output-Associative Relevance Vector Machine (OA-RVM) for dimensional and continuous prediction of emotions based on automatically tracked facial points. The feature set consisted of 20 points corresponding to the corners of the eyebrows, eyes, nose, mouth and chin (see Figure 2.4). Their work poses the dimensional labeling problem as regression rather than classification. Their proposed regression framework exploits the inter-correlation between the valence and arousal dimensions by including in their model the initial output estimation together with their input features. In addition, OA-RVM regression attempts to capture the temporal dynamics of the output by employing a window that covers a set of past and future outputs.

Scherer et al. (2013) use Automatic nonverbal descriptors to identify indicators of psychological disorders such as depression, anxiety, and post-traumatic stress disorder. They created a dataset called Distress Assessment Interview Corpus (DAIC) composed of 167 dyadic interactions between a confederate interviewer and a paid participant. The behavior descriptors they analyzed were vertical head gaze, vertical eye gaze, smile intensity, and some cues as hands and legs fidgeting. They found statistically significant differences in the intensity of smile in persons with and without psychological disorders, also an increased overall downwards angle of the gaze.

2.2.3 Vision-Based Using Color Information

Yamada and Watanabe (2004, 2005, 2007) examined discrete emotions (fear, happiness, and anger) based on skin color changes and temperature, using a video camera and a thermal camera. Their subjects were female students between the ages of 18 and 21. Those students were asked to watch some movie trailers as stimuli in a dark room under constant light and with their head being held in place. Initial skin color was used as the baseline to be compared to a small area on the left cheek. The findings revealed that subjects showed skin color changes during the experiments. Later, they synthesized the skin color changes into images and showed these images to 10 evaluators who found that the color changes made the human expression richer.

Nagaraj et al. (2010) used a visible and infrared camera (hyperspectral) to record subjects while placed in psychologically stressful situations. Their results showed that hyperspectral imaging may potentially serve as a non-invasive tool to detect if a subject is under stress. Their experiments suggest a high correlation between the intensity of the near-infrared band and the stress level.

Poh et al. (2010, 2011) reported experiments to measure the heart rate remotely in a non-invasive way. They used a basic webcam to record the videos of 12 participants (10 males and 2 females) between the ages of 18 and 31 years with varying skin colors. Their experiments were conducted indoors and with a varying amount of sunlight. Their subjects were seated in front of a computer while they were recorded by the webcam. Two kinds of videos of one minute each were recorded for each participant. During the first video, the participants had to sit still and stare at the webcam; for the second video recording, participants could move naturally as if they were interacting with the computer, but avoiding rapid motions. For their experiments, they considered the facial regions of each participant and used independent Component Analysis (ICA) over the raw RGB channels of the face region (see Figure 2.5). They also used band-pass filter based on the Fourier transform to remove artifacts. Their results showed that they were able to measure heart rate using a simple RGB camera with a RMSE of 2.29 bpm.

Lajevardi and Wu (2012), and Sai Pavan and Rajeswari (2013) use a novel technique for facial expression recognition called tensor perceptual color framework (TPCF). This method is based on information contained in color facial images and used for accentuating the facial expressions. A tensor is considered as a higher order generalization of a vector. The TPCF allows multilinear image analysis in different color spaces. The color images represented in different color spaces are unfolded to obtain 2D tensors which are used for feature extraction and classification. Their results demonstrate that color components provide additional information for robust facial expression recognition.

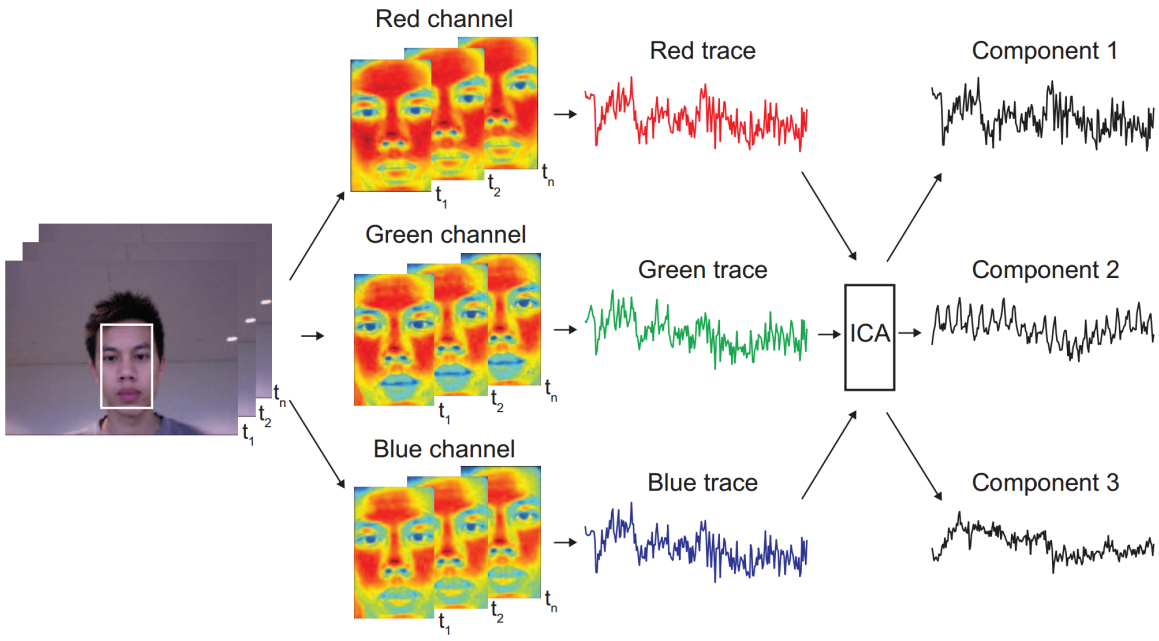


Figure 2.5: High level diagram of the system of Poh et al. (2010) to measure heart rate. Multiple raw traces t_x per channel are shown in false color (Image extracted from Poh et al. (2010)).

2.2.4 Combine Audio and Vision Features

Nicolaou et al. (2010) present experiments for classification of spontaneous affect based on audiovisual features using coupled Hidden Markov Models (cHMM) that allow to model temporal correlations between different cues and modalities. They also show the benefits of using the likelihoods produced from separate cHMMs as input to another classifier, rather than picking the label with a maximum likelihood for audiovisual classification of affective data. Interestingly, their experiments show that visual features contribute more than audio features in spontaneous affect classification in the valence dimension. For audio they used 15 features corresponding to Mel-frequency Cepstrum Coefficients and prosody features, while for video a total of 25 points corresponding to eyebrows, eyes, nose, mouth, shoulders and torso were used.

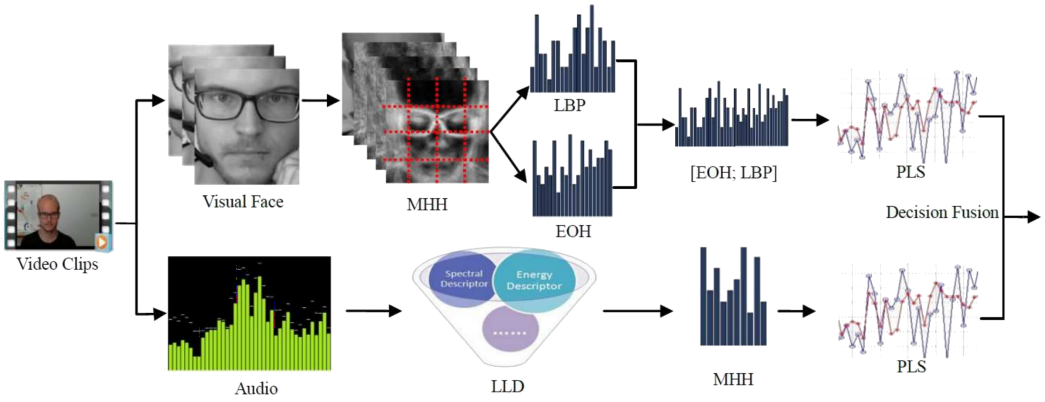


Figure 2.6: High level diagram of the system of Meng et al. (2013) for depression recognition (Image extracted from Meng et al. (2013)).

Eyben et al. (2011) fuse both visual (head motion, facial action units) and audio modalities in order to analyse human affect in valence and expectation dimensions. Their results show improved performance when using high-level event-based features such as smiles, head shakes or laughter rather than low-level signal-based ones such as facial feature points or spectral information when predicting affect from audiovisual data in valence and expectation dimensions.

Meng et al. (2013) present a method for modeling the vocal and visual modalities for depression recognition. They processed modalities individually but fuse the prediction in the last stage (see Figure 2.6). For the visual modality, they used Motion History Histograms (MHH) to capture the motion of each pixel on the face. Later they used Edge Orientation Histograms (EOH) and Locally Binary Patterns (LBP) to highlight the MHH. Them they use a Partial Least Square (PLS) regression to determine the level of depression. For audio modality, they used a set of Low-level descriptors (LLD) and also MHH and PLS. The fusion is based on the linear opinion pool method using weighted sum rule.

Chapter 3

Face Detection

3.1 Introduction

The face detector is the first step for the extraction of visual features. The problem of finding faces on arbitrary images is more difficult than other object detection problems due to the variability of face appearance. I developed a novel face detector method with following innovations. First I introduced the use of asymmetric Haar features that provide a rich feature space, which allows to build classifiers that are accurate and much simpler than those obtained with basic Haar features. The second innovation is the use of a genetic algorithm to search efficiently in the extremely large parameter space of potential features. The third innovation is the application of a skin color-segmentation scheme to reduce the search space.

In a series of papers, Viola and co-workers advocated an approach to object detection based on Haar features, which are equivalent to the difference of the sum of the intensity levels of two o more contiguous equal-sized rectangular image regions. They presented an algorithm for computing these features in constant time, which makes them suitable for real-time detection. Using Haar features, a cascade of classifiers based on the Adaboost algorithm was constructed, yielding accurate classification, albeit at the expense of long training times. Successful applications of this methodology were presented in face detection (Viola and Jones, 2001), image retrieval (Tieu and Viola, 2004), and pedestrian detection (Viola et al., 2005).

I propose three extensions to the work by Viola and co-workers. First, I introduce the use of *asymmetric Haar features*, eliminating the requirement of equal-sized positive and

negative regions in a feature. The Haar features with asymmetric regions can have regions with either different width or height, but not both. This results in a more expressive feature space, which, as shown later, allows to build classifiers that are much simpler than those obtained with the standard features. While the number of symmetric Haar features is large (around 180,000 in Viola's work), it is still feasible to perform exhaustive evaluation of these features in order to build a classifier. On the other hand, using asymmetric features, the number of potential features grows to over 200 million for a 24×24 pixel window, which makes exhaustive evaluation impossible. The second contribution of this work is the use of a genetic algorithm to search efficiently in the parameter space of potential features. Using this genetic algorithm, it is possible to generate a small feature set that allows exploitation of expressive advantage of asymmetric features. The third contribution is the application of a skin-color segmentation scheme to reduce the search space. The color segmentation can speed-up the classification process by eliminating from consideration all windows that do not contain regions that are similar to skin.

I present experimental results showing the application of this method to two sets that have been used commonly in the literature, the CMU profile test set, and BioID frontal test set. The experiments show that this method can attain similar results to other methods while generating simpler classifiers with fewer Haar features.

3.2 Haar Features

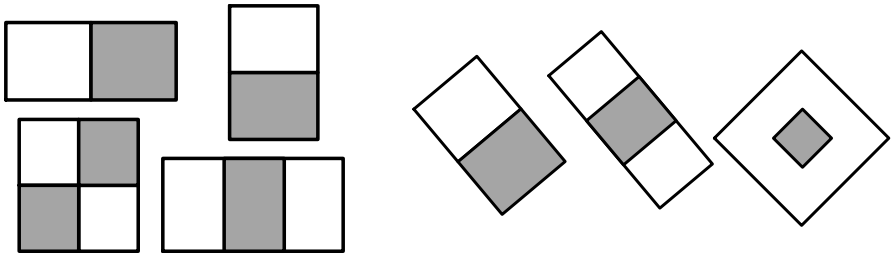
Haar features are based on Haar wavelets, which are functions that consist of a brief positive impulse followed of a brief negative impulse. In image processing, a Haar feature is the difference between the sum of all pixels in two or more adjacent regions. Papageorgiou et al. (1998) were the first to use Haar features for face detection. They used three types of Haar features of size 2×2 and 4×4 pixels, for a total of 1,734 different features in a 19×19 face image. Viola and Jones (2001) proposed the basic set of four types of Haar features that are shown in Figure 3.1a. The value of Haar feature is given by the sum of intensities of the

pixels in the light region minus the sum of intensities in the dark region. Using all possible sizes, they generate around 180,000 features for a 24×24 pixel image. Lienhart and Maydt (2002) presented an extension to the basic set with rotated Haar features as sown in Figure 3.1b. Using a straightforward implementation, the time required to perform the sum of pixels increases linearly with the number of pixels. Viola and Jones (2001) proposed to use the integral image as preprocessing to compute the sum of regions of any size in constant time. Each element of the integral image contains the sum of pixels in the original image that are above and to the left of that pixel; using this idea allows to compute a two-region Haar feature using only six memory accesses and a three-region Haar feature with only eight.

I propose an extension for basic Haar features, which I call *asymmetric Haar features*, and are shown in Figure 3.2. In contrast with basic Haar features, these new features can have regions with different width or height, but not both. It will be shown that these features are able to capture defining characteristics of objects with fewer Haar features than traditional ones, allowing the development of simpler and faster classifiers. By allowing asymmetry, the number of possible configurations for Haar features grows exponentially with respect to the number of parameters and is an overcomplete set. For the 6 Haar features shown in Figure 3.2, there are around 200 million possible configurations for a 24×24 image. Using all the possible configurations is unfeasible, therefore, to deal with this limitation I propose to use a Genetic Algorithm to select a subset of features. Details will be presented in the next section.

3.3 Selecting Haar Features with a Genetic Algorithm

Training a specialized detector using all the possible configurations of the asymmetric region Haar features would be impractical. For instance, using an initial training set of 3,000 examples and all the possible Haar configurations, we would need about 4 months on a 2.5GHz Intel Xeon computer for only one specialized detector. Therefore, to reduce the



(a) (b)

Figure 3.1: (a) Haar features introduced by Viola and Jones (2001). (b) Extension to the basic set proposed by Lienhart and Maydt (2002).

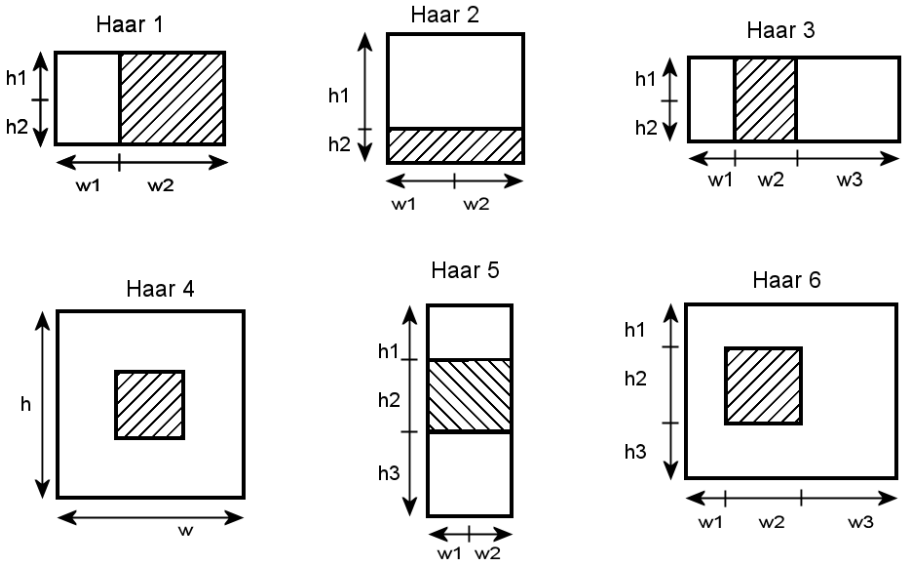


Figure 3.2: Asymmetric Haar features used.

Table 3.1: Individuals parameters for each type of Haar feature presented in Figure 3.2.

	Parameters
Haar 1	$\{h_1, h_2, w_1, w_2\}$
Haar 2	$\{h_1, h_2, w_1, w_2\}$
Haar 3	$\{h_1, h_2, w_1, w_2, w_3\}$
Haar 4	$\{h_1, w_1\}$
Haar 5	$\{h_1, h_2, h_3, w_1, w_2\}$
Haar 6	$\{h_1, h_2, h_3, w_1, w_2, w_3\}$

number of possible Haar features, I use a Genetic Algorithm (GA) to select the width and the height of a feature (Mitchell, 1998). In the GA, one individual is a weak classifier (WC) of faces based on only one Haar feature and is trained with the C4.5 algorithm (Quinlan, 1993). An individual represents the Haar parameters, as shown in Table 3.1. The parameters are respect the feature location. The fitness of each individual corresponds to the classification accuracy on an initial training set. The GA computes a WC for each place on the image and for each type of feature, for a total of 2,431 Haar features (see Table 3.2). Instead of using a binary representation of an individual (WC parameter), as it is suggested for optimization problems using GA (Mitchell, 1998), I opted for using a decimal representation that avoids the creation of invalid individuals. Also, I used a two point crossover with a deterministic crossover model presented in Kuri-Morales (2003). This model consists of combining the best and the worst individuals in the population, then the second best with the second worst, and so on. The best result was obtained with a 10% mutation rate. Using the GA to select a subset of Haar features, it is possible to reduce training time to 6 hours on our 2.5GHz Intel Xeon computer; this corresponds to a 99.8% reduction in time.

Table 3.2: Number of WCs for each location in a image of 24 × 24 pixels.

Haar	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
X **range**	[2, 20]	[1,21]	[3,19]	[3,19]	[1,21]	[2,21]	
Y **range**	[1,23]	[3,23]	[1,23]	[3,21]	[3,21]	[2,23]	
WCs	437	441	391	323	399	440	2431

3.4 Specialized Detectors

I use specialized detectors in frontal, left profile and right profile poses. In addition, I use 12 specialized detectors, one every 30 degrees to cover the 360 degree in-plane rotation of each pose. Therefore, there is a total of 36 specialized detectors for all the face poses. Each specialized detector consists of a cascade of strong classifiers created with the AdaBoost algorithm used by Viola and Jones (2001). The weak classifiers are based on the C4.5 rule induction algorithm Quinlan (1993) that is associated with only one Haar feature. Since Haar features can be rotated 90 degrees and horizontally inverted (mirror operator), it is possible to build a specialized detector by rotating 90 degree or by inverting the Haar features of an already trained detector (Wu et al., 2004). As shown in Figure 3.3, it is only required 2 specialized detectors in 0 and 30 degrees to build the remaining 9 detectors by applying the mirror or rotation operator. In the case of specialized detectors in profile faces, it is only required 3 specialized detectors in 0, 30, and 330 degrees as shown in Figure 3.4.

3.5 Training a Specialized Detector

A specialized detector is a cascade of strong classifiers (SC). To create a cascade of SC I used a variation of the algorithm presented by Wu et al. (2004). My algorithm is presented in Figure 3.5. After obtaining a subset of weak classifiers (WC) with the genetic algorithm,

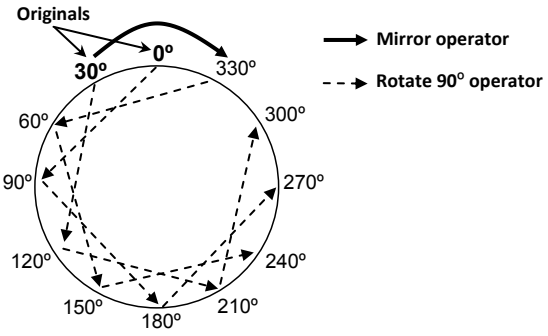


Figure 3.3: Specialized detector for frontal faces.

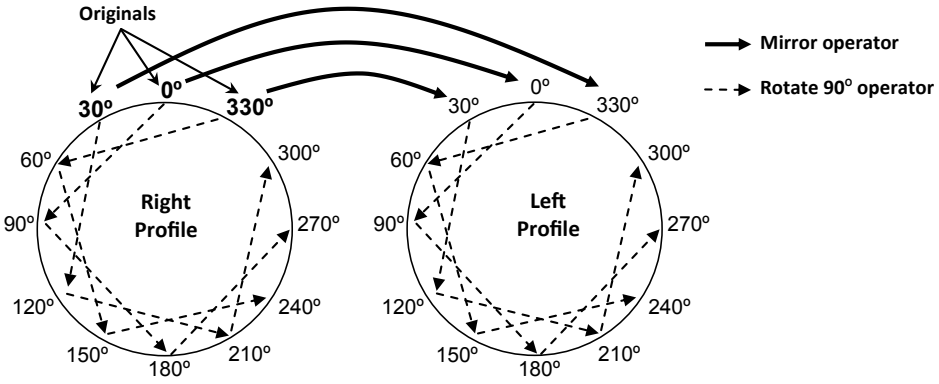


Figure 3.4: Specialized detectors for profile faces.

new examples are added to the training set and then the AdaBoost algorithm is used to create a new SC. AdaBoost ends when the minimum face detection rate and the maximum false positive rate per layer in the cascade are attained. If the target false positive rate is achieved, the algorithm ends. Otherwise all examples of non-faces correctly classified are eliminated and the training set is balanced adding non-face examples using a bootstrapping technique. With the training set updated, all the WCs are retrained and then used in AdaBoost. When the algorithm of Figure 3.5 ends, we will have a specialized detector with a face detection rate D and a false positive rate F. D and F can be computed for a specific cascade of SCs using equations 3.1 and 3.2, where d is the minimum face detection rate per

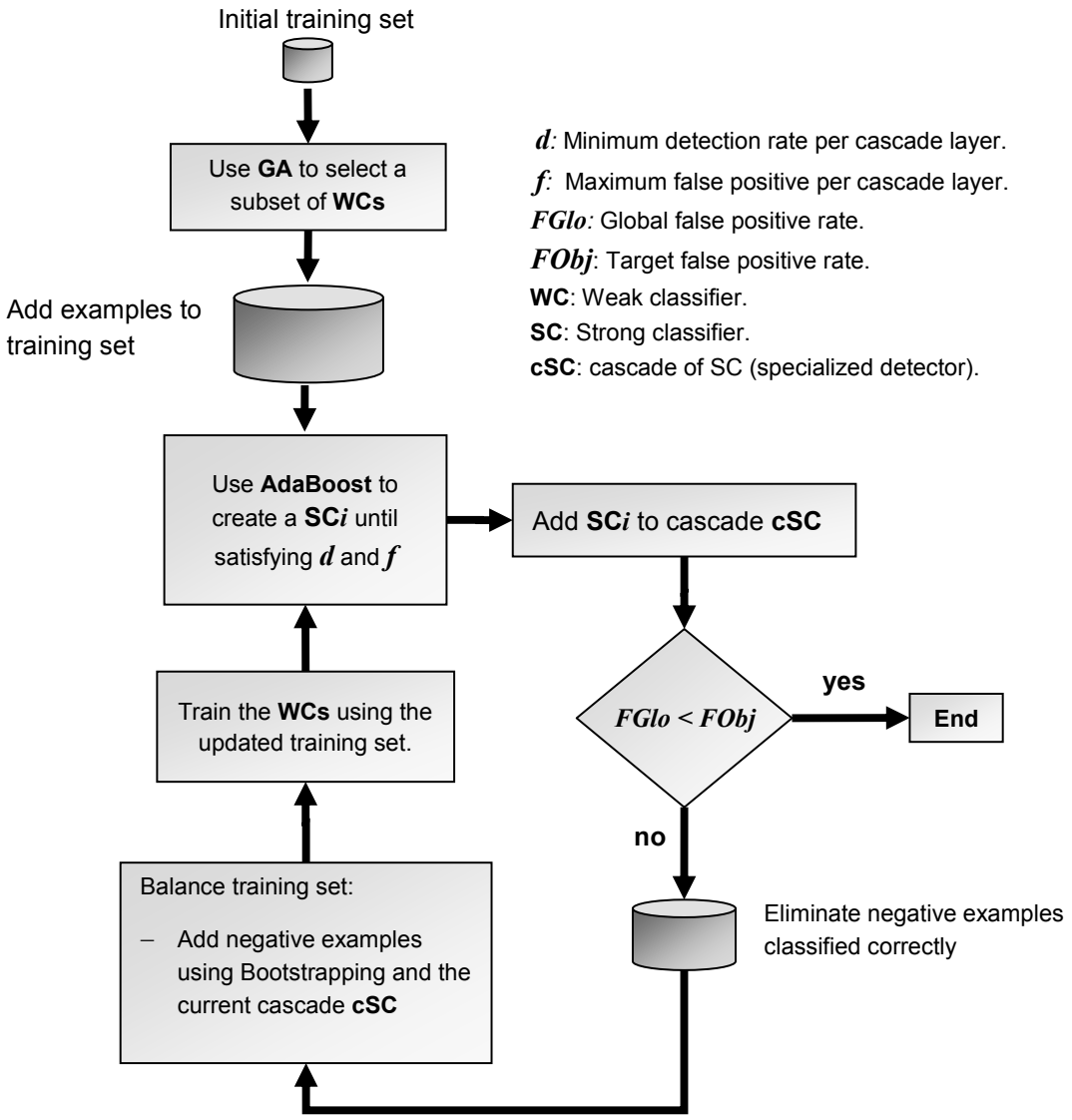


Figure 3.5: Algorithm to train a specialized detector.

layer and f is the maximum false positive rate per layer. For example, if we have $d = 0.999$ and $f = 0.35$ for a cascade of 8 layers, we will have a D of at least 0.992 and an F at most of at 2.25×10^{-4}.

$$D = \prod_{i=1}^{L} d_i \tag{3.1}$$

$$F = \prod_{i=1}^{L} f_i \tag{3.2}$$

3.6 Skin Color Segmentation

Skin color segmentation is a technique that has previously been used for face detection (Kovac et al., 2003, Chai and Bouzerdoum, 2000, Valaparla and Asari, 2003). For my approach, the skin color segmentation is used only to reduce the search space in color images. The skin color segmentation is based on the $YCbCr$ color space. I manually segmented a set of 33 images with 81 persons of different skin tones under different lighting conditions to define a rule to classify every pixel as skin or non-skin. Assuming that skin color has a normal distribution, a pixel is classified as belonging to skin if both its Cb and Cr components are within two standard deviations of the mean values for skin pixels found in the training images. This is illustrated in Equation 3.3, where $\bar{y}$ is the mean and σ is the standard deviation for a given channel. The resulting rule is shown in Equation 3.4.

$$\bar{y} - 2\sigma \leq \mathrm{C} \leq \bar{y} + 2\sigma \tag{3.3}$$

$$img_{x,y} = \begin{cases} 1, \quad if \quad (i) & 80 \leq \mathrm{Cb_{x,y}} \leq 140 \\ \qquad\quad (ii) & 135 \leq \mathrm{Cr_{x,y}} \leq 170 \\ 0, \qquad otherwise & \end{cases} \tag{3.4}$$

The skin color segmentation method was tested using 271 color images under different lighting conditions and with persons with different skin tones. The method can eliminate about 65% of all the pixels in the images, preserving the face region. The skin color segmentation method does not eliminate all the regions that do not contain skin, but it

(a) (b)

Figure 3.6: (a) Original color image. (b) Resulting image after skin color segmentation using Equation 3.4.

reduces the search space by 65% with only 4 comparisons per pixel. In Figure 3.6 we can see the result of applying the skin color segmentation method.

3.7 Results

For frontal detectors I used the same training set to select a subset of Haar features and to train the specialized detectors. For profile detectors, I used an initial training set to select a subset of Haar features and an extended training set to create the specialized detectors. The training sets were created with a careful selection of images that represents, as much as possible, the variation of faces. It includes faces of males and females, with different ages, of different races, with and without structural components such as glasses and beard, and different lighting conditions. Each face image was manually normalized to the specific degrees of rotation in the plane for each specialized detector. To increase the size of the training set, I mirrored each image. Since each specialized detector will be trained to cover a range of 30 degrees, I generate variations of each image rotating it in the range of ± 15 degrees. Non-face images were created by randomly selecting regions in images without faces. Table 3.3 shows the number of training images used for detectors in frontal and profile poses.

Table 3.3: Number of images used for frontal and profile detectors.

Pose	**Original images**	**Variations** (initial/extended)	**Total faces**	**Non-faces**	**Total**
Frontal	593	19	23,720	23,720	47,740
Profile	458	8/17	4,122/8,244	4,693/16,227	8,815/24,471

A subset from the CMU profile test set was used to test the profile detectors (Schneiderman and Kanade, 2000). This subset consists of 137 images with 214 profile faces with a rotation angle of ±30 degrees. I performed two experiments. The first experiment consisted of using two specialized detectors to cover only the left and right profile in a range of ±15 degrees in the image plane. The second experiment consisted of using six specialized detectors to cover the left and right profile in a range of ±45 degrees. As shown in Table 3.4, more specialized detectors can increase the detection rate by 10 percentage points, however the number of false positives is also increased in a similar proportion as number of classifiers used, in this case 3 times more. In comparison with other works (see Table 3.5), my approach with 6 specialized detector has a very similar performance than the approach of Wu et al. (2004).

To test the frontal detectors I used the BioID test set that consist of 1,521 grayscale images of 384×288 pixels, with a frontal view of 23 different persons under a high variety of lighting conditions, backgrounds, and face sizes (Jesorsky et al., 2001). The results of tests with different parameters are shown in Table 3.6. A higher detection rate implies a higher number of false positive.

Table 3.7 presents a comparison of my results with other works that used the BioID dataset. My method has a higher detection rate than the works of Jesorsky et al. (2001), Kirchberg et al. (2002), and Hamouz et al. (2004), this is a unfair comparison since they do not report the number of false positives. In comparison with the work of Fröba and Ernst (2004), my method has a lower detection rate and also more false positives.

Table 3.4: Results for the CMU test set.

	Test 1. 2 specialized detector for left and right profile in ±15°.	**Test 2.** 6 specialized detector for left and right profile in ±45°.
Number of WC in each SC	0°:[16, 34, 162, 200, 200]	0°:[16, 34, 162, 200, 200] 30°:[27, 38, 164, 200, 200] 330°:[25, 38, 200, 200, 200]
Detection rate	**82.2%**	**92.5%**
False positives	148	432

Table 3.5: Comparison with other works on CMU profile test set.

	Detection rate	**False positives**
Schneiderman and Kanade (2000)	92.8%	700
Wu et al. (2004)	91.3%	415
Test 2	92.5%	432

Table 3.6: Results for the BioID test set using different parameters.

	Detection rate	**False positives**	**Number of Haar features per layer**
Test 1	80.47%	92	[17, 39, 200, 200, 200]
Test 2	83.89%	126	[17, 39, 200, 200]
Test 3	93.68%	432	[11, 18, 37, 300, 300]

Table 3.7: Comparison with other works that used the BioID test set.

	Detection rate	**False Positives**
Jesorsky et al. (2001)	91.80%	Not reported
Kirchberg et al. (2002)	92.80%	Not reported
Hamouz et al. (2004)	91.30%	Not reported
Fröba and Ernst (2004)	97.75%	25
Ramirez and Fuentes (2005)	93.23%	2236
My method test 3	93.68%	432

Table 3.8 presents a comparison between the specialized detector in frontal faces and the detectors presented in Viola and Jones (2001) and Wu et al. (2004), which are based on basic (symmetric) Haar features. We can see in Table 3.8 that my detector uses fewer Haar features than the others. This suggest that asymmetric Haar features represent better the face appearance than symmetric Haar features. Additionally, the genetic algorithm used for feature selection enables us to use a much richer feature space without increasing computational costs.

To the best of my knowledge, a standard test set for multi-pose face detection is not currently in use. Therefore, I created a multi-pose test set. This test set consists of 45 color images with 60 profile faces and 101 frontal faces. The faces have a rotation of up to ± 45 degrees. I performed a experiment without the skin color segmentation and

Table 3.8: Comparison of my frontal face detector with other works based on Haar features.

	Number of stages in cascade	**Number of Haar features in all the cascades**
Viola and Jones (2001)	32	4,297
Wu et al. (2004)	16	756
My detector	5	666

Table 3.9: Multi-pose test set results.

	Without skin color segmentation	**With skin color segmentation**
Detection rate	91.93 %	95.03 %
False positives	110	40

another experiment with the skin color segmentation. Nine specialized detectors to cover the rotation range of ±45 degrees for faces in frontal and profile pose were used for the two experiments. Table 3.9 shows the results of the 2 experiments. As we can see in Table 3.9, the skin color segmentation reduces the number of false positives and increases the detection rate.

Figures 3.7, 3.8 and 3.9 show some representative results for the CMU profile set, the BioID set, and the multi-pose set, respectively.

3.8 Conclusions

I developed a detection system that introduces three extensions to previous state-of-the-art systems. First, I introduced asymmetric Haar features as a generalization to the basic set of Haar features. Experimental results show a competitive performance to other previous

approaches. It seems that asymmetric Haar features are better to describe objects with asymmetric appearance such as profile faces. Second, I reduced the training time using a genetic algorithm that allows to exploit the expressive advantage of asymmetric features to 0.02% of the time that would be required using the full feature set. Lastly, using the proposed skin color segmentation schema it is possible to reduce the search space resulting in faster processing and fewer false positives. My system can detect faces in different poses with a detection rate of up to 92.5% for profile faces and up to 93.68% for frontal faces.

Figure 3.7: Some results from the CMU profile test set.

Figure 3.8: Some results from the BioID test set.

Figure 3.9: Some results from the multi-pose test set.

Chapter 4

Emotion Recognition From Geometric Information

4.1 Introduction

This chapter presents results about experiments for emotion recognition using high-level features from geometric information. I performed experiments using the AVEC dataset (Schuller et al., 2011). I computed a set of high-level features including vertical and horizontal eye gaze, head tilt, smile intensity, and eyebrow motion. The high-level features were used for creating a classifier based on the Latent-Dynamic Conditional Random Fields method. This method has the advantage of learning the sub-structure of the affective signals as well as the dynamics between emotion labels. The proposed solution is based on a feature extractor that uses a face detector and a facial component tracker to create high-level features, as shown in Figure 4.1.

4.2 Dataset

One of the main problems in emotion recognition is the lack of enough data with natural affective expressions. Many of the previous works are based on databases with posed expressions that do not reproduce natural human behavior (Gunes and Pantic, 2010). One database that includes natural behavior of the participants is the SEMAINE dataset (McKeown et al., 2010). This database consists of sessions where a participant is interacting with a virtual character that shows a specific stereotyped emotional behavior (see Figure

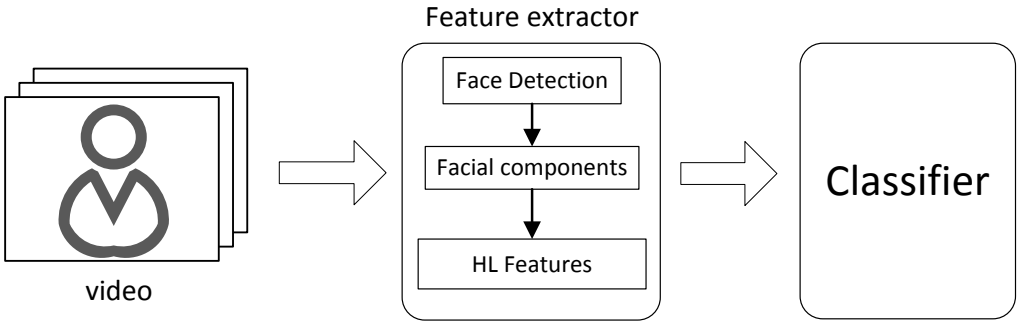


Figure 4.1: General diagram for emotion recognition using high-level (HL) features.

4.2). All the sessions meet the guidelines of the Sensitive Artificial Listener (SAL) scenario (Douglas-Cowie et al., 2008). Under the SAL conditions, the responses of the virtual characters are just stock phrases selected by a human operator with the goal of creating a conversational environment rich in emotional expressions. In this dissertation a subset of the SEMAINE database provided by Schuller *et al.* for the First International Audio/Visual Challenge (AVEC) was used (Schuller et al., 2011). The dataset consists of 95 videos of dyadic interaction sessions including the upper body of subjects. Labels for the affective dimensions activation, expectation, power and valence for each frame are included.

The data is divided into 3 subsets: training, development, and testing. The training set consists of 31 sessions, while the development set consists of 32 sessions that were used for validation of the model parameters. The test set consists of 11 video-only sequences, 11 audio-only sequences, and 10 audiovisual sequences. Table 4.1 shows a summary of the distribution of the AVEC dataset.

4.3 Facial Features

The face is the most important component in visual emotion recognition (Zeng et al., 2009). Many efforts have focused on the analysis of the facial expression components such

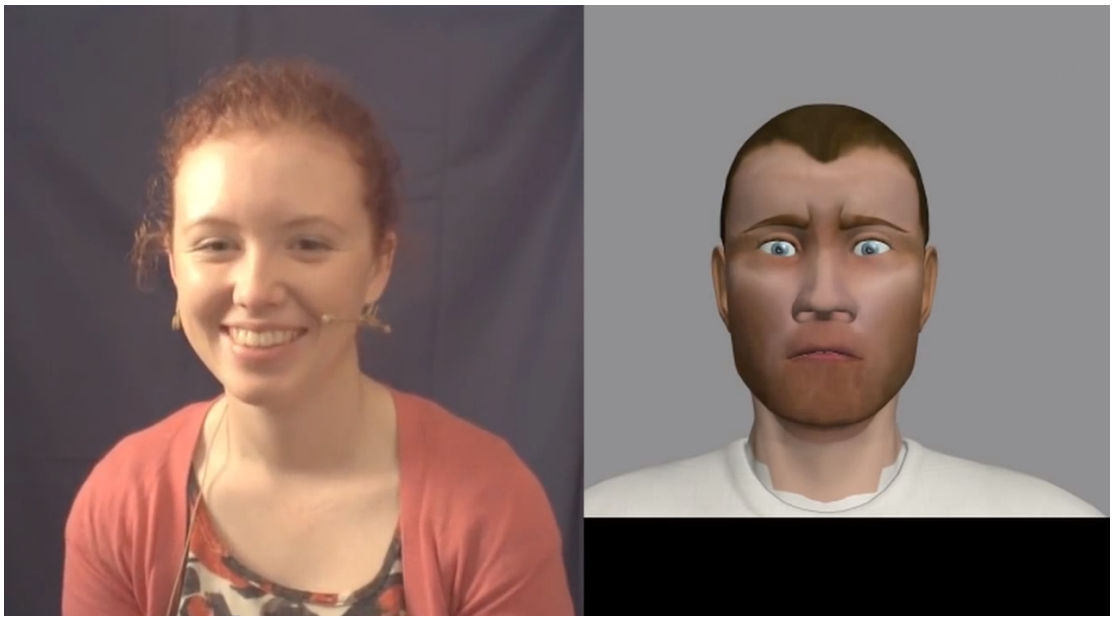


Figure 4.2: Single frame of a video from the SEMAINE dataset (McKeown et al., 2010) where a user interacts with a virtual character controlled by a human operator. (Image extracted from http://www.semaine-project.eu/).

Table 4.1: AVEC dataset distribution.

	Train	**Develop**	**Test**	**Total**
Number of Sessions	31	32	32	95
Number of Frames	501,277	449,074	407,772	1,358,123
Number of Words	20,183	16,311	13,856	50,350
Ave. word duration [ms]	262	276	249	263

as eyes, eyebrows and mouth. In many cases, the analysis of facial expression consists of the extraction of hundreds of low-level features that describe the face appearance as a whole. Recent approaches have performed this analysis by coding the interaction of the facial components or using the relative location of the components respect to other body components such as shoulders. However, it requires a high density of those low-level features to model the variability and expressiveness of a face. One of the goals on this dissertation is to design a set of high-level features that can encode more relevant information about the face. Visual high-level features include the head pose and orientation, horizontal and vertical eye gaze direction, eyebrows shape, and smile level. To determine which features or combinations of features are useful for emotion recognition a set of experiment was performed.

I selected a subset of visual communicative signals based on geometric information, which were shown to be useful when analyzing dyadic interactions (Bavelas et al., 2000, Krämer, 2008, Argyle and Dean, 1965). I used the Omron OKAO Vision software library (OKAO, 2011) to automatically extract the following facial features: horizontal eye gaze direction (degrees), vertical eye gaze direction (degrees), smile intensity (from 0-100) and head tilt in degrees. In addition, I used the Seeing Machines Face API (FaceAPI, 2011) to automatically track the location of the eyebrows.

4.4 Experimental Setup

An emotion is expressed over time and it changes slowly. Thus a given frame of a video is likely to have the same emotional state as the previous frame. This trend is exploited for some discriminative machine learning algorithms that attempt to model the dynamic of a class over time. Such is the case of Latent-Dynamic Conditional Random Fields (LDCRF) algorithm where the hidden dynamics among input features (e.g., gaze and smile) is explicitly learned (Morency et al., 2007) (see Figure 4.3). LDCRF offers several advantages over previous discriminative algorithms. In contrast to Conditional Random Fields

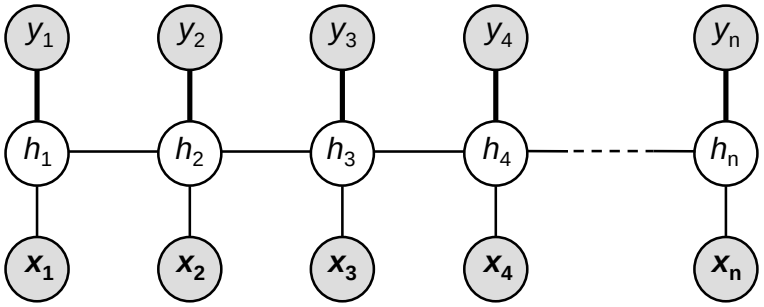


Figure 4.3: Graphical representation of the LDCRF model. x_j represents each observation of the sequence, h_j is a hidden state assigned to x_j , and y_j the class label of x_j (i.e. positive or negative). Gray circles are observed variables.

(CRFs) (Lafferty et al., 2001), LDCRF incorporates hidden state variables which model the sub-structure of affective sequences. The CRF approach models the transitions between affective states (e.g., lethargy to excitement), thus capturing extrinsic dynamics, but lacks the ability to represent internal sub-structure of input features (e.g., looking away and eyebrow-raise). LDCRF can learn the dynamics between affective labels and can be directly applied to labeled unsegmented sequences.

As described in Morency et al. (2007), the task of the LDCRF algorithm is to learn a mapping between a sequence of observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for each observation in a sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{\texttt{positive}, \texttt{negative}\}$. Each frame observation x_j is represented by a feature vector $\phi(x_j) \in \mathbf{R}^d$, for example, the eye gaze, smile and head tilt at each frame. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Given the above definitions, we define a latent conditional model:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h} \mid \mathbf{x}, \theta). \tag{4.1}$$

where θ are the parameters of the model.

Given a training set consisting of n labeled sequences $(\mathbf{x_i}, \mathbf{y_i})$ for $i = 1...n$, training is done following Lafferty et al. (2001) to learn the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$ using this objective function:

$$L(\theta) = \sum_{i=1}^{n} \log P(\mathbf{y}_i \mid \mathbf{x}_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2 \tag{4.2}$$

The first term in Eq. 4.2 is the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp\left(\frac{1}{2\sigma^2}||\theta||^2\right)$.

For testing, given a new test sequence $\mathbf{x}$, we want to estimate the most probable label $\mathbf{y}^*$ for a sequence that maximizes the conditional model using the optimize parameters θ^*:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{\mathbf{h}:\forall h_i \in \mathcal{H}_{y_i}} P(\mathbf{h} \mid \mathbf{x}, \theta^*) \tag{4.3}$$

For a more detailed discussion of LDCRF training and inference see Morency et al. (2007).

4.5 Algorithms

In addition to the LDCRF algorithm, Conditional Random Fields (CRF) were used (Lafferty et al., 2001), which have been already used in affective dimension classification tasks by Wöllmer et al. (2008). The CRF algorithm has a similar structure as the LDCRF algorithm but without the hidden variables. No latent dynamics is explicitly learned with the CRF algorithm. Also, the following non-temporal models from the Weka toolkit (Hall et al., 2009) are used for comparison:

- **Support Vector Machine (SVM).** SVM builds hyperplanes that represent the largest separation between classes of training dataset (Platt, 1999).

- **Decision Trees (DT).** Implementation of C4.5 algorithm for decision trees that uses the information entropy of a training dataset to build a classifier (Quinlan, 1993).

Table 4.2: Visual feature evaluation: comparison between the use of Local Binary Patterns features (LBP) Schuller et al. (2011) and my set of high-level features (HLF) described in Section 4.3 on the development set.

Accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**	**Average**
SVM + LBP (Schuller et al., 2011)	**60.2**	58.3	**56.0**	**63.6**	**59.5**
SVM + HLF	58.7	**60.3**	54.0	**63.6**	59.1

- **Multinomial Logistic Regression (MLR).** MLR is a variation of Logistic Regression that uses a ridge estimator for multiclass problems (le Cessie and van Houwelingen, 1992).

The following model parameters were automatically tuned: for CRF and LDCRF the L2-norm regularization term was validated with values $10^k, k = -2..3$, for LDCRF the number of hidden states (2-4) was validated, no validation was performed for DT and MLR. For training CRF and LDCRF I used the freely available hCRF library [1]. For SVM a radial basis function kernel SVM was used.

4.6 Experiments with Four High-Level Features

The goal in the experiments is the evaluation of the selected visual features with different machine learning algorithms. I performed a comparison of my set of high-level features (horizontal and vertical gaze, smile and head tilt) with the Local Binary Patterns (Ojala et al., 2002) features used in Schuller et al. (2011) on the same dataset. For fair comparison, I trained a Support Vector Machine (SVM) with a radial basis function (RBF) kernel as was done by Schuller et al. (2011). They used a feature vector of 5,900 elements that correspond to the low-level features computed with LPB's over the face of the subject. In comparison, my feature vector was only 4 elements corresponding to my high-level set of features. The performance of using my high-level features is similar to that reported by

[1] `http://sourceforge.net/projects/hcrf/`

Table 4.3: Classification results for different algorithms using Local Binary Patterns features (LBP) and my set of high-level features (HLF) on the development dataset.

Accuracy	**Video**				
(%)	**Activation**	**Expectancy**	**Power**	**Valence**	**Average**
SVM + LBP (Schuller et al., 2011)	60.2	58.3	56.0	63.6	59.5
DT + HLF	62.1	55.4	49.3	64.1	57.7
MLR + HLF	64.8	56.9	48.5	67.1	59.3
CRF + HLF	72.3	53.8	46.2	69.5	60.5
LDCRF + HLF	**74.5**	**60.0**	**60.3**	**72.9**	**66.9**

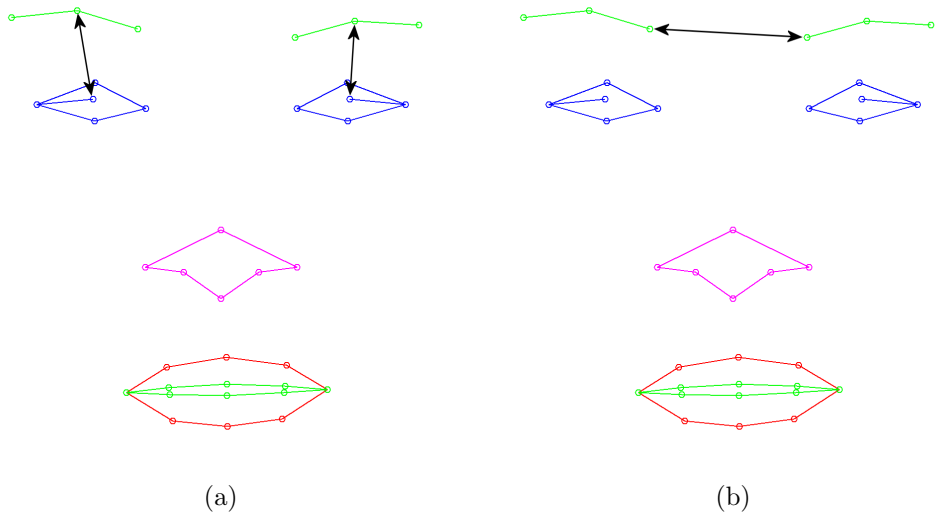
Schuller et al. (2011) as we can see in Table 4.2, showing that my selected features are at least as good as theirs. It seems that my set of high-level features can represent as much information for emotion recognition as 5,900 low-level features.

It can be seen from Table 4.3 that Decision Trees have a very similar behavior to the baseline based on Support Vector Machine and Local Binary Patterns (Schuller et al., 2011). Conditional Random Fields has a significant improvement over the baseline but only on the activation and valence dimensions. However, Latent-Dynamic Conditional Random Fields outperforms all of them in all of the affective dimensions, having a increase of more than 17 percentage points over the baseline in average (Schuller et al., 2011).

4.7 Experiments with Six High-Level Features

In addition to gaze, smile and head tilt, two new high-level features based on the eyebrow motion were added. The first was defined as the average distance between the eyebrows and the eyes called *eyebrows up-down* (see Figure 4.4a). The second one was defined as the distance between the inner corners of eyebrows called *eyebrows distance* (see Figure 4.4b).

First, I performed experiments with each high-level feature individually to determine how much each feature can contribute to the overall process. Table 4.4 shows the accuracy for each dimension using only LDCRF with the 6 high-level features individually. For



(a) (b)

Figure 4.4: High-level features based on eyebrows. (a) Average distance between center of eyebrows and center of eyes. (b) Distance between inner corners of eyebrows.

the activation dimensions, the head tilt feature obtained the best performance with an accuracy of 70%. Eyebrow features outperformed all the results presented in Table 4.3 for expectancy and power dimensions with an accuracy of 65.5% and 66.2% respectively. In the case of the valence dimension, the smile level feature obtained the highest accuracy of all the individual features and almost the same as LDCRF with 4 features presented in Table 4.3. Also it is interesting to note that, as we can expect, the smile level is highly related to the valence dimension. Moreover, it seems that the eyebrow movements are the most relevant features to use for the expectancy and power dimensions. On average the smile feature obtained the best results with an accuracy of 64.8% and the horizontal gaze the lowest with a accuracy of 60.0%.

My second set of experiments was the 6 high-level features using MLR and LDCRF. We can see in Table 4.5 that combining the 6 features has a better accuracy for all the dimensions than using only 4 features. The average performance was increased by 2.4 percentage points. Despite the slight improvement in accuracy for the expectancy and power dimension using 6 features instead of 4, it is possible to have a better accuracy using

Table 4.4: Classification results for each high-level feature for video modality using only LDCRF for video modality.

Accuracy	**Video**				
(%)	**Activation**	**Expectancy**	**Power**	**Valence**	**Average**
Head tilt	**70.0**	60.7	59.8	65.2	63.9
Horizontal gaze	60.8	61.8	57.9	59.6	60.0
Vertical gaze	67.2	61.5	61.3	67.4	64.3
Smile	65.5	62.3	58.8	**72.7**	**64.8**
Eyebrows distance	58.4	61.1	**66.2**	70.0	63.9
Eyebrows up-down	56.7	**65.5**	65.7	64.2	63.0

only the eyebrows distance for expectancy and the eyebrows up-down for power. It seems that there is a strong correlation between these 2 high-level features and the expectancy and power dimensions.

4.8 Audio and Audiovisual Data Experiments

Since the AVEC dataset (Schuller et al., 2011) includes audio data, I performed experiments to infer the emotional state using only audio data. I used the 1,941 features provided with the dataset. Each of the features was sampled over a duration of a single word (the mean word length is 263ms). As the dimensionality of the feature set is very high, I applied Correlation-based Feature Selection (CFS) to select a subset of features relevant for the task (Hall, 1999). Due to memory limitations of the Weka toolkit (Hall et al., 2009), a subsample of the audio training set was taken for feature selection (every third word). On the resulting subset a 10-fold cross validation CFS was performed on each of the four emotion labels (activation, expectancy, power, and valence) independently, that is, leaving a tenth of the training set out and running CFS on the remaining data. Features that

Table 4.5: Classification results for MLR and LDCRF using 4 and 6 high-level features for video modality.

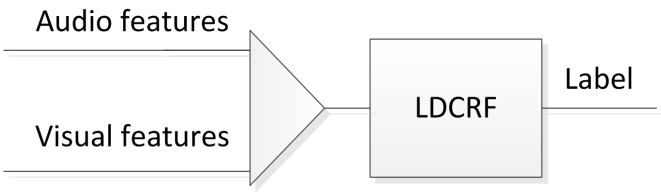
Accuracy (%)	**Video**				
	Activation	**Expectancy**	**Power**	**Valence**	**Average**
MLR					
gaze, smile and head tilt	**64.8**	56.9	48.5	67.1	59.3
gaze, smile, head tilt and eyebrows	63.2	**57.9**	**52.7**	**68.1**	**60.5**
LDCRF					
gaze, smile and head tilt	74.5	60.0	60.3	72.9	66.9
gaze, smile, head tilt and eyebrows	**77.0**	**61.5**	**64.2**	**74.4**	**69.3**

Table 4.6: Classification results for different algorithms on the development dataset for audio modality.

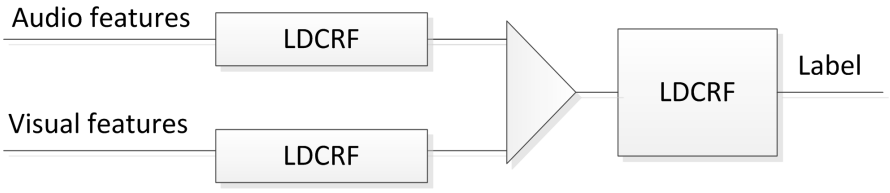
Accuracy (%)	**Audio**				
	Activation	**Expectancy**	**Power**	**Valence**	**Average**
Baseline Schuller et al. (2011)	63.7	63.2	65.6	58.1	62.7
Decision trees	60.8	65.9	63.1	**64.3**	63.5
CRF	62.9	67.3	67.0	44.6	60.4
LDCRF	**74.9**	**68.4**	**67.0**	63.7	**68.5**

were chosen in at least 5 of 10 folds were chosen as input features for our model, with the exception of activation where only the features selected in all of the ten folds were chosen as the 5 out of 10 approach lead to 91 features. This was done in order to keep the dimensionality low, and have a similar number of features across the different affective dimensions. This resulted in 19 features for activation, 7 for expectancy, 22 for power, and 15 for valence dimensions (see Table 4.9).

Similarly to the visual data experiments, the approach was evaluated using different algorithms on the development set as is shown in Table 4.6. LDCRF outperforms all other approaches with the exception of Decision Trees for the valence dimension.



(a) Early (feature level) fusion



(b) Late (decision level) fusion

Figure 4.5: The two of multimodal fusion techniques used in the experiments.

4.8.1 Audiovisual Data Experiments

To fuse visual and audio modalities it is necessary to align the visual features with audio features. Visual features are commonly sampled per frame and these audio features are sampled per word. Consequently, the audio features correspond to a longer and varying period of time in comparison with video features. To align visual features with audio features, the visual features are resampled at the word level using the mean values of all the frames happening during a specific word. With this process the sequences of audio and video features have the same length. For all the experiments on audiovisual data, only 4 high-level visual features (horizontal and vertical gaze, head tilt, and smile level) were used.

How to fuse different modalities is still an open research question. In this dissertation two approaches for fusion were explored (see Figure 4.5). The most straightforward one is to concatenate the audio and visual features and train a classifier on them (Figure 4.5a, Early fusion). An alternative to that is to concatenate the marginal probabilities output from the unimodal models (Figure 4.5b, Late fusion) and use that as an input to another

Table 4.7: Fusion methods using LDCRF classifiers on the development set.

Accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**	**Average**
Early Fusion (LDCRF)	79.3	63.4	66.9	62.8	68.1
Late Fusion (LDCRF)	**81.7**	**73.1**	**73.3**	**73.5**	**75.4**
Late Fusion (SVM)	75.4	69.4	65.3	72.1	70.5

classifier.

For audiovisual fusion three experiments were performed, comparing the fusion methods and evaluating them on the development dataset. The first was using early fusion, the second one was using later fusion, and the third one was using late fusion but instead of using LDCRF as the last classifiers, a linear SVM classifier trained on the development set was used. From Table 4.7 it can be see that the late fusion using LDCRF as a model to fuse the outputs of uni-modal classifiers performed best in all of the affective dimensions.

The same validation and testing methodology described earlier was also used for the audiovisual data. In the case of late fusion, the training was done using the training dataset and the validation using the development set (the same technique used for the unimodal models). In each case, a separate binary classifier was trained for each of the dimensions rather than one giving multiple-label outputs.

4.8.2 AVEC Challenge

The results of the experiments with the test set of the AVEC dataset (see Table 4.8) were submitted to the First International Audio/Visual Emotion Challenge and Workshop (Schuller et al., 2011). Please note that only 4 high-level visual features were used for these experiments. It can be see in Table 4.8 that high-level visual features outperform the baseline for all of the affective dimension by 14.8 percentage points on average for the video sub-challenge. Also for the audiovisual sub-challenge, the results based on LDCRF and late

Table 4.8: Official results on the test set.

Accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**	**avg.**
Video sub-challenge					
Baseline (Schuller et al., 2011)	42.2	53.6	36.4	52.5	46.2
LDCRF	**65.5**	**61.7**	**47.1**	**69.8**	**61.0**
Audio sub-challenge					
Baseline (Schuller et al., 2011)	55.0	**52.9**	**28.0**	44.3	**45.1**
LDCRF	**55.8**	50.1	19.8	**46.5**	43.0
Audiovisual sub-challenge					
Baseline (Schuller et al., 2011)	**67.2**	36.3	62.2	**66.0**	57.9
LDCRF	65.6	**53.4**	**62.9**	59.5	**60.3**

fusion outperform the baseline for the expectancy and power dimensions. For the audio sub-challenge, LDCRF performance is similar to the SVM baseline. The low performance of both SVM and LDCRF approaches (e.g. the best performance on the power labels is 28%) on this test set suggests a significant difference in the data distribution of the audio-only sub-challenge test set.

4.9 Conclusions

The results show that the approach presented based on visual high-level features and LDCRF instead of low-level features outperforms the previously published approaches for all four affective dimensions on the AVEC dataset, for both the development and test sets. By using Latent-Dynamic Conditional Random Fields it is possible to model the temporal information and the interaction between the high-level perceptual features. According to the experiments presented in this chapter, some high-level features contribute more than

others to recognize emotions depending on the evaluated affective dimension. It seems that the head tilt is the most relevant feature for the activation, the eyebrows for the expectancy and power, and the smile level for the valence. As a result of merging audio and visual modalities, it is possible to improve the performance over the unimodal approaches. In addition, LDCRF seems to be suitable for late feature fusion, outperforming the SVM for fusion.

Table 4.9: Selected audio features from the AVEC dataset for each dimension.

Activation (19 features)	**Power (22 features)**
23. audspec_lengthL1norm_sma_stddevRisingSlope	26. audspec_lengthL1norm_sma_meanSegLen
25. audspec_lengthL1norm_sma_stddevFallingSlope	179. pcm_Mag_spectralRollOff25.0_sma_amean
151. pcm_Mag_fband1000-4000_sma_stddevFallingSlope	254. pcm_Mag_spectralRollOff75.0_sma_quartile1
172. pcm_Mag_spectralRollOff25.0_sma_quartile3	593. pcm_Mag_harmonicity_sma_iqr1-2
212. pcm_Mag_spectralRollOff50.0_sma_quartile1	597. pcm_Mag_harmonicity_sma_percentile99.0
634. mfcc_sma[1]_quartile3	643. mfcc_sma[1]_rqmean
639. mfcc_sma[1]_percentile99.0	710. mfcc_sma[2]_lpgain
932. mfcc_sma[8]_percentile1.0	765. mfcc_sma[4]_percentile99.0
962. mfcc_sma[8]_lpgain	927. mfcc_sma[8]_quartile2
972. mfcc_sma[9]_iqr2-3	928. mfcc_sma[8]_quartile3
1016. mfcc_sma[10]_percentile1.0	932. mfcc_sma[8]_percentile1.0
1054. F0final_sma_quartile3	933. mfcc_sma[8]_percentile99.0
1102. voicingFinalUnclipped_sma_peakMeanRel	944. mfcc_sma[8]_peakMeanRel
1251. audspec_lengthL1norm_sma_de_percentile99.0	1061. F0final_sma_amean
1290. pcm_Mag_fband250-650_sma_de_quartile1	1271. pcm_zcr_sma_de_iqr2-3
1336. pcm_Mag_spectralRollOff25.0_sma_de_quartile1	1336. pcm_Mag_spectralRollOff25.0_sma_de_quartile1
1337. pcm_Mag_spectralRollOff25.0_sma_de_quartile2	1337. pcm_Mag_spectralRollOff25.0_sma_de_quartile2
1428. pcm_Mag_spectralFlux_sma_de_quartile1	1358. pcm_Mag_spectralRollOff25.0_sma_de_risetime
1756. mfcc_sma_de[8]_percentile1.0	1623. mfcc_sma_de[2]_rqmean
	1711. mfcc_sma_de[6]_percentile99.0
	1771. mfcc_sma_de[8]_upleveltime90
	1798. mfcc_sma_de[10]_quartile3
Expectancy (7 features)	**Valence (15 features)**
95. pcm_Mag_fband250-650_sma_amean	44. pcm_zcr_sma_quartile1
171. pcm_Mag_spectralRollOff25.0_sma_quartile2	45. pcm_zcr_sma_quartile2
212. pcm_Mag_spectralRollOff50.0_sma_quartile1	50. pcm_zcr_sma_percentile1.0
826. mfcc_sma[5]_minSegLen	172. pcm_Mag_spectralRollOff25.0_sma_quartile3
927. mfcc_sma[8]_quartile2	224. pcm_Mag_spectralRollOff50.0_sma_stddev
935. mfcc_sma[8]_amean	248. pcm_Mag_spectralRollOff50.0_sma_lpgain
1942. F0final_sma_f0p_segLenStddev	549. pcm_Mag_psySharpness_sma_quartile2
	932. mfcc_sma[8]_percentile1.0
	944. mfcc_sma[8]_peakMeanRel
	1345. pcm_Mag_spectralRollOff25.0_sma_de_flatness
	1363. pcm_Mag_spectralRollOff50.0_sma_de_iqr2-3
	1692. mfcc_sma_de[5]_rqmean
	1739. mfcc_sma_de[7]_stddev
	1758. mfcc_sma_de[8]_pctlrange0-1
	1761. mfcc_sma_de[8]_rqmean

Chapter 5

Emotion Recognition From Color Information

5.1 Introduction

Some researchers have proposed that the reason humans evolved color vision is to detect each others emotional and physical state from subtle skin color hue changes, which are due to different levels of hemoglobin and oxygenation under the skin (Changizi, 2009, Yuen et al., 2009). Primates with color vision, including humans, tend to have bare faces, while colorblind primates tend to have faces covered with fur. Different levels of hemoglobin and its oxygenation generate different skin color hues: colors range from blue to yellow depending on the levels of hemoglobin, and green to red depending on oxygenation. Skin tones, regardless of ethnicity, reflect light in a similar matter, therefore skin colors changes are still present and visible. Since human visual perception is based on red, green and blue color, analyzing RGB images of facial emotional might be useful for detecting emotional states.

For emotion recognition from color we created a dataset of spontaneous behavior that was created using a DSLR camera under controlled lightning conditions. It includes persons of different ethnicity, age, and gender. All subjects were recorded in a resolution of 1920×1080 pixels at 30 frames per second.

My interest was on exploring the facial skin color as reliable feature to determine the emotional state of a person. The first challenge is to create a dataset of spontaneous facial expressions of persons of different ages, culture, and genders. The second one is to

Figure 5.1: Experiment setup. The subject was in front of a monitor and a couple of lamps with light diffuser. The DSLR camera was mounted just behind the monitor.

determine the usefulness of the facial skin color as a feature to infer the emotional states of people.

5.2 Dataset

Since no suitable datasets were available, a new dataset was created to ensure spontaneous emotions recorded in high resolution and quality with consistency in lighting condition. This dataset was created under a controlled lab setting to ideally capture the skin color changes that occur at different emotional states. The subjects' reactions were recorded with a Canon EOS T4i DSLR camera in a quiet and isolated room under constant light. All videos were recorded using fixed parameters: ISO, focal length, and lens aperture. Also, all videos were recorded at 30 frames per second with a resolution of 1920×1080 pixels and stored in H.264 format. The subjects were asked to take a seat in front of a computer

Figure 5.2: Some examples of subjects in the dataset.

monitor and two lamps supplied constant lighting (see Figure 5.1). For capturing the most authentic emotions possible, a set of video clips with emotional content was used. The videos consist of short scenes from movies, television shows, or homemade videos and were used as stimuli to elicit positive, negative, or neutral emotion on subjects. Each video clip lasted 40 seconds. The content of positive videos ranges from comedy movies to homemade videos with funny situations. The content of negative videos includes scenes of movies with explicit physical violence and nauseating scenes. The neutral videos consist of trivial conversation of 2 o more persons. Between video stimuli, an additional intertrial

video clip with peaceful nature scenes was shown to help the subjects to relax and go back to their baseline emotional state. After watching each video clip, subjects were required to answer a short survey to rate their current emotional state. The survey consisted of a set of basic and discrete emotion classes (pleasant, unpleasant, disgust, fear, anger, happiness, sadness, excitement, and relaxation) on a scale between one and seven, where one was for a low intensity and seven to a high intensity. The dataset contains videos of 56 subjects with ages from 18 to early 40's, both male and female, and different ethnicities: Caucasian, African American, Hispanic, and Asian (see Figure. 5.2). 4 videos per stimulus category per subject were collected. However, for this first attempt, only one video per stimulus category was used and subjects wearing glasses were also excluded. The experimental subset was of 48 subjects, 23 females and 25 males. The total number of videos was of 144 videos, that is 3 videos per subject.

5.3 Feature Evaluation

I focused on 3 regions of interest (ROIs) to analyze the skin color of the face, corresponding to the forehead and both cheeks. To keep track of the ROIs along the entire video, a facial feature detector and tracker proposed by Saragih et al. (2009) was used. Their tracker is able to detect and track the location of the eyes, eyebrows, nose, mouth and face contour as shown in Figure 5.3. The raw RGB values were normalized using the color descriptor index proposed by Richardson et al. (2007). The indices are based on a color opponent model where each index represents the difference between the color of interest and the other color components. The indices are computed using equations 5.1, 5.2 and 5.3 for the red, green and blue indices respectively, where $\mathbf{p}$ is each pixel in a ROI and R, G and B the red, green and blue component of each pixel respectively.

$$RedX(\mathbf{p}) = 2\mathbf{p}_R - \mathbf{p}_G - \mathbf{p}_B \tag{5.1}$$

$$GreenX(\mathbf{p}) = 2\mathbf{p}_G - \mathbf{p}_R - \mathbf{p}_B \tag{5.2}$$

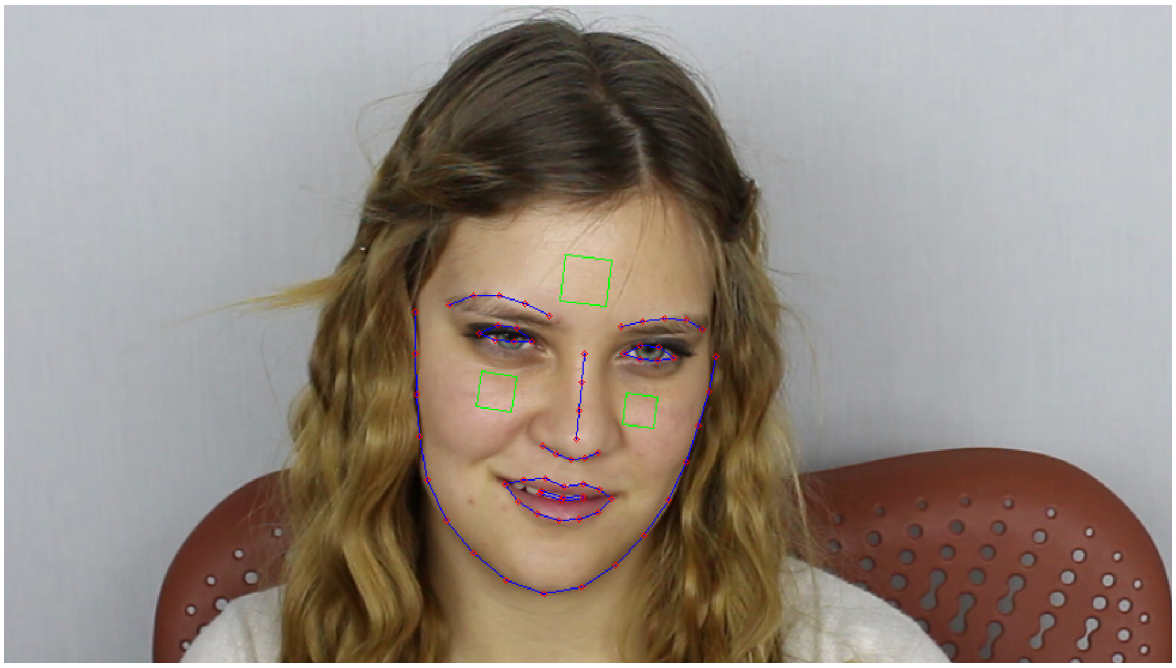


Figure 5.3: Facial features tracker. The green rectangles correspond the the region of interest (ROI's) were all the experiments were focus to analyze skin color.

$$BlueX(\mathbf{p}) = 2\mathbf{p}_B - \mathbf{p}_R - \mathbf{p}_G \tag{5.3}$$

The main goal is to evaluate the skin color as a reliable feature to infer the emotional state of a person in the valence dimension. Therefore, I performed a comparison of 5 machine learning algorithms to see whether it is possible to find a similar trend classifying the valence. The following algorithms were used: Decision Trees (DT) (Quinlan, 1993), Locally Weighted Regression (LWR) (Cleveland, 1979), K-Nearest Neighbors (KNN) (Altman, 1992), Multinomial Logistic Regression (MLR) (le Cessie and van Houwelingen, 1992), and Latent-Dynamic Conditional Random Field (LDCRF) (Morency et al., 2007). LWR and KNN are instance based learning algorithms; their main advantage is a zero training time and ther ability to learn complex functions. DT, MLR, and LDCRF were described in Chapter 4. I also performed experiments with different combination of ROI's to determine what ROI is the most relevant to the problem.

5.4 Experimental Setup

Each video stimulus lasts 40 seconds, but a sequence of interest (SOI) of 5 seconds was defined corresponding to the climax of each video stimulus. The baseline skin color for each subject was computed from one intertrial video of the same subject to ensure an emotionless sample. All pixels of each ROI were processed using equations 5.1, 5.2 and 5.3. After that, the mean value of each ROI was computed. To compute the change of color in ROI's, the baseline skin color was subtracted for each frame in the SOI. The valence of each subject was evaluated as a binary and a ternary classification, that is positive vs. negative emotion, and positive vs. neutral vs. negative, respectively. The total number of sequences was 144. Figures 5.4 and 5.5 show the behavior of the 9 indices and the SOI for a positive and negative stimulus respectively.

After computing the features for each SOI, the frame rate was resampled from 30 fps to 3 fps to reduce the noise caused by the subject movements, and also to reduce the size of the training data. Since one of the goals was to determine what ROI is the most relevant to infer the valence of a person, a set of experiments with different combinations of ROI's was performed.

All the experiments were performed using 10-fold cross validation. The classification was performed per sequence instead of per frame. For DT and MLR algorithms, the WEKA tool kit was used (Hall et al., 2009). For KNN and LWR, we implemented a custom version based on Altman (1992) and Cleveland (1979), respectively. The number of neighbors for KNN was defined by performing tests with different numbers of neighbors, resulting in fixing 7 neighbors as the best parameter. Experiments wiht LDCRF were performed using the hCRF library[1]. LDCRF was tuned for L2-norm regularization parameter with values of 0.01, 0.1, 0, 10, 100 and 1000. Also, the number of hidden states was tuned with values of 2, 3 and 4. In the case of LDCRF, the training sequences in each iteration of the cross validation was split in two-thirds for training and one-third for tuning parameters.

[1] `http://sourceforge.net/projects/hcrf/`

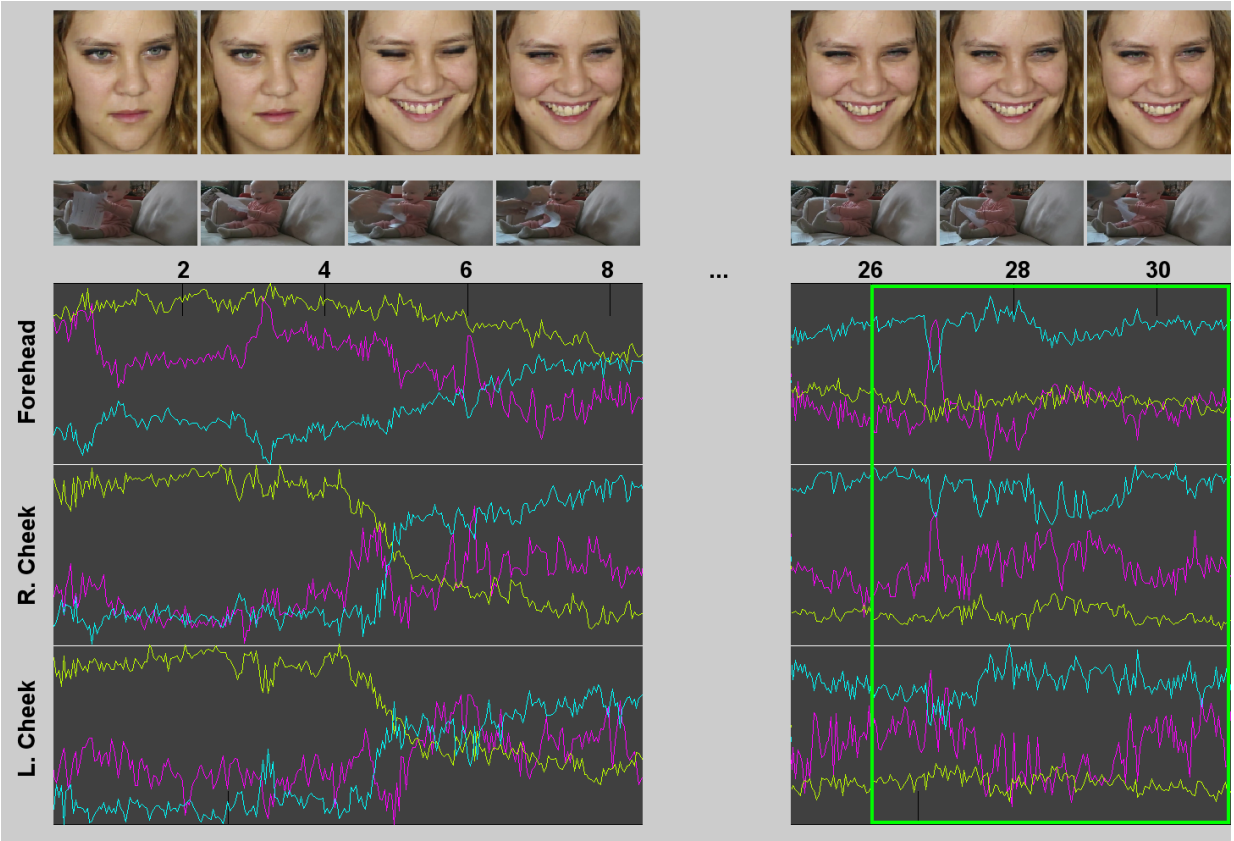


Figure 5.4: The 9 feature indices while the subject is watching a positive video. The values change as the subject changes her valence from neutral to positive. The green rectangle corresponds to the 5 seconds of the sequence of interest (SOI).

5.5 Results

Since the actual emotion experienced by the subjects is impossible to determine, two different target functions as potential representatives of the ground truth to define a label for each sequence were used. The first target function is the emotion that the video presented to the subjects was intended to elicit (denoted as *stimulus* in the results tables). The second target function is the emotion that the subject reported experiencing while watching the video according to a survey filled out immediately after the experiment (denoted as *survey* in the results tables).

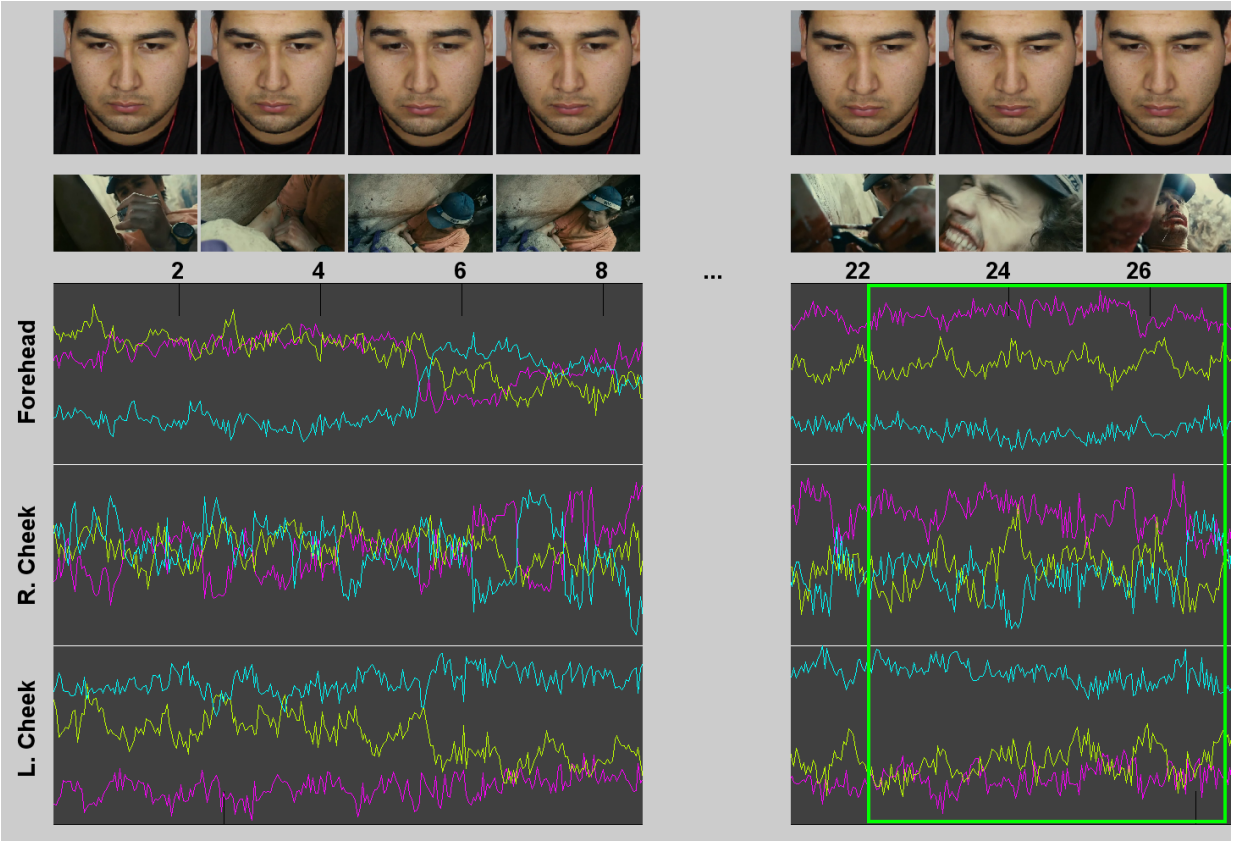


Figure 5.5: The 9 feature indices along 16 seconds while the subject is watching a negative video. The values change as the subject changes his valence from neutral to negative. The green rectangle corresponds to the 5 seconds of the sequence of interest (SOI).

A comparison between the class labels predicted and the stimulus and survey labels was performed. Also, the experiments included binary (positive or negative) and ternary (positive, neutral, or negative) classification.

As we can see in Table 5.1, the best result for binary classification was obtained with LDCRF using the survey answers as class labels with an accuracy of 77.1%. In terms of algorithms, LDCRF outperforms the other algorithms, while MLR also provided very good results. The best overall results were obtained using FH+LC (forehead and left cheek) as ROI's and LDCRF as the learning algorithm for both stimulus and survey labels.

Table 5.1: Results for Positive vs. Negative. Using as class labels the stimulus and the survey. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy(%)	**Positive vs. Negative**				
Stimulus	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	63.5	63.5	60.4	**68.8**	66.7
RC	52.1	57.3	59.4	58.3	**65.6**
LC	**62.5**	**62.5**	60.4	**62.5**	54.2
FH+RC	63.5	63.5	64.6	68.8	**70.8**
FH+LC	61.5	68.8	62.5	**75.0**	**75.0**
RC+LC	**58.3**	57.3	56.3	**58.3**	54.2
FH+RC+LC	66.7	67.7	68.8	68.8	**71.9**
Survey	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	**68.8**	65.7	65.7	67.7	60.4
RC	56.3	53.1	53.1	53.1	**61.5**
LC	**57.3**	56.3	**57.3**	54.2	45.9
FH+RC	64.6	62.5	58.3	**67.7**	66.7
FH+LC	65.6	67.7	64.6	70.8	**77.1**
RC+LC	55.2	53.1	53.1	60.4	**63.5**
FH+RC+LC	64.6	65.6	65.6	**71.9**	67.7

There is a similar behavior in the case of ternary classification, as shown in Table 5.2. As expected, the accuracy is lower than in the binary problem for all combinations of learning algorithms and ROIs. For most combinations of ROIs, MLR yields the best performance, while LDCRF is second.

Regarding of using the stimulus as class labels, the best results are obtained using MLR and FH+RC+LC, with 57.6% accuracy, closely followed by MLR and FH+LC. In the case of using the survey as class label, LDCRF with FH+LC yields the best accuracy, 56.3%.

Similarly as in the binary classification case, we can notice a consistency with the combination of FH+LC as best combination of ROI's for the 5 classification algorithms

Table 5.2: Results for Positive vs. Neutral vs. Negative. Using as class labels the stimulus and the survey. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy(%)	**Positive vs. Neutral vs. Negative**				
Stimulus	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	**47.9**	33.3	35.4	43.8	43.0
RC	**45.8**	31.9	35.4	43.1	43.8
LC	45.8	35.4	34.0	**50.0**	41.0
FH+RC	40.3	40.3	37.5	**51.4**	43.8
FH+LC	45.8	38.9	42.4	**56.3**	52.1
RC+LC	**42.4**	32.6	36.1	**42.4**	41.0
FH+RC+LC	45.1	39.6	44.4	**57.6**	53.5
Survey	**J48**	**LWR**	**KNN**	**MLR**	**LDCRF**
FH	41.7	44.4	40.3	**50.7**	33.3
RC	**46.5**	42.4	43.8	44.4	38.2
LC	38.2	47.2	42.4	**47.9**	**47.9**
FH+RC	47.9	44.4	41.0	**49.3**	42.4
FH+LC	50.0	47.2	47.2	51.4	**56.3**
RC+LC	44.4	43.1	43.1	**47.9**	**47.9**
FH+RC+LC	45.8	45.1	42.4	46.5	**53.5**

when the survey label is used.

It seems that the left and right cheek, used independently, provide a similar performance in the binary and ternary classification. However, the forehead, as an independent feature, was better that the left and right cheek alone for the binary classification, but it has a similar performance for the ternary classification.

The combination of left and right cheeks seems to have the same performance as either alone feature; this could be due its similar concentration of color.

Although the differences in accuracy are small for the various combinations of ROIs chosen, and more detailed analyses are necessary to rule out the effects of non-uniform

lighting, the better results obtained when using FH+LC as opposed to FH+RC could be due to physiological reasons. There is biological evidence that supports asymmetry in emotion expression, with the left side of the body being consistently rated as the more expressive (Roether et al., 2008, Gainotti, 2012). It appears that the data provided by RC is mostly redundant with FH, thus FH and FH+RC lead to very similar accuracies, while LC appears to provide additional information, thus there is a higher improvement when comparing FH+LC to FH alone.

5.6 Conclusions

In this chapter I presented results for detection of the valence emotional state from color changes in facial skin. The experiments were performed using a new spontaneous human emotion dataset with wide ranges of human subjects of different ages and ethnicities. To elicit spontaneous emotions, three different types of stimulus video clips were used: neutral, negative, and positive. The facial skin color is a reliable feature for detecting the valence of the emotion, with an accuracy of 77.1% for a binary label, and 56.3% for a ternary label. Experiments with the 3 face regions (forehead, left cheek and right cheek) show that the forehead is more relevant than the cheeks as a sole feature for recognizing the valence. However, the best results were obtained using the combination of the forehead and the left cheek, which is consistent with research in neuropsychology. Future work will be oriented to try to predict labels given from a human rater, which are a better approximation of the ground truth. Also, more experiments will be performed using richer color descriptors and alternate color spaces. Also more experiments with other learning algorithms can be done, particularly deep recurrent networks, as they have shown to be effective in similar problems.

Chapter 6

Emotion Recognition from Geometry and Color Information

6.1 Introduction

This chapter presents results about experiments combining geometric and color information for emotion recognition. As presented in Chapter 4, it is possible to recognize the emotional state of a person using a small set of high-level features from the geometric appearance of the face. In addition, as presented in Chapter 5, color information can also be useful to recognize the emotional state for the valence dimension.

6.2 Experimental Setup

For all the experiments, the AVEC dataset was used (Schuller et al., 2011). In addition to geometric data used for experiments described in Chapter 4, color information for the AVEC dataset was extracted as described in Chapter 5. It is important to note that the quality of videos from the AVEC dataset is significatively lower than the quality of videos used for the color features experiments. Furthermore, some subjects in the AVEC dataset have their forehead covered with hair or they are wearing glasses, as shown in Figure 6.1. These occlusions add a significant amount of noise to the color features.

The first set of experiments corresponds to the use of color information only from the forehead, right cheek, and left cheek, as described in Chapter 5. The second set of experiments corresponds to combined geometric and color information. For all these

Figure 6.1: Example of subjects with the forehead covered with hair or wearing glasses.

experiment, I used the Multinomial Logistic Regression (MLR) algorithm (le Cessie and van Houwelingen, 1992) and the Latent-Dynamic Conditional Random Field (LDCRF) algorithm (Morency et al., 2007). For MLR, the Weka tool kit was used (Hall et al., 2009). Experiments wiht LDCRF were performed using the hCRF library[1]. LDCRF was tuned for L2-norm regularization parameter with values of 0.01, 0.1, 0, 10, 100 and 1000. The frame rate was resampled from 50 fps to 3 fps to reduce the noise caused by the subject movements, and also to reduce the size of the training data. Since the AVEC dataset includes labels for 4 dimensions, all the experiments were performed for activation, expectancy, power, and valence affective dimensions.

6.3 Results

In Table 6.1 we can see the results of using MLR for color features. The combination of FH+RC obtained the highest accuracy for all the dimensions with an average of 59.6%. In comparison with the results presented in Table 4.5 for 6 geometric features and MLR, the color information was better than geometric information for expectancy and power with an improvement of 3.5 and 4.6 percentage points respectively. On average, there is just a slight difference of 0.9 percentage points between using geometric and color features. Color information appears to be almost as useful as geometric information using MLR.

[1] `http://sourceforge.net/projects/hcrf/`

Table 6.1: Classification results using MLR for color features with different combinations of ROI's for the AVEC dataset. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy (%)	**MLR**				
Features	**Activation**	**Expectancy**	**Power**	**Valence**	**Avg.**
FH	47.7	60.0	56.5	63.0	56.8
RC	53.0	**61.4**	52.8	64.0	57.8
LC	51.6	61.0	48.4	63.6	56.2
FH+RC	**51.8**	**61.4**	**57.3**	**68.0**	**59.6**
FH+LC	50.0	59.4	51.1	67.3	57.0
RC+LC	51.7	58.7	49.1	62.3	55.5
FH+RC+LC	49.8	58.7	53.5	66.1	57.0
6 geometric features (Table 4.5)	**63.2**	57.9	52.7	**68.1**	60.5

The results for LDCRF are shown in Table 6.2. The RC region obtained the highest accuracy for expectancy and power. The combination of FH+LC obtained the highest accuracy for activation and also for power, while the LC region obtained the highest accuracy for the valence dimensions. However, it is possible to reach a higher accuracy using the 6 geometric features (Table 4.5, LDCRF) for the activation, power, and valence. The only configuration where color features are better than geometric features is for the expectancy dimension and is consistent with results for MLR (Table 6.1). On average, the combination of FH+RC obtained the highest accuracy with 62.3%, but it is still 7 percentage points lower than the average accuracy using only geometric information, 69.3%.

It is interesting to notice that in contrast with the experiments presented in Chapter 5, where the LC and FH+LC were the best regions to recognize the valence, in these experiments there was no significant difference between RC and LC, or for combinations of FH+RC and FH+LC. The slightly better performance of the RC region could be due to the low quality of videos combined with the non-uniform lighting in the AVEC videos, where the right sides of the faces are receiving more light than the left side, as we can see in Figure 6.1. We speculate that since the color differences between LC and RC is very small,

Table 6.2: Classification results using LDCRF for color features with different combinations of ROI's for the AVEC dataset. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy (%)	**LDCRF**				
Features	**Activation**	**Expectancy**	**Power**	**Valence**	**Avg.**
FH	58.9	59.5	58.4	63.6	60.1
RC	59.5	**63.3**	**60.4**	64.7	62.0
LC	61.6	63.2	59.0	**65.2**	**62.3**
FH+RC	60.6	59.8	60.1	64.9	61.4
FH+LC	**63.7**	61.0	**60.4**	63.3	62.1
RC+LC	59.4	62.2	59.9	63.9	61.4
FH+RC+LC	62.5	61.3	59.6	63.9	61.8
6 geometric features (Table 4.5)	**77.0**	61.5	**64.2**	**74.4**	**69.3**

it can only be reliable detected and exploited under controlled illumination conditions.

For the second set of experiments, geometric and color information were straightforwardly combined in a single vector of features. The results using MLR are shown in Table 6.3. The combination of Geometric+Color outperforms the results using only geometric information for the expectancy, power, and valence. In comparison to using only color, the combination of Geometric+Color outperform the results for the activation and valence. It is interesting to see that the combination of Geometric+Color for valence has an accuracy of 72.4%, which is 4.3 percentage points better than using either only geometric or only color. On average, G6+FH+RC have the highest accuracy with 61.3%, that is better than the average of either only geometric or only color. However, using geometric and the per-dimension best combination of color regions, the average accuracy grows to 62.3%.

The results of experiments for LDCRF combining geometric and color features are shown in Table 6.4. In comparison with MLR (Table 6.3), the performance was improved for all the dimension with the exception of the valence. However, using only geometric with LDCRF is possible to reach a higher accuracy than using geometric and the per-dimension best combination. It seems that combining geometric and color information using LDCRF

Table 6.3: Classification results using MLR for geometric features and color features with different combinations of ROI's for the AVEC dataset. G6: Geometric features. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy (%)	**MLR**				
Features	**Activation**	**Expectancy**	**Power**	**Valence**	**Avg.**
G6+FH	59.2	57.3	53.1	**72.4**	60.5
G6+RC	**62.6**	57.8	52.8	68.6	60.5
G6+LC	**62.6**	**59.2**	49.2	68.7	59.9
G6+FH+RC	61.0	57.8	**54.8**	71.6	**61.3**
G6+FH+LC	60.7	57.9	49.8	**72.4**	60.2
G6+RC+LC	61.9	58.7	49.8	68.2	59.7
G6+FH+RC+LC	60.9	58.9	51.4	70.4	60.4
Best per dimension	62.6	59.2	54.8	72.4	**62.3**
6 geometric features (Table 4.5)	**63.2**	57.9	52.7	**68.1**	**60.5**
Best color features (Table 6.1)	51.8	**61.4**	**57.3**	68.0	59.6

can only slightly improve the accuracy respect to using only color.

6.4 Conclusions

Geometric and color features have shown to be useful features for emotion recognition. In this Chapter I presented results for emotion recognition combining features from geometric and color information. The experiment were performed using the AVEC dataset using Multinomial Logistic Regression (MLR) and Latent-Dynamic Conditional Random Field (LDCRF). As experimental results show, the average performance of combining geometric and color features outperform individual set of features using MLR. However, the average performance of LDCRF combining geometric and color outperform only the color features.

There is no a significant difference between regions or combination of regions (as was for experiments in Chapter 5); this could be due to low quality of the AVEC videos in conjunction with a non-uniform lightning of faces.

Table 6.4: Classification results using MLR for geometric features and color features with different combinations of ROI's for the AVEC dataset. G6: Geometric features. FH: forehead, RC: right cheek, and LC: left cheek.

Accuracy (%)	**LDCRF**				
Features	**Activation**	**Expectancy**	**Power**	**Valence**	**Avg.**
G6+FH	63.5	59.7	58.5	**69.9**	62.9
G6+RC	**67.1**	59.4	59.9	68.5	63.7
G6+LC	65.7	59.1	59.3	68.4	63.1
G6+FH+RC	62.3	**60.4**	60.7	69.5	63.2
G6+FH+LC	65.2	59.4	59.5	68.4	63.1
G6+RC+LC	65.3	59.6	59.8	67.8	63.1
G6+FH+RC+LC	65.7	58.6	**61.4**	69.8	**63.9**
Best per dimension	67.1	60.4	61.4	69.9	**64.7**
6 geometric features (Table 4.5)	**77.0**	61.5	**64.2**	**74.4**	**69.3**
Best color features (Table 6.2)	63.7	**63.3**	60.4	65.2	63.2

It seems that a straightforward combination of features is not taking advantage of the additional information obtained from color information. A possible solution could be use more sophisticated fusion techniques such as those presented in 4.8.1, that could lead to better merging of data from different sources.

Chapter 7

Conclusions

The goal of this dissertation was to explore the problem of automatic emotion recognition based on geometric and color information. Instead of a categorical view of emotions, a representation based on a small number of continuous latent dimensions was used. This dimensional model is more suitable for analyzing more complex and subtle affective states including shame, pleasure, anxiety, and depression.

For vision-based emotion recognition, the face is the most important source of information and is the first step for gathering visual features. I computed a set of high-level features from geometric information such as vertical and horizontal eye gaze, head tilt, smile intensity, and eyebrow motion. The high-level features were used for creating a classifier based on different machine learning algorithms, including Multinomial Logistic Regression (MLR) and Latent-Dynamic Conditional Random Fields (LDCRF). In addition to geometric features, I explored facial skin color changes as high-level features to determine the emotional state of a person. A set of experiments was performed using a new dataset of spontaneous behavior videos that includes persons of different ethnicities, ages, and genders.

7.1 Contributions

7.1.1 Face Detection

I developed a face detection method with three extensions to state-of-the-art systems. First, I introduced the asymmetric Haar features as a generalization to the basic set of Haar features. Experimental results show a competitive performance to other previous

approaches. These new features are better to describe the asymmetric appearance of objects such profile faces. The second improvement is a method based on a genetic algorithm to reduce the training time. This method allows to explore huge set of asymmetric Haar features. The last improvement is the application of a skin color-segmentation scheme to reduce the search space, resulting in faster processing and fewer false positives. The face detection approach presented in this dissertation has been used in related problem including the automatic morphing of face images (Zanella et al., 2009) and face detection in low-resolution color images (Zheng et al., 2010). In addition, the same approach has been used for a different object detection problem, street detection in satellite images (Ramirez and Fuentes, 2012).

7.1.2 Emotion Recognition from Geometric Information

I presented results about experiments for emotion recognition using high-level features from geometric information. I performed experiments using the dataset for the First International Audio/Visual Emotion Challenge (AVEC). The results show that the approach presented based on high-level features instead of low-level features outperforms previously published approaches for all four affective dimensions on the AVEC dataset. By using Latent-Dynamic Conditional Random Fields it is possible to model the temporal information and the interaction between the high-level perceptual features. According to the experiments presented, some high-level features are more useful to recognize depending on the evaluated affective dimension. It seems that head tilt is the most relevant feature for activation, the eyebrow position for expectancy and power, and smile for valence. Also I found that it is possible to improve the performance of unimodal approaches by merging audio and visual modalities. Partial results were published in the First International Audio/Visual Emotion Challenge and Workshop (Schuller et al., 2011), and these won the first place for the video sub-challenge (Ramirez et al., 2011).

7.1.3 Emotion Recognition from Color Information

I presented results for detection of the valence emotional state from color changes in facial skin. A new spontaneous human emotion dataset with wide ranges of human subjects of different ages and ethnicities was created. To elicit spontaneous emotions, three different type of stimulus video clips were used: neutral, negative, and positive. Facial skin color change is a reliable feature for detecting the valence of the emotion. LDCRF seems to be suitable to recognize the emotions due to its ability to model the sub-structure of feature sequences and hidden structure between features. Experiments with three face regions (forehead, left cheek and right cheek) show that the forehead is more relevant than the cheeks as a sole feature to recognize the valence. However, the best results were obtained using the combination of the forehead and the left cheek, which is consistent with research is neuropsychology. These findings were published in (Ramirez et al., 2014).

7.1.4 Emotion Recognition from Geometry and Color Information

Geometric and color features have shown to be useful features for emotion recognition. I presented results for emotion recognition combining features from geometric and color information. The experiments were performed using the AVEC dataset. Preliminary results using a straightforward combination of geometric and color information outperforms the average accuracy of individual features when using MLR. Using more sophisticated fusion techniques could lead to higher accuracy by taking advantage of additional information.

7.2 Future Work

This dissertation explored the problem of automatic emotion recognition. There are several related problems that could benefit from the findings in this dissertation, but also there are many other research possibilities to explore and suggestions for improvements.

Face Detection

For face detection, future work could be oriented in the following directions:

- Perform experiments using other boosting algorithms such as Float-Boost and Real AdaBoost to improve the accuracy. Other algorithms could help to create classifiers with a higher accuracy and faster detections.

- Use other optimization algorithms such as particle swarm optimization to improve the selection of Haar features. Due to the high density of possible configurations of asymmetric Haar features, it is difficult to determine the optimal set of features. A better optimization algorithm could lead to a better approximation to the ideal set.

- Take advantage of high resolution images for a more precise localization of the face. New cameras with high resolution are already available and a more precise detection could in application such as surveillance of biometric recognition.

- Test the approach on other object-detection problems. Face detection is a subproblem of object detection. The next step in the approach presented in this dissertation is apply it as a generic object detector.

Emotion Recognition

For emotion recognition, future work could be oriented in the following directions:

- Extracting more high-level features from different modalities such as audio and body gestures. There are still more cues that humans use to infer the emotional state of a person. For instance, the interaction of the hands with the face could tell us more about subtle emotions as shame or confusion. Also, physiological signals as heart rate or sweat can help to a better estimation of emotions

- Use a more accurate face tracker that can work under occlusions and noisy images. The ability to gather high quality facial features depends on the quality of the estima-

tion of face landmarks. Current face trackers are very sensitive to noise and depend on a good initialization.

- Explore more sophisticated methods for fusing different modalities and different kinds of features. The optimal approach to fuse modalities and features from different sources is still an open problem. In addition, some modalities are gathered at different sampling rates, which complicates the fusion. Thus, the exploration of new techniques to synchronize modalities could help to improve emotion recognition.

- Explore richer color descriptors and alternate color spaces. Better color descriptors could improve the performance of emotion recognition based on color. Also, different color spaces including L*a*b* or HSV could be useful.

- Take advantage of the temporal information included in the expressions of emotions. The duration and frequency of some high-level features such as the smile can bring extra information to infer the emotional state. A deeper study is required to determine the impact of temporal information in emotion recognition.

- The application of other learning algorithms can improve the performance of emotion recognition. In particular we will experiment with deep recurrent networks, as they have shown to be effective in similar problems.

- Use context information to discriminate emotions. Some expressions are ambiguous, for instance, crying could be related to a positive emotion but also to a negative emotion. Context information could help to reduce the error in ambiguous emotions. The problem here is related to how to gather the context information and how to exploit it.

- Extend the approach to a multi-user environment where the emotions of one subject can affect the emotions of other users. Humans use emotions to interact with other humans, hence emotion recognition should address the problem of multiple subjects interacting.

References

N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

M. Argyle and J. Dean. Eye-contact, distance and affiliation. *Sociometry*, 28:233–304, 1965.

J.P. Arias, C. Busso, and N.B. Yoma. Energy and F0 contour modeling with functional data analysis for emotional speech detection. In *Interspeech 2013*, pages 2871–2875, Lyon, France, August 2013.

Magda B. Arnold. *Emotion and Personality: Psychological aspects*, volume 1. Columbia University Press, 1960.

J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.

Yoshua Bengio and Yann Le Cun. Word-level training of a handwritten word recognizer based on convolutional neural networks. In *Proc. of the International Conference on Pattern Recognition*, pages 409–413. IEEE, 1994.

George Caridakis, Kostas Karpouzis, and Stefanos Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomputing*, 71(13-15):2553 – 2562, 2008. Artificial Neural Networks (ICANN 2006) / Engineering of Intelligent Systems (ICEIS 2006).

D. Chai and A. Bouzerdoum. A Bayesian approach to skin color classification in YCbCr color space. In *TENCON*, volume 2, pages 421–424, 2000.

Mark Changizi. *The vision revolution: How the latest research overturns everything we thought we knew about human vision.* Bella Books, Inc., Dallas, Tx, USA, 2009.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

C. Darwin. *The expression of the emotions in man and animals.* New York: Oxford University Press, 3rd. edition, 1998.

Ellen Douglas-Cowie, Cate Cowie, Roddy ans Cox, Noan Amir, and Heylen Dirk. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Corpora for Research on Emotion and Affect Workshop*, pages 1–4, 2008.

P. Ekman and W.V. Friesen. *Manual for the Facial Action Coding System.* Palo Alto: Consulting Psychologists Press, 1977.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200, 1992.

F. Eyben, M. Wollmer, M.F. Valstar, H. Gunes, B. Schuller, and M. Pantic. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.

Seeing Machines FaceAPI. Face API. 2011. `http://www.seeingmachines.com/product/faceapi/`.

J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

Bernhard Fröba and Andreas Ernst. Face detection with the modified census transform. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 91– 96, Erlangen, Germany, May 2004.

Guido Gainotti. Unconscious processing of emotions and the right hemisphere. *Neuropsychologia*, 50(2):205 – 218, 2012.

C. Garcia and M. Delakis. A neural architecture for fast and robust face detection. In *IEEE IAPR International Conference on Pattern Recognition*, pages 40–43, Quebec City, 2002.

D. Grandjean, D. Sander, and K. R. Scherer. Conscious emotional experience emerges as a function of multi level. *Consciousness and Cognition*, 17(2):484–495, 2008.

H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotion*, 1(1):68–99, 2010.

Mark Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1999.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletters*, 11:10–18, November 2009.

M. Hamouz, J. Kittler, J.-K Kamarainen, P. Paalanen, and H. Kälviäinen. Affine-invariant face detection and localization using GMM-Based feature detector and enhanced apparence model. In *6th International Conference on Automatic Face and Gesture Recognition*, pages 67–72, 2004.

C. E. Izard. *Human emotions*. Plenum Press, New York, 1977.

O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the Hausdorff distance. In *Third International Conference on Audio- and Video- based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 90–95. Springer, 2001.

K. J. Kirchberg, O. Jesorsky, and R. W. Frischholz. Genetic model optimization for Hausdorff distance-based face localization. In *International Workshop on Biometric Authentication*, pages 103–111. Springer, 2002.

J. Kovac, P. Peer, and F. Solina. Illumination independent color-based face detection. In *Third International Symposium on Image and Signal Processing and Analysis*, volume 1, pages 510–515, 2003.

N. C. Krämer. *Human behavior in military contexts*, chapter Nonverbal Communication, pages 150 – 188. Washington: The National Academies Press, 2008.

Angel F. Kuri-Morales. Efficient compression from non-ergodic sources with genetic algorithms. In *Fourth Mexican International Conference on Computer Science*, pages 324–329, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

S. M. Lajevardi and H. R. Wu. Facial expression recognition in perceptual color space. *IEEE Transactions on Image Processing*, 21(8):3721–3733, August 2012.

Steve Lawrence, C. Lee Giles, C. L., Ah Chung Tsoi, and Andrew D. Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1): 98–113, January 1997.

Richard Lazarus. *Emotion and adaptation*. New York: Oxford University Press, 1991.

S. le Cessie and J.C. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

Yann Lecun, Patrick Haffner, Lón Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Contour and Grouping in Computer Vision*. Springer, 1999.

Chul Min Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293 – 303, march 2005.

Stan Z. Li and ZhenQiu Zhang. Floatboost learning and statistical face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1112–1123, 2004.

Yong-Min Li, Shao-Gang Gong, Jamie Sherrah, and Heather Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 22 (5):413–427, May 2004.

Rainer Lienhart and Jochen Maydt. An extended set of Haar-like features for rapid object detection. In *International Conference on Image Processing*, volume 1, pages I–900–903, 2002.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1079 –1084, july 2010.

Hongying Meng and Nadia Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 378–387. Springer, 2011.

Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. Depression recognition based on dynamic facial and vocal expression features

using partial least square regression. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 21–30, New York, NY, USA, 2013. ACM.

Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, USA, 1998.

Louis-Philippe. Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, pages 1 –8, june 2007.

S. Nagaraj, S. Quoraishee, G. Chan, and K. R. Short. Biometric study using hyperspectral imaging during stress. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7674, pages 76740K–76740K–13, Orlando, Florida, April 2010.

M.A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *20th International Conference on Pattern Recognition (ICPR), 2010*, pages 3695–3699, aug. 2010.

M.A. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, 2011.

T. Ojala, 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*, 24(7):971–987, 2002.

OKAO. Vision software library. 2011. `http://www.omron.com/r_d/coretech/vision/okao.html`.

C.E. Osgood, W.H. May, and M.S. Miron. *Cross-cultural universals of affective meaning*. University of Illinois Press, Urbana, 1975.

M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

Constantine P. Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, pages 555–562, Washington, DC, USA, 1998. IEEE Computer Society.

Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

John C. Platt. Advances in kernel methods. chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18, 2010.

Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, January 2011.

J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

Geovany A. Ramirez and Olac Fuentes. Face detection using combinations of classifiers. In *The second Canadian Conference on Computer and Robot Vision, 2005*, pages 610–615, Victoria, British Columbia, Canada, 2005.

Geovany A. Ramirez and Olac Fuentes. Multi-pose face detection with asymmetric Haar features. In *IEEE Workshop on Applications of Computer Vision, 2008. WACV 2008.*, pages 1 –6, January 2008.

Geovany A. Ramirez and Olac Fuentes. Street detection with asymmetric haar features. In Luis Alvarez, Marta Mejail, Luis Gomez, and Julio Jacobo, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7441 of *Lecture Notes in Computer Science*, pages 398–405. Springer Berlin Heidelberg, 2012.

Geovany A. Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *1st International Audio/Visual Emotion Challenge and Workshop in conjunction with Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 396–406. Springer Berlin Heidelberg, 2011.

Geovany A. Ramirez, Olac Fuentes, Stephen L. Crites, Maria Jimenez, and Juanita Ordoñez. Color analysis of facial skin: Detection of emotional state. In *Computational Models for Social Interactions and Behavior (CMSI): Scientific Grounding, Sensing and Applications in conjunction with CVPR 2014*, 2014.

Andrew D. Richardson, Julian P. Jenkins, Bobby H. Braswell, David Y. Hollinger, Scott V. Ollinger, and Marie-Louise Smith. Use of digital webcam images to track spring green-up in a deciduous broadleaf forest. *Oecologia*, 152(2):323–334, 2007.

Claire L. Roether, Lars Omlor, and Martin A. Giese. Lateral asymmetry of bodily emotion expression. *Current Biology*, 18(8):329–330, April 2008.

Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998a.

Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Proceedings of 1998 IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, Santa Barbara, CA, June 1998b.

Paul Rozin and Adam B. Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3(1):68–75, 2003.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273 – 294, 1977.

S. Sai Pavan and C. Rajeswari. Emotion recognition for color facial images. *International Journal of Emerging Trends in Engineering and Development*, 2(3), May 2013.

Jason M Saragih, Simon Lucey, and Jeffrey Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009.

K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, pages 92–120, 2001.

Stefan Scherer, Giota Stratou, Marwa Mahmound, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. Automatic behavior descriptos for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, Shanghai, China, April 2013.

Henry Schneiderman and Takeo Kanade. A statistical model for 3-D object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2000.

Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. AVEC 2011 The first international audio/visual emotion challenge. In *First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*. Springer LNCS, 2011.

Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6): 803 – 816, 2009.

Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.

Sylvan S. Tomkins. *Affect Imagery Consciousness: The Positive Affects*, volume 1. New York: Springer, 1962.

Sylvan S. Tomkins. *Affect Imagery Consciousness: The Negative Affects*, volume 2. New York: Springer, 1963.

Ming-Jung Seow; D. Valaparla and V.K. Asari. Neural network based skin color model for face detection. In *32nd Applied Imagery Pattern Recognition Workshop*, pages 141–145, 2003.

Michel F. Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, ICMI '07, pages 38–45. ACM, 2007.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of 2001 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

Paul Viola, Michael Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

Martin Wöllmer, Florian Eyben, Stephan Reiter, Bjrn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, pages 597–600. ISCA, 2008.

Bo Wu, Haizhou Ai, Chang Huang, and Shihong Lao. Fast rotation invariant multi-view face detection based on real adaboost. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 79–84, 2004.

Rong Xiao, Ming-Jing Li, and Hong-Jiang Zhang. Robust multipose face detection in images. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):31–41, 2004.

T. Yamada and T. Watanabe. Effects of facial color on virtual facial image synthesis for dynamic facial color and expression under laughing emotion. In *13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 341–346, September 2004.

T. Yamada and T. Watanabe. Analysis and synthesis of facial color for the affect display of virtual facial image under fearful emotion. In *Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005).*, pages 219–224, May 2005.

T. Yamada and T. Watanabe. Virtual facial image synthesis with facial color enhancement and expression under emotional change of anger. In *The 16th IEEE International Symposium on Robot and Human interactive Communication.*, pages 49–54, August 2007.

Peter Yuen, Tong Chen, Kan Hong, Aristeidis Tsitiridis, F Kam, James Jackman, David James, Mark Richardson, L Williams, William Oxford, Jonathan Piper, Francis Thomas, and Stafford Lightman. Remote detection of stress using hyperspectral imaging technique. In *3rd International Conference on Crime Detection and Prevention (ICDP 2009)*, pages 1–6, 2009.

Vittorio Zanella, Geovany A. Ramirez, Hctor Vargas, and LornaV. Rosas. Automatic morphing of face images. In Mikko Kolehmainen, Pekka Toivanen, and Bartlomiej Beliczynski, editors, *Adaptive and Natural Computing Algorithms*, volume 5495 of *Lecture Notes in Computer Science*, pages 600–608. Springer Berlin Heidelberg, 2009.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

C. Zhang and T. Chen. *The Handbook of Video Database Design and Applications*, chapter From low level features to high level semantics. CRC Press, 2003.

Jun Zheng, Geovany A. Ramirez, and Olac Fuentes. Face detection in low-resolution color images. In Aurlio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, volume 6111 of *Lecture Notes in Computer Science*, pages 454–463. Springer Berlin Heidelberg, 2010.

Feng Zhou, F. De la Torre, and J.F. Cohn. Unsupervised discovery of facial events. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574 –2581, 2010.

Curriculum Vitae

Geovany Abisaí Ramírez García was born in Puebla, Mexico. The second son of Abisai Ramirez and Gloria Garcia. He earned his Bachelor in Computer Engineering from the UPAEP University at Puebla, Mexico in 2003. In 2006 he received his Master of Science degree in Computer Science from the National Institute of Astrophysics, Optics and Electronics at Puebla, Mexico. In 2006, he joined the doctoral program in Computer Science at The University of Texas at El Paso. While pursuing his doctoral degree, Geovany worked as visiting research assistant at the Institute for Creative Technologies of University of Southern California for the summers of 2010 and 2011. His main research was focused on the development of an artificial intelligence system for emotion recognition. From 2011 to 2014 he worked as research assistant at the System Ecology Lab of The University of Texas at El Paso, where he designed and developed systems for automatic phenology and remote sensing. Geovany has published six articles related to his doctoral research and also he participated in several international conference meetings and workshops, including the First International Audio/Visual Emotion Challenge where he and his team won the first place.