Optical flow for dynamic facial expression recognition

Athanasios Psaltis 3780198

Supervisors :
dr Robby T Tan
dr Nico van der Aa

October 17, 2013

Abstract

Facial expressions are critical in human face-to-face communication because they reflect emotion and often show a person's mental state. Usually the most distinctive emotions are conveyed through facial expressions. In fact, they are a form of nonverbal communication, which helps to tell the story, while describing the character. The face of a human has the foremost relevant visual characteristics of personality and emotion. Current existing methods fail to recognize facial expressions in an automatic way, because they require the manual detection of points in neutral face. In this master thesis project we have developed and analyzed a fully automatic facial expression classification system that uses sequences of frames describing the dynamics of the expressions. This includes face detection, face registration using landmark points of the face, feature extraction using optical flow and finally classification between six basic expressions. We present our fully integrated system which performs these operations accurately and in real time and represents a major step forward in our aim for achieving automatic human-to-human interaction recognition.

After detecting the face using Viola-Jones, an off-the-shelve face tracking library is used to get the landmark points of the face. The face is aligned in a stable window using an affine transformation technique. Finally, dense optical flow methods are used to get a motion representation of which magnitude and angle parameters are derived. Classification is implemented using a multi-class technique that extends the binary SVM classifier. Captured video streams of posed facial expressions of several subjects are used to train the pipeline and test its performance. Extensive experiments on the Cohn-Kanade database illustrate that this approach is effective for facial expression analysis. Furthermore, comparative recognition results indicate the advantage of our system over existing recognition systems.

The contributions of this paper are twofold: (1) to have a fully automated way to measure spontaneous facial expression and (2) to incorporate optical flow to cope with temporal information to enhance the overall recognition results. The results of the classification indicate that the correct alignment has a significant effect on the accuracy of the system. Furthermore, we investigate the possibility of combining information of both subjects to learn a structured SVM classifier for an interaction set, whose output is their interaction labels. On the other hand, still remains a challenge to recognize transitions between consecutive expressions. The techniques treated in this paper are useful for anyone working with facial feature extraction or facial recognition, as well as related fields that desire a way of model training. The study provides insights into the potential and limitations of automated measurement of facial motion.

Contents

Chapter 1

Introduction

1.1 Motivation

Everyday millions of people interact, and express their feelings in many ways. An essential part of face-to-face communication in human relationships is the interaction of expressions, which makes it a difficult task. Humans have the ability to read emotions from someone's face, but for computers this still is a challenging task. Facial expressions provide an important behavioral measure for the study of emotion, and social interaction. An efficient measurement of facial expression is instrumental for understanding the functionality of face-to-face communication. The Facial Action Coding System(FACS) [6] is an objective method for quantifying local facial movement in terms of component action. Using FACS and viewing video-recorded facial behavior at low frame rate and slow motion, coders can manually code nearly all possible facial expressions, which are decomposed into Action Units (AUs). FACS consists of 44 AUs, including those for head and eye positions. AUs are anatomically related to contraction of specific facial muscles. They can occur either singly or in combinations.

Despite attempts to recognize the interaction between humans and machines, recognizing human-to-human interaction has not been studied extensively. Furthermore, current existing methods fail to recognize facial expressions in an automatic way, because they require the manual detection of points in neutral face. Thus, the necessity of an automatic algorithm that could measure the facial expressions and recognize the label of a human-to-human interaction arises. In this work we investigate the possibility of measuring the intensities of spontaneous facial motion in an automated way. The results of such research could be used as an effective tool in behavioral science and social interaction applications. Imagine a world, in which machines would be able to understand the mood of a crowd in open spaces and use their knowledge to provide help where needed.

Automating the analysis of facial motion and face-to-face interaction would be beneficial for fields as diverse as security, medicine and marketing. In security contexts, facial expressions play a crucial role in the detection of the mood of the crowd in hardly controlled environments such as demonstrations. In medicine, facial expressions are the direct means to identify the medical state of the patient, for example when specific mental processes (e.g., pain, depression) are occurring. In marketing, facial expressions inform the publisher of the need to adjust the advertising message, by analyzing the mood of the crowd in public places. As far as interfaces between humans and machines are concerned, where computers take on a social role, it may enhance their functionality to recognize facial expressions of the user. It

is obvious that this is going to have an impact on our everyday life by enhancing the way we interact with computers or in general, our surrounding living and work spaces.

1.2 Background

Currently, most research dealing with facial feature extraction is focused on understanding and categorizing certain facial expressions. In a real face-to-face communication, we deal with facial expressions, typically occurring in uncontrolled conditions and combined with head pose variation and head movement. Ekman and Friesen devised the FACS to categorize every type of facial expressions, where movements on the face are described by a set of AUs. Contemporary research has indicated that using AUs can do more detailed classification [6]. These units indicate a certain motion of some muscle groups in the face. A specific muscle status of one of these groups is called an AU. Using FACS, trained human experts can manually code nearly any anatomically possible facial expression, deconstructing it into the specific AUs and their temporal segments that produced the expression. A specific combination of AUs can tell us what kind of emotion the subject is expressing in a more detailed manner. An example of such combinations exhibits the list of Table 1.1. Combining AUs from different subjects in a simple interacting scenario, we can draw conclusions about the label of the interaction. A more robust representation of the interaction could be acquired by including temporal information of the facial motion.

Table 1.1: AU combinations

Emotion	Action Units
Surprise	1+2+5B+26
Sadness	1+4+15
Happiness	6+12
Fear	1+2+4+5+20+26
Anger	4+5+7+23
Disgust	9+15+16

Psychologists and engineers alike have tried to analyze facial expressions. Although recognition of different emotions is a field of research of psychology studies, rather than computer vision, Ekmans' work inspired many researchers to analyze facial expressions by means of image and video processing. By tracking facial features and measuring the amount of facial movement, they attempt to categorize different facial expressions. Computer vision has the potential to objectively document continuous changes in human behavior, and automatically measure the subtle changes in motion of facial muscles. In particular, the AUs are said to be more objective and refer to what was actually performed by the subject rather than its interpretation. As AUs are independent of any interpretation, they can be used for any higher order decision-making process including recognition of basic emotions, or pre-programmed commands for an ambient intelligent environment. Figure 1.1 shows a sample of FACS coding of a fear expression.

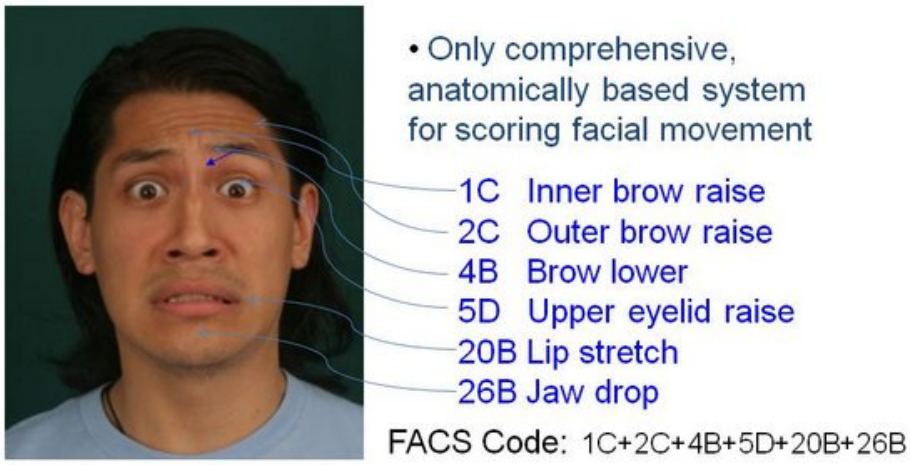

Figure 1.1: AUs and Fear expression

1.3 Goals

In a real face-to-face communication, we deal with non-posed facial expressions. Posed facial expressions are those that are created by asking subjects to make specific facial actions or expressions. Although our system is trained with posed facial expressions, it is still a challenge for us to test it under unconstrained conditions, such as lighting, as well as variation within the faces of different people and different poses.

We are interested in modeling human interactions, focusing on the recognition of interactions between two people in simple scenarios. The approach we follow is person-centric. First we track and analyze the facial motion of each person separately in a multi-camera video system. Combining the facial expression temporal graphs of the two person's simultaneously, we might be able to infer the types of the interaction.

The objective of this work is to develop a fully integrated system which detects faces on an image, extracts and tracks facials features, extracts the information conveyed in facial features and classifies the expressions accurately in real time. Our results will be assessed based on the ability of measuring the intensities and labeling the expression of spontaneous facial motion. As secondary goal would be to investigate the possibility of combining information of both subjects in a structured learning way. It is required to find the appropriate techniques to implement each step of the process and test the behavior, results and performance of each of these steps. Finding the limitations of this method and their possible explanations is a first step in improving it. Finally, in this study we evaluate the results of our method, by comparing the labels acquired in a subject interaction scenario using FaceReader system by Noldus Information Technology [4].

1.4 Thesis layout

The remainder of this paper is organized as follows. Chapter 2 reviews approaches to related to our project. Chapter 3 will provide an overview and analysis of the basic pipeline to get from an input video to an output classification. Moreover it presents an in-depth explanation of the techniques that were used in this master thesis project. In Chapter 4, a number of

experiments will be described and the results will be discussed. Finally, Chapter 5 concludes the research and list possibilities for future work.

Chapter 2

Related work

For many decades, developing a fast, accurate and robust automated system for recognizing a face and facial expressions has been a goal in computer vision. Initial research tried to represent all possible expressions in a robust coding scheme. Since the classification into AUs is based on facial anatomy, practically all expressions can be represented by FACS [6] without being associated to specific emotions. For action units that vary in intensity, a scale of 5 levels is used to measure the degree of muscle contraction. Although the number of atomic AUs is small, more than 7,000 combinations of AUs have been observed. FACS provides the necessary detail with which to describe facial expression. Other systems use emotion labels to describe an expression, by doing one-to-one mapping between facial expression and emotion. Such dynamic approaches detect emotions using spatio-temporal information extracted from the image sequences, that consists of the changes in the landmark configuration as well changes in the textures appearance of the faces.

Many papers in automatic facial expression recognition literature perform the analysis or recognition of expressions by considering both approaches [15] [38]. Several researchers have tried to recognize AUs [17] or[31]. The system of Lien et al.[17] used dense-ow, feature point tracking and edge extraction to recognize 6 upper face AUs or AU combinations and 9 lower face AUs and AU combinations. Examples of works which deal with emotion expression feature analysis are Lyons et al.[20] or Wen et al.[37]. The system of Lyons et al. synthesizes aspects of two major streams of research on face processing: the labeled elastic graph matching approach and Eigenface algorithms. There are also papers which consider the classification of facial expressions using both AUs and emotion expressions [18]. By relying only on FACS features, we may miss relevant behavioral patterns that are not captured by the coding system. Hence, our method uses both the AUs and emotion expression feature patterns extracted from facial expressions in video streams.

To train and evaluate the proposed systems, it is necessary to develop robust algorithms bearing in mind the appropriate databases containing variabilities of subjects with respect to races, ages, genders, illumination conditions and poses. Some examples of relevant facial expression databases for such task are: CohnKanade database [14], Japanese woman database (JAFFE) [13], MMI database [21]. Different survey papers [30] [15] [26] summarize the main facial expression analysis techniques and algorithms. Generally, many different approaches in this field consider three characteristic phases [30] which are face detection and registration, feature extraction and expression classification. They differ mainly in the features extracted from the video images and in the classifiers used to distinguish between the different emotions.

Emotions are displayed over time and therefore we believe a more robust representation of the interaction could be acquired by including temporal information of the facial motion. Tian et al.[30] classifies the facial features that model the facial changes into: (a) deformation features and (b) movement features. Deformation features do not have the information of the pixel movement into account, and they can be obtained from static images. Movement features are focused on the facial movements and they are applied to video sequences. A relevant technique to obtain such features is dense optical flow [8]. Although this paper requires manual detection of points in the neutral face, the integration of optical flow with image alignment techniques increases the system stability and improves the facial movement interpretation and related facial expression analysis. In our method we remove the necessity of manual labeling, but utilize the advantage of optical flow to incorporate stable features over time.

Mase [22] used optical flow to recognize facial expressions and estimate the activity of 12 facial muscles. He was one of the first to use image processing techniques to analyze facial expressions. Rosenblum et al.[27] also computed optical flow of regions on the face, where a radial basis function network was applied to classify expressions. Essa and Pentland [8] used an optical flow region-based method to recognize expressions. Donato et al.[5] tested different features for recognizing facial AUs and inferring the facial expression in the frame. Otsuka et al.[25] first computed optical flow, then computed the 2D Fourier transform coefficients, which were used as feature vectors for a hidden Markov model (HMM) to classify expressions. A similar approach, using dense optical flow, feature point tracking and high gradient component analysis in the same hybrid system, was proposed by Lien et al.[16], where hidden Markov models (HMMs) were used for the recognition of 15 AUs.

Dense sampling has shown to improve results over sparse interest points for image classification [9],[24]. The same has been observed for action recognition in a recent evaluation by Wang et al. [35], where dense sampling at regular positions in space and time outperforms state-of-the-art space-time interest point detectors. Also related to our approach, different papers have recently proposed dense optical flow techniques to increase the robustness of the motion tracking of facial points in video sequences. In the work of Wang [36], motion trajectories are obtained by tracking densely sampled points using optical flow fields between successive frames. Sanchez [28] used the displacements of the facial points which are densely and uniformly placed on a grid centered on the central face region. They compute the displacements of different sets of points along the sequence of frames describing each facial expression (i.e. from neutral to apex). By combining holistic spatial analysis and dense optical flow with local features in a hybrid system, Sanchez et al. increased accuracy to 95.4% correct. Yet, their system requires the manual detection of the considered pre-processing and feature points in the first (neutral) frame of the sequence.

Chapter 3

Theory

3.1 Basic pipeline

This section explains the basic pipeline to get from an input video to a final facial expression classification. In general, each of the different steps can be combined and interchanged with other methods. Each step is influenced by the previous step. Furthermore, different alignment methods can be combined, to get more accurate results.

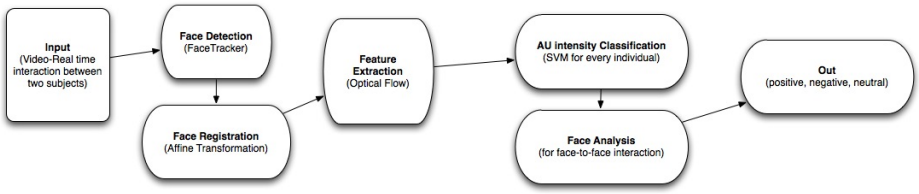

Figure 3.1: Basic Pipeline

As we stated in the previous chapter, most facial expression recognition approaches have three main phases: a) face detection and registration, b) feature extraction and c) expression classification. At the first stage, we accurately detect the subject's face in the image to facilitate a correct initialization of the feature extraction and to reduce the amount of pixels for further processing. Once the face is detected, we locate sixty-eight landmark points on the surface and pass it to the next step of registration. The more accurate the landmark points are, the more robust the alignment process becomes. Then, given a set of annotated frames, the registration step ensures stabilization of a video for a automatically defined rectangle that the system calculates based on landmark points. At the second phase we apply dense optical flow, where motion paths are generated and facial features acquired. We want to extract the spatio-temporal information from the frame sequences and get a robust representation of the motion as our feature descriptors. Finally, we use these features to distinguish between basic facial expression and estimate the activity of facial muscles (AUs). For this purpose, we build a multi-class SVM classification model from these descriptors. Results of every frame are given as an input into the next frame and the the combination of those lead us to more accurate conclusions. Classification results from two subjects could be combined to get the interaction classification. All of the methods discussed in previous chapter require

the manual detection of points in the neutral face. Our method does not suffer from this limitation, because it is based on a fully automated detection technique.

3.2 Face detection and registration

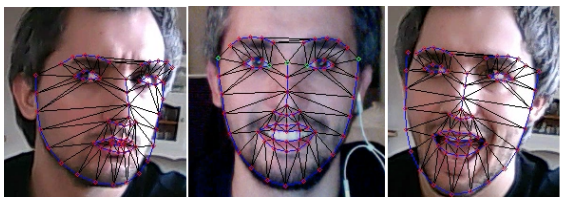

Figure 3.2: Landmark point detections from FaceTracker.

Each frame extracted from the video contains irrelevant information (visual data) for facial expression extraction. Face detection is used in order to remove the irrelevant visual data and isolate the face from the full image for further processing. The output of this phase is crucial, since the video should be examined and analyzed in real-time. Non-rigid face alignment and tracking is a common problem in computer vision. Despite recent advances, accurate face detection and alignment remains a challenging task in dynamic environments, such as video, where noise conditions, illumination, occlusion, and the subject's location and pose can vary significantly from frame to frame. A method that exploits the spatial relationships between feature points could provide robustness in real-world conditions.

For the face detection and registration we use the Face tracking library (FaceTracker) [29]. FaceTracker is trained on a dataset from CMU called the Multi-PIE database [11]. This database has a huge collection of images of faces that are all marked up with information about the position of the face and where the features are exactly located. Multi-PIE database is constrained to real world situation, containing a lot of different environmental conditions. Particularly, it contains images with people in daylight under directional lighting or indoor with lighting above their head. Hence, FaceTracker has been trained to those situations and is best suited for tracking. Note that face detection fails when the system tries to analyze situations that are really unusual in lighting conditions or in facial features.

The face detection in FaceTracker is done by applying the Viola-Jones algorithm [34], a widely used method for real-time object detection. It learns the detection classifier from labeled training data (faces and non-faces images). The algorithm adapted the idea of using Haar wavelets and developed the so-called Haar-like features. A Haar-like feature considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel

Table 3.1: Detection results produced by FaceTracker in different situations.

Indoors	95
Outdoors	87
Face-like objects	70
Occluded	40
Faces-Angle	81
Average	75

intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. FaceTracker [29] uses the Viola-Jones function. When the face is found, FaceTracker runs a model fitting algorithm to register a parameterized 3D face model onto the target face. All of the different points on the face are being detected using small patches of the image. FaceTracker returns the final landmark points after making iterative refinements to the estimations.

Figure 3.3: FaceTracker detects faces outdoors.

The question is how well this face and landmark point detector performs in uncontrolled situations. Applying FaceTracker on video streams which incorporate faces at an angle, or with open mouths or other non-neutral facial expressions, show a stable landmark point detection result in most cases. Despite the visually good results, we spotted a divergence from frame to frame at landmark point locations. We assume that this divergence is due to illumination changes. Of course, we could also set the points manually but it was not the purpose of this project, because we wanted our method to be based on automated detection and be able to work at real-time.

Image registration or image alignment algorithms can be classified into intensity-based and feature-based. One of the images is referred to as the reference or source and the others are referred to as the target or sensed images. Image alignment consists of moving, and possibly deforming, a template to minimize the difference between the template and (part of) an image. Intensity-based methods compare intensity patterns in images via correlation metrics, while feature-based methods find correspondences between image features such as points, lines, and contours.

The image registration method we apply is the inverse compositional algorithm [2], which finds an affine transformation to match the image to the template. An affine transformation is any transformation that preserves collinearity (i.e. all points lying on a line initially still lie on a line after transformation) and ratios of distances (e.g. the midpoint of a line segment remains the midpoint after transformation). An affine transformation is a composition of rotations, translations, scaling. This affine transformation is found by minimizing the sum of squared error of the template and the image warped back onto the coordinate frame of the template. The inverse compositional algorithm is a computationally efficient method to find these warp parameters. Our choice is motivated by the ability of the proposed algorithm to deal with faces with unseen intra-class appearance variation, while achieving significant improvement in computational efficiency. Figure 3.4 shows an example of this image registration process.

The inverse compositional algorithm can be described as follows : First we pre-compute the Hessian matrix, then we warp the Input image based and the current estimated warp and

compute the new warped image. Next we compute its local warp parameters, while updating the current estimated warp with the inverted incremental warp. Step 1 is done only once, Step 2 to 4 are iterated until warp is smaller than a threshold. The algorithm is described in details in the work of Baker [2]. Our choice is motivated by the ability of the proposed algorithm to deal with faces with unseen intra-class appearance variation, while achieving significant improvement in computational efficiency. Figure 3.4 shows the result of inverse compositional algorithm applied to an image.

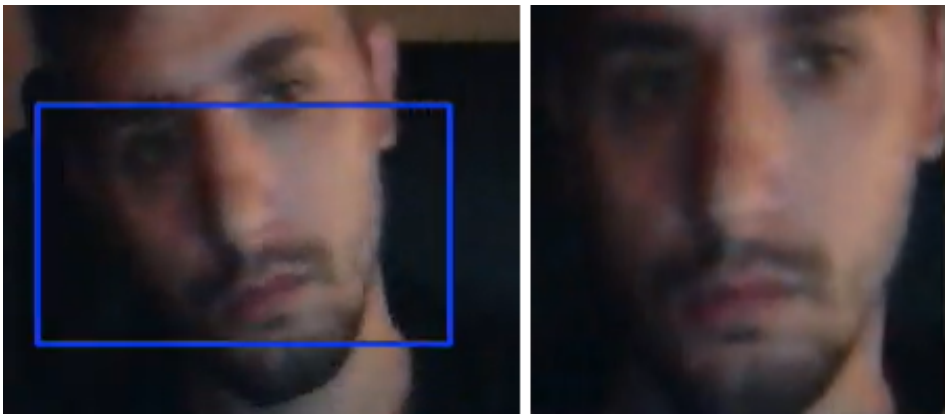

Figure 3.4: Face alignment example

3.3 Feature extraction

Feature extraction is applied to the face image to obtain interesting features that allows us to locate facial motion changes and distinguish between different intensities of AUs. Determining an adequate facial image representation for effectively measuring the intensity of facial expressions and action units is a challenging problem. Optical Flow [22] is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene. The concept of optical flow was introduced by the American psychologist James J. Gibson in the 1940s to describe the visual stimulus provided to animals moving through the world. It is a very challenging task to estimate optical flow for videos in which either foreground, or background exhibits remarkable motion variation, or those with low resolution due to artifacts like noise, occlusion and motion blur. A warping-based method as a pre-processing step could increase optical flow accuracy.

3.3.1 Optical flow - basics

Sequences of ordered images allow the estimation of motion as either instantaneous image velocities or discrete image displacements. The optical flow methods try to calculate the motion between two image frames. These methods are called differential (or gradient-based) since they are based on local Taylor series [10] approximations of the image signal. A common starting point for optical flow estimation is to assume that pixel intensities are translated from one frame to the next,

A motion flow field can be defined as the ideal representation of 3D motion as it is projected onto an imaging plane. To calculate the motion flow, we must determine the motion of each image point. There is no unique solution(ill-posed) for the computation of optical flow. Horn and Schunck begin with the initial hypothesis that the intensity structures of local time-varying image regions are approximately constant under motion for at least a short duration [12]. Formally, if $I(x, y, t)$ is the image intensity function for pixel x, y at time t, then:

$$I(x, y, t) \approx I(x + \delta x, y + \delta y, t + 1),$$

where $w = (\delta x, \delta y, 1)T$ is the displacement vector between images at time t and $t + 1$. This is the grey value constancy assumption. Linearizing this equation yields the optical flow constraint equation:

$$I_x \delta x + I_y \delta y + I_t = 0$$

where subscripts denote partial derivatives (see [12] for the full derivation). This constraint is not sufficient to compute both x, y components of the velocity vector, and demonstrates what is known as the aperture problem [32]. Only at image locations where there is sufficient intensity structure (or curvature) can the motion be fully estimated. For the same reason, the velocity of homogeneous surfaces cannot be recovered. Specular highlights are problematic because they remain static in position through motion. Illumination changes have the reverse effect and can cause optical flow to be registered when there is none. Shadows and shading will vary causing intensity values to change in the image and break the grey value constancy assumption.

It is useful, then, to sometimes ignore the grey value constancy assumption. Uras et al. introduce the gradient constancy assumption in [33]. Simply put, the gradient can be assumed not to vary due to the displacement:

$$\nabla I(x, y, t) \nabla I(x + \delta x, y + \delta y, t + 1),$$

To help solve the aperture problem, smoothness assumptions are almost always made. For much of the image, pixels will move similarly to their neighbors. In cases where there is no gradient or insufficient local structure, adding smoothness can produce more accurate results. Smoothing can be performed spatially or spatio-temporally. Smoothness assumptions are often violated at object boundaries, and so it is sensible to attempt generation of a piece-wise smooth flow field.

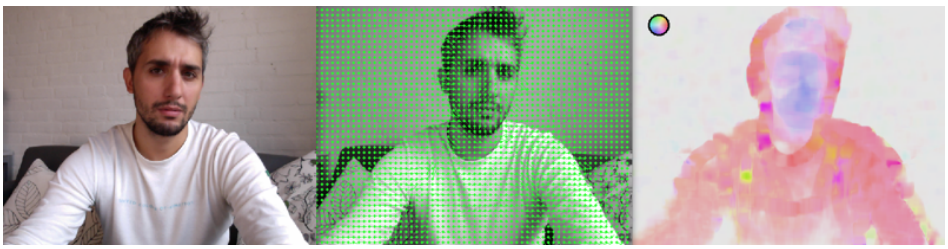

Figure 3.5: Dense optical flow (color encodes the direction of moving pixels, intensity their speed).

3.3.2 Lucas-Kanade optical flow

The LucasKanade algorithm [3] [19] for computing the optical flow has been used to estimate the displacement of facial points. This algorithm is one of the most popular differential (or gradient-based) methods for motion estimation computing in a video sequence. It approximates the motion between two frames which are taken at times t and $t + \delta t$ at every pixel position assuming a brightness constancy. The algorithm also assumes that the motion between two consecutive frames is small(points do not move very far) and that the points move the same way as their neighbors. Affine transformation at the previous stage, makes our model more robust as it eliminates large displacements.

Given two neighboring frames I_{t-1}, I_t, for a point $p = (x, y)^T$ in I_{t-1}, if the optical flow is $f = (u, v)^T$, then the corresponding point in I_t is $p+f$. Lucas-Kanade optical flow aims to find the offset f to minimize the match error between the local appearances of two corresponding points. That is, we can define a cost function upon $N(p)$, the local area of p:

$$e(f) = \sum_{x \in N(p)} (w(x)(I_t(x + f)I_{t-1}(x))^2$$

where $w(x)$ is the weight function. Optimize the above equation and we can get the solution:

$$f = G^{-1}H$$

where $G = \sum_{x \in N(p)} w(x)\nabla I_t(\nabla I_t)^T, H = \sum_{x \in N(p)} w(x)\nabla I_t \Delta I, \Delta I = I_{t-1} - I_t, \nabla I_t = \frac{dI_t}{dx}$.

Lucas-Kanade optical flow demands feature points with salient local appearances, or it can not track accurately. The global video sequence displacement vector for each considered point in any type of facial expression is computed by applying the LucasKanade algorithm between pairs of consecutive frames. These corresponding inter-frame displacement vectors are then added to obtain the global displacement vector corresponding to each point along the expression (i.e. from neutral face to apex).

In the dense flow tracking method, face normalization requires to locate a rectangular bounding box containing the central face and to compute the face angle normalization. This bounding box becomes more stable when points around the eyes and the symmetric axis are used as reference. Dense optical flow uses a grid of uniformly distributed points around the central facial region. Since the images in the considered databases have a certain spatial resolution (640 x 480 in Cohn-Kanade database), and the neighboring pixels in the face along consecutive frames present a high temporal correlation, the application of the Lucas-Kanade algorithm to each point of the considered facial region becomes computationally very expensive and not useful at all. Therefore, we first applied a two-level Gaussian pyramid to reduce 1/16 the whole number of points and also deal with larger movements. Hence, the facial movement vectors between frames now become smoother and equally representative, but the workload is reduced significantly. Now, the global displacement vector of each analyzed point is computed, and their corresponding modules and angles are obtained using Lucas-Kanade algorithm. In consequence, a global displacement vector of 1201 features (i.e. the module and angle of the 600 tracked points with the labeling of the expression) is extracted as pattern for each expression video sequence. Using Lucas-Kanade algorithm, we can achieve real-time computation of dense optic flow fields on a standard PC. Our choice is motivated by the ability of the proposed algorithm to capture subtle motions on faces, while being robust under noise. Figure 3.5 shows the result of optical flow algorithm applied to an image.

3.4 SVM classification

Support Vector Machines (SVMs) are a supervised classification technique. Moreover, they have been successfully applied to facial expression classification. They have gained prominence because they are robust, accurate and are effective even when using a small training sample. SVM is a supervised learning technique used for both classification and regression problems and it is derived from the statistical learning theory. By their nature SVMs are essentially binary classifiers, however, they can be adopted to handle the multiple classification tasks. The two approaches commonly used are the One-Against-One and One-Against-All techniques [23]. The problem that SVM tries to solve is finding the maximal margin separating hyperplane with respect to the training test that correctly classifies the data points, using specific kernel functions, by separating the points into two classes as much as possible.

The One-Against-All approach represents the earliest and most common SVM multi-class approach and involves the division of an N class dataset into N two-class cases. If say the classes of interest in a face image include eyes, mouth and eyebrows, classification would be affected by classifying eyes against non-eye parts i.e. (mouth and eyebrows) or mouth against non-mouth parts i.e. (eyes and eyebrows). The One-Against-One approach on the other hand involves constructing a machine for each pair of classes resulting in $N(N-1)/2$ machines. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes. This approach can be further modified to give weighting to the voting process. From machine learning theory, it is acknowledged that the disadvantage the One-Against-All approach has over One-Against-One is that its performance can be compromised due to unbalanced training data-sets (Gualtieri and Cromp, 1998), however, the One-Against-One approach is more computationally intensive since the results of more SVM pairs ought to be computed.

In SVM classification, it is possible to choose the specific kernel mapping suited to the application. However, in many works, some common types of kernel functions are used, in special the Gaussian, linear, or polynomial ones. Furthermore, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Based on the kernel mapping approach, every inner product is replaced by a nonlinear kernel function $K(x, y) = \phi(x) \cdot \phi(y)$ where x and y are two input data sets. There are different types of kernel mapping such as the polynomial kernel and the RBF kernel. Because of the highest performance in classifying the intensity of action units RBF kernel is used. It appeared to be easy to use and computationally efficient.

3.4.1 LibSVM

We used LibSVM package [1] for our facial classification experiments, as its tools were suitable for our classification purposes. LIBSVM is an integrated software library for support vector classification, regression and distribution estimation (one-class SVM). Besides that, it supports an efficient multi-class classification. In addition, it provides a simple interface where users can easily link it with their own programs. Some more features of LIBSVM are cross validation for model selection, probability estimates, various kernels, and weighted SVM for unbalanced data. Note that RBF kernel is already implemented in LibSVM. As usual, LibSVM requires from a training and a testing stage. During training, a set of the SVM were systematically adjusted. Once the SVM model has been trained, we use the test set of facial expression sequences to get the performance measurements. Since we wanted to know

the percentage of presence of each type of expression against all the others, we adopted the one-versus-all multi-class classification technique implemented in LibSVM. In this approach, binary classifiers are trained to distinguish each type of expression from the other ones, and the binary classifier with the highest output determines to which class the expression is assigned. Parameter selections of SVM have been systematically adjusted by line search for linear, polynomial and Gaussian kernels, and those parameters producing best classification results were selected. In general, we used approximately 75% of the image sequences to train the SVM and the remaining 25% for testing purposes.

Chapter 4

Experimentation

In this section, we will present the experiments performed to obtain more knowledge about the facial feature extraction. We will also investigate the performance of the our framework. All of our experiments have been performed on a Core i5 2.7-GHz MacOS 10.8.4 machine with 4 GBytes of memory.

4.1 Database description

In principle, facial expressions should be analyzed on spontaneously collected face videos. However, most of the spontaneous facial expression databases are not open for public use and manual labeling of spontaneous and naturally displayed facial emotions is difficult and error prone. Hence, we had to evaluate our facial expression recognition method on the well known and widely used Cohn-Kanade database [14]. The Cohn-Kanade database is actually the most widely used for recognizing facial expressions and it provides variability of subjects with respect to races, ages, genders as well as small changes in illumination conditions. The facial expression database consists of approximately 500 frame-sequences of about 100 different subjects ranging in age from 18 to 30 years who enacted among many others, the six prototypical emotions anger, disgust, fear, joy, sadness, and surprise. Image sequences of frontal faces showing posed facial expressions always begin with the emotion from neutral state to highest intensity level or apex as described in Table 4.1. These sequences were captured at 12 frames per second into 640 x 480 resolution with 8-bit precision for gray scale values.

There is a considerable number of subjects for which the number of expression sequences is reduced (i.e. only two expressions per subject). When some of the classes are represented by a significantly less number of instances than the other ones, hinder the learning performance of classification algorithms. This situation has been described as the class imbalance problem in Ertekin et al.[7]. Due to this, in our experiments we have considered the same numbers of samples per class, both in the training and in test sets. A total of 246 expression sequences were selected from the database. These sequences came from 41 random subjects with the 6 emotions per subject.

Table 4.1: AUs intensity description

Value	Intensity Name
0	Absent
1	Trace
2	Slight
3	Marked
4	Extreme

4.2 Face registration, Experiment 1

How accurate must the alignment be or can you still classify emotions with a less perfect face alignment? To answer this question we vary the alignment method each time on video streams captured in our lab. Looking in our basic pipeline, registration step takes as an input the location of the face in the image and tries to deform a template to minimize the difference between the template and the face. We compared three different alignment techniques: (1)Alignment based on region of interest(ROI), (2)alignment based on horizontal and vertical face axis and (3) inverse compositional algorithm for image alignment.

First method's basic idea, is to locate the center of the face and track it between successive frames. The tracking windows always shows the region of interest, which in our case is the face. This alignment method is robust under translation transformation but fails when the face turns at a certain angle (see figure 4.1).

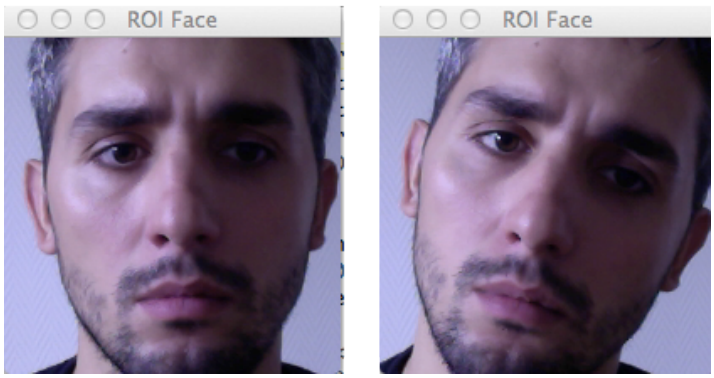

Figure 4.1: Alignment based on ROI

In our second alignment method we tried to eliminate the previous limitations and develop a robust method under translation, rotation and scale transformations. For this purpose, again we locate the center of the face and translate it in that way that both centers (face, template) match. Then, we simply measure the angle and the distance between the two eyes based on horizontal axis and rotate the image by that angle. We repeat the process for the vertical axis by measuring the angle between the nose and the mouth center. Last step, we use the measured distances to apply a scale factor on the final result. With this, we ensure that the alignment is robust under both rotations and scales (see figure 4.2).

Despite the good results of the previous alignment technique, we encounter a number of

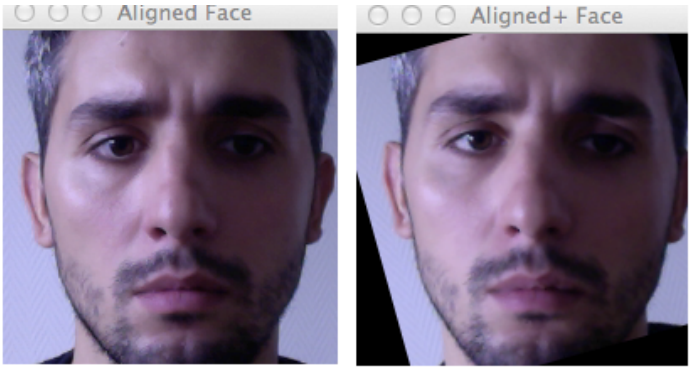

Figure 4.2: Alignment based on Horizontal and Vertical angles.

artifacts while varying the position of the face in the image and the roll angle in real time. These small artifacts had as a result the incorrect classification of basic expressions, even when the face had a neutral expression. Our third alignment technique was extensively described in the Theory chapter 3. We observed significant changes in the stability of the alignment window, as well as in the classification of spontaneous facial expressions.

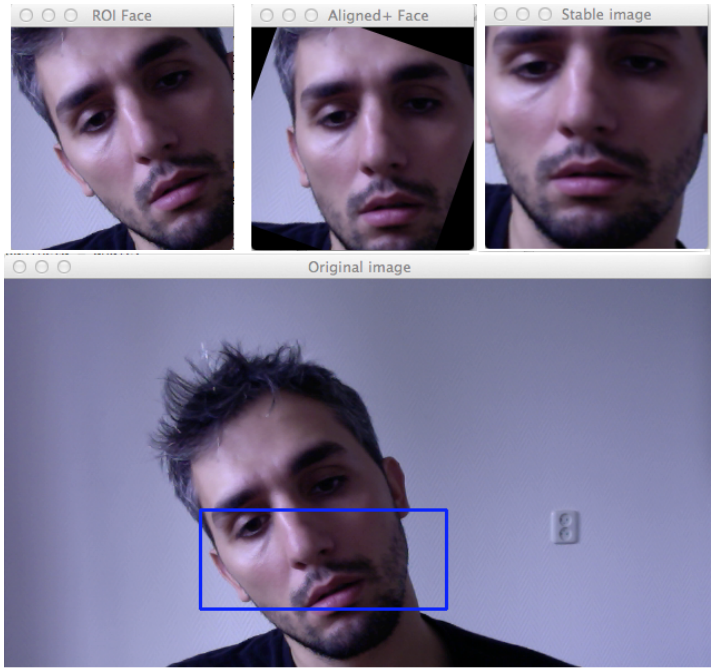

Figure 4.3: Comparison of three alignment techniques.

Since our face alignment is based on FaceTracker, it will never be perfect, since FaceTracker is not perfect as we described in our previous experiment. Hence, failure modes of our tracking system are reflected in our alignment step and of course in our classification rates. Despite that, we can observe that using a more accurate method, the final classification result increases significantly (4.2). We can explain this by looking at the motion vectors of the optical flow method. When the training and testing data are perfectly aligned, every basic expression

activates the same motion vectors each time it occurs, regardless the randomness and duration of it.

Table 4.2: Average classification results for three alignment methods.

-	ROI	Horiz. and Ver. face axis	Inverse compositional algorithm
Constrained Env.	70	78	89
Unconstrained Env.	58	70	84

4.3 Face features, Experiment 2

How important is temporal information (optical flow descriptors) in recognition of facial expression? For this purpose, we evaluate our algorithm on the Cohn-Kanade database for the six prototypic expressions namely surprised (Sr), scared (Sc), happy (H), sad (Sd), disgust(D) and angry (A). Sequences of frontal images representing posed facial expression always start with the neutral face and finish with the expression at its apex. In this experiment we have considered the same number of samples per class, both in the training and in testing sets. Finally, the dataset has 270 videoclips that were obtained by letting 41 subjects perform each of the 6 expressions once. This experiment was evaluated was used on offline recognition. The SVMLib tool was trained using 2 common types of kernels (polynomial, Gaussian). The best results were achieved using the Gaussian function, the results presented in this evaluation report on Cohn-Kanade dataset are all obtained with this type of kernel.

The resulting motion recognition system was evaluated using a modified cross validation. In our case, 75 % of the dataset (32 subjects) for training and the rest 25% for testing (9 subjects). In each fold of the LOOCV- Group, all video-clips performed by this group of subjects were kept as test video-clips and the classifier was built using the other video-clips. The results of this cross validation on Cohn-Kanade dataset are summarized in 4.3 below, where the expression error rate $ERexpr$ is defined as :

$$ERexp = NVmisclassified/NVtest,$$

where $NVmisclassified$ is the number of misclassified test video-clips and $NVtest$ is the number of test video-clips. In addition to the expression error rate, we used the image frame error rate $ERimage$, which is defined as :

$$ERimage = NFmisclassified/NFtest,$$

where $NFmisclassified$ is the number of misclassified frames in test video-clips and $NFtest$ is the number of frames in test video-clips. It can be seen from 4.3 that the expression error rate is smaller than the image error rate. We have proved that, in the ideal case, we can correctly classify an expression using majority voting as long as more than half of its frames are classified correctly(incomplete descriptors). This observation is especially important since it can make the system robust to local similarities between different types of expression (for example, fear and surprise) and also robust against noise.

4.3.1 FaceReader

In general, it is difficult to fairly compare the results of our facial expression recognition methods with other related works. We performed the same experiment on FaceReader system by Noldus Information Technology [4] using Cohn-Kanade dataset, and we compare the labels acquired with our previous results. FaceReader is a program for facial analysis. It has been trained to classify the 6 basic expressions, and offers a number of extra classification such as AUs activity, facial states, head orientation, subject characteristics, gaze and head tracking. FaceReader performs in real-time either in online or offline classification systems. As it can be noticed, our average expression recognition result for the incomplete video is quite similar to the FaceReader ones. Besides, our approach produced better recognition rates for the complete video descriptors (see Table 4.4). This can be explained as FaceReader bases its decision on still image classification. It performs frame to frame recognition, thus the ratio of incomplete descriptors is competitive to our methods. On the other hand, FaceReader misses the temporal information, and may fail to classify correctly when it comes to the final decision. Figure 4.4 illustrates the temporal evolution of 'disgust' expression. FaceReader classifies all 5 frames as neutral, but our method leads incrementally to the correct label, using temporal information.

Figure 4.4: Temporal evolution of expression : disgust .

Table 4.3: Fold Classification

Fold	Average Optical Flow	Average Face Reader
ERimage	14/69 = 20% error	20/69 = 29 % error
ERexp	80/700 = 11 % error	85/700 = 12% error

Table 4.4: Comperative recognition results produced by our algorithm and FaceReader for each basic expression using random number of complete and incomplete descriptors.

Expression	Surprised	Scared	Happy	Sad	Disgust	Angry	Average
Complete descriptors	97	80	97	84	91	85	0.89
Incomplete descriptors	80	70	79	70	81	73	0.75
FaceReader							
Complete descriptors	94	78	93	82	87	82	0.85
Incomplete descriptors	82	69	81	69	78	74	0.76

4.4 Discriminant power of the six basic facial expressions, Experiment 3

How well our system distinguish between different facial expressions? For this purpose, we evaluate our algorithm on the Cohn-Kanade database for the six basic expressions. In our case, 25 % of the dataset was used for testing (72 expression sequences, 12 subjects). Table 4.5 shows the corresponding confusion matrix of the six basic expression using the 72 test video streams. By analyzing the results according to the expression being recognized, we observe that for two expressions (Happy, Surprised) the system detected correctly all 12 corresponding test sequences. The 'Disgust' expression is correctly recognized in 11 out of 12 test video streams, while the rest 'Angry', 'Sad' and 'Scared' got worse results 10/12, 9/12 and 7/12 respectively. As comparative framework, we used optical flow-based approaches which work on sequence of images and which use this same Conh-Kanade database. In this context, a work by Sanchez et al. [28] also applied spatial analysis and dense optical flow tracking. Their recognition was performed using SVM and they reported and average recognition rate of 95%. As it can be noticed, our expression recognition results are quite similar to the corresponding ones reported by the Sanchez's method. Besides, our approach produced better recognition rates for "scared" and "angry" expressions. This could be explained as our expression classification is based on temporal information while Sanchez's method classifies based on still images. In cases that our method fails to correctly label the expression, we assume that the results could be even better if we exclude erroneous cases where the ground of truth label of expression miss-match with the expression performed in the video sequence.

Table 4.5: Confusion matrix.

Expression	Surprised	Scared	Happy	Sad	Disgust	Angry
Surprised	12	0	0	0	0	0
Scared	3	7	1	0	1	0
Happy	0	0	12	0	0	0
Sad	0	0	0	9	2	1
Disgust	0	0	0	0	11	1
Angry	0	0	0	1	1	10

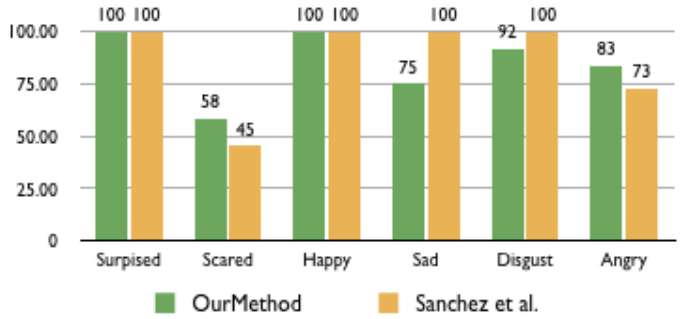

Figure 4.5: Discriminant power of respective subsets of the six basic expressions compared to another dense optical flow approach.

Chapter 5

Conclusions

We present a framework for facial expression recognition, where the facial motion temporal graphs are measured by using video sequence analyzed by optical flow. Experiments were conducted at each stage of the framework, identifying the possibilities. The advantage of this approach is that it accurately processes a video input, and draws conclusions for the facial expressions of the subject in real-time.

Our dense optical flow descriptors are robust under pose changes and noise. They capture the facial motion information in the videos streams efficiently and show adequate performance with state-of-the-art approaches for facial expression classification. We have also introduce an efficient solution to remove background motion and increase the stabilization by aligning the face during the video. We show that using a more accurate method, the final classification result increases significantly. Our descriptors combine appearance, and temporal information. Such a representation has shown to be efficient for action classification. We have proved that, in the ideal case, we can correctly classify an expression using majority voting as long as more than half of its frames are classified correctly. From our tests, we can conclude that our dense optical flow method using SVM as classifier produced on the Cohn-Kanade database quite similar results with an already existing system[4] and a state of the art approach [28].

However, a natural disadvantage of our method is that it presents a high processing time tracking and registering the face in a video sequence. As an extension to this, our system is unable to handle more than one faces in the image. We have also noticed that our classifier fails to recognize transitions between consecutive expressions, while skips short-length expressions.

We further explore the possibility of combining information of two subjects in a structured learning way. We are interested in modeling human interactions, focusing on the recognition of the interactions between two people in complex scenarios. In particular we deal with three interaction classes: positive, negative, neutral. These interactions classes are both symmetric and asymmetric since the people involved in it may perform the different combinations of facial expressions each time. An interaction descriptor has to include both local (appearance, motion) and global (spatio-temporal information) cues. We combine the extracted AUs intensities to learn a structured SVM classifier for an interaction set, whose output is their interaction label. The lack of an annotated interaction database hindered training and evaluation of the method we developed. Due to the simplicity of our structure (AUs intensities only) we come up with the conclusion that such structured learning techniques are applicable to more complex scenarios where the combination of local and global context cues is required (i.e. AUs and relative spatial location of people).

5.1 Future work

The face expression research community is shifting its focus to the recognition of spontaneous expressions. The major challenge that the researchers face is the non-availability of spontaneous expression data. Capturing spontaneous expressions on images and video is one of the biggest challenges ahead. Another major challenge is labeling of the data that is available. In addition, differences do exist in facial features and facial expressions between cultures (for example, Europeans and Asians) and age groups (adults and children). Face expression recognition systems must become robust against such changes. Another area where more work is required is the automatic recognition of expressions and AUs from different head angles and rotations.

Although this project provides a starting point for facial expression and human-to-human interaction recognition it is not yet complete. The focus has mainly been on aligning the face because these seem to play a large part in automatic recognition. The challenge was to make the system fully automatic. The first step to extend this method would be to improve the alignment part of the method with a more robust implementation. The next step would be to improve the dense optical flow algorithm with more efficient and accurate feature extraction algorithms. Another extension could be to combine video data from more than one cameras to get more accurate input to base a decision on.

Bibliography

[1] LibSVM, http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[3] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005.

[4] M.J. den Uyl and H. van Kuilenburg. The facereader: Online facial expression recognition. *Sentient Systems B.V.*, 2005.

[5] G. Donato M.S. Bartlett J.C. Hager P. Ekman and T.J. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Anal. Machine Intell.*, 1999.

[6] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[7] S. Ertekin, J Huang, and C. L. Giles. Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

[8] I. Essa and A. Pentland. Coding, analysis, interpretation, recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[9] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[10] J. Feldman. Taylor expansions in 2d. 2011.

[11] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition Intelligence*, 2008.

[12] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185203, 1981.

[13] M. Lyons S. Akamatsu M. Kamachi and J. Gyoba. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[14] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[15] B.V. Kumar. *Face expression recognition and analysis: the state of the art*. PhD thesis, Internal Report, Columbia University, 2009.

[16] J. Lien. *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity*. PhD thesis, Carnegie Mellon University, 1998.

[17] J.J. Lien, T. Kanade, J.F. Cohn, and C.C. Li. Automated facial expression recognition based on facs action units. In *Proceedings of the Third IEEE International Conference on Face and Gesture Recognition*, 1998.

[18] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615 – 625, 2006.

[19] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Seventh International Joint Conference on Articial Intelligence*, 1981.

[20] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[21] M. Pantic M.F. Valstar R. Rademakerand L. Maat. Web-based database for facial expression analysis. *in: Proceedings of the 13th ACM International Conference on Multimedia*, 2005.

[22] K. Mase. Recognition of facial expression from optical ow. *IEICE Transactions on Information and Systems*, E74-D(10):3474–3483, 1991.

[23] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42:1778 1790, 2004.

[24] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[25] T. Otsuka and J. Ohya. Recognizing multiple persons facial expressions using hmm based on automatic extraction of signicant frames from image sequences. *ICIP*, 1997.

[26] M. Pantic and L.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[27] M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Network*, 7(5):1121–1138, 1996.

[28] A. Sánchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernández, and J. J. Pantrigo. Differential optical flow applied to automatic facial expression recognition. *Neurocomput.*, 74(8):1272–1282, March 2011.

[29] J M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[30] Y. Tian, T. Kanade, and J.F. Cohn. *Handbook of Face Recognition*, chapter Facial expression analysis. Springer, 2004.

[31] Y.L. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[32] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, 1979.

[33] S. Uras, F. Girosi, A. Verri, and V. Torre. A computational approach to motion perception. *Biological Cybernetics*, 60:79–87, 1988.

[34] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2001.

[35] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, , and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[36] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA, 2011. IEEE Computer Society.

[37] Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *ICCV*, 2003.

[38] Z. Zeng, M. Pantic, G.I. Roisman, and T.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.