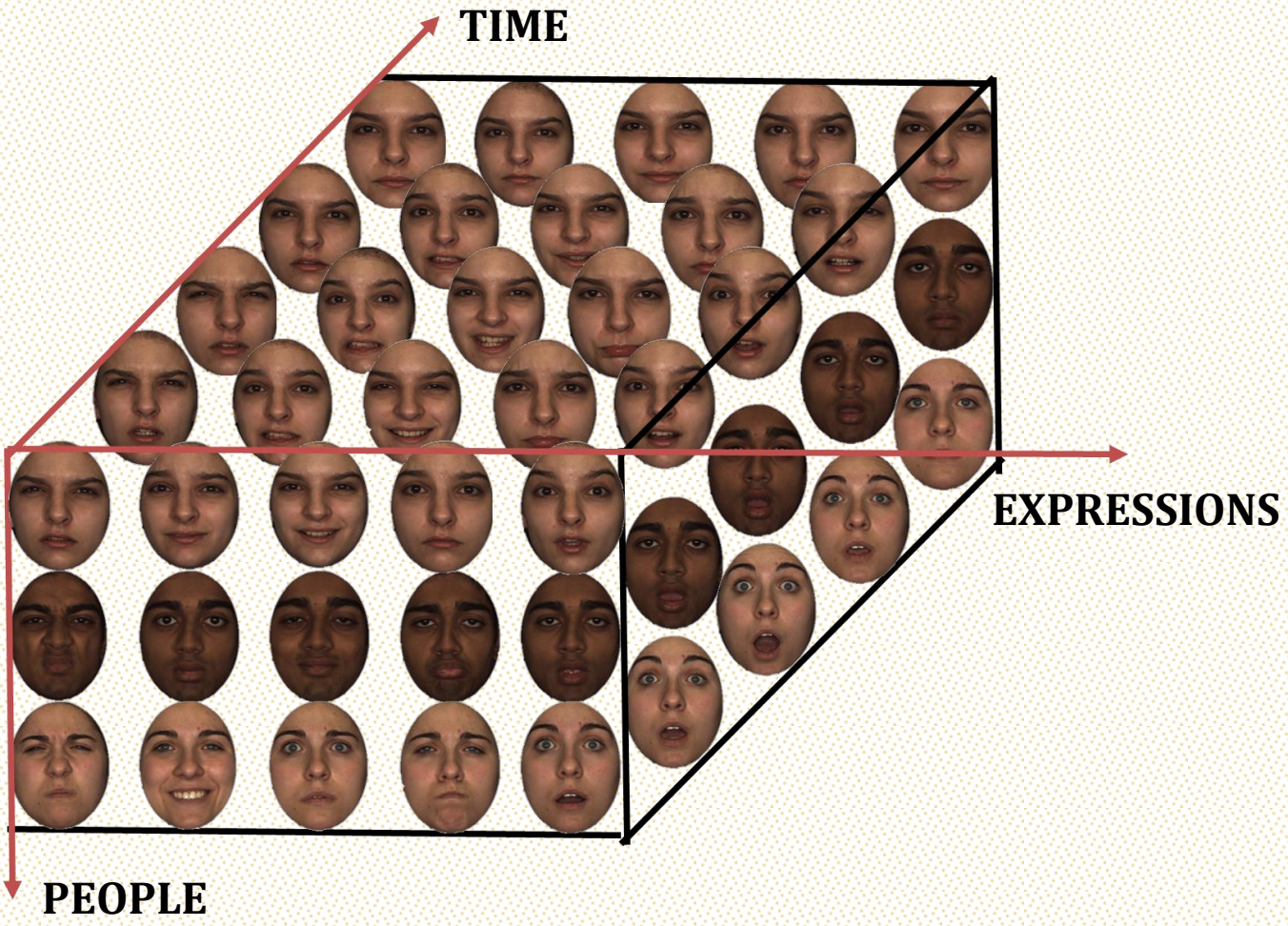

Carolina Raposo

Facial Dynamics for Identity and Expression Recognition

September 2012

· U C ·

Universidade de Coimbra

FACULTY OF SCIENCES AND TECHNOLOGY

UNIVERSITY OF COIMBRA

Integrated Master in Electrical and Computer Engineering

Facial Dynamics for Identity and Expression Recognition

Carolina Raposo

Juri: Professor Doutor Paulo Peixoto
Professor Doutor Jorge Batista
Professor Doutor Nuno Gonçalves

September 2012

This thesis was done under the supervision of Dr. Jorge Manuel Pereira Batista
Department of Electrical and Computer Engineering, University of Coimbra

Acknowledgements

Firstly, I would like to express my gratitude to my advisor Dr. Jorge Batista, for accepting me as his student and for all the help and support given throughout the past months.

I would like to thank all my colleagues in the Computer Vision Laboratory for their availability in aiding me.

Finally, I thank all my friends and family, especially my parents, for the great support.

Resumo

Estudos em Psicologia mostram que a dinâmica facial é um sistema biométrico, i.e., pode ser usada para reconhecimento de identidade. Com base nesta informação, a presente tese pretende demonstrar que o movimento facial por si só é suficiente para a identificação de pessoas, através de um conjunto de experiências. Em oposição à análise estática, a pesquisa relacionada com movimentos faciais é uma área de investigação relativamente recente. Para além do reconhecimento de identidade, o reconhecimento de expressões é também efectuado, usando diversas técnicas. O trabalho desenvolvido inclui o uso de diferentes descritores faciais que utilizam informação de forma, como é o caso dos Modelos de Forma Activa (ASM), e informação textural, como nos Padrões Binários Locais (LBP). No entanto, visto que é desejável analisar apenas as componentes dinâmicas, são usadas técnicas para remover informação de forma e textura, como a subtracção de imagens e os campos de fluxo óptico. As experiências são desenvolvidas usando estes descritores e a análise dos dados é efectuada usando duas técnicas principais: deformação dinâmica temporal (DTW) e análise de tensores. Foram usadas quatro bases de dados diferentes para avaliar a eficácia dos diferentes procedimentos. Estas bases de dados incluem diferentes números de indivíduos, sequências de expressões faciais com ou sem repetições e variedade de etnias e condições de luminosidade. Uma destas bases de dados foi criada como parte do presente trabalho, de modo a que sejam efectuadas experiências usando um indivíduo com alterações significativas na sua aparência. A novidade deste trabalho inclui o desenvolvimento de alguns procedimentos que combinam técnicas existentes de novas maneiras. Isto é importante pois foi verificada a superioridade destes novos métodos em relação a métodos já existentes. Em geral, todas as experiências produziram bons resultados, demonstrando que a dinâmica facial é um sistema biométrico conveniente. Foi concluído que, para bases de dados com poucos indivíduos, usar informação textural e de forma origina melhores resultados em reconhecimento de identidade. No entanto, com o aumento do número de pessoas é preferível usar apenas a dinâmica facial para o reconhecimento de identidade e expressões, pois as diferenças interpessoais também aumentam. Para além disto, usando a base de dados criada neste trabalho, foi mostrado que os procedimentos desenvolvidos são capazes de identificar o indivíduo cuja aparência é significativamente diferente da original, tornando-o quase irreconhecível. Esta experiência prova que a dinâmica facial é, de facto, um sistema biométrico conveniente e mostra a importância e relevância do presente estudo.

Palavras-Chave: Reconhecimento de Identidade, Reconhecimento de Expressões, Análise de Tensores, Padrões Binários Locais, Fluxo Óptico.

Abstract

Psychological studies indicate that facial dynamics is a biometric, i.e., it can be used for identity recognition. Based on this information, the present thesis attempts to demonstrate that facial motion alone is sufficient for performing person identification, through a series of experiments. As opposed to static analysis, research related to facial motion is a relatively new area of study. Besides identity recognition, expression recognition is also performed, using several techniques. The work developed includes the usage of different facial descriptors which make use of shape information, as is the case with Active Shape Models (ASM), and texture information, as in Local Binary Patterns (LBP). However, since it is desirable to analyse only the dynamic components, techniques are used for removing shape and texture information, such as image subtraction and optical flow fields. Experiments are conducted using these descriptors and data analysis is performed with two major techniques: dynamic time warping (DTW) and tensor analysis. Four different databases were used for assessing the efficacy of the different procedures. These databases include different number of individuals, facial expression sequences with or without repetitions and a variety of ethnicity and lighting conditions. One of these databases was created as part of the present work, so that experiments using an individual with significant changes in the appearance can be performed. The novelty in this work includes the development of some procedures which combine existing techniques in new ways. This is important since the superiority of these new methods over existing methods has been verified. In general, all the experiments yielded good results, demonstrating that facial dynamics is a proper biometric. It has been concluded that for databases with a small number of individuals, using shape and texture information leads to better identity recognition results. However, as the number of individuals increases, it is preferable to use only the facial dynamics for both identity and expression recognition, as the interpersonal differences also increase. Moreover, using the database created in this work, it has been shown that the developed procedures are capable of identifying an individual whose appearance is significantly different from the original, making him/her almost unrecognisable. This experiment proves that facial dynamics is in fact a proper biometric and shows the importance and relevance of the present study.

Key Words: Identity Recognition, Expression Recognition, Tensor Analysis, Local Binary Patterns, Optical Flow.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Identity and expression recognition has been a branch of computer vision with growing importance in the past decades. The use of still images of the face of an individual as a biometric (static analysis) has been a well-researched area. The term "biometrics" refers to the measurable biological characteristics which are used to quantify the physical features of an individual for use as a means of identification. Biometrics need to be universal, distinctive and repeatable. In respect to the human face, it is universal in the sense that it is the same for all people; it is distinctive since all faces are different, except in special cases; and it is repeatable because it does not change significantly in a short period of time, except when considering mechanisms such as facial motion (expressions), ageing, gaining weight or even cosmetic surgery. A relatively new area of study is the dynamics of facial expression. The term "dynamics", in this context, can be defined as the changes in facial motion over sequential time. One of the advantages over static analysis is that facial dynamics are less affected by physical changes such as ageing, gaining weight, wearing glasses, growing a beard, etc. Studies show that body and facial movements support person identification. There is considerable evidence that dynamic information is not redundant and may be beneficial for various aspect of face processing, including age, gender, and identity recognition. Based on this knowledge, the main objective of this work is to corroborate the following hypothesis: facial expressions can be used as an effective biometric for person identification.

1.1 Motivation

Determining the identity of a person automatically is a continually growing subfield of computational intelligence. Face recognition systems are used to verify the identity of an individual by matching a given face against a database of known faces. It has become an alternative to the traditional identification and authentication methods such as the use of keys, ID cards

and passwords, allowing secure identification and personal verification to be performed. Thus, face recognition technology can be applied to a wide variety of application areas including access control for PCs, airport surveillance, private surveillance, criminal identification and for security in ATM transaction. Moreover, the face recognition system is moving towards the next-generation smart environment where computers are designed to interact more like humans. It has become the most convenient tool for human interaction with machines, home automation systems, and intelligent robots. Because of its natural interpretation (human visual recognition is mostly based on face analysis) and low intrusiveness (unlike finger print), face-based recognition is one of the most important biometric traits.

In 1972, Paul Ekman identified six basic emotions: anger, disgust, fear, happiness, sadness and surprise [3], being these the main target of research recently. Psychological studies show that emotions, in the form of facial expressions, are more important than spoken words, during communication. Analysis of this means of interaction between people is currently subject of attention and thus, an automatic, efficient and accurate facial expression recognition system is a powerful tool. It is useful in areas such as anthropology, clinical psychology, psychiatry and neurology since emotions automatically estimated by computers are considered to be more objective than those labelled by people. Facial expression recognition can also be used by service providers in order to obtain implicit user feedback from the customers' facial expressions. In the computer graphic area, facial expressions estimated from real images can be used to animate synthetic characters and produce high quality computer animation. Recent research has shown that it is not only the expression itself, but also its dynamics that are important when attempting to decipher its meaning. Ekman et al. [4] suggest that the dynamics of facial expression provides unique information about emotion that is not available in static images.

The growing importance of facial dynamics as a field of research, in both identity and expression recognition, is due to the many applications presented above and to the fact that it constitutes a relevant biometric, providing an interesting topic for the development of this work.

1.2 Related Work

Psychological studies show that facial dynamics are relevant when performing recognition of identity, gender and expressions. In [5] , an experiment demonstrated that moving displays of the six prototypical expressions (happiness, sadness, fear, surprise, anger and disgust) were recognised more accurately than static displays of the face at the apex of the expressions. This indicated that facial expressions can be recognised in the absence of information about facial features. Experiments described in [6] show that people are capable of discriminating between

individuals and between genders from motion-based information alone. In [7], the results of identifying moving and still videotaped faces of famous and unknown people indicated that the first are significantly better recognised, proving again that facial dynamics are relevant in identity recognition. Conclusions obtained in [8] reveal that dynamic information contributes more to recognition in poor viewing conditions such as poor illumination, low-image resolution, recognition from distance etc. and that with increasing viewer's experience, dynamic information becomes more relevant. Despite the evidences from psychological studies already presented about the value of facial dynamics in face and expression recognition, only recently have researchers started to pay an important attention to the use of facial dynamics in automatic face analysis . The role of facial dynamics in facial expression recognition has been the focus of research and important contributions have already been made [9].

In [10], Local Binary Pattern (LBP) features are computed and Supervised Locality Preserving Projections (SLPP) are used to derive a generalised low dimensional expression manifold. Then, a Bayesian temporal model of the manifold is formulated in order to represent the facial expression dynamics. Results show that using dynamic information is advantageous when compared to using static information alone. Instead of using SLPP, Isomap embedding is used in [11] for obtaining the low dimensional expression manifold. A Gaussian Mixture Model (GMM) is then applied to cluster data and an Active Shape Model (ASM) is learnt for each cluster. Probabilist tracking is then performed to account for face motion. In [12], Locally Linear Embedding (LLE) is used for estimating the manifold of texture variation due to facial expression and it has been show that texture information provides better results than shape information alone. LLE is also used in [13] as a dimensionality reduction technique and Support Vector Machines (SVM) are then used for classification of expressions, involving the lower face. More recently, Hidden Markov Models (HMM) have been used to model temporal dynamics of expressions [14] and spatiotemporal Gabor Motion Energy filters (GME) as a biologically inspired representation for dynamic facial expressions [15]. All of the examples presented use methods that include an initial dimension reduction, followed by the classification. In this work, a similar approach is described.

As mentioned before, facial dynamics is not only used for facial expression classification but also for identity recognition, being this the main focus of the present work. Many dimensionality reduction techniques have been used, such as Laplacian EigenMaps (LE), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Locality Preserving Projections (LPP) [16, 17, 18]. Active Shape Models (ASM) and their extension Active Appearance Models (AAM) are used in video sequences for feature extraction [16, 19, 20, 21]. In [22] an Extended Volume LBP (EVLBP) operator is presented where texture analysis of video sequences is performed. Volume LBP is a special case of EVLBP because the difference is that the number of points forming the neighbourhood is variable. It has been shown that this

new method yields better results than Volume LBP. HMM are used in [16, 20] for modelling the dynamics of facial motion, providing a classification based on the probabilities obtained. In [23], a new kind of HMM is presented (adaptive HMM) in which each HMM is adapted during testing, resulting in better modelling over time. Another approach uses motion signatures (obtained from feature points which are tracked automatically) to create a tensor [2], which is a generalisation of the concept of a vector. High-Order Singular Value Decomposition (HOSVD) [24] is then applied for decomposing the tensor and Self-Organising Maps (SOM) are used for classification. Other methods include the use of Neural Networks such as the Multi-Layer Perceptron (MLP) [21] and Dynamic Time Warping (DTW), as well as the related methods: Continuous DTW (CDTW), Derivative DTW (DDTW) and Weighted derivative DTW (WDTW) [19]. In the latter case, DTW is used for computing the similarity between data vectors, being an efficient classification method. The present work analyses different feature extraction, dimensionality reduction and classification methods.

1.3 Overview

The organisation of this thesis was done with the objective of showing the reader that it is a sequence of procedures which attempt to overcome the disadvantages that arose throughout the implementation. Chapter 1 introduces the topic, presenting the research already developed in this area as well as its importance and main applications. Chapter 2 gives a theoretical background, describing and explaining all the main techniques used in this work as part of the different procedures. All of these techniques were obtained from available toolboxes, indicated in chapter 2. In chapter 3, the actual work developed is presented, where all six procedures are described, and results are shown. The discussion of the results provide a good evaluation of the performance of each procedure. The last subsection includes a description of the databases used, with example frames. Each of the remaining subsections includes a description of a procedure and its main disadvantages, which are attempted to be overcome. The first part of this chapter is an explanation of the application of each technique to facial dynamics. Concrete examples with visualisations of the results of each method are given with the objective of providing a better understanding for the reader. Lastly, chapter 4 gives all the important conclusions as well as the direction of future work.

Chapter 2

Background Theory

In this section, a theoretical introduction to the techniques used throughout the present work is given. This allows the reader to understand the general idea and main function of each method. Some techniques include feature extraction, pattern construction and classification, being used in different stages of the proposed procedures. All of these techniques were obtained from toolboxes available online.

2.1 Active Shape Models

Active Shape Models (ASM) [25] are statistical models, trained on a set of representative shapes, which iteratively manipulate themselves to fit an object in an image. In this work, the referred objects are faces. A set of landmarks (representative points, e.g., the corner of the left eye) forms a shape. The ASM method detects facial landmarks through a local-based search constrained by a global shape model, statistically learnt from training data. One shape is aligned to another with a similarity transform (allowing translation, scaling, and rotation) that minimises the average euclidean distance between shape points. The mean shape is the mean of the aligned training shapes (which in this case are manually landmarked faces). The ASM method starts the search for landmarks from the mean shape aligned to the position and size of the face determined by a global face detector. It then adjusts the locations of shape points iteratively until convergence. Four steps of the ASM search are shown in Figure 2.1 [1]. Object identification and location are robust because the models are specific in the sense that instances are constrained to be similar to those in the training set. They can be used to locate objects or as feature parameters to be passed to another system. In this work they are used for automatically locating feature landmarks. One of the ASM used in this work was downloaded from [1].

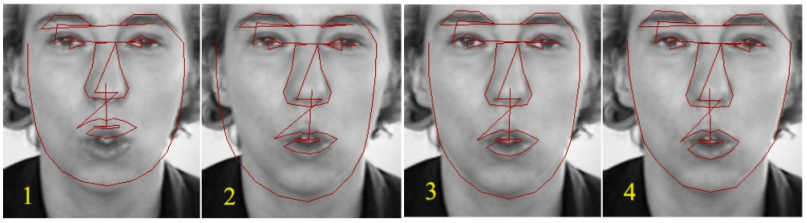

FIGURE 2.1: Steps of an ASM search. The shape model is shown in each picture. Figure taken/reprinted from [1].

2.2 Volume Local Binary Patterns

Volume Local Binary Patterns (VLBP) are an extension of the classic Local Binary Patterns (LBP) to perform dynamic texture analysis and represent temporal sequences as vectors. A brief theoretical explanation of these two methods is given in the following subsections. The MATLAB code used for computing LBP and VLBP was downloaded from [26].

2.2.1 Local Binary Patterns

The LBP texture analysis operator is defined as a grey-scale invariant texture measure since it tolerates monotonic grey-scale changes. Moreover, it has a high discriminative power and involves only simple computations. In general terms, the LBP operator forms labels for the image pixels and creates a histogram using these labels as a texture descriptor. Firstly, the neighbourhood of a pixel, composed of P equally spaced pixels on a circle of radius R, is thresholded with the value of that pixel by computing differences and a binary number is assigned to it, according to

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{2.1}$$

Note that the grey-scale invariance is a result of considering only the signs of the differences, instead of their values. Afterwards, for every pixel in the neighbourhood, a weight is defined. By summing the multiplied weights and binary values, an LBP code is obtained for each pixel. Let g_c be the grey level of the centre pixel and g_p the grey values of the neighbouring pixels, the LBP code of the centre pixel can be computed by

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p. \tag{2.2}$$

Finally, the occurrences of the LBP codes in the image are collected into a histogram which is the texture descriptor of the image.

2.2.2 Volume Local Binary Patterns

The LBP operator only deals with spatial information. To incorporate temporal information, VLBP can be used. The idea is that instead of considering a 2D neighbourhood, the face sequence is seen as a rectangular volume and the neighbourhood of each pixel is defined in three dimensional space. The neighbouring pixels in VLBP are P equally spaced pixels on a circle of radius R in the present frame (frame t) and $P+1$ pixels in the frames $t-L$ and $t+L$, where L is the time interval. Here, instead of considering P neighbouring pixels, a total of $3P+2$ pixels is used. Similarly to Equation 2.2, the VLBP code of a pixel is given by

$$VLBP_{L,P,R} = \sum_{p=0}^{3P+1} s(g_p - g_c)2^p \tag{2.3}$$

where g_p are the grey levels of the pixels in frames $t-L$, t and $t+L$. Figure 2.2 shows the computing procedure for $VLBP_{1,4,1}$ as an example.

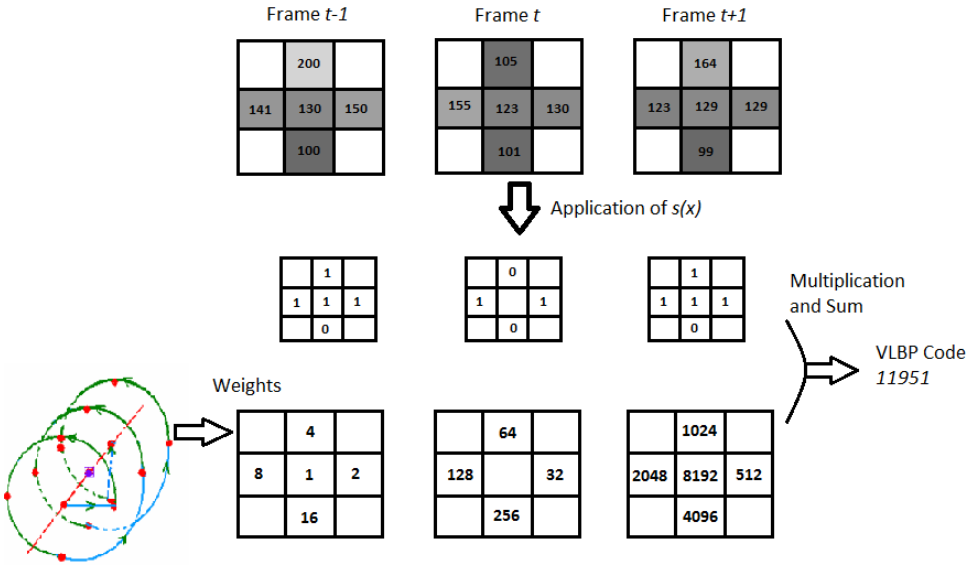

FIGURE 2.2: Steps of the computing procedure for $VLBP_{1,4,1}$.

2.3 Optical Flow

Optical flow is a method used for calculating the motion between two image frames which are taken at times t and $t+\Delta$. This work uses facial motion as a basis for people identification and thus determining optical flow fields can be of significant aid. Considering the optical flow as a velocity field associated with image changes, three constraints must be taken into account:

1. **Grey value constancy**: A moving point does not vary (instantly) its appearance

2. **Small motion**: observed points do not move very far between two consecutive images

3. **Spatial coherence**: if a point moves, its neighbours exhibit the same behaviour

Using constraints 1) and 2), the *brightness constancy equation* can be written as

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{2.4}$$

where $I(x, y, t)$ is the luminance level of pixel (x, y) at time t and the corresponding δ represent small variations. The Taylor expansion of $I(x + \delta x, y + \delta y, t + \delta t)$ results in

$$I(x + \delta x, y + \delta y, t + \delta t) \approx I + \frac{\delta I}{\delta x}\delta x + \frac{\delta I}{\delta y}\delta y + \frac{\delta I}{\delta t}\delta t \tag{2.5}$$

where $I = I(x, y, t)$. Considering equation 2.4 it follows that $\frac{\delta I}{\delta x}\delta x + \frac{\delta I}{\delta y}\delta y + \frac{\delta I}{\delta t}\delta t = 0$, which is equivalent to

$$\frac{\delta I}{\delta x}v_x + \frac{\delta I}{\delta y}v_y + \frac{\delta I}{\delta t} = 0 \Rightarrow \begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = -I_t \tag{2.6}$$

after dividing by δt and making $v_\lambda = \frac{\delta \lambda}{\delta t}$ and $I_\lambda = \frac{\delta I}{\delta \lambda}, \lambda = x, y, t$. Since this is an equation with two unknowns, it is necessary to use constraint 3) and assume that the neighbour pixels exhibit the same velocity, resulting in the following system of linear equations:

$$\begin{bmatrix} \Sigma I_x^2 & \Sigma I_x I_y \\ \Sigma I_x I_y & \Sigma I_y^2 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = - \begin{bmatrix} I_x I_t \\ I_y I_t \end{bmatrix}. \tag{2.7}$$

The solution of this system is the velocity vector $[v_x \quad v_y]$ which can be solved for every pixel in each frame. In this work, the MATLAB code with the implementation of Optical Flow was downloaded from [27].

2.4 Principal Component Analysis

Principal component analysis (PCA) is a mathematical procedure used for identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. This is a useful technique since it allows the reduction of the number of dimensions in the data, without much loss of information. It is used as a data reduction technique in order to describe each shape (face) by a point in a low-dimensional space. Performing PCA on a set of shapes results in a linear parametric model that gives the approximation of shape $\mathbf{s}$:

$$\mathbf{s} \approx \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i \tag{2.8}$$

where $\mathbf{s}_0$ is the mean shape, $\mathbf{s}_i$ are the eigenvectors which are linearly combined using the vector of shape parameters, $\mathbf{p} = [p_1, \dots p_n]^T$. n is the number of eigenvectors that holds a certain variance, being this the dimension of the low-dimensional space. In this work, the MATLAB code of PCA is part of the MATLAB Toolbox for Dimensionality Reduction ([28]).

2.5 Dynamic Time Warping

Dynamic Time Warping (DTW) is a method commonly used for computing the best alignment between two signals by warping their time axes. The two signals, which initially had different temporal dimensions (different duration), after the usage of DTW result in two other signals with the same duration: time warping. The DTW algorithm allows "elastic" transformation of time series in order to detect similar shapes with different phases. The alignment path built by DTW computes distances between points and must satisfy three conditions:

- **Boundary condition**: The starting and ending points of the warping path must be the first and the last points of aligned sequences, respectively;
- **Monotonicity condition**: Points must be order chronologically;
- **Step size condition**: The shifts in time of the warping path are limited.

Besides computing an alignment between time sequences, DTW also gives a measure of their similarity. This measure is commonly used for classification [19], as is the case in this work. Figure 2.3 shows the result of applying DTW to the two misaligned waves (time sequences) depicted in the first image. The second image shows the accumulated distance matrix and optimal path as a white line. Darker colours correspond to shorter distances and therefore it can be seen that the optimal path is the path that yields the shortest distance between the two sequences. Using this path it is possible to construct the warped signals, which are depicted in the third image. The connecting path is also shown (fourth image) for better visualisation of the similarities between these two signals. MATLAB code with the implementation of DTW was downloaded from [29].

2.6 Tensors

Tensors are multidimensional arrays of data and therefore can be used to analyse multivariate data. A vector is considered as a first-order tensor and a matrix as a second-order tensor. In this work tensors are used to describe dynamic sequences of different people performing different expressions. The notation used in the present chapter is the following: scalars are

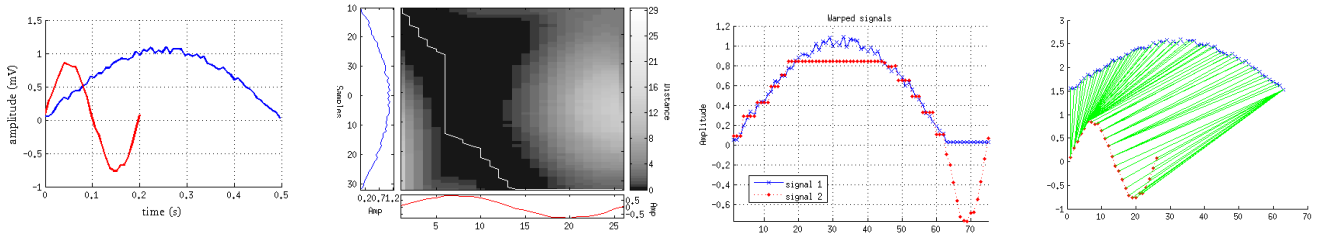

FIGURE 2.3: Usage of the DTW algorithm on two signals (1st), resulting in an accumulated distance matrix and optimal path (2nd). The warped signals (3rd) and the connecting path (4th) are also shown.

denoted by lower case letters (a, b, ...), vectors by bold lower case letters ($\mathbf{a}$, $\mathbf{b}$, ...), matrices by bold upper-case letters ($\mathbf{A}$, $\mathbf{B}$, ...), and higher-order tensors by calligraphic upper-case letters ($\mathcal{A}$, $\mathcal{B}$, ...). Tensor algebra is performed using TP Tool ([30]).

2.6.1 Tensor Algebra

The *order* of a tensor can be defined as the number of indices required to write that tensor and, therefore, matrices all have tensor order 2. The *rank* of a tensor $\mathcal{A}$, denoted $R = rank(\mathcal{A})$, is the minimum number of simple tensors necessary to express $\mathcal{A}$ as a linear combination. Tensor rank extends the notion of matrix rank since the latter is defined as the number of linearly independent rows or columns. An N^{th}-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is a simple tensor if it has rank 1, i.e., if it can be written as the outer product of N vectors: $\mathcal{A} = \mathbf{u}_1 \circ \mathbf{u}_2 \circ \ldots \circ \mathbf{u}_N$. Thus, a rank-$R$ tensor can be expressed as

$$\mathcal{A} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_1^{(r)} \circ \mathbf{u}_2^{(r)} \circ \ldots \circ \mathbf{u}_N^{(r)}. \tag{2.9}$$

A matrix singular value decomposition (SVD) can be expressed as a rank-R decomposition:

$$\mathbf{M} = \mathbf{U}_1 \mathbf{\Sigma} \mathbf{U}_2^T = \sum_{i=1}^{R} \sum_{j=1}^{R} \sigma_{ij} \mathbf{u}_1^{(i)} \circ \mathbf{u}_2^{(j)}, \tag{2.10}$$

where $\mathbf{U}_1$ is an orthogonal column-space, Σ is a diagonal singular value matrix and $\mathbf{U}_2$ is an orthogonal row-space. The mode-n vectors (mode-n space) of an N^{th}-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ are the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying index i_n while keeping the other indices fixed. In the case of a third order tensor, three mode spaces exist where mode-1 corresponds to column space, mode-2 to row space, and mode-3 to depth space.

2.6.2 Tensor Flattening

The mode-n vectors of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ are the column vectors of matrix $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times (I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N)}$ that result from *flattening* (or unfolding) the tensor $\mathcal{A}$. Tensor unfolding can be considered as splitting a tensor into mode-n vectors and rearranging these vectors column-wise to form a matrix. An example for a third-order tensor is show in Figure 2.4. The $n-rank$ of $\mathcal{A}$, as an generalisation of the definition of column and row rank of matrices

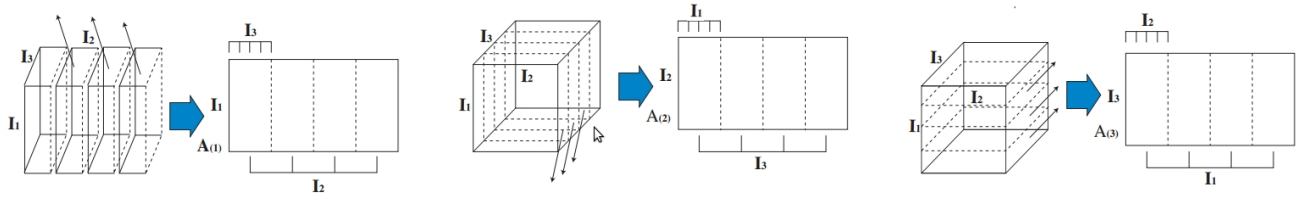

FIGURE 2.4: Flattening of a third-order tensor resulting in matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$. Figure taken/reprinted from [2].

and denoted R_n, is defined as the dimension of the vector space generated by the mode-n vectors:

$$R_n = rank_n(\mathcal{A}) = rank(\mathbf{A}_{(n)}). \tag{2.11}$$

The *mode-n product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{M} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{M}$, is a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N}$. In terms of flattened matrices, the mode-n product can be expressed as

$$\mathbf{B}_{(n)} = \mathbf{M}\mathbf{A}_{(n)} \tag{2.12}$$

and tensor $\mathcal{B}$ is found by folding matrix $\mathbf{B}_{(n)}$ back into tensor representation.

2.6.3 Tensor Decomposition

A matrix $\mathbf{D} \in \mathbb{R}^{I_1 \times I_2}$ is a two-mode mathematical object that has two associated vector spaces: a row space and a column space. SVD orthogonalises these two spaces and decomposes the matrix as $\mathbf{D} = \mathbf{U}_1 \mathbf{\Sigma} \mathbf{U}_2^T$. Using mode-$n$ products, $\mathbf{D}$ can be written without the need of a generalised transpose as $\mathbf{D} = \mathbf{\Sigma} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$. Extending the concept of matrix SVD, N-mode SVD or Higher Order SVD (HOSVD) decomposes an N-order tensor $\mathcal{D}$ into N orthogonal spaces $\mathbf{U}_1, \mathbf{U}_2 \ldots \mathbf{U}_N$ and expresses $\mathcal{D}$ as the mode-n product:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_N \mathbf{U}_N \tag{2.13}$$

where $\mathcal{Z}$ is the *core tensor* and is analogous to the singular value matrix that results from SVD. It governs the interactions between the orthogonal spaces obtained from HOSVD. Figure 2.5 illustrates the result of the HOSVD algorithm used in a third-order tensor.

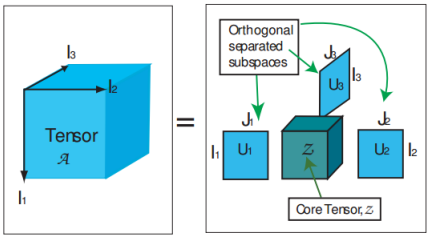

FIGURE 2.5: Decomposition of a third-order tensor resulting in matrices $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$, and core tensor $\mathcal{Z}$. Figure taken/reprinted from [2].

2.6.4 The HOSVD Algorithm

The presented theory allows the creation of an algorithm for decomposing tensor $\mathcal{D}$ [24]: For $n = 1$ to N:

- Unfold $\mathcal{D}$ along dimension n to find matrix $\mathbf{D}_{(n)}$;
- Perform SVD on $\mathbf{D}_{(n)}$ to compute $\mathbf{U}_n$ in (2.13) by setting it to be the left matrix of SVD.

Afterwards, solve for the core tensor: $\mathcal{Z} = \mathcal{D} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \ldots \times_N \mathbf{U}_N^T$.

HOSVD is beneficial since it allows a multi-factored space to be decomposed into its constituent modes. The different modes can then be analysed separately and important information about the data can be extracted.

2.7 Self-Organising Maps

Self-organising maps (SOMs) are a data visualisation technique [31] which reduces the dimension of data through the use of unsupervised artificial neural networks. This means that during learning, only input data is given, without the need for presenting target data, that is, without any external supervision. SOMs reduce data dimension by producing a map of usually 1 or 2 dimensions which plots the similarities of the data by grouping similar data items together. Thus, the main advantages of SOMs are the possibility of visualising N dimensional data in 2D and detecting similarities and degrees of similarity. Therefore, they can be used as classification techniques, which is the case in this work. A toolbox for creating SOM has been downloaded from [32].

Chapter 3

Identity and Expression Recognition

The present chapter is a detailed description of the work developed. With the purpose of demonstrating that facial dynamics is a biometric, several procedures were developed as attempts to overcome existing issues. Dynamic databases are necessary for testing the procedures created and thus, a description of all four databases used is given in section 3.7. Throughout the present chapter, the databases are referred to as *Database 1*, *2*, *3* and *4*. *Database 4* has been created as part of the present work, in order to investigate whether facial dynamics alone is an effective biometric. Each section of this chapter includes a description of the methods created, as well as the obtained results. A significant amount of existing techniques are used, and some novelty is introduced in the combination of these different techniques. The different procedures presented in this section describe the evolution of the work developed. In the first method, the first database was used and dynamic time warping was performed for identity and expression classification. The methods that follow try to overcome some disadvantages as is presented throughout this chapter. Firstly, an explanation of the application of the different techniques to the case of facial dynamics is given, with the aid of visual examples.

Active Shape Models

Each dynamic database comprises a set of facial expressions performed by one or more individuals. Each of these facial expressions is defined by a set of frames showing the individual's face. In this work, Active Shape Models (ASM), as introduced in section 2.1, are used for automatically locating feature landmarks in each of these frames. Thus, each image of the sequence becomes represented by a set of landmarks, i.e., its 2D coordinates. This kind of representation of the faces encodes shape information since it gives the relative position of the feature points in relation to each other.

Two models are used in this work. The first model was constructed as part of the work developed in [16], referred to as *ASM 1*, and contains 58 landmarks. It was created using several frames from *Database 1*, which were manually landmarked and afterwards used for training the ASM. The second model, *ASM 2*, was downloaded from [1], and contains 76 landmarks. This model uses the MUCT Face Database which consists of 3755 faces with great diversity of lighting, age, and ethnicity. Figure 3.1 shows the feature points of each model.

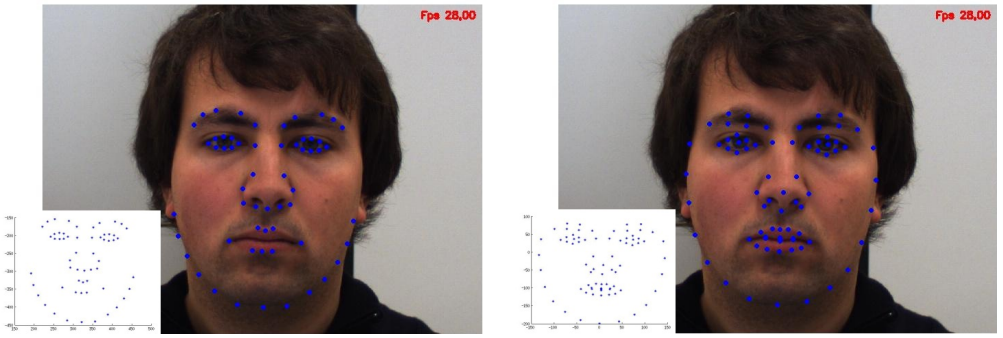

FIGURE 3.1: Feature points of *ASM 1* (left) and *ASM 2* (right) after fitting the face.

Volume Local Binary Patterns

Besides ASM, another facial descriptor has been used: Volume Local Binary Patterns (VLBP). This is a texture descriptor since it uses the sequences of grey-scale images for computing histograms, as described in subsection 2.2.2. Here, each face in each frame is detected, using a software for locating faces, and aligned to a reference frame. A more detailed description of this preprocessing step is given in section 3.4. Alignment of the faces is needed because VLBP computes the codes considering the neighbourhood of each pixel, in sequential frames, and thus translation, rotation and scale components must be removed. Figure 3.2 shows the histogram resultant from applying VLBP to a sequence of an individual performing a facial expression. It can be observed that the range of codes computed varies from 1 to over 16000, and so each facial sequence becomes represented by a very long vector. This causes the procedures to be more computationally costly than when ASM are used. The main advantage is that much more facial information is considered.

Optical Flow

When applied to faces, optical flow creates fields of vertical and horizontal velocities, as explained in section 2.3, which provide a description of the facial motion. Figure 3.3 shows the velocity vectors for each pixel, obtained from the grey-scale images. It can be observed

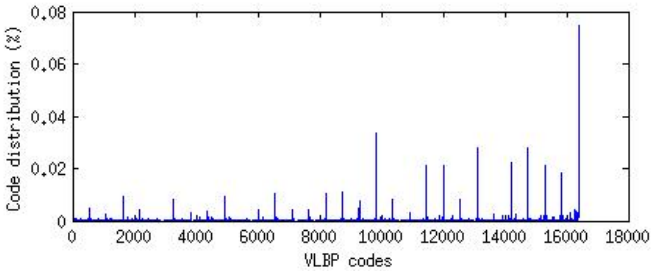

FIGURE 3.2: Normalised histogram of the VLBP codes of a facial sequence.

that motion is present more significantly around the mouth and the eye region, as expected. In this work, optical flow is applied to the facial sequences, using consecutive pairs of images. From each pair of images, a field of velocity vectors is obtained, and thus each new sequence becomes one frame smaller than the original one. Optical flow is used as a motion descriptor, and so texture and shape information are removed. Analysis is done by applying VLBP to the optical flow sequences, so that each sequence becomes represented by a vector (histogram).

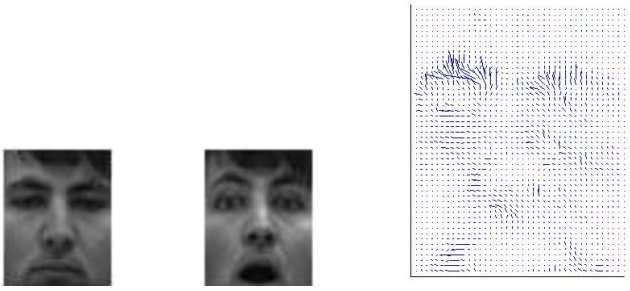

FIGURE 3.3: Optical flow field (right image) computed using the two grey-scale images.

Principal Components Analysis

As introduced in section 2.4, PCA is used in this work as a dimension reduction technique. Each shape, in this case, can be represented by a set of feature points, when ASM is used. The parametric model obtained from PCA (equation 2.4) can be rewritten as

$$\mathbf{s} \approx \mathbf{s}_0 + \Phi\mathbf{p} \tag{3.1}$$

where Φ is the matrix of the n lead Eigenvectors. Inverting this equation, the set of Eigen-coefficients for shape $\mathbf{s}$ can be extracted:

$$\mathbf{p} \approx \Phi^T(\mathbf{s} - \mathbf{s}_0). \tag{3.2}$$

Each n-element vector $\mathbf{p}$ is now used for representing each face, instead of the coordinates for all feature points. This representation of faces, or facial sequences, besides reducing the spacial dimension, leads to noise removal and reduction of intrasubject variations, which degrade the recognition performance. Figure 3.4 shows the first three Eigencoefficients of each shape of the facial expression sequence, obtained after performing PCA holding 97% variance.

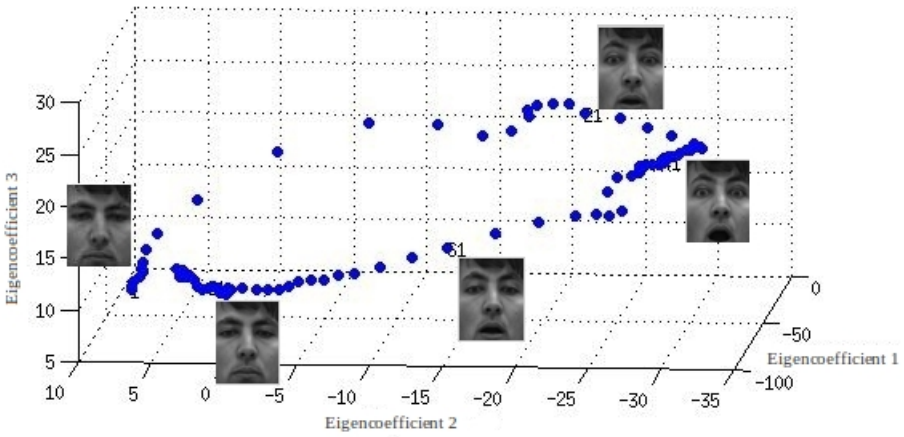

FIGURE 3.4: Surprise dynamics of an individual plotted in the subspace spanned by the first three Eigencoefficients.

Dynamic Time Warping

Creating references from a set of existing sequences or constructing tensors always require the facial expression sequences to have the same number of frames, i.e., to have the same duration. However, even for the same individual, it is not easy nor expected that he or she performs a facial expression twice in exactly the same time interval. This aggravates for the case of many individuals performing several facial expressions, where very different durations are obtained. The solution for this issue is using Dynamic Time Warping (DTW) between each pair of sequences so that two warped sequences are obtained, i.e., the resulting sequences have the same duration. Since DTW takes into account not only euclidean distances between points of the sequences but also their evolution (see section 2.5), it is only logical to perform DTW between repetitions of the same facial expression, in this case. The left and right images in figure 3.5 show two original sequences of a subject performing the happiness dynamics and the warped sequences, respectively. The same sequences are plotted with the same colour. The original sequences have two different durations (90 and 50 frames) and the warped sequences have the same duration (96 frames). It can be observed that the warped sequences maintain the same pattern, with the difference that they are extended, by repeating points, which represent frames, so that they become more similar to the other sequence. It is expected that, for warped sequences, the frames with the same index correspond to the same

instant in both sequences. For example, if the 50th frame in one sequence corresponds to the apex, which is the most expressive frame, of the facial sequence, the 50th frame in the other sequence should correspond to the apex as well. This can be observed in figure 3.5, because the points relative to the apex in the red sequence present more repetitions than the initial ones. Figure 3.6 shows the warping path between the two sequences. The pictures indicate the beginning and ending frames of the facial expression, as well as the most expressive one. It can be seen that in both sequences, the expression starts in the 20th frame, leading to a unitary slope because very few frames are repeated. Although the apex zone of sequence 1 lasts for about 40 frames, in sequence 2 it lasts only 15 frames. Thus, a null slope is obtained in some segments between frames 30 and 70 of sequence 1.

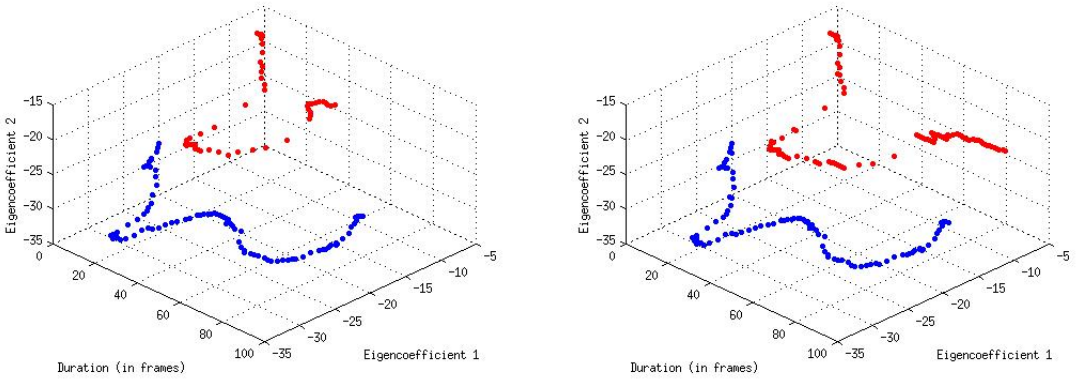

FIGURE 3.5: Warped sequences (right image) resultant from applying DTW to the sequences in the left image.

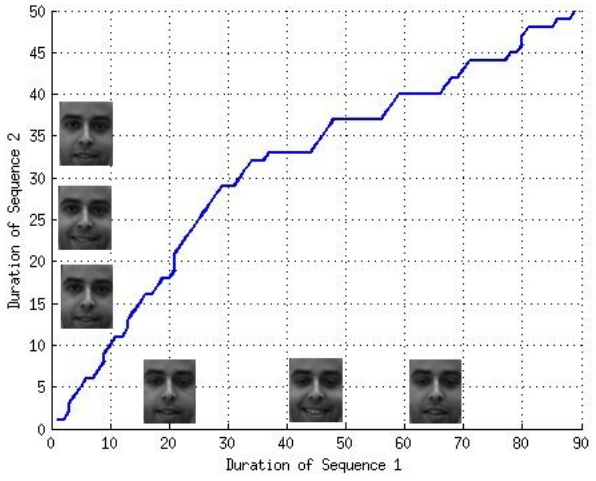

FIGURE 3.6: Warping path between the two sequences.

Tensors

A database with a significant amount of individuals may comprise a great variation of textures (due to different ethnicity), lighting conditions, poses, genders, facial expressions, etc. It is often desirable to analyse only a certain characteristic (e.g. the gender) in the presence of variations of other characteristics. Thus, a technique for separating the different characteristics is of great use, so that only the target one is considered. Tensor analysis can be used for solving this problem, since it provides a separation of the data in subspaces representative of each characteristic. In this work, identity and facial expressions are the two characteristics subject to analysis and thus, tensors are used in most of the procedures developed. Considering databases comprising repetitions of the facial expressions by the same individual, or comprising great lighting or appearance variations, the goal is to perform identity and expression recognition under these conditions. As described in section 2.6, tensors are multidimensional arrays of data constructed so that each dimension comprises one characteristic. For example, if the J facial expression sequences performed by I individuals are sets of L points in a P-dimensional space, which correspond to L frames, a $4D$-tensor $\mathcal{T} \in \mathbb{R}^{I \times J \times P \times L}$ which represents the entire data set may be constructed. Decomposing this tensor as explained in subsection 2.6.3 leads to the mode-n product

$$\mathcal{T} = \mathcal{Z} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{expressions} \times_3 \mathbf{U}_{eigencoef} \times_4 \mathbf{U}_t \tag{3.3}$$

where $\mathcal{Z}$ is the core tensor, and $\mathbf{U}_i, i = \{people, expressions, eigencoef, t\}$ are the subspace matrices which represent each of the characteristics. This decomposition provides a separation of the data so that only the relevant features are present in each of the subspaces. In this work, the subspace matrices $\mathbf{U}_{people}$ and $\mathbf{U}_{expressions}$ are used for performing identity and expression recognition, because their row vectors span the invariance of the characteristic in question across all the other characteristics. Depending on which facial descriptor is used (VLBP or ASM), $3D$ or $4D$ tensors are constructed. However, they have in common the first two dimensions which always represent the people in the database and the facial expressions. Figure 3.7 shows a visualisation of the $4D$ tensor previously described.

Self Organising Maps

Self Organising Maps (SOM) are a good way for representing multi-dimensional data as $2D$ maps. They are useful in the case where a significant amount of training data is available. In this case, not only full facial sequences, containing the onset-apex-offset, are used, but also streams of facial expressions. In this work, a facial expression *sequence* refers to a complete sequence containing the onset, apex and offset stages. A *stream* refers to part of a sequence, i.e., a set of sequential frames obtained from that sequence. The usage of streams leads to

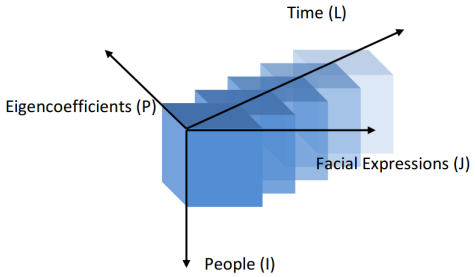

FIGURE 3.7: Visualisation of a $4D$ tensor.

a great amount of training data, and thus SOM are used for visualisation and classification. To each training stream, a label is assigned identifying the corresponding facial expression. SOM organises the data and extracts the similarities for creating a map where adjacent cells have the same label. Figure 3.8 shows the labels of the cells of a SOM which represents the six facial expressions (right image). The left image is a distance matrix of the SOM. It can be observed that the areas in blue, which correspond to smaller distances, cluster together and are located in the same regions of the SOM as the labels. The clusters are separated by cells which correspond to longer distances.

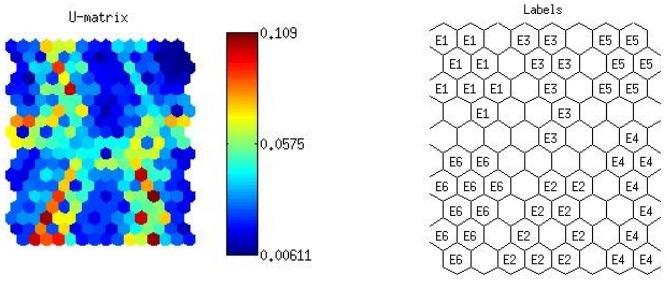

FIGURE 3.8: Distance matrix of the SOM (left image) and corresponding labels (right image).

Frame Subtraction

Since the main objective of this work is to demonstrate that facial dynamics is a biometric, experiments are performed not only using the original data sets but also modified data sets which attempt to eliminate or reduce the component related to facial shape and texture by subtracting the neutral face of each individual from every frame in the database. A similar procedure using grey scale images is used in [12], with the purpose of minimising undesirable variances.

Figure 3.9 shows the result of fitting ASM 1 to a neutral face (left image) and to a frame in the apex of a facial expression (middle image), after alignment to a reference frame. The image on the right shows the outcome of subtracting these two images, where it can be observed

that the shape information is completely removed. In this case, subtraction is performed between the corresponding feature points in each frame. Analogously, figure 3.10 shows the pixel by pixel subtraction of the grey levels of the two left images, in the right image. Most of the texture information is removed in this case.

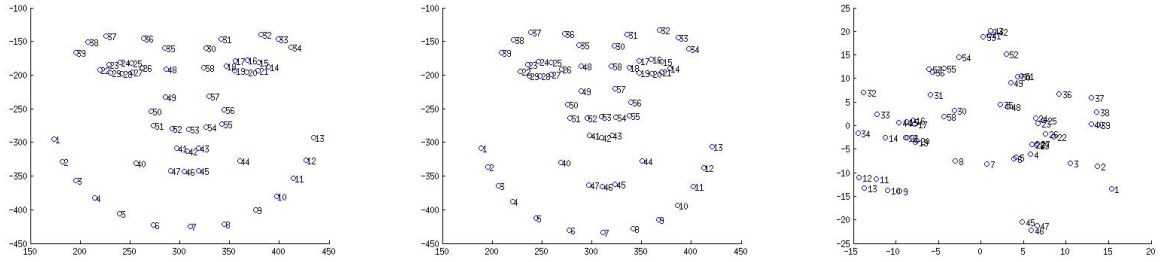

FIGURE 3.9: Subtraction of the left and middle set of landmarks, resulting in the set of points in the right image.

FIGURE 3.10: Subtraction of the left and middle grey-scale images, resulting in the right image.

Note: In the following procedures, all the experiments for both identity and expression recognition were performed with the same query sequences. This means that even if the person is wrongly identified, expression recognition is performed using the correct individual. In a real system, this would not be the case. However, for testing purposes, this method was adopted.

3.1 Dynamic Time Warped Shape Sequences

Based on the work developed in [16], where dynamic information is used for recognition, a new method was created with the same objective: recognise identity and facial expressions using facial motion. This procedure uses the feature points obtained from ASMs (see section 2.1), which fit the face in each frame of the database. Note that a database with repetitions of the same expression by the same individuals is required and thus, this procedure has only been tested with *Database 1*, which was constructed as part of the work developed in [16]. Initially, each frame of each sequence is represented by a set of landmarks, which can be

interpreted as a point in a high dimensional space. Thus, a preprocessing step is required for reducing the spacial dimension. In order to remove scale, translation and rotation components, all the frames are aligned to a reference frame using a Generalised Procrustes Analysis (GPA). The dimension of the space in which the faces are represented is reduced by using a Principal Components Analysis (PCA), as introduced in section 2.4. Instead of being represented by a point in a $2N$-dimensional space, where N is the number of landmarks, each shape (face) of the database is now represented by a point in a p-dimensional space, where p is the number of Eigenvectors that hold a user defined variance, in this case 97%. As mentioned in the beginning of this chapter, in order to remove shape components, each frame of the database is subtracted by the neutral face for each person, so that only dynamic components are present and each set of landmarks represents the motion relative to the neutral face. Experiments are performed for both the original and the subtracted data sets. The idea in which this method is based consists on comparing a query sequence with reference sequences so that identity and expression recognition are performed, sequentially. In [19] a similar procedure is used, with the difference that $3D$ shapes are used and tests are performed using databases of people uttering words.
The reference sequences for identity recognition are created as follows:

1. PCA (with 98 % variance) is performed on a new set constructed from $r-1$ out of the r repetitions for each person and expression, and the mapping is determined;

2. For each person and repetition, all the expression sequences are concatenated in the same order, resulting in $(r-1)Np$ longer sequences (Fig. 3.11), where Np is the number of people in the data set;

3. Dynamic Time Warping (see section 2.5) is performed between each of these new sequences and a reference sequence, resulting in warped sequences with a fixed length, which equals $len \times Ne$, where len is the user defined length of each expression sequence (80 frames in this case, corresponding to approximately 3 seconds for performing a full facial expression), and Ne is the number of expressions in the database;

4. For each person, the mean sequence is computed using the $r-1$ repetitions, using the arithmetic mean for each point of the sequence;

5. Lastly, for future computational simplicity, each reference for identity recognition - referred to as $ref_p, p = 1, 2, \ldots, Np$ - is the result of sampling len elements of the respective mean sequences. The result is Np new sequences of length len.

The reference sequences for expression recognition are created as follows:

1. PCA (with 99 % variance) is performed in each of the Np new sets, which consist of $r-1$ repetitions of every expression, one set per individual (fig. 3.12). The result is Np new mappings which will be used during testing;

2. Each of these new sequences is warped (by performing DTW) with a reference sequence (one of the repetitions), resulting in sequences of length len;

3. The mean sequence of the three repetitions of each expression is the reference for expression recognition - referred to as $refE_{pe}, p = 1, 2, \ldots, Np; \quad e = 1, 2, \ldots, Ne$. Therefore, for each person in the database there are Ne references.

As mentioned in section 2.5, DTW can be used as a classifier since it gives a measure of the similarity between two signals. For both identity and expression recognition, the assumption is that "closer" signals (frame sequences) correspond to the same person and facial expression, respectively. It is possible to use this method in this case because the sequences obtained after PCA comprise only the principal components and thus they are mapped in different locations in the p-dimensional space. It has been observed that similar sequences (repetitions of the same expression by the same individual) are mapped to similar locations, allowing distance measures to be used. This can be seen in Figure 3.11, which depicts three repetitions of the concatenated sequences for each individual for the original data set (left) and the subtracted data set (right). As expected, it can also be observed that the sequences of the subtracted data set all have a common point near the origin of the referential system. This is due to the fact that neutral faces in each sequence are transformed to zero in the p-dimensional space. However more evident in the case where the original database is used, in both cases it can be observed that different tones of the same colour (same person) are "closer" to each other than to other colours, which represent other individuals. Thus, during the testing phase, each query sequence, to which only one individual and expression are associated, is initially warped with $ref_p, p = 1, 2, \ldots, Np$ and Np measures (distances) are determined. The reference sequence that yields the minimum distance (e.g. ref_2) belongs to the same individual as the query sequence. Expression recognition is analogous to this procedure, but only the expression sequences associated with the person identified ($refE_{2e}, e = 1, 2, \ldots, Ne$) are used for distance measuring. Figure 3.12 shows the expression sequences associated with each individual of the database as a result of PCA. It can be observed that all sequences (represented by different colours) start and end at a common point, which corresponds to the neutral face. The point of each sequence which is located farther from this common point corresponds to the apex of each facial expression.

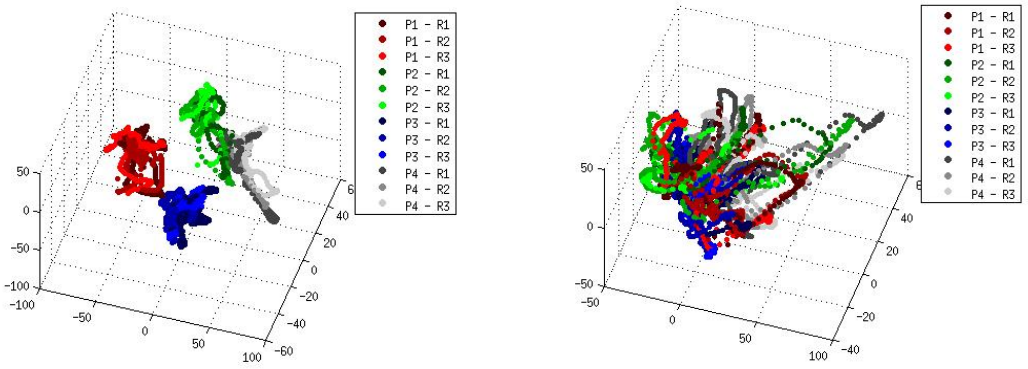

FIGURE 3.11: 3D representation of the sequences of concatenated expressions for each person using the original data set (left) and the subtracted data set (right).

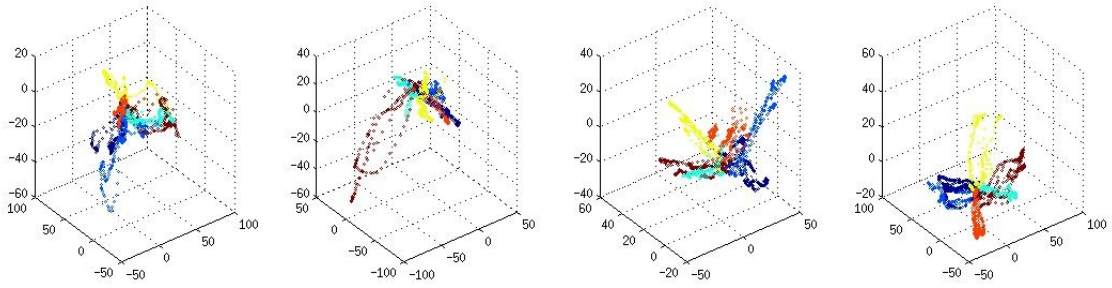

FIGURE 3.12: 3D representations of the different expressions performed by each individual. Each colour corresponds to one facial expression.

3.1.1 Experimental Results

This procedure was tested with *Database 1* and the results for identity recognition are shown in table 3.1. As in [16], the method is trained with 3 repetitions of each expression and person and tested with the remaining one, allowing direct comparison to be made. As expected, higher recognition rates are obtained when the original data set is used. From Fig. 3.11 it is clear that greater separation is obtained using the original data set and thus, it is easier to identify the individual to which a certain query sequence corresponds. However, even without shape information (when the subtracted data set is used), it is possible to obtain good results in identity recognition. Moreover, despite resulting in worse fits to the faces, better results were obtained using *ASM 2* due to the fact that this model comprises more landmarks than *ASM 1*, providing more shape and dynamic cues. When shape information is removed, the existence of more dynamic information results in significantly better recognition rates. Table 3.2 shows the results of expression recognition for each person in the database. General analysis demonstrates that the existence of shape information results in better recognition results. In expression recognition, using *ASM 2* leads to worse results due to the fact that the fitting to the faces is not perfect. In this case, greater precision is needed and thus flawed

Subtraction	**ASM**	**Average identity recognition**
No	1	100%
No	2	100%
Yes	1	80.21%
Yes	2	97.92%

TABLE 3.1: Overall identity recognition rates for different input parameters.

		Individual 1	**Individual 2**	**Individual 3**	**Individual 4**	**Overall**
Subt.	ASM	Exp.Recog.	Exp.Recog.	Exp.Recog.	Exp.Recog.	**Rate**
No	1	79.17%	87.50%	95.83%	95.83%	89.58%
No	2	75.00%	91.67%	95.83%	87.50%	87.50%
Yes	1	66.67%	75.00%	91.67%	83.33%	79.17%
Yes	2	41.67%	66.67%	95.83%	62.50%	67.71%

TABLE 3.2: Results on expression recognition for each individual in the database.

fittings result in errors. These errors are more evident when shape information is removed, leading to worse recognition results. Note that all the results presented were obtained by testing this method with each of the four repetitions of every expression, after training with the remaining three repetitions. These are average results and not the best results obtained. The results obtained with this method can be compared to the results obtained in [16]. Considering identity recognition using the original data set and *ASM 1*, 100% of the sequences were classified correctly with the present method. In [16], an identity recognition rate of 96.88% was obtained, which is a lower value. Moreover, an overall expression recognition rate of 89.58% was obtained with this method, which is a result about 4% better than the one obtained in [16]. Thus, the present method leads to overall better recognition rates, when compared to the existing method. Despite providing good results, the present method has the disadvantage that $Ne \times Np$ reference sequences have to be created for performing both identity and expression recognition. Moreover, performing DTW is computationally expensive when using data sets with a higher number of individuals. In order to try and solve these issues, a new procedure is presented using the idea that each individual and expression can be represented as a single point, increasing the separation between them, and simple euclidean distances can be used for classification.

3.2 Tensor Analysis of Shape Sequences

Previous work has shown that tensors can be used for analysing facial dynamics since their decomposition provides the separation of the different components in orthogonal subspaces [2]. Each subspace is represented by a matrix, whose row vectors span the invariance of a certain characteristic (identity, facial expression, etc.) across the remaining characteristics. Thus, each of these vectors can be interpreted as a point in a v-dimensional space, where v is the

length of the vector. Since each of these points represents one element of the corresponding subspace, tensor analysis can be used to solve the aforementioned issue. As in the first procedure, a preprocessing step is performed identically in this procedure, with the only difference that the variance of the PCA is 99% so that more principal components are used. Afterwards, for each person and facial expression, DTW is performed in $r-1$ repetitions, where r is the total number of repetitions, with a reference sequence, which is one of them. The outcome is $r-1$ new sequences with a fixed length from which a mean sequence is computed. Lastly, these newly determined $Np \times Ne$ sequences are used for constructing a 4D tensor.

Each frame of each sequence is represented by a point in a d-dimensional space, where d is the dimension obtained from PCA, and each sequence contains Nf frames. To define a fourth-order tensor which represents the entire data set, the sequences are organised so that identity, facial expressions, principal components and sequence length (frame order) are encoded in orthogonal dimensions. More specifically, for a 4D tensor $\mathcal{D} \in \mathbb{R}^{I \times J \times K \times L}$, I is the number of people in the data set, J is the number of facial expressions (in this case 6), K is equal to d, and L equals Nf. Using HOSVD, the tensor $\mathcal{D}$ is expressed as in equation 2.13, that is, as the mode-n product of a core tensor, $\mathcal{Z}$, and 4 orthogonal subspace matrices, $\mathbf{U}_{people}$, $\mathbf{U}_{expressions}$, $\mathbf{U}_{pc}$, and $\mathbf{U}_{length}$:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{expressions} \times_3 \mathbf{U}_{pc} \times_4 \mathbf{U}_{length}. \tag{3.4}$$

For the people subspace matrix $\mathbf{U}_{people}$, each row vector represents one person and the row vectors span the space of the people from the database across different expressions, principal components and sequence frames. The expression subspace matrix $\mathbf{U}_{expressions}$ row vectors span the space of facial expressions and describe the invariance across different people and shape and temporal information. In this work, the main focus of interest are these two subspace matrices since they encode the information which is relevant for analysis.

In this method, testing is performed as follows:

Tensors $\mathcal{B}_{exp}$ and $\mathcal{B}_{per}$ are computed by

$$\mathcal{B}_{exp} = \mathcal{Z} \times_2 \mathbf{U}_{expressions} \times_3 \mathbf{U}_{pc} \times_4 \mathbf{U}_{length}. \tag{3.5}$$

$$\mathcal{B}_{per} = \mathcal{Z} \times_1 \mathbf{U}_{people} \times_3 \mathbf{U}_{pc} \times_4 \mathbf{U}_{length}. \tag{3.6}$$

For every query sequence:

1. DTW with a reference sequence is performed, resulting in a sequence with length Nf;
2. A $4D$-tensor, $\mathcal{D}_{new}$, of size $1 \times 1 \times d \times Nf$, is constructed using the new sequence;

3. $\mathcal{D}_{new}$ is flattened along dimension 1, resulting in a matrix of size $1 \times d \cdot Nf$;

4. For every facial expression, the corresponding subtensor ($\mathcal{B}_{exp}i, i = 1, \ldots Ne$) is flattened along dimension 1;

5. These two new matrices are multiplied using the pseudo-inverse ($\mathbf{s}_{new} = \mathbf{D}_{new(1)}[\mathbf{B}_{exp}i_{(1)}]^{\dagger}$), resulting in a vector which represents the projection of the new sequence in the people subspace (Fig. 3.13) ;

6. The minimum euclidean distance between these vectors (projections) and the row vectors of the people subspace matrix provides the identity recognition;

7. Suppose that individual *idv* was recognised. A similar procedure is performed for expression recognition, using the matrix obtained from flattening tensor $\mathcal{B}_{per}idv$;

8. The shortest distance between the vector obtained (sequence projected in the expression subspace - Fig. 3.13) and the row vectors of the expression matrix provides the expression recognition.

The left and right images of figure 3.13 show the 2D representations of the row vectors of the identity and expression subspace matrices, respectively, as filled circles. Also, all the query sequences are projected into the people and facial expression subspaces (Fig. 3.13), represented as asterisks. Different people and facial expressions are represented by different colours and it can be observed that sequences representing the same element (and thus have the same colour) cluster together. This property allows distance between points to be a classifier.

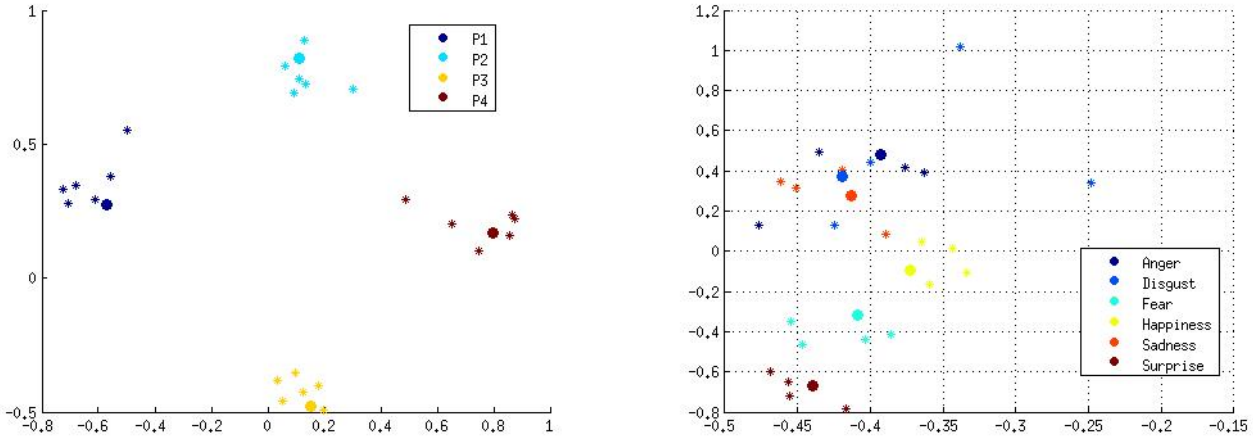

FIGURE 3.13: 2D representation of the points representing each person (left) and expression (right), and the projections of the test sequences.

Subtraction	**ASM**	**Average identity recognition**
No	1	100%
No	2	100%
Yes	1	90.62%
Yes	2	96.88%

TABLE 3.3: Overall identity recognition rates for different input parameters.

		Individual 1	**Individual 2**	**Individual 3**	**Individual 4**	**Overall**
Subt.	ASM	Exp.Recog.	Exp.Recog.	Exp.Recog.	Exp.Recog.	**Rate**
No	1	75.00%	100%	95.83%	91.67%	90.63%
No	2	70.83%	95.83%	91.67%	87.50%	86.46%
Yes	1	70.83%	95.83%	95.83%	100%	90.63%
Yes	2	66.67%	100%	87.50%	83.33%	84.38%

TABLE 3.4: Results on expression recognition for each individual in the database.

3.2.1 Experimental Results

Using this procedure, results for identity and expression recognition are shown in tables 3.3 and 3.4, respectively. Tests were performed in identical conditions as in section 3.1. General comparison between this and the previous procedure shows that overall results improved. In identity recognition, an improvement of 10% has been obtained when using *ASM 1* and the subtracted database. Expression recognition shows that results are coherent. As an example, individual 1 produced the worst results in both procedures. In addition, recognition rates increased significantly when using the subtracted data set: about 10% for *ASM 1* and 17% for *ASM 2*. This leads to the conclusion that tensor analysis provides better separation between individuals and facial expressions than PCA in conjunction with DTW. It can also be observed that similar results are obtained in expression recognition when using the original or the subtracted data sets. This demonstrates that it is possible to recognise correctly the facial expression when using only dynamic cues, which was not evident with the previous procedure. The results obtained show that, despite being computationally simple and inexpensive, this method is efficient. Comparing to the previous method, this one does not require the individual computation of a reference for each person and facial expression, since the separation of subspaces solves this problem. Tensor analysis is further explored throughout this work.

Up to this point, it has been shown that facial dynamics is indeed a proper biometric. However, since only a small database has been used, the conclusion is limited to a small number of individuals. Thus, it is necessary to perform experiments with larger data sets, which are more representative of real situations. Moreover, is it not easy to have access to databases with multiple repetitions of each facial expression. The next procedures are an attempt to overcome these issues.

3.3 Tensor Analysis of Shape Streams

Previous work [2] has shown that different streams of a certain facial expression, which are short sequences representing part of the facial expression, have similarities in the sense that when represented in the same referential, they cluster together. These streams are constructed by sampling each emotion sequence using a windowing technique such that multiple samples of each sequence are ascertained. Using this technique, a database with a single sequence of each facial expression performed by each individual can be used since both training and testing samples can be obtained in significant amounts. Moreover, some databases do not include the complete facial expressions (onset-apex-offset), making this technique advantageous because only parts of the facial expressions are needed. The procedure presented in this section makes use of this technique, being this the main difference when compared to the previous procedure. Since the streams all have the same length, it is not necessary to apply DTW. Besides using only euclidean distances for classification, Self Organising Maps (SOM) are also used in the testing phase, as in [2]. The present procedure is, in some aspects, similar to [2]. However, the main differences include the usage of techniques for tracking features, instead of ASMs, and the fact that the method is not used for identity recognition.

In this case, since each training sequence (a stream) represents a different part of the facial expression, it is not logical to construct the tensor using the mean sequence of the training sequences. Thus, here the $4D$-tensor is defined as $\mathcal{D} \in \mathbb{R}^{I \times (J \cdot R) \times K \times L}$ where R is the number of streams of each facial expression used for training and I, J, K, L have the same meaning as in section 3.2. Based on the knowledge that streams of the same facial expression cluster together, two classification methods are applied using the row vectors of the facial expression subspace matrix obtained from HOSVD of $\mathcal{D}$. One method creates a SOM using these vectors and the query stream is projected to the map, allowing label comparison to be made. Figure 3.14 illustrates the SOM created using *Database 1*, where $Ei, i = 1, \ldots, Ne$ represent the 6 basic expressions. It is clear that, for different people, streams of the same facial expression are located in adjacent cells in the SOM, so that $2D$ spacial location can be used for classification. The other method uses only euclidean distances to the mean point of each cluster of row vectors, as illustrated in Fig. 3.15. Each dot represents a row vector and each cross represents the mean point. The same facial expression performed twice by the same individual, or by different individuals, may have very different durations, depending on the velocity with which it is done. Thus, one stream from each sequence may not only represent a different part of the facial expression, but may also represent a different percentage of the facial expression. Despite these facts, it has been observed that, in fact, facial streams cluster

together, which is of great relevance in this work.

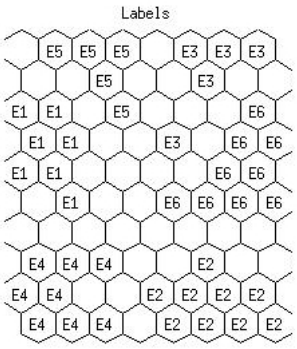

FIGURE 3.14: SOM representing the six basic facial expressions.

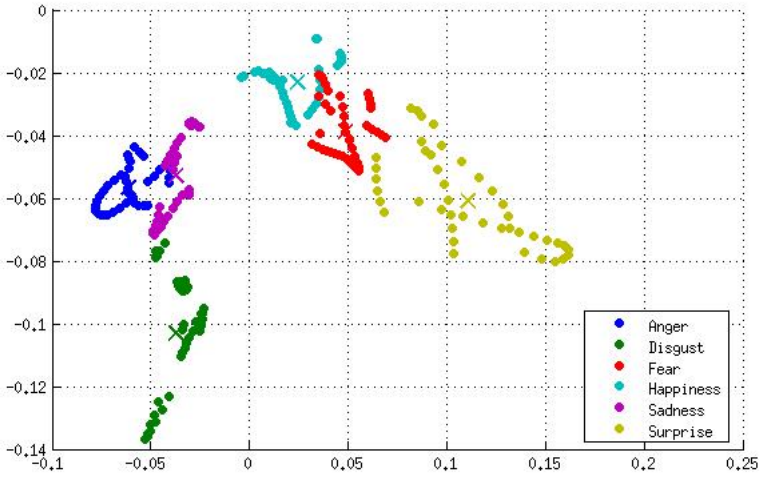

FIGURE 3.15: 2D representation of facial expression streams as well as the mean point.

3.3.1 Experimental Results

This new procedure is tested with databases 1 and 2, making use of its advantages. These are very different databases, and so allow a good evaluation of the method.

3.3.1.1 Database 1

Table 3.5 shows the overall identity recognition rates for *Database 1* under the same conditions as in sections 3.1 and 3.2. The training streams are sampled from three out of the four repetitions of each facial expression and individual, and the testing stream is determined using the central part of the remaining repetition, so that more dynamic information is

Subtraction	ASM	Average identity recognition
No	1	100%
No	2	93.75%
Yes	1	90.63%
Yes	2	87.50%

TABLE 3.5: Overall identity recognition rates for different input parameters.

present, and the query sequence is unseen (not part of the training set). The results were obtained using a window size of 50 frames and a step size of 5 frames. As in the previous procedures, using shape information (original data set) provides better recognition results. However, in this case, using *ASM 2* leads to poorer results (an average difference of about 8%), indicating that the misfits to the faces are more significant. This can be explained by the fact that using overlapping streams results in poor fits being present in more than one repetition.

Tables 3.6 and 3.7 show the results for expression recognition for each individual, using the two classification methods: distances and SOM, respectively. As in section 3.2, it can be observed that similar results are obtained either using using shape information or not. Slightly better results are obtained using SOM and it can be seen that the results are coherent in the sense that similar rates are obtained for the same people. This method leads to overall worse rates than the method in the previous section, with the difference ranging between 6% and 9%. The reason is that only a short part of each facial expression for each individual is shown during testing, increasing the ambiguity, especially if this part includes frames with the neutral expression. However, the results obtained are similar to the rates in [2].

Analysis of the influence of stream length and step size for both the original and the subtracted data sets is shown in table 3.8. *ASM 1* was used in these experiments. The table on the left shows the results for the subtracted data set. Under the same conditions, results for the original data set are shown in the table on the right. General analysis demonstrates that using longer streams leads to better results, which can be explained by the fact that these sequences include more dynamic information, necessary for the recognition of facial expressions, as opposed to short sequences which may only contain neutral faces. When shorter streams are used, it is advantageous to have small steps because it increases the number of training sequences. However, if the step is too small, there may exist more streams with a considerable amount of neutral faces, which are misleading sequences. Moreover, better expression recognition results are obtained with the subtracted data set. This can lead to the conclusion that better separation is obtained in this case due to the fact that subtracted sequences of the same expression performed by different individuals are more similar, having less interpersonal differences.

Subt.	ASM	**Individual 1** Exp.Recog.	**Individual 2** Exp.Recog.	**Individual 3** Exp.Recog.	**Individual 4** Exp.Recog.	**Overall Rate**
No	1	54.17%	95.83%	70.83%	91.67%	78.12%
No	2	62.50%	95.83%	79.17%	87.50%	81.25%
Yes	1	66.67%	95.83%	66.67%	91.67%	80.21%
Yes	2	62.50%	95.83%	75.00%	79.17%	78.13%

TABLE 3.6: Results on expression recognition for each individual in the database using distances.

Subt.	ASM	**Individual 1** Exp.Recog.	**Individual 2** Exp.Recog.	**Individual 3** Exp.Recog.	**Individual 4** Exp.Recog.	**Overall Rate**
No	1	66.67%	91.67%	79.17%	87.50%	81.25%
No	2	62.50%	95.83%	75.00%	87.50%	80.21%
Yes	1	79.17%	91.67%	75.00%	91.67%	84.38%
Yes	2	54.17%	95.83%	75.00%	83.33%	77.08%

TABLE 3.7: Results on expression recognition for each individual in the database using SOM.

		Step			
		1	5	10	15
Length	20	82.29%	71.88%	68.75%	57.29%
	30	70.83%	78.13%	82.29%	77.08%
	40	79.17%	83.33%	83.33%	80.21%
	50	82.29%	80.21%	85.42%	86.46%

		Step			
		1	5	10	15
Length	20	80.21%	75.00%	68.75%	54.17%
	30	73.96%	78.12%	79.17%	69.79%
	40	76.04%	77.08%	80.21%	82.29%
	50	78.12%	78.12%	82.29%	84.38%

TABLE 3.8: Results of overall expression recognition for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases.

3.3.1.2 Database 2

This procedure was also performed using *ASM 2*, and *Database 2*, which is a database with no repetitions. Thus, using each sequence, a set of streams is constructed and all but the central one are used for training. Testing is done with the central stream, which is an unseen sequence but can have a percentage of frames in common with streams from the training data set. Table 3.9 shows the results for identity and expression recognition (using two classification methods) for different stream lengths and step sizes. In each column, the values on the left were obtained using the original data set and the values on the right using the subtracted one. The values between parenthesis in the column with the step sizes indicate the maximum percentage of common frames between the query sequence and the training sequences. Thus, as expected, a higher percentage leads to better recognition rates for both identity and expression, since part of the query sequence is present in the training set. In this case, since there is only one sequence per person and expression, longer streams lead to a small amount of training samples, reducing identity recognition rates. With a reasonable amount of training samples, even without overlapping (or a very short percentage), it is possible to obtain good recognition results. The fact that slightly better recognition results can be obtained with the subtracted data set can be justified by the interpersonal differences, as mentioned above. It can also be observed that large step sizes lead to worse results when

Length	Step	% Id. Recog.	% Exp. Recog.	% Exp. Recog. (SOM)
25	14 (44%)	89.99 / 91.95	93.68 / 96.55	94.83 / 94.83
	22 (12%)	85.63 / 90.80	90.23 / 91.95	86.78 / 86.78
	25 (0%)	85.63 / 84.48	83.91 / 86.78	68.97 / 62.07
35	10 (71%)	91.95 / 90.80	94.25 / 94.25	95.40 / 94.83
	20 (43%)	86.21 / 84.48	86.21 / 87.36	86.21 / 86.78
	30 (14%)	77.01 / 77.01	79.31 / 81.03	44.25 / 73.56
45	5 (89%)	94.83 / 94.25	96.55 / 96.55	96.55 / 96.55
	15 (67%)	86.21 / 86.21	85.06 / 87.36	56.87 / 86.78
	24 (47%)	79.31 / 78.16	84.48 / 85.63	82.76 / 66.67

TABLE 3.9: Results of overall identity and expression recognition for different stream lengths and steps, and using both the original (left values) and the subtracted (right values) databases. The values between parenthesis indicate the maximum percentage of common frames between the query sequence and the training sequences.

using classification with SOM than with distance measures. A small amount of training samples can be the reason to this, since not enough samples are available for constructing the maps properly.

Good recognition results can be obtained using this procedure, even with databases which comprise many individuals. Thus, this is an efficient and useful method, with many possible applications nowadays. However, the usage of *ASM 2* produces misleading results and this procedure requires an ASM trained with images manually landmarked. Attempting to create a method which uses more facial information, such as texture cues, the following procedure uses grey images of the face and thus, does not require an ASM fully trained.

3.4 Tensor Analysis of Texture Streams

The main idea of this procedure is identical to the previous one, having the difference that instead of using ASMs for describing the faces, here each face is represented by a grey scale image of 40×40 pixels. Similarly to the previous method, here image subtraction is performed, i.e., for each pixel, its grey levels are subtracted and a new image is created. This subtraction attempts to remove or reduce texture and shape information. Figure 3.16 shows sampled frames of a subtracted sequence, where it can be seen that only the areas which exhibit greater dynamism are depicted. For creating the 40×40-pixel grey scale images, a software downloaded from [1] was used. This software locates faces in images, and creates a log with the face location, as well as the position of each eye. Since it "cuts" the images keeping the same area of the face, this algorithm removes scale and translation components. Using the positions of the eyes, it is possible to remove rotation components in relation to a reference face. Errors may occur if the eyes are totally or partially closed because the positions may be wrongly determined. Figure 3.16 shows an example of the outcome of this procedure. It can be seen that the different frames of the same sequence are properly aligned. In this

FIGURE 3.16: 40×40-pixel grey scale images sampled from a sequence of an individual performing a facial expression (happiness). The top and bottom rows show frames from the original and the subtracted database, respectively.

method, not only shape and dynamic information is used, but also texture information. VLBP is the used texture descriptor so that each stream is represented by an histogram of grey levels. Therefore, both texture and temporal information are encoded in this histogram, which is a vector containing the occurrences of each grey level. In [22], VLBP is used for face recognition. However, the method does not use tensor analysis, making the present procedure considerably different. This method can be considered innovative since no other method makes use of VLBP as descriptors for the construction of tensors. Each stream is created as follows:

1. Each frame of the sequence is divided into $M \times N$ non-overlapping blocks;
2. For each bloc, a volume with the depth equal to the stream length is defined;
3. For each volume, VLBP, as defined in subsection 2.2.2, is performed;
4. Each stream is created by concatenating the $M \times N$ histograms obtained: for every block in the same set of frames, a volume is defined.

Afterwards, all the training streams are used for computing PCA with 99% variance, so that dimensionality reduction is performed. This step is needed because the histograms contain tens of thousands of values, making this procedure computationally expensive. The resulting streams are used for constructing the tensor. Here, since the histograms encode texture and temporal information, $3D$-tensors were created, so that identity, facial expressions and VLBP codes are represented in orthogonal subspaces. More specifically, for a 3D tensor $\mathcal{D} \in \mathbb{R}^{I \times J \times K}$, I is the number of people in the data set, J is the number of facial expressions (in this case it is 6) multiplied by the number of repetitions of each expression and K is the dimension of each facial sequence, i.e., the dimension of the vector obtained after the PCA of the histogram resultant from the VLBP method. Using HOSVD, tensor $\mathcal{D}$ is expressed as in equation 2.13, that is, as the mode-n product of a core tensor, $\mathcal{Z}$, and 3 orthogonal subspace matrices, $\mathbf{U}_{people}$, $\mathbf{U}_{expressions}$ and $\mathbf{U}_{histVec}$:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{expressions} \times_3 \mathbf{U}_{histVec}. \tag{3.7}$$

$\mathbf{U}_{histVec}$ is the feature histogram subspace matrix, whose row vectors span the invariance of the texture and temporal information across the different people in the database and facial expression repetitions. For the people subspace matrix $\mathbf{U}_{people}$, each row vector represents one person and the row vectors span the space of the people from the database across different expressions and feature histograms. The expression subspace matrix $\mathbf{U}_{expressions}$ row vectors span the space of facial expressions and describe the invariance across different people and texture and temporal information. A $2D$ representation of the row vectors of $\mathbf{U}_{expressions}$ is depicted in Fig. 3.17. The classification method computes the distances to the mean of the row vectors (marked in Fig. 3.17 with crosses). This is the only classification method used because SOM did not produce significantly better results. The current procedure is tested with all the databases available, since there is not the need for specific training, as opposed to the case when ASMs are used. The following subsections present the results of the different tests.

3.4.1 Experimental Results

In the following experiments, VLBP with $L = 2, P = 4, R = 1$ has been used. This means that the neighbourhood of each pixel is defined considering 4 equally spaced pixels on a circle of radius 1 in the present frame (frame t) and 2 pixels in the frames $t - 2$ and $t + 2$. This procedure is tested with all four databases.

3.4.1.1 Database 1

Using *Database 1*, a set of experiments, identical to the ones described in the previous section, were conducted. Fixing the number of blocks in 5×4, and varying stream lengths and step sizes, the obtained results are shown in table 3.10. Using the original data set, for every stream length and step size, a 100% identity recognition rate was obtained. For the subtracted data set, high identity recognition rates were obtained (table on the left), demonstrating the efficiency of this method. In relation to the identity, higher separation is obtained if texture information is used, as expected. Thus, higher recognition rates are obtained in this case. The previous hypothesis that interpersonal differences cause expression recognition rates to decrease can be confirmed with this experiment since the results (shown in the middle and right tables) are better when the subtracted data set is used, for most of the stream lengths and step sizes. In order to evaluate the effect of the number of blocks used, with a combination of stream length and step size of 40 frames and 5 frames, respectively, results are shown in table 3.11. In theory, using more blocks leads to a more detailed description of the faces since the grey level histograms are created using smaller parts of the face. In practice, this can be confirmed by analysing the results which show that higher identity and expression recognition

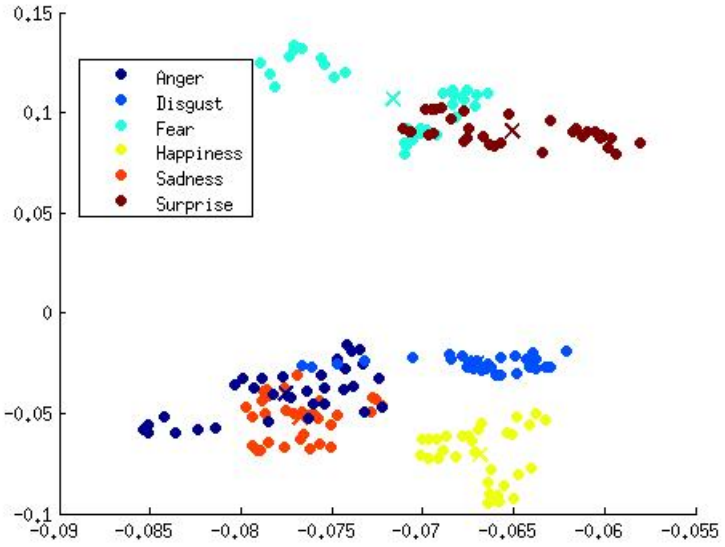

FIGURE 3.17: 2D representation of facial expression streams as well as the mean point.

		Length	
		40	50
Step	5	98.96%	97.92%
	10	98.96%	97.92%
	15	98.96%	97.92%

		Length	
		40	50
Step	5	89.58%	89.58%
	10	89.58%	89.58%
	15	89.58%	89.58%

		Length	
		40	50
Step	5	86.46%	87.50%
	10	90.62%	87.50%
	15	87.50%	87.50%

TABLE 3.10: Results of overall identity (table on the left) and expression (tables in the middle and on the right) recognition for different stream lengths and steps, and using both the subtracted (left and middle) and the original (right) databases. 5×4 blocks were used.

	Subtracted DB		**Original DB**	
Blocks	Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
1x1	76.04%	64.58%	89.58%	62.50%
3x2	91.67%	82.29%	100%	85.42%
3x3	94.79%	85.42%	100%	87.50%
5x4	98.96%	89.58%	100%	86.46%

TABLE 3.11: Recognition rates for different blocks sizes, using *Database 1* with stream length 40 and step size 5 frames.

rates are obtained as the number of blocks increases. Comparing to the previous method, this one produces significantly better results, due to the fact that much more facial information is used. In identity recognition, an improvement of about 9% has been achieved, when using the subtracted data set. Considering expression recognition, for both the original and the subtracted data sets, improvement ranges from 3% to 10%. The main disadvantage is that this procedure is computationally more expensive, so that using larger data sets with a great number of repetitions for each facial expression may be highly time consuming, during the training phase.

Length	Step	% Identity Recog.	% Expression Recog.
25	14 (44%)	100 / 100	100 / 100
	22 (12%)	100 / 100	91.38 / 99.43
	25 (0%)	98.28 / 98.38	83.91 / 97.13
35	10 (71%)	100 / 100	100 / 100
	20 (43%)	100 / 99.43	99.43 / 98.85
	30 (14%)	98.85 / 98.85	88.51 / 94.25
45	5 (89%)	100 / 100	100 / 100
	15 (67%)	100 / 100	100 / 100
	24 (47%)	100 / 100	100 / 97.70

TABLE 3.12: Results of overall identity and expression recognition for different stream lengths and steps, and using both the original (left values) and the subtracted (right values) databases. 3×2 blocks were used. The values between parenthesis indicate the maximum percentage of common frames between the query sequence and the training sequences.

	Subtracted DB		Original DB	
Blocks	Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
1x1	65.51%	83.91%	60.34%	49.43%
3x2	98.28%	97.13%	98.28%	83.91%
3x3	98.85%	97.70%	98.28%	89.08%
5x4	97.70%	98.28%	98.85%	93.68%

TABLE 3.13: Recognition rates for different blocks sizes, using *Database 2* with stream length 25 and step size 25.

3.4.1.2 Database 2

Experiments conducted with *Database 2* allow the previous hypothesis to be corroborated. The conditions in which the experiments were performed are the same as in the previous section, with the same database. General comparison shows that better results are obtained in this case (tables 3.12 and 3.13). Considering identity recognition, an improvement of up to 20% has been obtained, since using a significant number of blocks with this method, leads to identity recognition rates varying from 98% to 100%, as shown in table 3.12. Expression recognition rates also increased more than 10% in many cases, for both types of data sets. Firstly, keeping the number of blocks constant (3×2), stream length and step size were varied. Smaller steps, which correspond to higher percentages of common frames between training and testing sequences (overlapping), lead to higher identity and expression recognition rates, as already observed. In this case, even with very small or even non-existing overlapping, very high expression recognition results are obtained. As observed in the previous experiment, increasing the number of blocks leads to better recognition rates, for both identity and expression (table 3.13). It is clear that this method is significantly more efficient that the one which uses ASMs as face descriptors. Considering only expression recognition, experiments with *Database 2* show that the difference between rates obtained with the subtracted and the original data sets is higher than when using *Database 1*. This is due to the fact that larger intersubject differences are present, resulting in worse expression recognition results when

Blocks	Length, Step	Subtracted DB		Original DB	
		Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
3x2	15,10	73.33%	86.67%	88.33%	80.00%
	20,15	40.00%	55.00%	86.67%	33.00%
	25,15	71.67%	75.00%	96.67%	61.67%
5x4	15,10	88.33%	96.67%	98.33%	95.00%
	20,15	60.00%	73.33%	100%	58.33%
	25,15	90.00%	95.00%	100%	80.00%

TABLE 3.14: Recognition rates for different blocks sizes, stream lengths and step sizes, using *Database 3*.

using the original data set. This can be confirmed by observing figure 3.18, which depicts the differences between expression recognition rates obtained with the subtracted and the original data sets, as the number of individuals increases. A stream length and step size of 25 frames was used, as well as 3×2 blocks. Results clearly show that, as the number of people in the database increases, the difference becomes larger, corroborating the hypothesis presented.

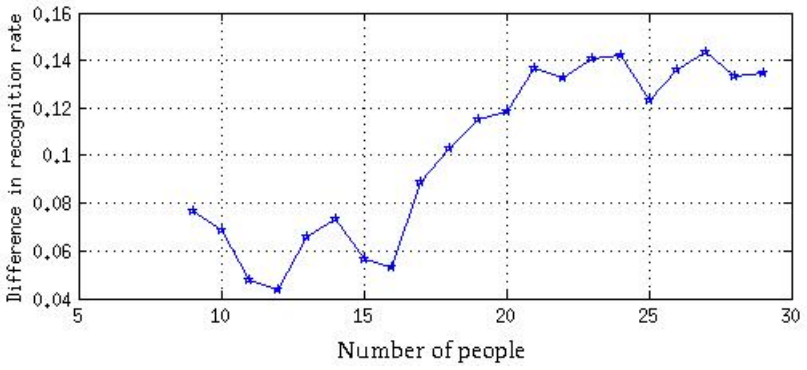

FIGURE 3.18: Difference between expression recognition rates obtained with the subtracted and the original data sets, for increasing number of people in the database.

3.4.1.3 Database 3

As mentioned before, the software used for locating the faces may produce poor results in certain conditions. In order to assure that the results obtained and previously presented are not influenced by poor registration, experiments are performed with *Database 3*, which is a registered database. Table 3.14 shows coherent results, since higher expression recognition rates are obtained when using the subtracted data set. Moreover, a greater amount of blocks, leads to better overall recognition results, as does a higher percentage of common frames between query and training streams, as already concluded.

3.4.1.4 Databases 1 and 4

Up to this point, it has been shown that facial dynamics is a proper biometric since good recognition results can be obtained even when shape and texture components are partially or totally removed. Theoretically, physical changes such as ageing, gaining weight, wearing glasses, growing a beard, etc., do not affect this kind of biometric, being this the main advantage and focus of this work. Thus, in order to test the efficacy of this procedure in the presence of significant appearance changes, a new database was created, *Database 4*, in which an individual performs the six basic expression, three times each. One of the repetitions is performed with the normal appearance, another one is performed with a painted face, with different colours, and, for the remaining one, the individual's face is covered with foam. The foam makes the individual almost unrecognisable, even for humans.
Experiments are performed using one sequence for every facial expression of each individual of *Database 1* and the sequences of *Database 4* with the normal appearance. This set of sequences is used for training, so that 5 individuals are considered. The remaining sequences of *Database 4* are used for testing. Since, in this case, it is desirable to assess the method as a biometric, and not an expression recognition system, only identity recognition results are shown. Figure 3.19 shows the $3D$ representations of the row vectors (filled circles) of the people subspace matrix for the subtracted (left picture) and the original (right picture) data sets, when using 40 frames as stream length, 20 frames as step size and one bloc. The projection of the query streams for the painted face (crosses) and for the face with foam (asterisks) are also depicted. Colours for the points relative to the query sequences indicate the result obtained, each colour corresponds to one of the 5 individuals. Since the brown circle is the target individual, it can be observed that for both data sets, streams relative to the painted face are located closer to the target circle than the ones relative to the fame with foam (note the scale of the graphics). This is explained by the fact that a greater appearance change is present in the latter case, and so the streams are projected to farther locations. Identity recognition results for different stream lengths and step sizes are shown in tables 3.15 and 3.16. The first two tables (3.15) correspond to the sequences performed with the painted face. It can be observed that even with the original data set, some streams were correctly classified, which indicates that the subject is not completely unrecognisable. However, as expected, much higher rates were obtained when using the subtracted data set because the identification was performed by using only dynamic information. In this case, very good recognition results were achieved, proving that facial dynamics is definitely a useful biometric. For the sequences filmed with the face covered with foam, the results are shown in table 3.16. In the case where the original data set is used, the significant change in appearance leads to incorrect classification for all the query streams. Although the dynamic information

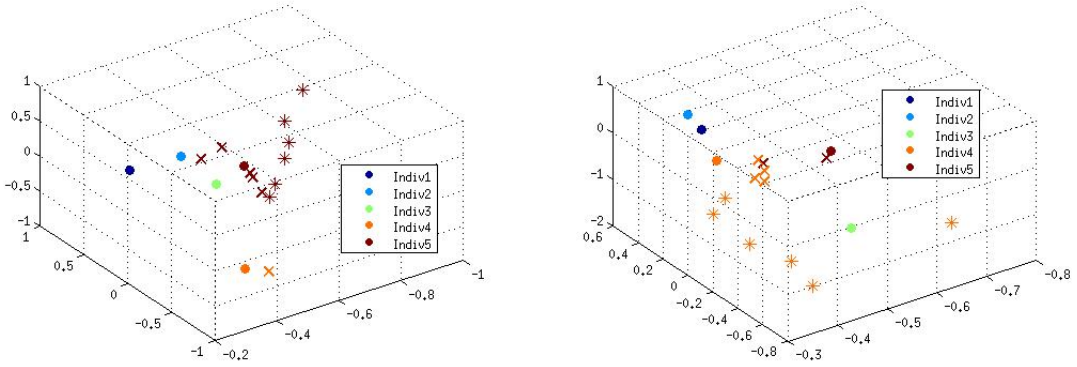

FIGURE 3.19: 3D representation of row vectors of the people subspace matrix using the subtracted data set (left) and the original data set (right). The projection of the query sequences is also shown.

		Step				
		2	5	10	15	20
Length	30	83.33%	83.33%	66.67%	66.67%	83.33%
	40	83.33%	83.33%	83.33%	83.33%	83.33%
	50	83.33%	83.33%	83.33%	83.33%	83.33%

		Step				
		2	5	10	15	20
Length	30	33.33%	33.33%	33.33%	33.33%	33.33%
	40	33.33%	33.33%	33.33%	33.33%	33.33%
	50	16.67%	16.67%	16.67%	16.67%	16.67%

TABLE 3.15: Results of overall identity recognition (painted face) for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases.

is present, the difference in appearance is such that it is impossible to correctly identify the individual. Reducing shape and texture components, leads to a great increase in the recognition rates. However, results in this case are slightly worse than for the painted face. This is due to the fact that the surface of the face is not smooth in this case and so, slight changes in the position of the face lead to significant changes in the subtracted images. Figure 3.20 shows frames of an original and subtracted sequence for both the painted face and the face with foam. It can be observed that the subtracted frames in the latter case contain areas which should have the grey level 128, indicating that there is no movement, but present other values (e.g. cheeks and forehead). This does not occur so significantly with the painted face, where there is greater definition of the face. Moreover, it has been observed that increasing the number of blocks leads to worse recognition results when using the subtracted data set for the face with foam. This can be explained by the fact that a more detailed description of the face is obtained, and so these areas that incorrectly present movement lead to errors.

It has been demonstrated that facial dynamics is an efficient biometric since it is clearly possible to perform identity recognition even when the individual is "masked". Based on this information, an attempt to emphasise facial dynamics is performed in the next section, where a new procedure which uses optical flow is presented.

		Step				
		2	5	10	15	20
Length	30	83.33%	66.67%	66.67%	83.33%	83.33%
	40	83.33%	66.67%	66.67%	66.67%	100%
	50	66.67%	83.33%	66.67%	66.67%	66.67%

		Step				
		2	5	10	15	20
Length	30	0.00%	0.00%	0.00%	0.00%	0.00%
	40	0.00%	0.00%	0.00%	0.00%	0.00%
	50	0.00%	0.00%	0.00%	0.00%	0.00%

TABLE 3.16: Results of overall identity recognition (face with foam) for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases.

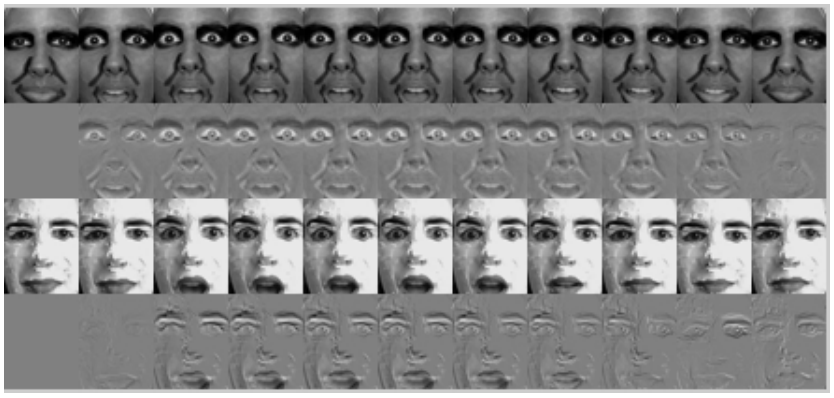

FIGURE 3.20: Example frames (original and subtracted) of the same individual performing a facial expression with different appearances.

3.5 Tensor Analysis of Optical Flow Streams

Comparing to the previous procedure, the only difference in the present one is that, instead of using grey images of the faces for computing VLBPs, this method calculates optical flow fields of the grey images (after alignment), both from the original and the subtracted data sets, as described in section 2.3. For each facial expression sequence, optical flow is computed for each pair of consecutive frames, so that the resulting sequence contains minus one frame than the sequence of images. This procedure is similar in some aspects to the one from [33], as it uses optical flow and LBP. However, it is used for gait recognition and does not use VLBP, as opposed to the present one. As explained in section 2.3, for every pixel of every frame there is a corresponding velocity vector $[v_x \quad v_y]$. Thus, for each sequence of images, the optical flow algorithm returns two sequences of velocity fields, one in the horizontal direction, and the other in the vertical direction. Figure 3.21 shows the velocity fields (bottom rows) in the horizontal (left) and vertical (right) directions for the two pairs of grey scale images. Notice that the pair of subtracted images is obtained from the pair of original images, by subtracting the first frame. Since optical flow only has non-zero values for pixels which exhibit motion, its computation using original data sets produces sequences which already greatly minimise shape components and remove texture information. However, since different results are obtained using the original or the subtracted data sets, experiments will be performed using both. The difference, as can be observed in figure 3.21, is mainly in the intensity of the pixels, which corresponds to velocity amplitudes, and not in the patterns obtained. In both pictures, the

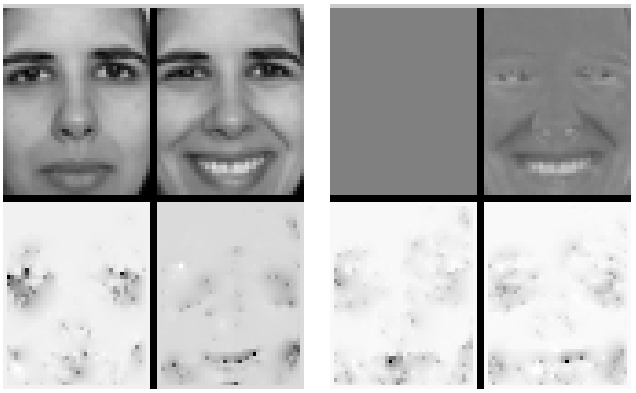

FIGURE 3.21: Example frames of *Database 4* (top rows) and corresponding directional velocity fields (bottom rows).

cheek and mouth regions are the ones which exhibit greater motion, as expected.
In order to use all the directional information, each sequence of optical flow fields (horizontal and vertical) is used for computing a histogram, as explained in section 3.4. For each sequence of images, the two histograms obtained are concatenated, producing a vector twice as long as the one obtained in the previous procedure. Thus, this method is computationally more costly, being this the reason for the choice of parameters in the following experiments.

3.5.1 Experimental Results

Experiments using this procedure are performed using databases 1, 3 and 4. *Database 2* was not used because it comprises a high number of individuals, making the training stage of this method computationally expensive. However, the remaining databases are sufficient for performing a proper evaluation.

3.5.1.1 Database 1

For the aforementioned reason, experiments with *Database 1* were performed similarly to the previous section, except for the cases where 5×4 blocks were used. Tables 3.17 and 3.18 show the results for different stream lengths and step sizes, using 3×2 blocks, for the original and the subtracted data sets, respectively. The values of the stream lengths used are 1 frame smaller than the ones used in the previous section because the complete sequences in the case are also one frame smaller. Considering identity recognition rates, it can be observed that using the original database slightly better results are obtained. This is due to the fact that static and dynamic areas are more distinguishable because the amplitudes of the velocity vectors have greater variation. However, in this case, the difference is smaller because, in both cases, texture and shape information is considerably minimised. Using the subtracted

		Length	
		39	49
Step	5	97.92%	100%
	10	98.96%	100%
	15	97.92%	100%

		Length	
		39	49
Step	5	79.17%	73.96%
	10	83.33%	73.96%
	15	71.88%	73.96%

TABLE 3.17: Results of overall identity (table on the left) and expression (table on the right) recognition for different stream lengths and steps, and using the original database. 3×2 blocks were used.

		Length	
		39	49
Step	5	98.96%	98.96%
	10	97.92%	98.96%
	15	96.88%	98.96%

		Length	
		39	49
Step	5	82.29%	83.33%
	10	81.25%	83.33%
	15	78.12%	83.33%

TABLE 3.18: Results of overall identity (table on the left) and expression (table on the right) recognition for different stream lengths and steps, and using the subtracted database. 3×2 blocks were used.

	Subtracted DB		**Original DB**	
Blocks	Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
1x1	56.25%	47.92%	79.17%	51.04%
3x2	92.92%	81.25%	98.96%	83.33%
3x3	98.96%	85.42%	100%	76.04%

TABLE 3.19: Recognition rates for different blocks sizes, using *Database 1* with stream length 39 and step size 10 frames.

data set leads to better overall expression recognition results. Direct comparison cannot be made since, in this case, only 3×2 blocks are used, as opposed to the previous case where there were 5×4 blocks, leading to more detailed descriptions and thus, better results. In table 3.19, the number of blocks is varied, maintaining the stream length and step size in 39 and 10 frames, respectively. It can be confirmed that increasing the number of blocks, leads to overall better identity and expression recognition results. Using a significant amount of blocks, good results can be obtained, demonstrating the effectiveness of this method.

3.5.1.2 Database 3

Identical experiments as the ones performed in the previous section using *Database 3*, were conducted with the present method. Table 3.20 shows the results. Direct comparison shows that using the subtracted data set, better results are obtained, leading to the conclusion that this method extracts more information from the same set of images. In identity recognition, up to 21% increase in the rates was achieved. Considering expression recognition, the improvement is not so significant. However, for many combinations of stream length and step size, an increase of over 5% was obtained. With the original database, expression recognition results are considerably better in this case, which is expected since most of the texture and

Blocks	Length, Step	**Subtracted DB**		**Original DB**	
		Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
3x2	15,10	66.67%	78.33%	80.00%	56.67%
	20,15	55.00%	60.00%	85.00%	50.00%
	25,15	85.00%	85.00%	93.33%	88.33%
5x4	15,10	90.00%	90.00%	95.00%	95.00%
	20,15	81.67%	73.33%	96.67%	71.67%
	25,15	100%	100%	100%	100%

TABLE 3.20: Recognition rates for different blocks sizes, stream lengths and step sizes, using *Database 3*.

shape information is not present, yielding less interpersonal differences. One important conclusion is that even without this texture and shape information, very good identity recognition results are obtained with this method.

3.5.1.3 Databases 1 and 4

Lastly, the present procedure is tested using *Databases 1* and *4*, identically to the previous section. Here, since using both the original and the subtracted data sets leads to the removal of texture and shape components, it is expected that both data sets yield good results in identity recognition. Figure 3.22 depicts the row vectors of the people subspace matrix (obtained from applying HOSVD to the tensor) as filled circles, and the projections of all the query sequences. Crosses represent the streams performed with the painted face and asterisks refer to the face with foam. For both the subtracted (picture on the left) and the original (picture on the right) data sets, it can be observed that all the query streams are projected to locations close to the target individual (brown circle). This does not happen when using the previous method (section 3.4) with the original data set, as already explained. Experiments performed with this method have shown that a 100% identity recognition rate is obtained for the streams performed with the painted face, whether using the original or the subtracted data set, and for many combinations of stream lengths, step sizes and number of blocks. Table 3.21 shows the identity recognition results for the streams performed with the individual's face covered with foam, using one bloc. It can be observed that using the original data set yields better results, which can be explained by the fact that greater variation in the amplitudes of the velocity vectors is obtained in this case, allowing a better distinction between dynamic and static areas. Comparing to the previous experiment with these databases, slightly worse results are obtained in this case, where at most one more query stream was incorrectly classified. However, increasing the number of blocks to 3×3 leads to a 100% recognition rate for every combination of stream length and step size, when using the original data set. This is a very significant improvement, which demonstrates that optical flow is an efficient descriptor of facial motion. Table 3.22 shows the results when using

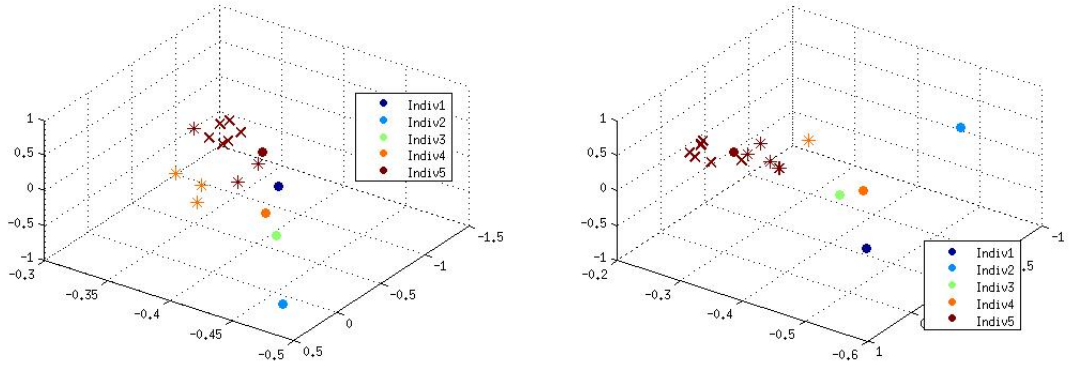

FIGURE 3.22: 3 representation of row vectors of the people subspace matrix using the subtracted data set (left) and the original data set (right). The projection of the query sequences is also shown.

		Step			
		5	10	15	20
Length	30	50.00%	50.00%	66.67%	66.67%
	40	66.67%	50.00%	66.67%	50.00%
	50	66.67%	66.67%	66.67%	66.67%

		Step			
		5	10	15	20
Length	30	66.67%	83.33%	50.00%	83.33%
	40	66.67%	66.67%	66.67%	83.33%
	50	66.67%	66.67%	66.67%	66.67%

TABLE 3.21: Results of overall identity recognition (face with foam) for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases. 1×1 blocks were used.

		Step			
		5	10	15	20
Length	30	100%	100%	100%	100%
	40	100%	100%	100%	100%
	50	100%	100%	100%	100%

		Step			
		5	10	15	20
Length	30	100%	100%	100%	100%
	40	83.33%	83.33%	83.33%	83.33%
	50	66.67%	83.33%	83.33%	83.33%

TABLE 3.22: Results of overall identity recognition for different stream lengths and steps, and using the subtracted database. The left table refers to the painted face and the right table refers to the face covered with foam. 3×3 blocks were used.

the subtracted data set, for both appearances of the individual's face, which are considerably better than the ones obtained without using optical flow. These experiments are very useful for demonstrating the superiority of the present method in relation to the previous one. It is evident that optical flow is a good descriptor of facial motion, being able to remove shape and texture information, which is crucial in experiments where the individual's appearance is significantly different from the original. A new procedure is proposed in the following section, making use of the properties of optical flow.

3.6 Optical Flow Tensors

The last procedure uses tensors for both subspace separation, as in the previous procedures, and sequence representation, as in [34]. Firstly, using the directional optical flow fields (horizontal and vertical), new sequences are created by calculating the angle of each velocity vector, in the interval $[-\pi/2, \pi/2]$, in the velocity fields. Figure 3.23 shows the phase component of the velocity field obtained from the two pairs of images. Darker areas (near black) correspond approximately $\pi/2$ rad. Light areas (near white) correspond to $-\pi/2$ rad, which are vertical velocities in the negative direction. Grey areas, near the grey level 128, correspond to null angles, which indicate horizontal velocities. For both the original and the subtracted data sets, it is possible to identify facial features in the phase image, such as the mouth and eye region. For the subtracted data set, the values obtained for the angles are more constant in the whole image. Afterwards, each stream of the phase of the velocity fields (part of a

FIGURE 3.23: Phase component of the velocity field obtained from pairs of frames from the original and the subtracted data set.

sequence) is used for creating a vector, as follows:

1. Each stream is divided into volumes, as explained in section 3.4, with the depth equal to the stream length;

2. For each volume, sub-volumes are created with the same width and height, and with depth equal to 3 frames. Each sub-volume is created by moving the window of size 3 frames with step 1 along the whole stream;

3. For each sub-volume, a histogram of the values of the velocity angles is created. The histograms have B bins, which is a user-defined value;

4. The histograms are used for creating a $4D$ tensor $\mathcal{T} \in \mathbb{R}^{M \times N \times V \times B}$, where $M \times N$ is the number of blocks, V is the number of sub-volumes in each volume and B is the number of bins. This step is slightly different from the one in [34], where $3D$ tensors are created because the volumes are not divided into sub-volumes. The goal here is to add a temporal component, as in VLBP;

5. Lastly, each tensor, which corresponds to a stream, is vectorised into a $M \cdot N \cdot V \cdot B$-element vector.

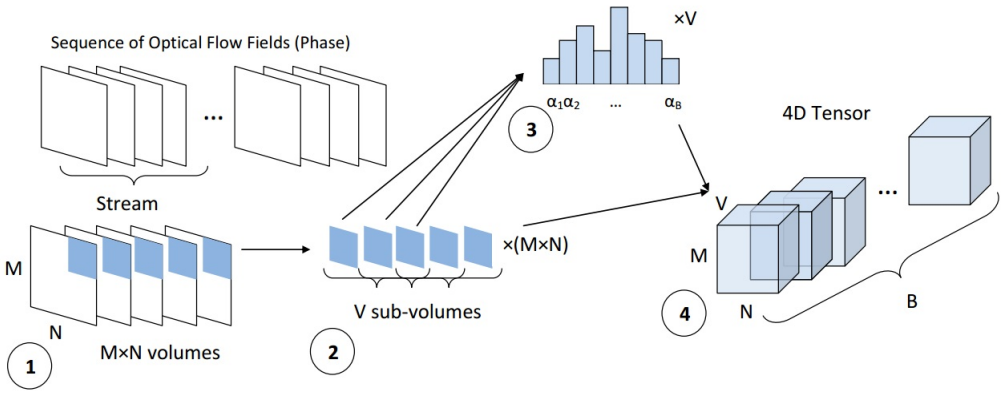

FIGURE 3.24: Steps of the creation of the vectors for constructing the tensor.

Figure 3.24 illustrates the first four steps of the creation of the vectors which represent the streams. The set of vectors created from the set of training streams are used for constructing the tensor, as in the previous procedures. Classification is performed in an identical manner.

3.6.1 Experimental Results

This last procedure is tested with databases 1 and 4, for assessing its performance. Recognition results are shown and discussed.

3.6.1.1 Database 1

Experiments with *Database 1* were performed for analysing the efficacy of this procedure. Tables 3.23 and 3.24 show the results for varying stream lengths and step sizes. It can be observed that using the subtracted data set leads to better identity and expression recognition results, which can be explained by the fact that less noisy information is present. From figure 3.23, it has been observed that more constant values of the angles are obtained with the subtracted data set. The great variations present when using the original data set constitute noise, leading to worse results. Comparing to the results obtained in the previous section with the original data set, better results can be achieved with the present method. This demonstrates that, even using only the phase component of the velocity fields, it is possible to obtain good recognition results. Results in table 3.25 demonstrate that increasing the number of blocks leads to better identity and expression recognition, as expected since a more detailed description of the facial motion can be obtained. Direct comparison with the experiments in the previous section show that the present method produces worse results, particularly if the number of blocks is small. Since the only information extracted from the optical flow fields is the phase (the amplitude is discarded), it is more difficult to perform

		Length	
		39	49
Step	5	83.33%	84.38%
	10	80.21%	84.38%
	15	81.25%	84.38%

		Length	
		39	49
Step	5	82.29%	75.00%
	10	77.08%	75.00%
	15	67.71%	75.00%

TABLE 3.23: Results of overall identity (table on the left) and expression (table on the right) recognition for different stream lengths and steps, and using the original database. 5×4 blocks and 36 bins were used.

		Length	
		39	49
Step	5	100%	96.88%
	10	98.96%	96.88%
	15	94.79%	96.88%

		Length	
		39	49
Step	5	88.54%	86.46%
	10	86.46%	86.46%
	15	77.08%	86.46%

TABLE 3.24: Results of overall identity (table on the left) and expression (table on the right) recognition for different stream lengths and steps, and using the subtracted database. 5×4 blocks and 36 bins were used.

	Subtracted DB		**Original DB**	
Blocks	Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
1x1	56.25%	30.21%	40.63%	22.92%
3x2	66.67%	69.79%	56.25%	63.54%
3x3	85.42%	72.92%	65.62%	61.46%
5x4	98.96%	86.46%	80.21%	77.08%

TABLE 3.25: Recognition rates for different blocks sizes, using *Database 1* with stream length 39, step size 10 frames and 36 bins.

	Subtracted DB		**Original DB**	
Bins	Identity Recog.	Expression Recog.	Identity Recog.	Expression Recog.
8	96.88%	83.33%	82.29%	73.96%
36	98.96%	86.46%	80.21%	77.08%
72	97.92%	87.50%	83.33%	77.08%
180	96.88%	89.58%	89.58%	82.29%

TABLE 3.26: Recognition rates for different blocks sizes, using *Database 1* with stream length 39, step size 10 frames and 5×4 blocks.

recognition in this case. A significant improvement with the increase in the number of blocks used is an indication that using only the phase information interferes with the recognition. Increasing the number of bins of the histograms, leads to a more specific information since the amplitude of each bin decreases, and different angle values are more distinguishable. Table 3.26 shows that, in fact, better recognition rates are obtained when more bins are used. In fact, very good recognition results are obtained in this case if the number of bins is high, showing that using only the phase components of the optical fields is enough for performing recognition. The drawback is that this makes the procedure more computationally costly.

		Step				
		2	5	10	15	20
Length	30	50.00%	50.00%	50.00%	50.00%	50.00%
	40	50.00%	33.33%	66.67%	50.00%	50.00%
	50	50.00%	50.00%	50.00%	50.00%	50.00%

		Step				
		2	5	10	15	20
Length	30	83.33%	100%	100%	100%	100%
	40	100%	100%	100%	100%	100%
	50	100%	100%	100%	100%	100%

TABLE 3.27: Results of overall identity recognition (face with foam) for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases. 1×1 blocks and 36 bins were used.

3.6.1.2 Databases 1 and 4

The present procedure was also tested with databases 1 and 4, as in the previous sections. Figure 3.25 shows a $3D$ representation of the row vectors relative to each individual as well as the projection of the query streams for both the painted face (crosses) and the face with foam (asterisks). Comparison with figures 3.19 and 3.22 show that, in this case, the distance between the target vector and the projections of the query sequences is considerably smaller when using the original data set. This indicates that this method is able to discard the noisy information that is present in this experiments, where the individual's appearance is altered. Results show that for the query streams sampled from the sequences filmed with the painted face, varying stream lengths, step sizes and the number of blocks, 100% identity recognition rate is always obtained, for both the original and the subtracted data sets. When testing with the face covered with foam, slightly worse results are obtained, especially with a small number of blocks, due to the fact that not enough detail is used. This can be observed in tables 3.27 and 3.28, which show the results when using only one block or a set of 3×3 blocks, respectively. Very good identity results are achieved when using the original data set: 100% identity recognition rates were obtained for most of the combinations of stream length and step sizes. Comparing to the results obtained in the previous section, this method yields better results, showing that the phase component is sufficient for performing recognition. For the subtracted data set, the present method produces worse results than the previous one, for the experiments performed with the face covered with foam. With the subtracted database, the grey levels of the faces are more constant. In this case, since altered faces are used, it is important to have a greater texture distinction between the more static areas of the face and the dynamic ones. This is obtained with the original data set. For the subtracted data set, static areas present variations in the grey levels of the same amplitude as the variations in the dynamic areas. Thus, the phase component of the optical flow fields in the static areas constitute noisy information, leading to worse results. Figure 3.26 shows the phase component of the optical flow fields computed for the four pairs of images. It can be observed that using the original data set leads to a greater distinction of the angles between static and dynamic areas. The present method demonstrates that using only partial information extracted from the optical flow fields is sufficient for performing identity and expression recognition.

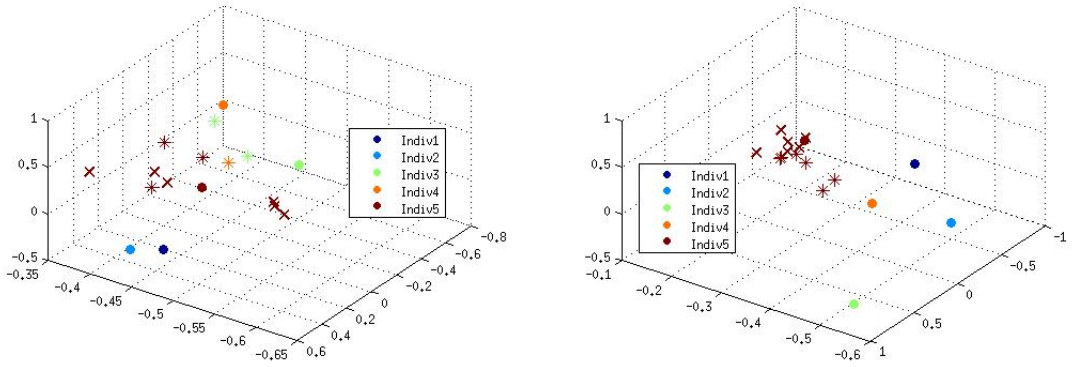

FIGURE 3.25: 3D representation of the row vectors of the people subspace matrix using the subtracted data set (left) and the original data set (right). The projection of the query sequences is also shown.

		Step				
		2	5	10	15	20
Length	30	50.00%	50.00%	50.00%	50.00%	50.00%
	40	66.67%	66.67%	66.67%	83.33%	83.33%
	50	100%	66.67%	66.67%	66.67%	66.67%

		Step				
		2	5	10	15	20
Length	30	100%	100%	100%	100%	100%
	40	100%	100%	100%	100%	100%
	50	100%	100%	100%	100%	100%

TABLE 3.28: Results of overall identity recognition (face with foam) for different stream lengths and steps, and using both the subtracted (left) and the original (right) databases. 3×3 blocks and 36 bins were used.

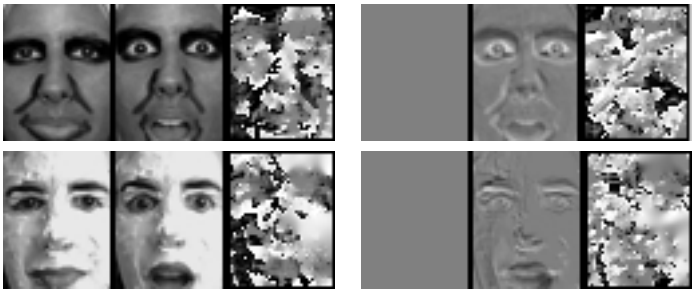

FIGURE 3.26: Angular component of the optical flow fields of the pairs of images from the original (left column) and the subtracted (right column) data sets.

This procedure yielded better results than the previous one when using the original data set, for the case when the individual's appearance is altered.

3.7 Description of the Databases

3.7.1 Database 1

The first data set, referred to as *Database 1*, is comprised of 112 facial expression sequences as it includes 4 people performing the 6 basic emotions (anger, disgust, fear, happiness, sadness

and surprise) plus the neutral one, 4 times each. Each sequence includes the onset, apex and offset of the facial expression. All four people are of the male gender. This database was created as part of the work developed in [16].

3.7.2 Database 2

This data set includes 29 individuals performing the six basic emotions, one time each. Not all facial expressions include the onset or the offset. Of these 29 people, 14 are female and 15 are male. This database includes a high variety of ethnicity. This database is a subset of the BU-4DFE Database [35].

3.7.3 Database 3

Database 3 includes 10 people (7 male, 3 female), each performing the six basic facial expressions once. The frames of this data set were created from $3D$ models, and thus the frames are registered, i.e., all the frames present zero rotation, translation and scale components in relation to a reference frame. Each sequence includes the onset, apex and offset of the facial expression. This database was created in the Computer and Robot Vision Laboratory of the Institute of Systems and Robotics at the University of Coimbra.

3.7.4 Database 4

This is the only database constructed as part of the present work. It includes one individual performing the six basic facial expressions, 3 times each. One repetition is performed with the normal appearance, another one is performed with the individual having her face painted and and, for the remaining one, the individual's face is covered with white foam, changing her appearance significantly. Each sequence includes the onset, apex and offset of the facial expression. Figure 3.27 shows example frames of the individuals of this database.

FIGURE 3.27: Example frames of *Database 4*.

Chapter 4

Conclusion

The main focus of this thesis is to demonstrate that facial dynamics - in the form of facial expressions such as happiness, surprise, and anger - constitute a biometric, i.e., can be utilised for performing identity recognition. Several identity and expression recognition methods were developed, mostly based on tensor analysis, due to its great properties in subspace separation. Different approaches in relation to data description have been considered, including Active Shape Models (ASM), which provide shape information, and Local Binary Patterns (LBP), which produce texture descriptors.

Throughout the implementation of this work, important conclusions and observations have been made. Firstly, it has been observed that, in the presence of a significant number of individuals, using dynamic information alone, by minimising or even removing shape and texture information, yields better recognition results. This indicates that the interpersonal differences, more evident in shape and texture cues, constitute disadvantageous information when encoding the different facial expressions. Moreover, it has been observed that performing data analysis using tensors produces better results than using Dynamic Time Warping (DTW) in conjunction with Principal Component Analysis (PCA). Since texture information gives a much more detailed description of the faces, significantly better results are obtained when using it, instead of using shape information alone. Even if the subtracted data sets are used, dynamic information is considered in every pixel of every frame, as opposed to case of ASM where only the N feature points which belong to the ASM are considered. Results obtained from experiments using Optical Flow fields have produced better results than using the grey scale images without processing. This demonstrates that extracting the dynamic cues using this method leads to a more informative data in the sense that the facial motion is described more accurately. This fact is clearly observed from the experiments performed with the database which includes the individual's appearance considerably altered, because very good recognition rates have been obtained. The relevance of the good results obtained

with this database is that they evince the fact that facial dynamics is indeed a proper biometric, because the individual cannot be identified using its texture and shape information alone, being this the main focus of this work.

Improvement in recognition rates could be achieved by using an ASM with more feature points, located in areas where motion is present, such as the top of the cheeks. The ASM must be trained with landmarked images of the individuals of the considered database for better fitting to the faces, which was not the case with one of the ASMs used in this work. Moreover, as already mentioned, the software for locating faces may return poor results in the presence of total or partial occlusion of the eyes. This leads to incorrect registration of the faces, which is undesirable when performing Volume LBP. Thus, another possible improvement would be to use a set of three points which would be tracked in each frame and afterwards used for registering the faces by rotating, translating and scaling in relation to a reference frame. However, the efficiency of the presented procedures shows that they are adequate for inclusion in automatic systems. It is important to notice that most of the computational cost is due to the training stage of the procedures, and testing is not significantly time consuming. Some novelty is introduced because the proposed methods constitute innovative aggregations of existing techniques.

In the future, similar experiments with the proposed procedures could be made, but incorporating 3D data and using both shape and texture descriptors. Moreover, it would be interesting to test the procedures with a registered database containing a significant number of individuals performing the facial expressions repetitively.

Bibliography

[1] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV*, 2008. http://www.milbo.users.sonic.net/stasm.

[2] Lisa Gralewski, Neill W. Campbell, Edward Morrison, and Ian Penton-Voak. Analysis of facial dynamics using a tensor framework. *Journal of Multimedia*, 1(6):10–21, 2006. URL http://dblp.uni-trier.de/db/journals/jmm2/jmm1.html#GralewskiCMP06.

[3] P Ekman. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 335 (1273):63–69, 1992. URL http://www.ncbi.nlm.nih.gov/pubmed/1348139.

[4] Paul Ekman, Wallace V Friesen, and Joseph C Hager. *Facial Action Coding System*, volume 11. Consulting Psychologists Press, Stanford University, Palo Alto, 2002. URL http://linkinghub.elsevier.com/retrieve/pii/S1462901107001268.

[5] J N Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37 (11):2049–2058, 1979. URL http://www.ncbi.nlm.nih.gov/pubmed/521902.

[6] H Hill and A Johnston. Categorizing sex and identity from the biological motion of faces. *Current Biology*, 11(11):880–885, 2001. URL http://www.sciencedirect.com/science/article/B6VRT-438BMM3-T/1/f9a50f21d5025d25e1c8f2be7d5b214d.

[7] B Knight and A Johnston. The role of movement in face recognition. *Visual Cognition*, 4(3):265–273, 1997. URL http://discovery.ucl.ac.uk/1341099/.

[8] Alice J OToole, Dana A Roark, and Herv Abdi. Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences*, 6(6):261–266, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12039608.

[9] Vinay Kumar Bettadapura. Face expression recognition and analysis: The state of the art. *Emotion*, pages 1–27, 2009. URL http://www.cc.gatech.edu/~vbettada/files/FaceExpressionRecSurvey.pdf.

[10] Caifeng Shan, Shaogang Gong, and Peter W. Mcowan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *In Proc. British Machine Vision Conference*, pages 297–306, 2006.

[11] Changbo Hu, Ya Chang, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5 - Volume 05*, CVPRW '04, pages 81–, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2158-4. URL http://dl.acm.org/citation.cfm?id=1032636.1032972.

[12] Jane Reilly and John McDonald. Modelling the manifold of facial expression using texture. *2008 International Machine Vision and Image Processing Conference*, pages 63–68, 2008. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4624386.

[13] Jane Reilly, John Ghent, and John McDonald. Investigating the dynamics of facial expression. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Paolo Remagnino, Ara Nefian, Gopi Meenakshisundaram, Valerio Pascucci, Jiri Zara, Jose Molineros, Holger Theisel, and Tom Malzbender, editors, *Advances in Visual Computing*, volume 4292 of *Lecture Notes in Computer Science*, pages 334–343. Springer Berlin / Heidelberg, 2006. ISBN 978-3-540-48626-8. URL http://dx.doi.org/10.1007/11919629_35.

[14] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *FG*, pages 406–413, 2011.

[15] Tingfan Wu, Marian S Bartlett, and Javier R Movellan. Facial expression recognition using gabor motion energy filters. *Neural Computation*, pages 42–47, 2010. URL http://mplab.ucsd.edu/~ting/pdfs/wu2010CVPR4HBposter.pdf.

[16] P Martins and J Batista. *Identity and expression recognition on low dimensional manifolds*, pages 3341–3344. 2009.

[17] Stefanos Zafeiriou. Facial behaviometrics : the case of facial deformation in spontaneous smile / laughter university of twente. *Analysis*, pages 13–19, 2011.

[18] J. Shermina. Application of locality preserving projections in face recognition. In *(IJACSA) International Journal of Advanced Computer Science and Applications*, volume 1, pages 82–85. 2010.

[19] L Benedikt, Darren Cosker, P L Rosin, and D Marshall. Assessing the uniqueness and permanence of facial actions for use in biometric applications. *IEEE Transactions on Systems Man and Cybernetics Part A Systems and Humans*, 40(3):449–460, 2010. URL http://dx.doi.org/10.1109/TSMCA.2010.2041656.

[20] Adam E. Gaweda and Eric Patterson. Individual identification based on facial dynamics during expressions using active-appearance-based hidden markov models. In *FG*, pages 797–802, 2011.

[21] Hyun-Chul Choi and Se-Young Oh. Facial identity and expression recognition by using active appearance model with efficient second order minimization and neural networks. *2007 International Symposium on Computational Intelligence in Robotics and Automation*, pages 131–136, 2007. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4269901.

[22] Abdenour Hadid, Matti Matti Pietikinen, and Stan Z Li. Learning personal specific facial dynamics for face recognition from videos. *Learning*, 4778:1–15, 2007. URL http://www.springerlink.com/index/n000007027576324.pdf.

[23] Xiaoming Liu and Tsuhan Chen. Video-based face recognition using adaptive hidden markov models. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003 Proceedings*, 1(18-20 June 2003):I–340–I–345, 2003. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1211373.

[24] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION*, pages 447–460, 2002.

[25] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995. ISSN 1077-3142. doi: 10.1006/cviu.1995.1004. URL http://dx.doi.org/10.1006/cviu.1995.1004.

[26] Zhao G and Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. URL http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab.

[27] Jason Chi. Optical flow, 2008. URL http://chi3x10.wordpress.com/2008/06/09/optical-flow-in-matlab/.

[28] Laurens van der Maaten. Matlab toolbox for dimensionality reduction, November 2010. URL http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html.

[29] Pau Mico. Continuous dynamic time warping, 2007. URL http://www.mathworks.com/matlabcentral/fileexchange/16350-continuous-dynamic-time-warping.

[30] S. Nagy, Z. Petres, and P. Baranyi. TP tool – a MATLAB toolbox for TP model transformation. In *Proceedings of the 8th International Symposium of Hungarian Researches on Computational Intelligence and Informatics*, pages 483–495, Hungarian, 2007. URL http://tptool.sztaki.hu.

[31] Tuevo Kohonen. The self-organizing map. In *PROCEEDINGS OF THE IEEE*, volume 78, 1990.

[32] Som toolbox, jun 1991. URL http://www.cis.hut.fi/somtoolbox/.

[33] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, De Zhang, and James J. Little. Incremental learning for video-based gait recognition with lbp flow. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, June 2011.

[34] Xinbo Gao, Yimin Yang, Dacheng Tao, and Xuelong Li. Discriminative optical flow tensor for video semantic analysis. *Comput. Vis. Image Underst.*, 113(3):372–383, March 2009. ISSN 1077-3142. doi: 10.1016/j.cviu.2008.08.007. URL http://dx.doi.org/10.1016/j.cviu.2008.08.007.

[35] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *FG*, pages 1–6, 2008.