



# *Lip Analysis for Person Recognition*

*Usman Saeed*

A doctoral dissertation submitted to:

TELECOM ParisTech

in partial fulfillment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Major subject: Image Processing, Biometrics

Approved by the following examining committee:

Supervisor:	Prof. Jean-Luc Dugelay
President of the jury:	Prof. Andrzej Drygajlo
Examiner:	Dr. Christophe Garcia
Examiner:	Prof. Alice Caplier
Member:	Prof. Christian Wellekens

12<sup>th</sup> February 2010



## Abstract

The human face is an attractive biometric identifier and face recognition has certainly improved a lot since its beginnings some three decades ago, but still its application in real world has achieved limited success. In this doctoral dissertation we focus on a local feature of the human face namely the lip and analyse it for its relevance and influence on person recognition. In depth study is carried out with respect to various steps involved, such as detection, evaluation, normalization and the applications of the human lip motion.

Initially we present a lip detection algorithm that is based on the fusion of two independent methods. The first method is based on edge detection and the second one on region segmentation, each having distinct characteristics and thus exhibit different strengths and weaknesses. We exploit these strengths by combining the two methods using fusion. Then we present results from extensive testing and evaluation of the detection algorithm on a realistic database. Next we give a comparison of the visual features of lip motion for their relevance to person recognition. For this purpose we extract various geometric and appearance based lip features and compare them using three feature selection measures; Minimal-Redundancy-Maximum-Relevance, Bhattacharya Distance and Mutual Information.

Next we extract features which model the behavioural aspect of lip motion during speech and exploit them for person recognition. The behavioural features include static features, such as the normalized length of major/minor axis, coordinates of lip extrema points and dynamic features based on optical flow. These features are used to build client model by Gaussian Mixture Model (GMM) and finally the classification is achieved using a Bayesian decision rule. Recognition results are then presented on a text independent database specifically designed for testing behavioural features that require comparatively more data.

Lastly we propose a temporal normalization method to compensate for variation caused by lip motion during speech. Given a group of videos for a person uttering the same sentence multiple times we study the lip motion in one of the videos and select certain key frames as synchronization frames. We then synchronize these frames from the first video with the remaining videos of the same person. Finally all the videos are normalized temporally by interpolation using lip morphing. For evaluation of our normalization algorithm we have devised a spatio-temporal person recognition algorithm that compares normalized and un-normalized videos.

## Acknowledgments

I would like to pay my heartfelt gratitude to my supervisor, Prof. Jean-Luc Dugelay and my closest colleagues, Fedrico Matta, Jihene Bennour, Carmelo Velardo, Nesli Erdogmus, Antitza Dantcheva. I would like to thank my fellow students and colleagues from Eurecom, who have made my time at Eurecom a really memorable one: Umer, Najam, Rizwan, Sabir, Angela, Hajer, Benoit, Nicolas, Simon, Marco, Antony, Zuleita, Sara, Carina, Randa, Giuliana, Ikbal, Erhan, Turgut, Daniel, Konstantinos, Antonio, Shakti, ...

I would like to appreciate the jury members, for having dedicated a part of their time reading and evaluating this thesis. My thanks also to the research grants from Similar, BioBiMo, ActiBio, for having funded my research activities and several pleasurable trips to conferences.

Most of all, I would like to express my deepest gratitude to my family, their sacrifice and support was vital for the completion of this thesis.

# Table of Index

<b>ABSTRACT</b> .....	<b>3</b>
<b>ACKNOWLEDGMENTS</b> .....	<b>4</b>
<b>TABLE OF INDEX</b> .....	<b>5</b>
<b>LIST OF FIGURES</b> .....	<b>9</b>
<b>LIST OF TABLE</b> .....	<b>10</b>
<b>CHAPTER I. INTRODUCTION</b> .....	<b>12</b>
1. MOTIVATIONS.....	12
2. ORIGINAL CONTRIBUTIONS.....	13
3. OUTLINE .....	14
<b>CHAPTER II. INTRODUCTION TO BIOMETRICS</b> .....	<b>17</b>
1. INTRODUCTION .....	17
2. TYPES OF BIOMETRIC IDENTIFIERS.....	17
3. OPERATIONAL MODES.....	19
3.1. <i>Verification</i> .....	19
3.2. <i>Identification</i> .....	19
4. ARCHITECTURE.....	20
4.1. <i>Enrolment</i> .....	20
4.2. <i>Recognition</i> .....	21
4.3. <i>Adaptation</i> .....	21
5. PERFORMANCE EVALUATION.....	21
5.1. <i>Measures for Verification</i> .....	21
5.2. <i>Measure for Identification</i> .....	23
5.3. <i>Types of Errors</i> .....	24
6. LIMITATION AND ISSUES.....	25
6.1. <i>Accuracy</i> .....	25
6.2. <i>Scale</i> .....	25
6.3. <i>Privacy</i> .....	26
7. CONCLUSIONS.....	26
<b>CHAPTER III. STATE OF ART</b> .....	<b>28</b>
1. INTRODUCTION .....	28

---

2.	PRE-PROCESSING .....	30
2.1.	<i>Speech Segmentation</i> .....	30
2.2.	<i>Face &amp; Lip Detection</i> .....	30
3.	FEATURE EXTRACTION .....	30
3.1.	<i>Audio Feature Extraction</i> .....	30
3.2.	<i>Video Feature Extraction</i> .....	32
4.	CLASSIFICATION.....	32
4.1.	<i>Template Matching</i> .....	32
4.2.	<i>Stochastic Models</i> .....	33
4.3.	<i>Neural Networks</i> .....	35
5.	FUSION.....	35
5.1.	<i>Early Integration</i> .....	35
5.2.	<i>Intermediate Integration</i> .....	36
5.3.	<i>Late Integration</i> .....	36
6.	EXAMPLES OF LIP BASED PERSON RECOGNITION .....	37
6.1.	<i>Audio - Video Lip Biometric</i> .....	38
6.2.	<i>Video only Lip Biometric</i> .....	41
6.3.	<i>Conclusions</i> .....	45
7.	AUDIO-VIDEO SPEECH DATABASES .....	45
7.1.	<i>Introduction</i> .....	45
7.2.	<i>VALID Database</i> .....	46
7.3.	<i>Italian TV Database</i> .....	46
7.4.	<i>Other Databases</i> .....	48
<b>CHAPTER IV. LIP DETECTION &amp; EVALUATION .....</b>		<b>51</b>
1.	INTRODUCTION.....	51
2.	STATE OF ART: FACE DETECTION.....	51
2.1.	<i>Feature Based Techniques</i> .....	52
2.2.	<i>Image Based Techniques</i> .....	54
3.	STATE OF ART: LIP DETECTION .....	56
3.1.	<i>Image Based Techniques</i> .....	56
3.2.	<i>Model Based Techniques</i> .....	57
3.3.	<i>Hybrid Techniques</i> .....	58
4.	STATE OF ART: VISUAL LIP FEATURE .....	59
4.1.	<i>Static</i> .....	59
4.2.	<i>Dynamic</i> .....	60
5.	PROPOSED LIP DETECTION.....	61

---

5.1.	<i>Edge Based Detection</i> .....	62
5.2.	<i>Segmentation Based Detection</i> .....	63
5.3.	<i>Error Detection and Fusion</i> .....	64
5.4.	<i>Experiments and Results</i> .....	66
5.5.	<i>Conclusions</i> .....	69
6.	EVALUATION OF LIP FEATURES .....	69
6.1.	<i>Introduction</i> .....	69
6.2.	<i>Previous Work on Feature Selection</i> .....	69
6.3.	<i>Proposed Feature Extraction</i> .....	70
6.4.	<i>Feature Selection</i> .....	72
6.5.	<i>Experiments and Results</i> .....	73
6.6.	<i>Conclusions</i> .....	75
<b>CHAPTER V. APPLICATION OF LIP FEATURES</b> .....		<b>77</b>
1.	INTRODUCTION .....	77
2.	LIP FEATURES FOR PERSON RECOGNITION .....	77
2.1.	<i>Introduction</i> .....	77
2.2.	<i>Behavioural Lip Features</i> .....	77
2.3.	<i>Person recognition</i> .....	78
2.4.	<i>Results and experiments</i> .....	83
2.5.	<i>Conclusions</i> .....	86
3.	LIP FEATURES FOR HCI.....	86
3.1.	<i>Introduction</i> .....	86
3.2.	<i>Head gesture recognition</i> .....	87
3.3.	<i>Lip Reading</i> .....	90
3.4.	<i>Conclusions</i> .....	92
4.	LIP FEATURES FOR GENDER .....	92
4.1.	<i>Related Works</i> .....	92
4.2.	<i>Proposed Method</i> .....	94
4.3.	<i>Experiments and Results</i> .....	99
4.4.	<i>Conclusion</i> .....	102
<b>CHAPTER VI. LIP FEATURE NORMALIZATION</b> .....		<b>104</b>
1.	INTRODUCTION .....	104
2.	SYNCHRONIZATION .....	105
2.1.	<i>Synchronization Frame Selection</i> .....	105
2.2.	<i>Synchronization Frame Matching</i> .....	107
2.3.	<i>Person Recognition</i> .....	108

---

2.4.	<i>Experiments and Results</i> .....	109
2.5.	<i>Conclusions</i> .....	110
3.	NORMALIZATION.....	111
3.1.	<i>Optimal Number of Frames</i> .....	111
3.2.	<i>Transcoding</i> .....	111
3.3.	<i>Person recognition</i> .....	113
3.4.	<i>Experiments and results</i> .....	115
3.5.	<i>Conclusions</i> .....	117
<b>CHAPTER VII. CONCLUSIONS</b> .....		<b>119</b>
1.	CONCLUDING SUMMARY.....	119
2.	FUTURE WORKS.....	120
3.	PUBLICATIONS.....	122
<b>CHAPTER VIII. APPENDICES</b> .....		<b>124</b>
FACE AND EYE FEATURES FOR PERSON RECOGNITION.....		124
1.	FACIAL FEATURE EXTRACTION.....	124
1.1.	<i>Face Angle</i> .....	124
1.2.	<i>Face Symmetry</i> .....	125
2.	EYE DYNAMICS.....	126
3.	PERSON RECOGNIZER MODULE.....	127
4.	EXPERIMENTAL RESULTS AND DISCUSSIONS.....	127
5.	CONCLUSIONS AND FUTURE WORKS.....	128
<b>REFERENCES</b> .....		<b>130</b>

---

## List of Figures

FIGURE 1: DISTRIBUTIONS OF NORMALIZED SIMILARITY SCORES.[1] .....	22
FIGURE 2: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE. [1] .....	23
FIGURE 3: CUMULATIVE MATCH SCORES (CMSS). [1] .....	24
FIGURE 4: GENERIC AUDIO –VISUAL PERSON RECOGNITION SYSTEM .....	29
FIGURE 5: AN HMM WITH THREE STATES .....	34
FIGURE 6: FIRST 7 FRAMES FOR SOME OF THE TV SPEAKER.....	47
FIGURE 7: OVERVIEW OF LIP DETECTION. ....	62
FIGURE 8: A) MOUTH ROI, B) COLOR TRANSFORM, C) EDGE DETECTION.....	63
FIGURE 9: A) MOUTH ROI, B) COLOR TRANSFORM, C) REGION DETECTION.....	64
FIGURE 10: HISTOGRAMS FOR SEGMENTATION ERRORS.....	67
FIGURE 11: EXAMPLE OF IMAGES WITH 15 % SEGMENTATION ERROR .....	68
FIGURE 12: PIXEL INTENSITY PROFILES. ....	71
FIGURE 13: IDENTIFICATION RESULTS COMPARING DCT AND PCA COEFFICIENTS. ....	83
FIGURE 14: RECOGNITION RESULTS A) IDENTIFICATION RESULTS B) VERIFICATION RESULTS.....	85
FIGURE 15: A) DETECTED FACE B) TRACKING POINTS.....	89
FIGURE 16: LIGHT VARIATION.....	89
FIGURE 17: SCALE VARIATION .....	90
FIGURE 18: LIP DETECTION SEQUENCE AND SUPERIMPOSED IMAGE.....	91
FIGURE 19. ARCHITECTURE OF THE MULTIMODAL RECOGNITION SYSTEM.....	95
FIGURE 20. GENDER RECOGNITION RESULTS.....	102
FIGURE 21: (A) MOUTH ROI. (B) LK OPTICAL FLOW. (C) MEAN VECTOR.....	106
FIGURE 22: MEAN OPTICAL FLOW $OF_T$ FOR VIDEO .....	106
FIGURE 23: LIP FEATURE IMAGE .....	107
FIGURE 24:(A) EXISTING FRAMES (B) LIP ROI (C) MORPHED LIP ROI (D) MORPHED FRAME .....	112
FIGURE 25: ORIGINAL FRAMES AND TEMPORAL X-RAY IMAGE. ....	113
FIGURE 26 : CORRECT IDENTIFICATION RATES (CIR) .....	116
FIGURE 27 : VERIFICATION RATES (EER) .....	116
FIGURE 28: FACIAL FEATURE POINTS WITH FACE ANGLE.....	124
FIGURE 29: A) BACKGROUND REMOVED FACIAL IMAGE B) RIGHT FACIAL IMAGE C) LEFT FACIAL IMAGE D) LEFT-RIGHT OVERLAID IMAGE. ....	125
FIGURE 30: OPTICAL FLOW OF EYE MOTION.....	126

---

## List of Table

TABLE 1 : LIP BASED PERSON RECOGNITION SYSTEMS USING VIDEO INFORMATION.....	44
TABLE 2: TECHNICAL DETAILS OF ITALIAN TV DATABASE. [1].....	47
TABLE 3: ERRORS AND OR FUSION .....	66
TABLE 4: LIP DETECTION RESULTS.....	68
TABLE 5: RANK RESULTS FOR MRMR, BHATTACHARYA DISTANCE, MUTUAL INFORMRATION.....	73
TABLE 6: RANK FUSION .....	74
TABLE 7: IDENTIFICATION RATE FOR DIGIT RECOGNITION.....	91
TABLE 8: PERSON RECOGNITION RESULTS.....	110
TABLE 9: PERSON RECOGNITION RESULTS.....	115
TABLE 10: RESULTS FOR FEATURE VECTORS IN IDENTIFICATION RATES. ....	127



## *Chapter I. Introduction*

---

### 1. Motivations

Homeland security has become a central issue in the 21<sup>st</sup> century life and biometrics/video surveillance form the corner stone of our security apparatus. Amongst the various biometric identification systems such as fingerprint, DNA, which are deployed today, face based person identification despite being not the most performant, does provide certain attributes that make it a good choice; such as being non intrusive, easy to collect, and well-accepted by the general public.

Several trends can be observed in face recognition research. The first trend was initiated by the rapid decrease in the cost of webcams and video surveillance cameras. Thus recognizing people using video sequences instead of images has attracted the attention of the research community. Videos have certain advantages, they not only provide abundant data for pixel-based techniques, but also record the temporal information. The other notable trend in the field of face recognition has been the use of only physical information, thus ignoring the behavioural aspect. Behavioural information has been proven to be useful for discriminating identities and hybrid system combining physical and behavioural techniques not only improve recognition results but also offer robustness to variation, such as caused by illumination. This brings us to another trend in face recognition which is the development of spatial normalization techniques to reduce the variation caused by illumination, pose, etc. but the effects of temporal variation in facial videos have been completely ignored to the best of our knowledge by the research community.

Thus the first motivation of our work was to explore the behavioural aspect of face recognition by using lip motion. We believe that the speech pattern is a unique behavioural characteristic of an individual that is acquired over time, which can be used as a biometric identifier. Thus we have analysed text independent speech videos from individuals collected over an extended period of time (almost 2 years), to study their utility for person recognition. Some research has already been carried out to recognize people based on lip motion, but there have been several limitations. Firstly it is mostly focused on using physical features, not the behavioural ones; secondly they have experimented on comparatively short length, text dependent videos, which in our belief are inadequate to learn the underlying behavioural pattern.

The second motivation of our work has been to study the negative effects of lip motion as a source of variation. Degraded performance in face recognition has mostly been attributed to three main sources of variation in the human face, these being pose, illumination, expression and appropriate spatial normalization techniques have been proposed. But there is another mode of variation that has been conveniently neglected by the research community which is caused by speech. The deformation caused by lip motion during speech can be considered as a major cause of low recognition results, especially in videos that have been recorded in studio conditions where illumination and pose variations are minimal. Therefore we have studied the effects of lip motion and propose a temporal normalization algorithm that given several videos of a person uttering the same phrase, analysis the lip motion and attempts to reduce the effect of speech variation.

## 2. Original contributions

In this section we underline the original contributions of this thesis.

The first major contribution is related to detection and evaluation of the outer lip contour. Lip detection is achieved by the fusion of two independent methods, one based on the edge detection and the second one on region segmentation. The novelty lies in the fusion of two methods, which have different characteristics and thus exhibit different strengths and weaknesses. The other significance of this study lies in the extensive testing and evaluation of the detection algorithm on a realistic database. Next we also provided a comparison of visual features from lip motion for their relevance to person recognition. We extracted various geometric and appearance based lip features and compared them using three feature selection measures; Minimal-Redundancy-Maximum-Relevance (mRMR), Bhattacharya Distance and Mutual Information.

The second major contribution is based on the intuitive notion that people have a unique visual speech pattern. Thus we extracted static and dynamic features which model the behavioural lip motion during speech, special attention was paid not to include any physical attributes of the lip shape and appearance. Based on the outer lip contour that has been detected, behavioural features were extracted, which include static features, such as the normalized length of major/minor axis, coordinates of lip extrema points and dynamic features based on optical flow. These features were then modelled using a Gaussian Mixture Model (GMM) and finally the classification was done using a Bayesian decision rule.

The third major contribution is a temporal normalization algorithm to compensate for variation caused by visual speech. We propose a temporal normalization method that, given a group of videos for a person studies the lip motion in one of the videos and selects certain key frames as synchronization frames based on a criterion of significance (optical flow). Next a search algorithm compares these synchronization frames from the first video with the remaining videos of the same person, within a predefined window. Finally all the videos are normalized temporally using lip morphing. Evaluation of our normalization algorithm was carried out using a spatio-temporal person recognition algorithm from video information. By applying discrete video tomography, the algorithm summarizes the facial dynamics of a video sequence into a single image, which is then analyzed by a modified version of the eigenface.

Minor effort was made for application of lip features in the domain of Human Computer Interaction (HCI). We have designed a response registration system with two options for user interactions; the first one is based on gesture recognition elaborated by a single choice question for which the user can respond by nodding of the head. The second interface based on lip reading is illustrated by a multiple choice questions system for which the user only articulates the lip motion of the digit of choice. Another minor contribution was the application of lip features for Gender Recognition. We have proposed a hybrid system that combines the physical and behavioural information from the human face for gender recognition. The physical information is in the form of appearance and the behavioural in the form of head and lip motion. A unified probabilistic framework is then used to combine the physical and behavioural features for gender recognition.

### 3. Outline

This doctoral dissertation is organised as follows:

In Chapter II we provide an introduction to the fundamental notions of biometrics.

In Chapter III we review the literature on audio-visual person recognition using lip information.

In Chapter IV we present a novel lip detection algorithm, with an evaluation of the detection process and an evaluation of various lip features for person recognition.

In Chapter V we present applications of lip features with main focus on person recognition, and minor work in Gender Recognition and HCI.

In Chapter VI we propose a temporal normalization method for visual speech and investigate its effects on person recognition.

In Chapter VII we conclude this dissertation with a summary and some comments on future perspectives.

In Chapter VIII we present some initial results for person recognition using face and eye features. We also present a Gesture Recognition system elaborated by a single choice question for which the user can respond by nodding of the head.



## *Chapter II. Introduction to Biometrics*

---

### 1. Introduction

Biometric person recognition has gained vast interest in the scientific community due to several developments in the past few decades. The foremost has been their utility to the law enforcement agencies for public security and access to sensitive facilities, specially their application to air travel. Another reason has been their application in the electronic commerce and finance, where the need for secure access to restricted areas and resources is paramount. To a lesser extend it has also been applied for personalisation, where it can be used to adapt advertisements and products for identified clients.

The term biometrics has a Greek origin from the words bios (life) and metron (measure) and means “measure of life”, and is defined as

“The measurement and analysis of unique physical or behavioural characteristics as a means of verifying personal identity.”

The following properties are desirable in a biometric identifier.

*Universal:* Every one should possess it

*Permanent:* It should not vary over time.

*Distinctive:* It should be as different as possible for individuals.

*Robust:* It should be as similar as possible for same individual.

*Collectable:* It should be easy to collect, as non intrusive as possible.

*Acceptable:* Be accepted by the public as a biometric trait.

*Safe from attacks:* It should be difficult to alter or be reproduced by an impostor.

### 2. Types of Biometric Identifiers

Biometric identifiers are generally classified as being physical or behavioural, but it is not always possible to classify them with a clear distinction between physiological and behavioural, identifiers that exhibit both characteristics are classed as hybrid biometric identifiers.

The most important physiological biometric identifiers are the following:

*Fingerprint:* Fingerprints possess good discriminatory power and were the first biometric identifiers to be used in real recognition systems. But has negative connotation as it usually associated with crime.

*Iris:* The complex iris texture carries very distinctive information. Although iris recognition is a very promising, with regards to accuracy and speed, but it requires considerable user cooperation.

*DNA:* Deoxyribonucleic acid (DNA) contains the genetic information about individual. It represents the ultimate unique code for one's individuality, except for the fact that identical twins have identical DNA patterns. However, its practical application has been limited due to complex chemical analyses.

*Retina:* Retina is a light sensitive tissue lining the inner surface at back of the eyeball, the structure of retinal blood vessels are used as a biometric, which is characteristic of each individual and each eye. Although this biometric identifier is considered as one of the most secure, its intrusive nature has restricted its use.

*Hand and finger geometry:* Hand geometry recognition systems are based on a number of measurements taken from the human hand and fingers; the geometry of the hand is an inexpensive technique, well accepted and easy to collect, but not one of the most discriminating.

Then, a few examples of behavioural biometric identifiers are:

*Gait:* Gait is the particular way one walks, it is generally not as distinctive as other more accepted biometrics and may not remain constant over time, but it is well accepted by the population and is not intrusive.

*Keystroke dynamics:* It is hypothesized that people can be identified by the patterns of keystrokes. It is not expected to be unique and one may assume to observe large variations in typing patterns. Typing behaviour acquired by specialized training may also adversely influence recognition.

Finally, some examples of hybrid biometrics:

*Voice:* The acoustic patterns used in speaker recognition reflect both anatomy (size and shape of the throat and mouth) and behavioural patterns (voice pitch and prosody). Voice suffers from the presence of background noise and may not remain invariant over time.

*Face:* Facial appearance is a physiological trait, whereas facial motion can be characterized as behavioural. Face is a non intrusive, easy to collect and well accepted, but at the moment its accuracy is quite low, due to the variation caused by illumination, pose and expression.

*Signature:* The way a person signs his name is known to be a characteristic of that individual. The shape of the signature is typically a physiological pattern, while the speed and the inclination during the signature are behavioural. Negative aspects are that variation may exist between signatures of the same person and that it can be reproduced by professional forgers.

### 3. Operational Modes

Operational modes describe the set of the conditions which are themselves defined by the scenario under which a biometric recognition system works. A biometric recognition system has two main operational modes, verification and identification.

#### 3.1. Verification

In a verification scenario, a user presents his biometric identifier and claims an identity, the recognition system then verifies his claim and decides to accept it or reject it. The authentication process is done through a one-to-one comparison between the biometric pattern presented and the claimed model pattern stored in the system.

We can formally describe the verification problem as follows. If we consider an input feature vector  $\mathbf{x}$ , and a claimed identity  $\varpi$ , then a verification system must determine if the pair  $(\varpi, \mathbf{x})$  belongs to class  $k_\varpi$  or  $\bar{k}_\varpi$ , where  $k_\varpi$  is the client class (claim is true) and  $\bar{k}_\varpi$  is the impostor one (claim is false). If we represent the stored model pattern for user  $k_\varpi$  as  $\Theta_{k_\varpi}$ , the decision rule is the following:

$$(\varpi, \mathbf{x}) \in \begin{cases} k_\varpi & \text{if } S^{(VER)}(\mathbf{x}, \Theta_{k_\varpi}) \geq \theta \\ \bar{k}_\varpi & \text{otherwise} \end{cases}$$

where  $\theta$  is a predefined threshold, and  $S^{(VER)}(\mathbf{x}, \Theta_{k_\varpi})$  is the similarity score between the test  $\mathbf{x}$  and the model  $\Theta_{k_\varpi}$ .

#### 3.2. Identification

In an identification scenario, a user presents his biometric identifier but makes no claim about his identity, the system then matches this pattern to all models in the database to find the most likely identity, in a one-to-many comparison. It is more convenient for user as he needs not make a claim, but it can be more complex and less robust for the recognition system, due to a more difficult one-to-many scenario.

We can formally describe the identification problem as follows. If we consider an input feature vector  $\mathbf{x}$ , then an identification system must determine the identity of the user,  $k \in \mathbf{N}$ , where  $\{k | k = 1, \dots, K\}$  are the clients enrolled in the system and  $k = K + 1$  represents the reject case. If we denote the stored model pattern for identity  $k$  as  $\Theta_k$ , and with  $\theta$  the predefined threshold, the decision rule is the following:

$$\mathbf{x} \in \begin{cases} k & \text{if } \max_k \{S^{(ID)}(\mathbf{x}, \Theta_k)\} \geq \theta \quad k = 1, \dots, K \\ K + 1 & \text{otherwise} \end{cases}$$

where  $\theta$  is a predefined threshold and  $S^{(ID)}(\mathbf{x}, \Theta_k)$  is the similarity (or matching) score between the test  $\mathbf{x}$  and the model  $\Theta_k$ .

## 4. Architecture

A typical recognition system is composed by two modules, for the enrolment and recognition tasks, and can optionally have a third one, for the adaptation of user models.

### 4.1. Enrolment

The objective of the enrolment phase is to register new users in the recognition system. The biometrics of the user is first acquired using a capture device. Usually this is followed by a pre-processing step that enhances and normalizes the acquired information. Next, the quality of the pre-processed signal is checked to estimate if the acquisition was adequate or not. After that, the feature extraction step transforms the signal, trying to isolate the significant features that characterize the individual and to discard the irrelevant and redundant information. In most cases, the feature extractor computes a reduced representation of the pre-processed signal. In the end, the enrolment module estimates a model of the client, representing the potential range of biometric features for that user, and stores it in its internal database.

## 4.2. Recognition

The recognition module verifies and/or identifies users, the user presents his biometric identifier to an acquisition device, pre-processing and feature extraction phases are then carried out as in the enrolment module. Afterwards, the classification step compares the discriminative features of the test user with the model patterns retrieved from the database, and computes one or more similarity score(s) depending on the operational mode. The final decision of the system is determined by the operational mode in question: in a verification task the user claim is confirmed or rejected, while in identification the user identity is established.

## 4.3. Adaptation

The adaptation module is optional and it is useful for updating the user models stored in the database. Most of the biometrics are not permanent and vary over time, especially the behavioural and hybrid ones like: gait, voice, signature and face. Consequently, the actual biometric identifier of a client gradually differs from the original acquisitions, and may eventually lead to degradation in performance. Therefore, the adaptation module is meant to cope with those variations and to provide an updated representation of each user, it progressively adds new acquisitions and adapts the stored model using this new data.

# 5. Performance Evaluation

Automatic biometric person recognition is affected by several variations for example, variable or ambiguous acquisition protocol, changes in the user's physiological and behavioural state over time, ambient conditions such as lighting, etc. It is thus fundamental to evaluate the performance of a biometric system and to understand its strengths and limitations, this section is dedicated to performance evaluation measures and their uncertainties.

## 5.1. Measures for Verification

A system operating in a verification (or authentication) mode can make two major types of decision errors:

*False rejection:* occurs when a client makes a true identity claim is erroneously rejected.

*False acceptance:* occurs when an impostor who makes a false identity claim is erroneously accepted.

It is then possible to define the following four decision error measures:

*False rejection rate (FRR)*: the expected proportion of transactions with true claims incorrectly denied.

*False acceptance rate (FAR)*: the expected proportion of transactions with false claims incorrectly confirmed.

*Correct acceptance rate (CAR)*: the complementary measure to the FRR, and represents the expected proportion of transactions with true claims correctly confirmed. Mathematically:  $\eta_{\theta}^{(CAR)} \equiv 1 - \xi_{\theta}^{(FRR)}$  for  $\forall \theta$ .

*Correct rejection rate (CRR)*: it is the complementary measure of the FAR, and represents the expected proportion of transactions with false claims correctly denied. Mathematically:  $\eta_{\theta}^{(CRR)} \equiv 1 - \xi_{\theta}^{(FAR)}$  for  $\forall \theta$ .

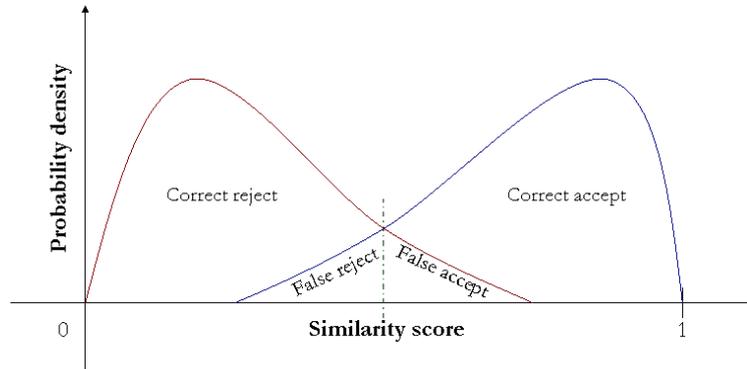


Figure 1: Distributions of normalized similarity scores.[1]

Figure 1 shows an example of client (blue curve) and impostor (red curve) distributions of normalized similarity scores; a client similarity score is calculated by matching a test pattern of a client with its model pattern, while an impostor similarity score is computed by matching a test pattern of a user (the impostor) with a different model pattern (the claimed client).

In an ideal situation where these distributions are disjoint, classification would be straight forward, but in real world these distributions overlap and a decision threshold (green dotted line) has to be selected. Therefore, because of this threshold create two regions of errors for false rejects and false accepts. When the threshold value is increased, the system becomes more secure with higher FRR and lower FAR; on the other hand if the threshold value is decreased, the system becomes less secure with lower FRR and higher FAR. The choice of a threshold value should be made carefully, by evaluating the context and requirements of the application in question.

A receiver operating characteristic (ROC) curve as shown in Figure 2, gives a full performance overview of a verification system, from low to high security configurations. It is practical way in which FARs are plotted as a function of FRRs. For drawing this graph, it is necessary to compute several pairs of FRRs and FARs at various threshold values.

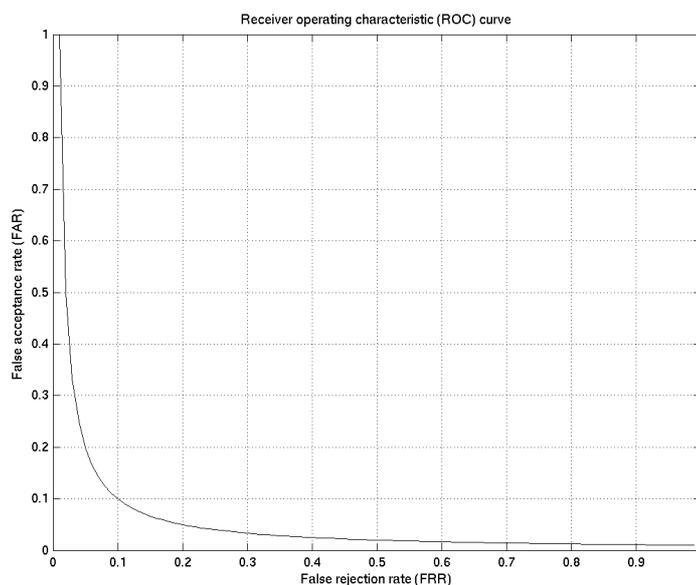


Figure 2: Receiver operating characteristic (ROC) curve. [1]

Even if it is preferred to present verification results by plotting the ROC curve, some authors just report a single error measure; it is the equal error rate (EER), a precise point on the ROC curve at which FRR and FAR are equal.

## 5.2. Measure for Identification

The most common measure for systems operating in the identification mode is the correct identification rate (CIR) which is defined as the percentage of test patterns matched to the correct model.

In an ideal case the correct match always has the highest similarity score, thus most authors compute the CIR by considering only the highest similarity score in each test i.e. the best match. However, in real cases the correct match can be lower than the top score, thus for a better insight on the recognition capabilities of a system.

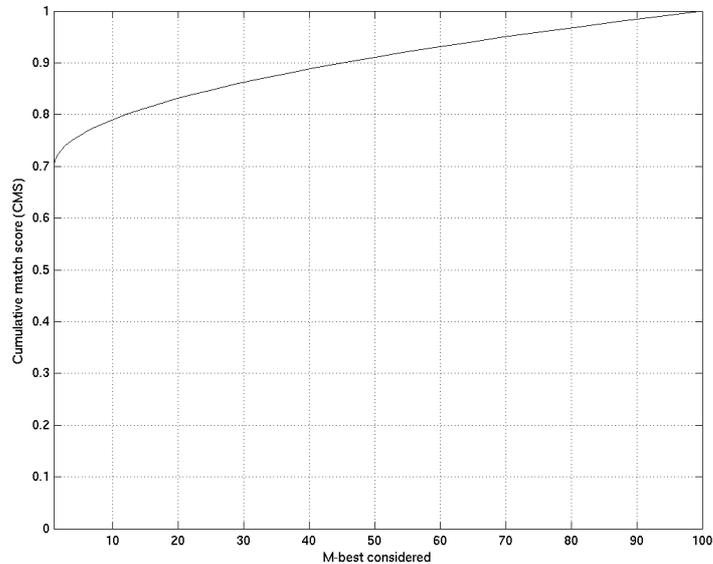


Figure 3: Cumulative match scores (CMSs). [1]

A cumulative match score (CMS) has been proposed, that is defined as the percentage of test patterns for which the correct match is among the highest  $M$  values. Figure 3 shows an example of CMSs plotted as a function of the  $M$  best similarity values retained.

### 5.3. Types of Errors

The performance evaluation of recognition system can be affected by two kinds of errors: systematic errors and random errors.

#### 5.3.1. *Systematic errors*

These errors are caused by the bias in the test procedure. One possibility is the uneven representation of certain classes, under or over representation problem, thus affecting the recognition rates. Solution for this lies in equal representation of classes in testing and training. Another bias may arise when client enrolment and recognition are done using the same data set. This causes over fitting in the models thus rendering the system vulnerable to changes in the operational environment and surely overoptimistic on the actual recognition capabilities of the system. To avoid this problem it is strongly recommended to use disjoint datasets for enrolling users and testing the system.

### 5.3.2. *Random errors*

These errors are caused by the natural variations in the clients and acquired samples, and their effect can only be reduced by increasing the number of tests. In fact, the number of users and the number of test samples affects how accurately we can interpret the measured error rates: the larger the number of test, the more accurate results are likely to be. However, increasing the number of tests is not always a viable option, thus it can be useful to collect multiple biometric identifiers per person or increasing the number of people tested, rather than the total amount of attempts.

## 6. Limitation and Issues

Automatic person recognition is a challenging pattern recognition problem, and several evaluation campaigns [2] have been carried out to evaluate various systems, some biometric recognition techniques such as fingerprint are mature enough for real world applications, but for the rest the present market is quite restricted. In this section we analyze the main issues and limitations which negatively affect the performances of biometric devices.

### 6.1. Accuracy

Accuracy in a biometric system is affected by several limitations. First is the lack of discriminative power of the acquired biometric pattern. This limitation could be further worsened due to the feature extraction process by which the discriminative power of the features maybe lost or redundant information retained. Another is that the biometric identifier may not be universally available in the entire population.

Accuracy could be further affected by variation in the environment, considering the example of a face recognition system, which is affected by illumination, pose and facial expressions, facial hair, presence or absence of eyeglasses, aging, etc.

### 6.2. Scale

In verification scale does not really matter, because it requires only a one-to-one comparison between the test biometric pattern and the claimed model pattern stored in the system. On the contrary, in an identification scenario scaling has a definitive effect when the number of clients is very large, because each unidentified pattern is matched up to all the model patterns present in the system, in a one-to-many comparison. Typical approaches to scaling include using multiple hardware units in parallel and coarse to fine pattern partitioning.

### 6.3. Privacy

Individual's right to privacy is taken very seriously in the modern society and is closely related to biometric security. The aim of biometric security is to enhance the privacy by providing secure access but in fact, it may be possible to track clients, infringing the individual right to privacy, also personal biometric data may be abused for unintended or criminal purposes.

The first major concern is identity theft, i.e. legitimate ownership of an identifier biometric or otherwise. It should not be possible for an impostor to spoof the biometric trait of a client, and then use it to be recognized in his place.

The second security issue is related with the integrity of the models enrolled in the database. In most cases, the enrolment of a new user is supervised by an operator, who can check the identity of the client and assures that the captured patterns are authentic. Though, when the enrolment can be unsupervised or when there is an adaptive procedure to update the client's model, it might be possible for an attacker to inject counterfeit biometric samples into the system, in order to corrupt the final decision results. The privacy of the individual requires both clever legislation and design of reliable and secure recognition applications.

## 7. Conclusions

In this chapter we have defined the fundamentals of biometric identification, starting with the basic definition and ideal properties that a biometric identifier should exhibit. Then we describe the basic types of biometric (i.e. physical and behavioural), with examples from modern biometric systems. After that we detail the two universally used operational modes and the generic architecture that is common to all biometric systems. We then examine various performance evaluation measures and the common errors of a biometric system. Finally we discuss some issues regarding biometrics such as privacy and scalability that have risen due to the use of biometric identification on a large scale in public places.



## *Chapter III. State of Art: Lip Based Audio-Visual Person recognition*

---

### 1. Introduction

This thesis covers an array of diverse topics from person and gender recognition to video normalization, thus to present a comprehensive state of the art on all the involved topics is impossible. Therefore in this chapter we provide a comprehensive state of the art on Audio-Visual (AV) person recognition using lip features. A point that we would like to emphasize is that only those works will be presented here that involve AV features extracted from the lips (mouth), e.g. works that have combined speech with complete face will not be presented. Where ever the need arises smaller and specific state of the arts will be presented in relevant chapters. AV person recognition mostly consists of three parts, pre-processing, feature extraction and classification, which will be presented in the next sections. We will also present some commonly used data fusion techniques and databases and at the end we will present examples of AV person recognition systems. Figure 4 describes the steps commonly involved in an audio-visual person recognition system.

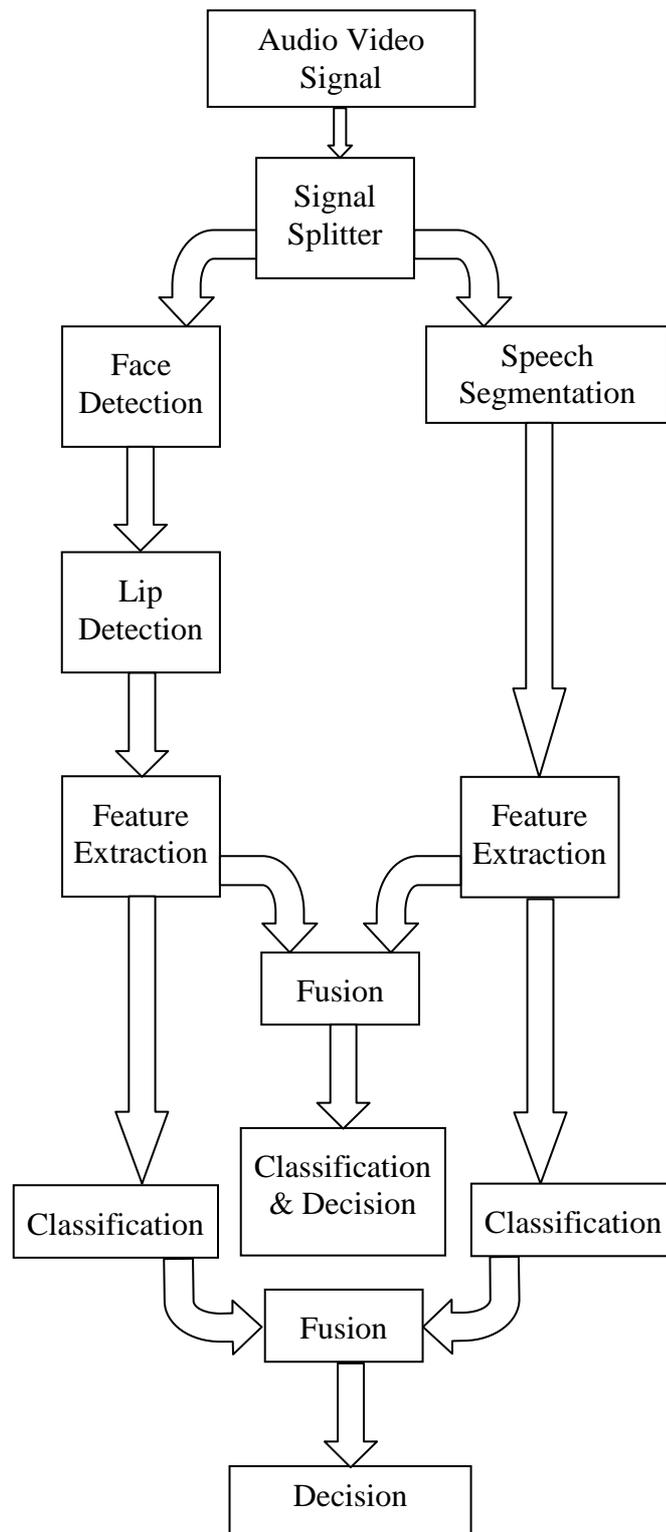


Figure 4: Generic Audio –Visual Person Recognition System

## 2. Pre-processing

The pre-processing step in an audio-visual person recognition system generally consists of segmentation of the audio signal, face detection and lip detection/segmentation.

### 2.1. Speech Segmentation

Segmentation in audio domain normally consists of separating the speech and non speech segments. This can be achieved by energy thresholding [3] zero crossing rate and periodicity measures [4]. [20] have proposed using a combination of features, which are spectral centroid, spectral flux, zero-crossing rate, 4Hz modulation energy, and the percentage of low-energy frames for discriminating between speech and various types of music.

Mel-frequency cepstral coefficients (MFCCs) [6] and perceptual linear prediction (PLPs) cepstral coefficients [7] which were originally designed for speech and speaker recognition, have also been successfully applied to speech/ non speech segmentation [8] in combination with Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs).

Conventional pattern recognition techniques can also be employed to learn the difference and hence segment the videos into speech and non-speech, typically techniques such as vector quantization [9] and HMM [10] have been employed.

### 2.2. Face & Lip Detection

The task of face detection consists of localizing a face in a complex scene. It is mostly achieved either by low/high level image analysis or by using pattern recognition techniques on images. Once the face has been detected in a complex scene, lip detection has to be carried out, by first localizing the mouth region and then segmenting the lips. Further details regarding face and lip segmentation will be provided in Chapter IV.

## 3. Feature Extraction

This step involves the extraction of features from audio and video signal, which will be used in the classification stage for recognition.

### 3.1. Audio Feature Extraction

Despite the fact that we have not specifically used audio features in our work but we will still provide a state of the art on the most commonly used audio features for the sake of completion. Audio features are mostly categorized as basic, derived and model based.

---

### 3.1.1. *Basic Features*

The most widely used features in speech and speaker recognition are the Cepstral features which are the Discrete Cosine Transform (DCT) of the logarithm of the short-term spectrum. DCT transform results in two advantages, the first is that it de-correlates the features, and also allows reducing the dimensionality of the features for classification.

The most widely used variants of the standard cepstrum is the Mel frequency cepstral coefficients (MFCCs) [14]. The speech signal is first divided into blocks called frames by applying a windowing function, such as the Hamming window. Next short-term log-power spectrum is computed and subsequently smoothed by a bank of triangular filters, the pass bands of which are based on a frequency scale known as Mel scale. The Mel scale effectively reduces the contribution of higher frequencies to the recognition. Finally, DCT is applied. Due to the low cross correlation of MFCCs, their covariance can be modelled with a diagonal matrix.

### 3.1.2. *Derived Features.*

Speech features can also be derived from the basic features by taking first or higher order derivatives, of which delta or delta-delta cepstral coefficients [15] are commonly used. The delta cepstrum is usually computed by applying a linear regression over the neighbourhood of the current cepstral vector.

Other derived features are the long-term spectral mean, variance/standard deviation, and the covariance of basic features [16]. Some derived features are computed by reducing the dimensionality of the basic features by applying a transformation (PCA, LDA), thus concentrating the most of the variance into few features.

### 3.1.3. *Model Based*

Several mathematical models have been proposed that try to understand and mimic the human vocal system. Excitation-Modulation (EM) model [11] replicates the human vocal system as a time-varying filter excited by a wide-band signal. The time-varying filter represents the acoustic transmission characteristics of the vocal tract and nasal cavity, glottal pulse shape and lip radiation. The drawback of this model is that it assumes that the excitation is produced from the lower end of the vocal tract thus they may be unsuitable for certain speech sounds, such as fricatives.

The Linear Prediction (LP) [12] models the audio signal as a linear combination of past values and a scaled present input. Features are constructed from the speech model parameters after being transformed in a perceptually meaningful domain. Some feature domains useful for speech coding and recognition include reflection coefficients (RC's), log-area ratios (LAR's), line spectral pairs (LSP) [13].

### 3.2. Video Feature Extraction

Features extracted from lips have been studied for A-V Speech and Speaker recognition and HCI. They are generally classified as static and dynamic feature, the static features characterize the physical aspect of the lip and the dynamic features which characterize the motion of the lip during speech represent the behavioural aspect of lip. Further details regarding their extraction will be provided in Chapter IV.

## 4. Classification

Classification consists of creating representative models of the classes and then calculating a similarity measure between the test vectors and the class models. There are two phases, enrolment and recognition; enrolment consists of extracting features from the audio-visual speech signal and storing them as class models. Then, to authenticate a user, the recognition algorithm compares the test signal with stored user models.

Several techniques have successfully been applied for the classification of A/V speech signals, the two main are template matching methods and stochastic method. In template matching the classification is deterministic. The test feature is assumed to be an imperfect replica of the template, and the alignment of test features to template is selected to minimize a distance measure. The template method can be time dependent or time independent. In stochastic models, the pattern matching is probabilistic and results in a measure of the likelihood or conditional probability, of the observation given the models. Some more complex template matching methods are presented below.

### 4.1. Template Matching

The simplest template model can be created by computing the centroid of a set of training feature vectors and classification can be carried out using a distance measure between the test feature and the centroid model. Most commonly used distance measures are Euclidean, Mahalanobis and city block.

#### 4.1.1. *Dynamic Time Warping*

DTW [18] is a popular method to compensate for speaking-rate variability in template based systems. Due to timing inconsistencies in human speech the text dependent template model may be temporally desynchronized with a test feature. Given reference and input signals, the DTW algorithm does a constrained, piecewise linear mapping of one (or both) time axis(es) to align the two signals while minimizing asymmetric match score. At the end of the time warping, the accumulated distance is the basis of the match score.

---

#### 4.1.2. *Vector Quantization Source Modeling*

Vector Quantization (VQ) source modelling [17] uses multiple templates; a code book is first created for each speaker class using his training data by a clustering algorithm. A pattern match score can then be computed as the distance between a test vector and the minimum distance codeword in the claimant's VQ code book.

The clustering procedure used to form the code book averages out temporal information from the code words. Thus, there is no need to perform a time alignment. The lack of time warping greatly simplifies the system; however, it neglects speaker dependent temporal information that may be useful for classification.

#### 4.1.3. *Nearest Neighbors*

The Nearest Neighbor (NN) [20] distance is defined as the minimum distance between a test feature and the enrolment feature. The NN distances for all the test frames are then averaged to form a match score. Similarly, the test features are also measured against a set of stored reference "cohort" speakers to form match scores. The match scores are then combined to form a likelihood ratio approximation [19].

NN combines the strengths of the DTW and VQ methods. Unlike the VQ method, the NN method does not cluster the enrolment training data to form a compact code book. Instead, it keeps all the training data and can, therefore, use temporal information.

### 4.2. Stochastic Models

The pattern-matching problem can also be formulated as measuring the likelihood of an observation given the speaker model. The observation is a random vector with a conditional probability density function (pdf) that depends upon the speaker. The conditional pdf for the claimed speaker can be estimated from a set of training vectors and given the estimated density, the probability that the observation was generated by the claimed speaker can be determined. This probability is the match score. The estimated pdf can either be a parametric or a nonparametric model.

Features could be modelled using single-mode Gaussian probability density functions (pdfs) [27]. However, the most popular models are multimode mixtures of multivariate Gaussians [28], commonly known as Gaussian Mixture Models (GMMs). HMMs are widely used as models for both static and dynamic characteristics [29].

#### 4.2.1. *GMMs*

A GMM represents a probability distribution as a weighted aggregation of Gaussians

$$p(x) = \sum_{i=1}^m w_i N(x; \mu_i, \Sigma_i)$$

where  $x$  is the “observation” corresponding to a feature vector and the GMM parameters are the mixture weights  $w_i$ , the number of mixture components  $m$ , the mean  $\mu_i$ , and the covariance matrix of each component. GMM parameters are often estimated using the Expectation-Maximization (EM) algorithm [28]. Being iterative, this algorithm is sensitive to initial conditions and it may also fail to converge if the norm of a covariance matrix approaches zero.

#### 4.2.2. HMMs

HMMs are a popular stochastic model for modelling sequences. In an HMM, the observations are a probabilistic function of the state; i.e., the model is a doubly embedded stochastic process where the underlying stochastic process is not directly observable (it is hidden). The HMM can only be viewed through another set of stochastic processes that produce the sequence of observations [31].

The HMM is a finite-state machine, where a pdf  $p(x|s_i)$  is associated with each state  $s_i$ . The states are connected by a transition network, where the state transition probabilities are  $a_{ij} = p(s_i|s_j)$ . A three state HMM is given in Figure 5.

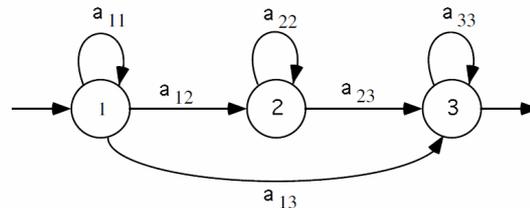


Figure 5: An HMM with three states

The most common HMM learning rule is the Baum-Welch algorithm, which is an iterative maximum likelihood estimation of the state and state-transition parameter [30]. Due to the iterative nature of the learning, the estimated parameters depend on their initial settings. Viterbi decoding [30] is typically used for efficient exploration of possible state sequences during recognition; it calculates the likelihood that the observed sequence was generated by the HMM.

### 4.3. Neural Networks

The key strength of neural networks is that their training is generally implemented as a nonparametric, nonlinear estimation, which does not make assumptions about underlying data models or probability distributions. However, compared to HMMs, neural networks are generally worse at modelling sequential data. The most popular models are the memory less type, such as MLPs [21], radial basis functions [21], neural tree networks [22], Kohonen's self-organizing maps [23], and learning vector quantization [24]. The models capable for capturing temporal information include time-delay neural networks [25] and recurrent neural networks [26].

## 5. Fusion

Humans use information fusion extensively, e.g. seeing and touching an object. Fusion provides various advantages such as complementary information for better understanding, increasing robustness in case one sensor might not be working properly. Information fusion, be it in the form of raw data, features or final decisions has been studied extensively, and borrows extensively from other fields such as statistics, signal processing and decision theory, thus only a short introduction will be provided here. It has commonly been divided in the following categories.

### 5.1. Early Integration

In early integration, fusion takes place before the classification process, either using the raw data directly or after the feature extraction process. Some [32] have argued that information is lost during the classification process as it moves from raw data through the feature extraction and final classification process, thus information integration should take place as early in the classification process as possible. On the other hand early integration does suffer from noise and requires larger amount of training data.

#### 5.1.1. *Raw Data*

Raw data [33] from the sensors has been combined in several ways, first is simple averaging e.g. audio from several microphones can be combined by averaging to reduce the effect of noise. Another method is called mosaicing, e.g. in [34] images from several cameras providing a different view of the same scene are combined.

### 5.1.2. *Feature Data*

Features extracted from raw data can be combined in various ways, the simplest being by weighted summation. Another, more commonly used is by concatenation [33], but it has its own drawbacks, such as the features must have the same frame rate and increase in dimensionality. Increase in dimensionality has commonly been dealt by using linear or non linear transforms, non linear transforms such neural networks have been used in [35]. Linear transforms such as PCA and LDA have also been used in [36].

## 5.2. Intermediate Integration

This is a relatively new concept in information fusion which exploits the temporal aspect of modalities during the classification process. Specific classifiers have been designed to handle data from several modalities simultaneously [37]. Factorial HMMs [38], Multistream HMMs [29] and coupled HMMs [39].

## 5.3. Late Integration

Here the fusion of information takes place after the classifier has provided either a final decision about the class or just an opinion, thus this category can be further divided on the basis that a classifier provides a decision or an opinion.

### 5.3.1. *Decision Fusion*

In decision fusion, the final decision by the classifier or group of classifiers is combined. Here it is essential to point out the classifiers might be similar but working with different features (audio and video), or different classifiers (ANN and SVM) working with same data.

Various methods have been proposed for decision fusion, one being *majority voting* [34] where a decision is reached by selecting the decision on which majority of classifiers agree. Contrary to the majority rule, some have proposed that it might be interesting to have a look at the second or third choice of each classifier, thus giving rise to the idea that each classifier provides a *ranked list* [40] of choices, which are then combined to form the final decision. In high security applications a good option is to use the *AND rule* [41] where the decision is only reached if all the classifiers agree on the same choice. This is very restrictive resulting in a high False Rejection Rate FRR, and low False Acceptance Rate FAR in a verification scenario. Contrary to the AND rule is the *OR rule* [41], where a decision is reached as soon as one of the classifier makes the decision. This is useful in a low security but high convenience scenario, resulting in a low FRR and high FAR.

### 5.3.2. *Opinion Fusion*

In opinion fusion [33],[34] the classifiers provide an opinion regarding the possible decisions. Outputs from various classifiers are in different units e.g. one classifier could give output in distances and another in probabilities, thus first a mapping has to be created to transform outputs into one unit.

Opinions from classifiers can be combined using weighted sum, product rule or using a post-classifier. In a weighted summation [42] the opinions are summed up after multiplying them with appropriate weights. Weights can be selected based on the discriminative power and reliability of the classifiers or if they are equal, the weighted summation reduces to simply an arithmetic mean. In a weighted product rule [42], the opinion  $o_{i,j}$  regarding a class  $j$  from classifier  $i$  can be combined using

$$f_j = \prod_{i=1}^N (o_{i,j})^{w_i}$$

Where  $w_i$  is the weight and  $N$  the total number of classifiers. The disadvantage of using weighted product rule is that one classifier can have a large influence over the fused opinion.

The opinion outputs from classifiers could themselves be concatenated and considered as features vectors, thus using a post-classifier for fusion and final decision. Various classifiers have been used as post- classifiers, [30] have compared kNN, decision trees and logistic regression, with the former providing the best accuracy and lowest computational cost. [43] has shown that median radial basis function network performs better than fuzzy k-means or fuzzy VQ. Comparison between SVM, Bayesian classifier, fisher linear discriminant, decision trees and MLP was shown in [44].

## 6. Examples of Lip Based Person Recognition

In this section we provide examples of AV biometric systems that use lip features in some regard. It is divided into two sections, the first one presents systems that uses both audio and visual information and the second one only video information.

## 6.1. Audio - Video Lip Biometric

Wagner et al. [52] presents an approach that combines lip motion with acoustic voice analysis using synergetic computers. Acoustic data was acquired using a microphone and digitized with a 16 bit A/D converter. Image data was acquired by a high speed camera at 256X256 resolution. Acoustic signal was pre-processed using a short time Fourier transform with 16 Hann window of size 2048 sample points. Video signal was sub-sample to 128X128 and optical flow calculated, the resultant was further sub sampled to 16X16 and a three dimensional Fourier transform was applied.

Evaluation was carried out by the MELT algorithm, which is a special form of a synergetic computer. A locally collected database was used which consisted of 101 persons. 10 samples per person have been collected, uttering there surname and a word. At low resolution the motion analysis yields a recognition rate of 92.1%, and when combined with audio using an AND rule, the misclassification rate reaches 0%. With higher resolution motion analysis the recognition rate alone reaches 99 %.

Pan et al. [53] presents a novel fused HMM to integrate audio and video features from speech. HMMs are first build individually and then fused together by exploiting the dependences between the audio hidden states and visual observations within a general fusion method which is optimal in the maximum entropy sense. The individual HMMs are first trained independently using the EM algorithm. Then the whole fused HMM is trained by fixing the parameters in the individual HMMs.

Experiments were carried out on a locally collected database with 5 people saying there names, each consisting of roughly 80 training sample, 40 testing sample, 10 samples in which this person claimed the identity of another person. Audio signal was first processed by cepstral Linear Prediction Coding (LPC) and then a 64-word vector quantizer (VQ) was applied. The mouth ROI is extracted and sent to a 16-word vector quantizer. The outputs of two VQs are classified using audio and visual HMMs and the proposed fused HMM, the best overall result is achieved by fused HMM with 0.51 % verification error.

Broun et al. [54] evaluates the effectiveness of using visual speech features for person verification. First a mouth ROI is detected using hue and saturation thresholding. This estimate is combined with a motion based mouth estimation method, called accumulated difference image ADI, which is calculated by taking difference between subsequent frames and summing over a series of frames. Features pertaining to appearance and shape are extracted; shape features consist of the mouth width, upper/lower lip width, lip opening, and the distance between the horizontal lip line and the upper lip. Intensity and shape features are then combined using an algorithm based on markov random fields. Audio features consist of 12 cepstral coefficients.

---

Classification is carried out using a third order polynomial based classifier and decision threshold on the XM2VTS database. The database includes 295 speakers, saying the sentence “Joe took father’s green shoe bench out”. First two sessions were used for training, third for evaluation and the last for testing. Late fusion is used for combining audio and video features, with a FAR 8.2 % and FRR of 4.4 %.

Jourlin et al. [55] describes a system that consists of two separate audio and video classifiers and a score integration method. Lip shape and intensity features are extracted, the lip is located and tracked in an image sequence using Active Shape Models (ASM), and its parameters are used as shape features of the inner and outer lip contour. Audio features consist of 39 components of Linear Cepstral Coefficients with first and second order derivatives.

Tests were carried out in the M2VTS database of 37 speakers, pronouncing in French, digits from zero to nine. The first 3 sessions were used for training, fourth for evaluation and the last for testing. Using acoustic features only the training set is segmented and one model is train for each digit and speaker resulting in an identification rate of 97.2 %. Labial features only results in 72.2% and the combination using a weighted summation results in 100 % correct identification.

Fox et al. [56] presents a speaker identification system using dynamic audio and video features, using HMMs with early and late integration. Audio was first segmented manually and then used to segment video sequences. Lip ROI was extracted by locating the lip corners manually. Audio features consisted of 8 dimensional Mel-Frequency Cepstral Coefficients and nine delta features were calculated and concatenated with the static features. Lip ROI were converted to grey level images and DCT was applied, then 15 coefficients were extracted using zigzag scan. Dynamic features were then extracted form these static features by taking difference of DCT coefficients over a certain number of frames.

Testing was carried out on the XM2VTS database using 252 out of the 259 subjects using the phrase “Joe took father green show bench out”. Three sessions were used for training and one for testing. Early integration was testing by concatenating audio and video features before classification and late integration was done using weighted summation rule. Using audio only resulted in 86.91 % identification rate and audio/video with early integration resulted in 80.61 %.

Wark et al. [57] investigates the use of speech and lip information for person identification using multi stream HMMs. The audio subsection consists of mel-cepstral coefficients, energy and delta coefficients, thus the audio feature vector consists of 26 dimensions. The lip tracking system is based on a chromatic-parametric approach. Lip features consisted of sampled chromatic information on paths normal to the tracked lip contour. Features were then reduced by PCA followed LDA, to form the final features of dimension 36.

Classification was carried out using a multi stream HMM. The audio stream states consist of a number of Gaussian mixtures, the video stream states consist of only a single Gaussian mixture, due to the nature of the data. Speaker identification experiments were performed on the M2VTS database consisting of 37 subjects counting from zero to nine in French, the first three sessions were used as training sessions, and the fourth as a test session. Audio only results although better than combined audio/video were seriously effected by noise and under noisy conditions in which case audio/video results were much better.

Kanak et al. [58] describes a system that fuses audio and video lip features and uses them for person identification. Lip regions are first cropped using hand labelled localization information and then a subset of the lips is used to create the eigenlip space. Finally lip images are projected in the eigenlip space and coefficients obtained are used as visual features. The audio feature vector consists of 13 cepstral coefficients including the 0<sup>th</sup> gain coefficient and the first and the second delta MFCC vectors.

Tests were carried out on a locally collected database of 38 people where each speaker utters her/his name and the same 6-digit number (3458-572). The speaker identities were modelled using a HMM, where state transitions model temporal correlations and Gaussian classifiers model signal characteristics. Each speaker in the database is modelled using a separate HMM. The system consists of two independent classifiers with audio-only and fused audio-visual features. For the final decision, a Bayesian classifier is incorporated to combine the two decisions. With the best result of 1.41 % EER was reported for the digit dataset and 2.58 % EER for the name dataset.

Ichino et al. [59] have proposed a multimodal biometrics system by using the kernel fisher discriminant analysis, to model nonlinearity of decision boundary between client and impostor. The lip ROI is first extracted using colour information from skin and lip using LDA , then P-Fourier descriptor is applied as a feature extractor. Audio features consist of LPC coefficients.

Classification is then carried out first separately for audio and video features using kernel mutual subspace method which combines the Mutual subspace method with kernel principal component analysis. The scores from audio and video subsystems are then concatenated and form a new feature vectors and kernel fisher discriminant analysis is used for classification of people in two classes, genuine and impostor class. Tests were carried out on XM2VTS database with best accuracy of 93.6%.

---

Faraj et al. [60] describes a system that combines acoustic and visual features at the feature level and were evaluated first by a Support Vector Machine (SVM) for speaker identification and then by a HMM for speaker verification. Visual features are extracted by calculating the optical flow by assuming that the lip ROI contains lines or edges, the computations can be carried out in 2D subspaces instead of the 3D spatiotemporal space. Next this dense velocity vectors are quantized by allowing only 3 directions ( $0^{\circ}$ ,  $45^{\circ}$ ,  $-45^{\circ}$ ), and only 20 values. The audio feature vector consists of 12 cepstral coefficients extracted from the Mel-frequency spectrum with normalized log energy, 13 delta coefficients (velocity), and 13 delta-delta coefficients (acceleration).

Tests were carried out on the XM2VTS database using the Lausanne protocol. SVMs were used for speaker identification with one-against-one method and SVM parameters deduced from empirical evaluation, resulting in 100% correct recognition. Speaker verification tests were carried out using HMM with Gaussian mixture density used as the probability density function for model observations in a state. Resulting in a 98 % correct recognition.

## 6.2. Video only Lip Biometric

Wark et al. [46] have proposed a lip tracking system that is based on chromatic information and does not require iterative optimization. First candidate pixels are selected whose Red-Green ratio lies within a certain limit. Next a series of morphological operations are carried out. The upper lip contour was modelled using a 4<sup>th</sup> order polynomial and lower by a 2<sup>nd</sup> order polynomial using least square approximation method. Features consisted of colour profiles along the normal of points on the outer lip contour. PCA was then applied to reduce the dimensionality and LDA to improve the discrimination power.

GMM were used to model subjects. Tests were carried out on M2VTS database which consists of 37 subjects counting from zero to nine repeated 5 times. First 3 sessions were used for training and the rest 2 for testing and a recognition rate of around 90% was reported.

Mok et al. [45] presents a study using lip shape and intensity features for person authentication. First the lip region was segmented using fuzzy clustering method in CIE-LAB colour space. Then a 14 point Active Shape Model (ASM) is used to describe the exact shape of the outer lip contour, the ASM parameters form the shape based features. Next intensity features along the horizontal and vertical axis of the outer lip contour are extracted. 15 points are sampled along vertical axis and 35 along the horizontal, these are then concatenated and PCA is applied to reduce the dimensionality.

Classification is carried out using continuous density, left to right HMM consisting of six states. Differential change is also studied. Testing was carried out on a database of 40 speakers each with an utterance of 3 seconds. The phrase consisted of the number in English “3725”, repeated ten times. Best results (98 %) were obtained when shape information was combined with first 8 modes of variation from the intensity profiles.

Cuesta et al. [47] focuses on using Motion History Image (MHI) of the lip movements for person recognition. First MHI are created for each person and each word by accumulating intensity changes of pixels, then phase only correlation filters are created to model movements for each class. So when a new MHI is filtered, depending if the filter belong to same or different classes a high or low correlation peak is observed.

A Bayesian classifier was used with maximum score of the correlation as the potential function. Tests were carried out on a small database of 9 people, pronouncing 9 digits, one to nine. Best results obtained were 100 % recognition rate and an FRR of 5 %.

Luettin et al. [48] present a lip-reading system, which has been used to identify people, lip boundary shape and intensity features are extracted and then used to identify people with HMMs. An ASM based approach is first used to locate, track and parameterize the lip in an image sequence. The lip shape is thus represented by a set of labelled points. The lip intensity around the lip contour is modelled by extracting grey level intensity vectors perpendicular to the contour at each ASM labelled point. PCA is then performed to reduce the feature space.

Speaker modelling is achieved using an HMM with Gaussian mixtures on shape and intensity features extracted frame by frame. As the system was tested both text independent and text dependent modes, thus for TD mode one HMM per word class and speaker was built, while for TI mode one HMM was built for each speaker representing all words. HMMs were initialized by linear segmentation of the training vectors onto the HMM states followed by iterative segmentation by k means clustering and Viterbi alignment. Tulip database was used which consists of 96 grey level image sequences of 12 speakers. Each uttering four digits in English twice. TD mode resulted in a recognition rate of 91.7 % and TI in 97.9 %.

Lucey et al. [49] have evaluated various static area based lip features for speech and speaker recognition using HMMs. The features that have been extracted from a mouth ROI, include classical PCA, SLDA which is LDA with specific prior knowledge of the speakers, MRPCA in which the mean is first removed and PCA is then applied, WLDA which is LDA with prior knowledge of the word classes.

Testing was carried out on M2VTS database, which consists of 36 subjects, four sessions each consisting of ten digits in French i.e. zero to nine, where the first 3 sessions were used for training and the last session for testing. HMMs were used for classification with SLDA performing better than the rest with an error rate of 19.71 %.

---

Faraj et al. [50] describes a novel feature extraction technique for person authentication based on orientation estimation in 2D manifolds. Lip motion estimation is carried out by computing the components of the structure tensor from which normal flows are extracted. By projecting the 3D spatiotemporal data to 2-D planes they obtain projection coefficients which are used to evaluate the 3-D orientations, thus increasing computational efficiency. Next this dense velocity vectors are quantized by allowing only 3 directions ( $0^{\circ}$ ,  $45^{\circ}$ ,  $-45^{\circ}$ ), and only 20 values resulting in a feature vector of 40 parameters. These quantization values were obtained by fuzzy c-means clustering.

Tests were carried out on the XM2VTS database according to the Lausanne protocol, 200 speakers were used for training, 70 as impostors for testing and 25 as impostors for evaluation. Verification was carried out within a GMM framework with HMMs. An error rate of 22 % was reported.

Cetingul et al. [51] study two features, the first lip features are dense motion vectors extracted over a uniform grid in the mouth ROI. Motion matrixes are then separately transformed using 2D DCT and a certain number of DCT coefficients are then selected using zigzag scan. The second feature set consists of lip shape and motion information. Outer lip contour is first extracted using the “Jumping snake” technique. Next motion vectors are extracted from pixels along the outer lip contour, and DCT is applied. The shape information is derived from parameterization of the lip contour into eight measures.

Feature analysis is then performed on the extracted features using Mutual Information within a Bayesian framework resulting in a reduced set of features. Next a temporal feature selection is applied on the already reduced feature set using LDA. The features selected from the Bayesian feature selection were concatenated over a window creating higher dimension feature vector, to which LDA was applied. Tests were carried out on MVGL-AVD database consisting of 50 speakers. Each speaker utters his/her name and a fixed digit password “348-572” ten times and a secret phrase. In the speaker identification scenario using grid based motion features to which both Bayesian and temporal feature discrimination was applied performed the best with an error rate of 5.2 %.

The lip based person recognition systems using video information only presented above are summerized in Table 1.

System	Features	Classification	Database	Results
Wark et al. 1998	PCA/LDA on Color profile vectors	GMM with Probability theory	M2VTS 37 Subjects Contents: 0-9	90% CIR
Mok et al. 2004	ASM Parameters PCA on Intensity profile vectors	HMM/GMM	Private 40 Subjects Contents: 3725	98% CIR
Cuesta et al. 2008	Motion History Image	Bayesian Classifier	Private 9 Subjects Contents: 0-9	100% CIR
Luetin et al. 1996	ASM Parameters PCA on Intensity profile vectors	HMM/GMM	Tulip 12 Subjects Contents: 4 Digits	97.9 % CIR
Lucey et al. 2003	PCA/LDA on mouth ROI	HMM	M2VTS 36 Subjects Contents: 0-9	19.71 % ER
Faraj et al. 2006	Quantized Motion Vectors	GMM/HMM	XM2VTS 200 Subjects 3 Phrase	78% CVR
Cetingul et al. 2006	DCT on Motion Vectors Distance measures for lip contour	HMM	MVGL-AVD 50 Subjects Contents: Name, 348- 572	5.2 % EER

Table 1 : Lip based person recognition systems using video information

### 6.3. Conclusions

In this section we have presented systems that attempt to recognize persons from lip features. Regarding the features used, it can be observed that there is no consensus on the type of features that are most performant. Each system extracts its own type of feature without providing any comparison with others. Another point regarding the features used is that they mostly consist of features modelling the physical aspect of the lip such as shape and appearance. Classification techniques used mostly originate from the speech community and are based on GMM and HMM. In regard to the databases used, they are quite diverse ranging from small private to large publicly available databases, but one point which is common to all is the contents which consist of short text dependent phrase/numbers. Lastly the results have been presented in a wide variety of measures such as CIR, EER, etc. so a direct comparison of the systems is impossible.

## 7. Audio-Video Speech Databases

### 7.1. Introduction

In this section we elaborate some of the AV databases already available and comment on their suitability for our experiments. In this thesis as we have developed several systems based on lip analysis with varied requirements on the database contents and configurations, there are two fundamental requirements for developing and assessing the performance of our recognition systems.

A data set with enough data per user to enable the enrolment and recognition using behavioural biometric identifiers; for our techniques we estimate that they require at least 3-4 minutes of video per person.

A data set with multiple sessions of individuals repeating the same sentence, with little or no illumination and pose variation.

Unfortunately, to the best of our knowledge there are no databases that fulfil both the requirements of our experiments, specifically the first requirement is even harder to fulfil. Thus we decided to use two different databases for our experiments, Valid Database and Italian TV Database, explained below.

## 7.2. VALID Database

The database [65] consists of five recording sessions of 106 subjects (77 males, 29 females) over a period of one month. This database consists of text dependent recordings. The database was recorded using a Canon 3CCD XM1 PAL digital video camcorder, with a colour sampling resolution of 4:2:0 with the audio captured using 16 bit stereo samples at a frequency of 32kHz with PCM encoding. The video frame rate is 25 fps with a pixel resolution of 576 x 720 (rows x columns) and 24 bit pixel depth.

The content of the VALID database was designed to supplement that of the XM2VTS database. Three utterances were recorded per session (in English), namely:

- 1: “<Subject’s full name>”
- 2: “5 0 6 9 2 8 1 3 7 4”
- 3: “Joe took father’s green shoe bench out”

Utterance 1 was used for subject identify, and is not distributed publicly. Utterances 2 and 3 are the same as those in XM2VTS. A head rotation sequence was recorded during Session 1, where the subject was asked to face four targets, placed approximately 15 degrees above, below, left, and right of the camera, resulting in images shots that were slightly off-frontal.

The five sessions were recorded over a period of one month, the first session was recorded under controlled acoustic/illumination conditions. For this session, a blue background screen was employed. The other four sessions were recorded in noisy real world scenarios with no control on illumination or acoustic noise. Acoustic noise/interference includes computer fan noise and speech/movement of third persons. The uncontrolled sessions were recorded indoors, predominantly in offices, but also in more open spaces. The illumination consisted of office and natural lighting.

## 7.3. Italian TV Database

This database was collected by [1] by recording the TV news from the Italian national channel RAI 1, over a period of 21 months. It consists of 208 video clips from 13 TV speakers (8 men and 5 women) of 13 seconds each. Figure 6 illustrates the data set by showing the first 7 frames for some of the speakers.



Figure 6: First 7 frames for some of the TV speaker.

The database is split into an enrolment subset and a recognition subset: 104 video sequences (8 for each of the 13 clients) are employed for the training of our systems (enrolment), and the remaining 104 (8 for each of the 13 clients) are used for their testing (recognition). We explicitly keep the two data subsets disjoint, in order to reduce the risk of systematic errors in the test procedure.

		WHOLE DATABASE	ENROLMENT SUBSET	RECOGNITION SUBSET
<b>OVERALL</b>	Number of individuals	13	13	13
	Number of men	8	8	8
	Number of women	5	5	5
	Number of videos	208	104	104
	Number of frames	68640	34320	34320
	Length of videos	45min. 49sec.	22 min. 54sec.	22 min. 54sec.
<b>PER PERSON</b>	Number of videos	16	8	8
	Number of frames	5280	2640	2640
	Length of videos	3min. 31sec.	1min. 46sec.	1min. 46sec.
<b>PER VIDEO</b>	Number of frames	330	330	330
	Length	13 sec.	13 sec.	13 sec.
<b>RESOLUTION</b>	Spatial	192 pixel rows or height	192 pixel rows or height	192 pixel rows or height
		224 pixel columns or width	224 pixel columns or width	224 pixel columns or width
	Temporal	24.97 frames/second	24.97 frames/second	24.97 frames/second
<b>COMPRESSION</b>	Compression rate	118 Kbits/second	118 Kbits/second	118 Kbits/second
	Format	Windows Media Video 9	Windows Media Video 9	Windows Media Video 9

Table 2: Technical details of Italian TV database. [1]

## 7.4. Other Databases

### *Face Database – CMU* [61]

This audio-visual database consists of 10 subjects (7 males and 3 females). The vocabulary includes 78 isolated words commonly used for time, such as, "Monday", "February", "night", etc. Each word was repeated 10 times. High quality video was acquired at a resolution of 720 X 480 with a blue screen background. Audio was acquired with a tie microphone in a soundproof room.

### *AVICAR Database* [62]

This database was specifically designed to model a car environment, four video cameras were placed on the dashboard and the video streams were multiplex into a single video stream to provide four views of subject. Data was recorded in a car with 3 different speeds levels, idling, 35 mph, 55 mph, with windows rolled up and down. Eight array microphones were mounted on the dashboard for audio. The database consisted of 100 speakers, with equal distribution of male and female. The subjects read a sequence of numbers, letters, phone numbers and TIMIT sentences.

### *M2VTS Database* [63]

This small database consists of 37 subjects and recorded in 5 sessions. During each session the subjects were asked to count from 0-9 in their native language, which was mostly French with a video resolution of 720\*576.

### *The Extended M2VTS Database - University of Surrey* [64]

This audio-video database and contains four sessions of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. The speech part consists of ten digit sequences and a seven word phrase, at high quality 720\*576 resolution video, 16-bit audio recorded at 32 KHz.

### *BANCA Database* [66]

This database is a multi-modal database captured in four European languages (English, French, Italian, Spanish) in two modalities (face and voice). For recording, both high and low quality microphones and cameras were used. The subjects were recorded in three different scenarios, controlled, degraded and adverse over 12 different sessions spanning three months. In total 208 people were captured, half men and half women. Subjects were asked to recite random 12 digit numbers along with their name, address, and date of birth resulting in a 20 second recordings on average. The banca evaluation protocol specifies the training/testing sets and experimental configurations also.

*Video Face Database - University of Texas [67]*

It contains both static images and videos of 284 subjects (208 female, 76 male). The recordings are at close range with neutral background and controlled lighting conditions. The clips are text independent and are responses of approximately 10 second duration to mundane questions. Audio has been removed for all clips. The videos are captured at 30 frame/sec and a resolution of 720 X 480.

*MyIDea Biometrics Database [68]*

This database includes audio and video face data. Inspired from BANCA, it also follows the same protocol of recording at 3 different quality levels. The subjects speak couple of predefined sentences and then rotate their heads for profile shots.

*BIOMET Biometrics Database [69]*

This database is similar to MyIDEA but includes data acquired in 3 sessions for a total of 227 subjects with a balanced age variation from 20 to 60 years. The other main difference is that subjects are French speaking.

*AV-TIMIT Database [70]*

The database consists of 223 speakers (117 males / 106 females ). Mostly native English speakers, reciting 20 TIMIT sentences, with first sentence being common to all. The recordings are taken in an office environment with controlled background, lighting and audio noise level. At a video resolution of 720\*480 at 30f/s, and audio at 16khz.

*VidMIMIT Database [71]*

The database consists of 43 subjects (19 females, 24 males), reciting short sentences. The data was recorded in 3 sessions with a gap of a week between sessions. Each person recites 10 sentences from the TIMIT corpus. The recording was in noisy office environment with professional digital video camera. The videos were stored as jpeg of resolution 512X384. The audio consisted of mono, 16 bit, 32 KHz wav files.



## *Chapter IV. Lip Detection & Evaluation*

---

### 1. Introduction

The main topic of this chapter is lip detection and the evaluation of the lip detection results. Due to the fact that significant amount of effort has already been made on lip detection and keeping in mind the main focus of this thesis, we concentrated on designing a lip detection algorithm that could achieve reasonable performance in real world conditions. We have proposed a lip detection algorithm based on fusion of two independent methods and the evaluation of the results on a database with real world conditions. At the end we extract static and dynamic features from the detected lip and evaluate their relevance for person recognition. But first we present a few state of the arts on face detection, lip detection and the commonly used visual features from the lip.

### 2. State of Art: Face Detection

Detecting a face maybe an effortless task for humans but teaching computers to perform the same task is highly complex. It consists of acquiring a video or image of a complex environment that may contain one or several faces, segmenting the scene, detecting the face(s) and then verifying the detection results. This problem is further compounded by the presence of illumination and orientation/pose variation. Initial works which began in the 70's focused more on an idealistic scenario but with the influx of several other domains of research for which face detection is a prerequisite, work today is more focused towards real world scenarios.

Face detection techniques have mostly been classified in two categories; feature based approaches, which makes use of low and high level understanding of facial images to detect faces. Image based techniques that use pattern recognition techniques on images or feature vectors extracted from images for face detection.

## 2.1. Feature Based Techniques

The first category of techniques exploits prior knowledge of face, and using image understanding techniques builds low and high level models of the face. They can be further classified based on the level of understanding; low level techniques use edges and colour information to build models that lack knowledge of facial features. Feature analysis based techniques model the facial features and the connections between the features. Model based techniques further increase the level of understanding by explicitly and accurately modelling facial features.

### 2.1.1. *Low Level Analysis*

*Edges:* Image representation using edges for face detection was first applied in [72] for face detection from line drawing of facial photographs. Sobel operator has been the most common technique that has been applied in several face detectors [73]. Canny edge detector combined with heuristics have been used in [74]. Marr Hildreth edge detector has been applied in [75]. Laplacian of Gaussian has been applied in [72]. Advance techniques such as steerable filtering that consists of edge detection, determination of filter orientation and tracking of neighbouring edges has been applied in [76].

*Colour:* Modelling skin colour has been one of the most widely used techniques for face detection. It has been shown that despite huge differences in skin colour among races and ethnic groups, skin colour occupies a small portion of the colour space [77]. RGB colour space has been mostly used due to its ability to handle luminance variation in [80]. [78] have modelled skin colour using the histograms of normalized red and green colour. In [81] HSI color model has shown good representation capabilities for facial feature such as lips and eyes. YIQ has also been applied in some face detection systems, where it has been able to enhance the skin region for Asian people [82]. Skin color has mostly been modelled using histograms [83], Gaussian Distributions [78] and Hu's moments [84]. Classifiers range from simple distance measure to multilayer perceptrons [84]. [85] have used an adaptive color mixture model to track faces under varying illumination using stochastic models of color distribution.

*Motion:* Using videos the face can be detected by modelling facial motion, the simplest being analysis of frame difference. In [86] frame differences are accumulated and then silhouettes are extracted and classified as faces. In [79] eye pair hypothesis is used on frame difference images to detect presence of a face. [81] have based their system on optical flow. Comparison between frame differencing and moving contours has been presented in [88].

### 2.1.2. *Feature Level Analysis*

In this category facial features are explicitly detected and their geometrical relations are analyzed for face detection. It can be further divided into two categories, feature searching and constellation analysis.

*Feature Searching:* This category starts by detecting a prominent facial feature and then other features are estimated or detected using anthropometric measure. Eyes [86] and face axis [89] face border [72] have been the most used due to their relative ease of extraction.

[90] begins with a hypothesized top of head location and scan the image until it reaches the eyes that are associated with unusually large number of edges, then the rest of the features are detected using the distance from top of head to eyes as a measure. Finally a template based method is used to detect faces. In [91] they detect the eyes first from a binarized pre-processed image and then search for nose and mouth. Detection of each feature increases the chances of the candidate image to be a face. Eye motion has been employed in [92] by modelling the eyes using Gabor responses.

*Relational Analysis:* In this category modelling of geometric constraints between facial features is done using statistical models. [93] use statistical shape models on facial features that have been detected using multi-scale Gaussian filters. In [94] images pre-processed with Gaussian filters are searched with a structural, texture, and feature models for face detection. In [95] images are first processed by a gradient operator and then a two stage face detector is applied which consists of a Hough transform and template matching.

Graph matching [96] is another popular technique for face detection, in which feature information is represented by nodes and geometrical constraints are modelled by edges.

### 2.1.3. *Active Model Based*

These techniques represent the highest level of understanding of local facial features, an active model when properly initialized, tries to deform to model the feature using image information such as edges and color. They can be further divided into the following categories

*Snakes:* Snakes or active contours first proposed by [97], must be first initialized near the head boundary, then using edge information it assumes the shape of the head. The deformation is guided by minimization of internal and external energies. Elastic energy [99] is commonly used as internal energy and external energy is derived from the images. Commonly used external energies include, image gradient [98], skin color [100]. Energy minimization is carried out using optimization techniques such as steepest gradient descent.

*Deformable Templates:* Snakes are especially prone to two problems, first is getting stuck in local minimas and the second is illumination variation causes problems in edges detection. [101] extended the idea of snakes by incorporating global information using templates. A template based on prior knowledge of the feature is initialized near the feature and like the snake it deforms under energies to assume the shape of the feature. [102] use a two stage approach, Hough transform is used first to get a rough localization and then a simple template is used for final modelling. Latest research [103],[104] has mostly focused on issues of time complexity, template adaptation and energy terms.

*Point Distribution Models:* These are compact parameterized models, the Point Distribution Models (PDM) is first trained by using manually labelled points. Next PCA is used to model the modes of variations. Finally a fitting algorithm is used to morph the model according to local image information. It was initially developed by [105] to model the appearance of the face, using 152 manually selected points, then a local grey level search strategy was used to fit the model on test images. [106] proposed a multi-resolution approach to address the problem of multiple face candidates.

## 2.2. Image Based Techniques

Image based techniques consider face detection as a pattern recognition problem, using images directly or features extracted from images, they try to train a classifier with face / non face classes. The advantage image based techniques have over the feature based techniques is that as they don't have to accurately model features, they are somewhat immune to environmental variations and incomplete facial information.

### 2.2.1. *Linear Subspace Techniques*

These techniques are based on the fact that facial images lie in a subspace much smaller than overall image space. [107] initially developed PCA for efficient representation of human faces. [87] extended the technique for person recognition, commonly known as eigenfaces. They also defined a Distance-From-Face-Space (DFFS), which is the residual error resulting from approximating a face with its principal components, to measure the "faceness" of an image. [108] have used PCA to model both faces and background clutter, i.e. eigenface and eigenclutter. [109] have proposed to use Factor Analysis (FA) which is quite similar to PCA but assumes that the observed data is from a well defined model. EM is used to learn the model from training images and then a window based search algorithm is used to detect faces. Fisher's linear discriminant analysis has also been used for face detection in [110] and LDA in [111].

### 2.2.2. *Naïve Bayes Classifier*

[112] have used a naïve bayes classifier at multiple resolutions using local appearance, such as intensity patterns near the eyes which are more distinctive. They have extended their approach with wavelet representation to detect profile faces in [113]. Similar methods using local features extracted by multiscale and multiresolution filters are reported by [114]. These features are then modelled using mixture of Gaussians and classified by a bayes classifier.

### 2.2.3. *Information Theory Based*

[114] have applied Kullback relative information to detect faces, first a training set is used to select Most Informative Pixels (MIP) to maximize Kullback relative information, which are then used to create linear feature for classification, finally DFFS is used to detect faces. [115] combine view based and model based methods, first a visual attention algorithm is applied to reduce the search space and then faces are detected using a combination of hierarchal Markov random fields and maximum a posteriori estimation.

### 2.2.4. *Neural Networks*

Initially simple neural networks such as MLPs were applied to the problem of face detection [116]. [117] reported the first comprehensive work on a large dataset of frontal images, and then extended it to faces with pose variation using router neural networks [118]. [119][223] have used convolution neural network and [120] a neural network based on constrained generative model. Other techniques include, [121] which use a probabilistic decision based neural network (PDBNN) and [122] which propose a new architecture named Sparse Network of Winnows (SNoW). The main drawback of neural network approach is that the network architecture parameters have to be tuned for good performance.

### 2.2.5. *Support Vector Machines*

SVM is a linear classifier which aims to maximize the distance between the separating hyper planes, thus minimizing the error on test images. The optimal hyperplane is defined by a weighted combination of a subset training vector, called support vectors. SVMs were first applied to the problem of face detection by [123]. SVMs have also been applied in the wavelet domain by [124].

### 2.2.6. *HMMs*

A face can be modelled as states of an HMM, by dividing the face into regions such as forehead, eyes, nose, mouth and chin. Then an HMM can be trained to detect a face if these features are observed in an appropriate order. [125],[126] have applied 1D and pseudo 2D HMMs, the images are first segmented uniformly from top to bottom but are later replaced by segmentation from viterbi segmentation. The parameters of HMMs are estimated using Baum-Welch algorithm. [127] have used HMM and KLT to detect and recognize faces. They use KLT coefficients instead of intensity values as input to detection.

## 3. State of Art: Lip Detection

Lip detection is still an active topic of research; the reason being the numerous applications where lip detection either serves as a pre-processing step or directly provides visual information to improve performance. It has been most successfully applied to Audio-Video Speech and Speaker recognition, where it has considerably improved recognition results, especially in the presence of noise. Another domain of application is gesture recognition for closely related fields of HCI, affective computing and expression recognition. It has also been used in the analysis and synthesis of lips for talking head in video conferencing applications. Lip detection literature can be loosely categorized in techniques which directly use image information, those which build models and hybrid techniques that combine the above mentioned techniques to increase robustness.

### 3.1. Image Based Techniques

Image based techniques use the pixel information directly, the advantage is that they are computationally inexpensive but the disadvantage is that they are adversely affected by variation such as illumination.

#### 3.1.1. *Color Based*

Several algorithms base the detection of lips directly on color difference between the lip and skin, but lack of contrast and illumination variation adversely affects these techniques. Some have also suggested color transforms that increase the contrast between skin and lip regions. [128] reported that difference between red and green is greater for lips than skin and proposed a pseudo hue as a ratio of RGB values. [129] have also proposed a RGB value ratio based on the observation that blue color plays a subordinate role so suppressing it improves segmentation.

Color clustering has also been suggested by some, based on the assumption that there are only two classes i.e. skin and lips, this may not be completely true if facial hair or teeth are visible. Fuzzy clustering was applied for lip detection in [130] by combining color information and spatial distance between pixels in an elliptical shape function. [131] have used expectation maximization algorithm for unsupervised clustering of chromatic features for lip detection in normalized RGB color space. Markov random fields have also been proposed to add spatial continuity [132] to segmentation based on color, thus making segmentation more robust.

[137] propose a lip modelling method that is based on the predictive validation technique. Instead of using clustering, GMMs are learned by using predictive validation and applied on normalized RGB pixel values for lip segmentation. [138] have proposed a lip segmentation method for video sequences. First a logarithmic color transform is performed from RGB to HSI space, next color and motion parameters are analyzed and extracted. Then spatiotemporal segmentation based on markov random fields is applied using red hue and motion parameters.

### 3.1.2. *Subspace Based*

[133] have proposed a lip detector based on PCA, firstly outer lip contours are manually labelled on training data, PCA is then applied to extract the principal modes of contour shape variation, called eigencontour, finally linear regression was applied for detection. LDA has been employed in [134] to separate lip and skin pixels. [135] have used manifolds to learn the “lip space”, which is then used for tracking and extracting the lips. [136] have proposed a method in which a Discrete Hartley Transform (DHT) is first applied to enhance contrast between lip and skin, then a wavelet multi scale edge detection is applied on the  $C_3$  component of DHT.

## 3.2. Model Based Techniques

Model based techniques are based on prior knowledge of the lip shape and can be quite robust if properly initialized. They are however computationally expensive as compared to image based techniques as they usually involve minimization of a cost function.

[139] have proposed a simple 3 state geometric model made of parabola to model the mouth. [140] have proposed a real time tracker that models the dynamic contours of lips using quadratic B-Splines learned from training data using maximum likelihood estimation algorithm. Tracking is then carried out using Kalman filtering for both frontal and profile view of the lips. [141] have proposed a model consisting of two parabolas for the upper lip and one for lower lip, [142] use quartics instead of parabolas.

Snakes have been commonly used for lip segmentation [143] and achieve reasonable results but need to be properly initialized. Another problem faced by snakes is their inability to detect lip corners as they are located in low gradient regions. [144] have proposed a jumping snake that removes the limitations present in classical snake. It can be initialized far from the lip edge and the parameter adjustment is easy and quite intuitive.

[145] have proposed ASM and AAM, which learn the shape and appearance of lips from training data that has been manually annotated. Next PCA is applied to reduce the dimensionality and using cost functions models are iteratively fitted to test images for lip detection. [146] have modified ASM by using a texture constrained shape prediction instead of the classical Gaussian models. Local texture is modelled using classifiers learnt from training sets by AdaBoost.

Deformable templates initially proposed by [101] has been extended and modified by several others. [147] have proposed a lip detection method based on PDM of the face.

### 3.3. Hybrid Techniques

These techniques combine image based and model based techniques for lip detection. Image based techniques are considered relatively computationally less expensive but not so robust to illumination and other types of variation. Model based techniques on the other hand are robust and accurate but are much more computationally complex. Thus majority of the hybrid techniques proposed in the literature use color based techniques for a quick and rough estimation of the candidate lip regions and then apply a model based approach to extract accurate lip contours.

[148] have proposed a hybrid technique that first applies a color transform to reduce the effect of lighting. Then horizontal and vertical projections of the lip are analyzed to detect the corner points and finally a geometric lip model is applied. [149] have combined a fuzzy clustering algorithm in CIELAB color space for rough estimation and then an ASM for accurate detection of lip contours. [150] have proposed a hierarchical segmentation method based on modelling of prior lip information using GMMs in HSI color space and geometric models.

[151] have proposed a hybrid system that models the lip area by expectation maximization algorithm after a color transform in RGB space has been applied. Then a snake is initialized, which is fitted on the upper and lower contours of the mouth by a multi level gradient flow maximization. [139] have proposed a lip tracking by combining lip shape, color and motion information. The shape has been modelled using two parabolas, lip and skin color is modelled by Gaussian distribution and motion by modified Lucas-Kanade tracking.

---

## 4. State of Art: Visual Lip Feature

Visual lip features can be broadly divided into two categories, static and dynamic. Static features provide a snapshot of the lip shape and appearance at a specific instant of time and characterize the physical aspect of the lip. The dynamic features represent the motion of the lip during speech or other labial activity and represent the behavioural aspect of lip. It has been established that static features provide relevant data for person recognition, where as dynamic features are more suitable for speech recognition.

### 4.1. Static

The static features can be further divided into appearance, shape and hybrid features.

#### 4.1.1. *Appearance*

Appearance features are based on the Region of Interest (ROI) around the detected lip, which is normally a rectangle which may or may not contain other facial features such as lower nose or chin. Color or grey level pixel values could be used directly [195],[196] but the dimensionality of the feature vector becomes prohibitively large for classification, thus a dimension transformation is highly useful. Examples of such transforms are PCA, generating eigenlips [197], 2-D Fourier Transform [195],[196], the Discrete Cosine Transform (DCT) [157], the Discrete Wavelet Transform (DWT)[157], Linear Discriminant Analysis (LDA) [198], Fisher Linear Discriminant (FLD), and the Maximum Likelihood Linear Transform (MLLT) [217]. LDA and FLD take separation between the pattern classes into account thus provide most discriminant features.

Other techniques include binarization of image pixel intensities [200], aggregation of pixel intensity differences between image frames [201], nonlinear image decomposition based on the “sieve” algorithm [202]. In [49] SLDA which is LDA with specific prior knowledge of the speakers is applied. MRPCA is used in [49] where the mean is first removed and PCA is then applied. Appearance based features suffer from head pose and lighting variation but are quite fast to compute.

#### 4.1.2. *Shape Based*

Shape based features assume that most of the discriminatory information is contained in the lip shape [203], thus several studies have extracted compact representations of the lip shape. Although they are invariant to pose and lighting but they are difficult to extract and computationally intensive. They are further divided as geometric and model based features.

*Geometric:* Geometric features, such as the height, width, perimeter of the mouth have been used for speaker recognition [204]. Distances or angles between key points located on the lip margins, mouth corners, or jaw have been used in [205],[206].

*Model Based:* Model-based visual features are normally used when a parametric or statistical facial feature extraction algorithm is used. In model-based approaches, the model parameters are directly used as visual speech features such as Deformable templates [208], Active Shape Models (ASM) [210], and snakes [211].

#### 4.1.3. *Hybrid*

Shape and appearance based features have been combined into a single model for improving performance. PCA appearance features are combined with ASMs to create AAMs [207]. In [29], the lip feature vector is formed by concatenating the Karhunen Løve Transform (KLT) coefficients of the inner-outer lip contour points with the texture information. Perez et al. [209] utilize a set of lip shape features extracted by ASM together with DCT coefficients of the grey-level appearance information.

## 4.2. Dynamic

Dynamic features that capture the behavioural rather than the physical aspect of the lip have been used for some time in the speech recognition community and now have recently been added to the person recognition domain. In [212] Motion History Images (MHI) have been used as visual feature, they are created for each person by accumulating intensity changes of pixels over the entire sequence. Before summation time information implicitly added by multiplying the intensity values in each frame by the frame number.

Optical flow is the most common and easy to extract visual feature. In [50] dense optical flow is first calculated, then these dense velocity vectors are quantized by allowing only 3 directions ( $0^{\circ}$ ,  $45^{\circ}$ ,  $-45^{\circ}$ ), and only 20 values resulting in a feature vector of 40 parameters. These quantization values were obtained by fuzzy c-means clustering. In [52] video signal was sub sample to 128X128 and optical flow calculated, the resultant was further sub sampled to 16X16 and a three dimensional Fast Fourier Transform (FFT) was applied to extract visual features.

Another method of extracting dynamic features is to first extract static features, then to derive dynamic features from these static features by taking derivative over a window. In [213], the radial magnitudes are measured from points around the circumference of the lip, stepping pixel by pixel to the mid point of the principal diagonal. The final lip signature is then derived by taking the DCT of the radial magnitudes. Dynamic features measuring the rate of change are then extracted by first and second order differentials, over a sliding window. In [56] Lip ROI was converted to grey level images and DCT was applied, then 15 coefficients were extracted using zigzag scan. Dynamic features were then extracted from these static features by taking difference of DCT coefficients over a certain number of frames.

## 5. Proposed Lip Detection

In this section we present a lip detection method to extract the outer lip contour that combines edge based and segmentation based algorithms. The results from the two methods are then combined by AND/OR fusion. The novelty lies in the fusion of two methods, which have different characteristics and thus exhibit different type of strengths and weaknesses. The other significance of this study lies in the extensive testing and evaluation of the detection algorithm on a realistic database. Most previous studies either never carried out empirical comparisons to the ground truth or sufficed by using a limited dataset. Some studies [220],[221] do exist that have presented results on considerably large datasets but these mostly consists of high resolution images with constant lighting conditions. Figure 7 gives an overview of the lip detection algorithm. Given a database image containing a human face the first step is to select the mouth Region of Interest (ROI) using the tracking points provided with the database. The next step involves the detection where the same ROI is provided to the edge and segmentation based methods. Finally the results from the two methods are fused to obtain the final outer lip contour.

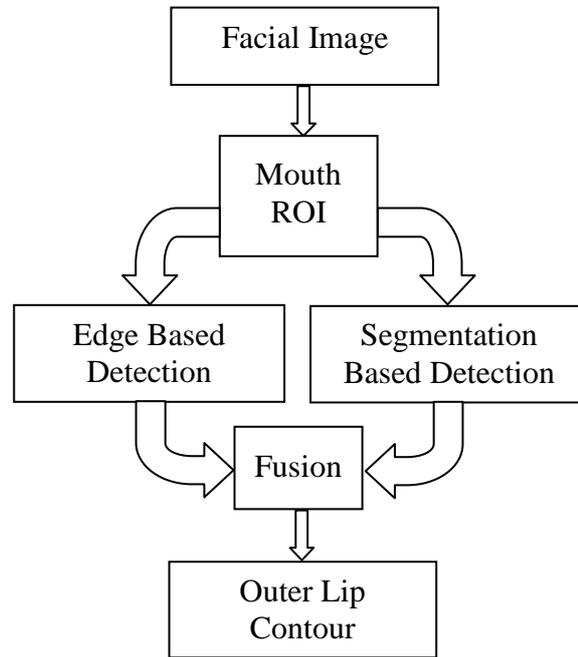


Figure 7: Overview of lip detection.

### 5.1. Edge Based Detection

The first algorithm is based on a well accepted edge detection method, it consists of two steps, the first one is a lip enhancing color transform and the second one is edge detection based on active contours. Several color transforms have already been proposed for either enhancing the lip region independently or with respect to the skin. Here, after evaluating several transforms we have selected the color transform proposed by [216]. It is based on the principle that blue component has reduced role in lip / skin color discrimination. It has been defined as:

$$I = \frac{2G - R - 0.5B}{4}$$

Where R,G,B are the Red, Green and Blue components of the mouth ROI. The next step is the extraction of the outer lip contour, for this we have used active contours [152]. Active contours are an edge detection method based on the minimization of an energy associated to the contour. This energy is the sum of internal and external energies; the aim of the internal energy is to maintain the shape as regular and smooth as possible. The most straightforward approach grants high energy to elongated contours (elastic force) and to high curvature contours (rigid force). The external energy models the edge of the object and is supposed to be minimal when the active contours (snake) is at the object boundary. The simplest approach consists of using regularized gradient as the external energy. In our study the contour was initialized as an oval half the size of the ROI with node separation of four pixels.

Since we have applied active contours which have the possibility of detecting multiple objects, on a ROI which may include other features such as the nose tip, jaw line etc. an additional cleanup step needs to be carried out. This consists of selecting the largest detected object approximately in the middle of the image as the lip and discarding the rest of the detected objects.

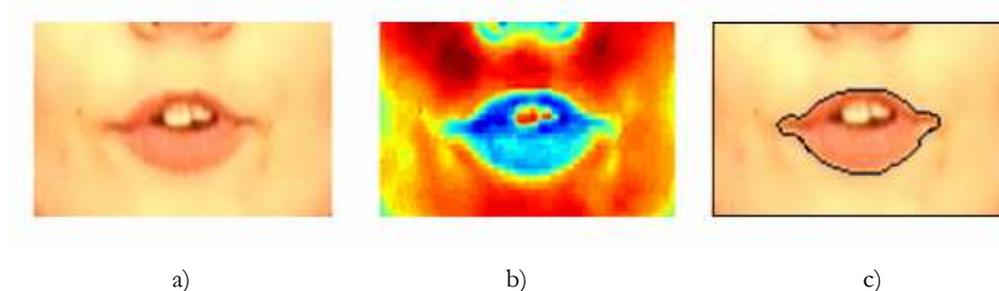


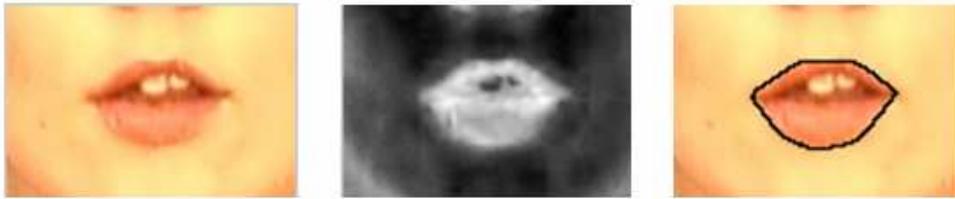
Figure 8: a) Mouth ROI, b) Color Transform, c) Edge Detection.

## 5.2. Segmentation Based Detection

In contrast to the edge based technique the second approach is segmentation based after a color transform in the YIQ domain. As in the first approach we experimented with several color transform presented in the literature to find the one that is most appropriate for lip segmentation. [153] have presented that skin/lip discrimination can be achieved successfully in the YIQ domain, which firstly de-couples the luminance and chrominance information. They have also suggested that the I channel is most discriminant for skin detection and the Q channel for lip enhancement. Thus we transformed the mouth ROI from RGB to YIQ color space using the equation below and retained the Q channel for further processing.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.31135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

In classical active contours the external energy is modelled as an edge detector using the gradient of the image, to stop the evolution of the curve on the boundary of the desired object while maintaining smoothness in the curve. This is a major limitation of the active contours as they can only detect objects with reasonably defined edges. Thus for the second method we selected a technique called “active contours without edges” [222], which models the intensities in different region of the image and uses it as the stopping term in active contours. More precisely this model [222] is based on Mumford–Shah functional and level sets. In the level set formulation, the problem becomes a mean-curvature flow evolving the active contour, which will stop on the desired boundary. However, the stopping term does not depend on the gradient of the image, as in the classical active contour models, but is instead based on Mumford–Shah functional for segmentation.



a) b) c)  
Figure 9: a) Mouth ROI, b) Color Transform, c) Region Detection

### 5.3. Error Detection and Fusion

Lip detection being an intricate problem is prone to errors, especially the lower lip as reported by [154]. We faced two types of errors and propose appropriate error detection and correction techniques. The first type of error, which was commonly observed, was caused when the lip was missed altogether and some other feature was selected. This error can easily be detected by applying feature value and locality constraints such as the lip cannot be connected to the ROI's boundary and cannot have an area value less than one-third of the average area value in the entire video sequence. If this error was observed, the detection results were discarded.

The second type occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect thus we proposed to use fusion as a corrective measure, under the assumption that both the detection techniques will not fail simultaneously.

---

The detection results from the above described methods were then fused using AND and OR logical operators. The outer lip contours are used to create binary masks which describe the interior and the exterior of the outer lip contour. These were then fused as;

#### AND Logical Operator

The first was the logical AND operator defined as

A	B	$\Lambda$
0	0	0
0	1	0
1	0	0
1	1	1

#### OR Logical Operator

The second was the logical OR operator defined as

A	B	V
0	0	0
0	1	1
1	0	1
1	1	1

Table 3 presents the commonly observed errors and the effect of OR fusion on the results.

	Type 1 Error	Type 2 Error	No Error
Segmentation Based			
Edge Based			
OR Fusion			

Table 3: Errors and OR Fusion

#### 5.4. Experiments and Results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on Valid Database [65] which consists of five recording sessions of 106 subjects using the third utterance. One image was extracted from each of the five videos to create a database of 530 facial images. The reason for selecting one image per video was that the database did not contain any ground truth for lip detection, so ground truth had to be created manually, which is a time consuming task. The images contained both illumination and shape variation; illumination from the fact that they were extracted from all five videos, and shape as they were extracted from random frames of speaker videos.

As already described above the database did not contain any ground truth with respect to the outer lip contour. Thus the ground truth was established manually by a single operator using Adobe Photoshop. The outer lip contour was marked using the magnetic lasso tool which separated the interior and exterior of the outer lip contour by setting the exterior to zero and the interior to one.

To evaluate the lip detection algorithm we used the following two measures proposed by [221], the first measure determines the percentage of overlap (OL) between the segmented lip region  $A$  and the ground truth  $A_G$ . It is defined as

$$OL = \frac{2(A \cap A_G)}{A + A_G} * 100$$

Using this measure, total agreement will have an overlap of 100%. The second measure is the segmentation error (SE) defined as

$$SE = \frac{OLE + ILE}{2 * TL} * 100$$

*OLE* (outer lip error) is the number of non-lip pixels being classified as lip pixels and *ILE* (inner lip error) is the number of lip-pixels classified as non-lip ones. *TL* denotes the number of lip-pixels in the ground truth. Total agreement will have an SE of 0%.

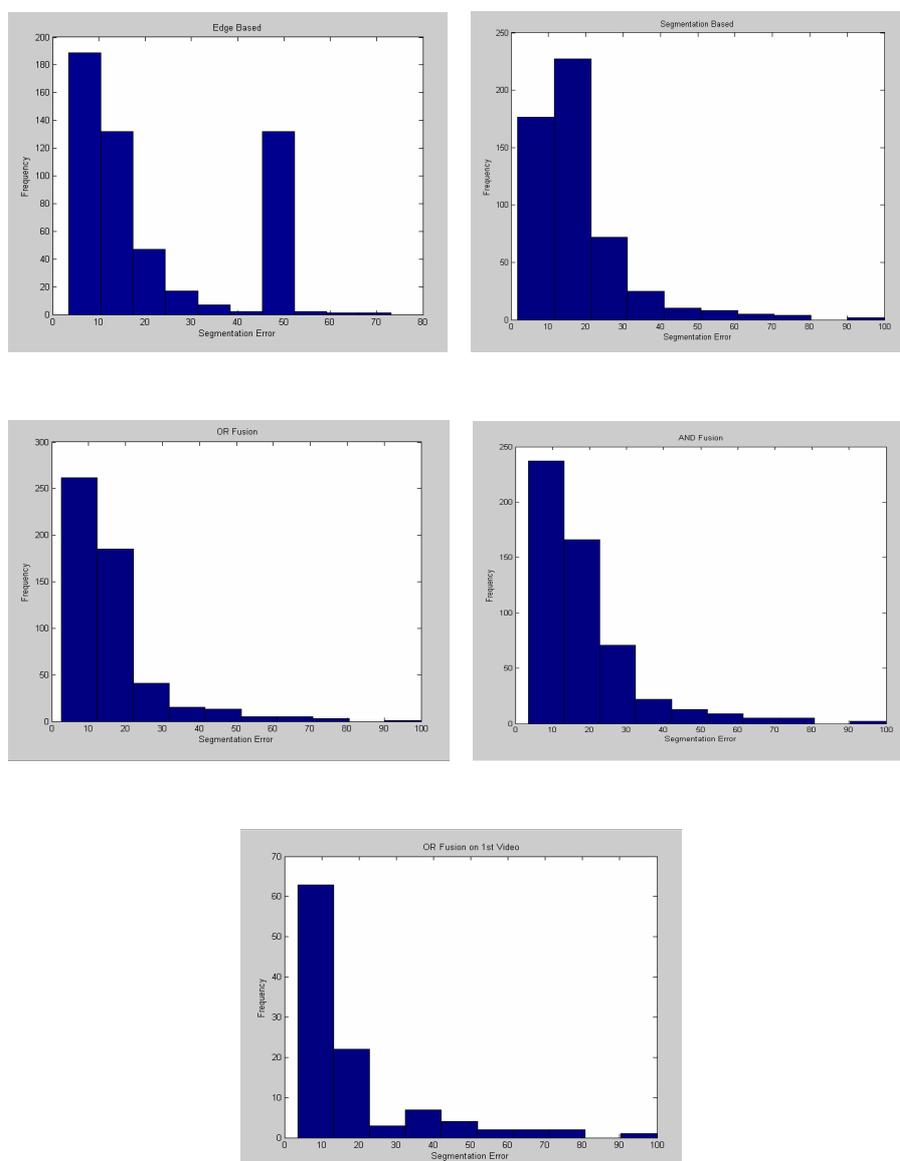


Figure 10: Histograms for Segmentation Errors

Initially we calculated the overlap and segmentation errors for edge and segmentation based methods individually, and it was visually observed that edges based method was more accurate but not robust and on several occasions missed almost half of the lip. This can also be observed in the histogram of segmentation errors; although the majority of lips are detected with 10% or less error but a large number of lip images exhibit approximately 50% of segmentation error. On the other hand segmentation based method was less accurate as majority of lips detected are with 20% error but was quite robust and always succeeded in detecting the lip.

Lip Detection Method	Mean Segmentation Error (SE) %	Mean Overlap (OL) %
Segmentation Based	17.8225	83.6419
Edge Based	22.3665	65.6430
OR Fusion	15.6524	83.9321
AND Fusion	18.4067	84.2452
OR Fusion on 1 <sup>st</sup> Video	13.9964	87.1492

Table 4: Lip detection Results

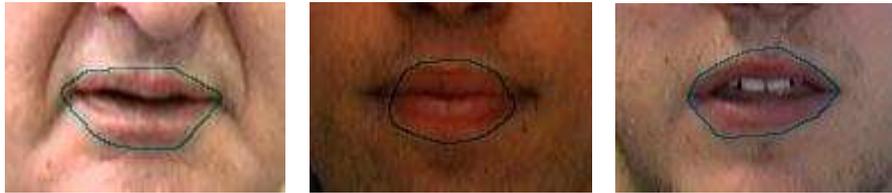


Figure 11: Example of Images with 15 % Segmentation Error

The fusion techniques were then applied and the best results were observed for OR fusion. The minimum segmentation error obtained was around 15%, which might seem quite large, but on visual inspection of Figure 11, it is evident that missing the lip corners or including a bit of the skin region can lead to this level of error. Another aspect of the experiment that must be kept in mind is the ground truth. Although every effort was made to establish an ideal ground truth but due to limited time and resources some compromises had to be made. “OR Fusion on 1<sup>st</sup> Video” are the results that were obtained when OR fusion was applied to only the images from the first video, which are recorded in studio conditions.

## 5.5. Conclusions

In this section we have presented a novel lip detection method based on the fusion of edge based and segmentation based methods, along with empirical results on a dataset of considerable size with illumination and speech variation. We observed that the edge based technique is comparatively more accurate, but is not so robust and fails if lighting conditions are not favourable, thus it ends up selecting some other facial feature. On the other hand the segmentation based method is robust to lighting but is not as accurate as the edge based method. Thus by fusing the results from the two techniques we achieve comparatively better results which can be achieved by using only one method. The proposed methods were tested on a real world database of considerable size and illumination/speech variation with adequate results.

## 6. Evaluation of Lip Features

### 6.1. Introduction

Visual lip features have been studied in speech and speaker recognition for some time now but a relatively new addition to the use of lip features is in the field of person recognition, where the aim is to use only visual lip features to recognize people. But this field is still nascent and lacks a comprehensive study to compare various lip features for person recognition. Some effort has been made in this regard but it has either been limited by the diversity of features extracted [215], or has sufficed by using a classifier as a feature selection method. In this section we aim to compare visual features from lip motion for their relevance to person recognition. For this purpose we have extracted various geometric and appearance based lip features and compared them using three feature selection measures.

### 6.2. Previous Work on Feature Selection

The objective of feature selection is to choose a subset of features from a much larger set of features, given an evaluation criterion to compare the performance of the new subset. Traditionally several factors have been the motivation behind feature selection, first and foremost has been to reduce the number of features to save time and computational complexity for the classification process, secondly to improve classification results by removing irrelevant parameters, and lastly to prevent over-fitting due to parameters which are mostly noise. Feature selection algorithms can be divided into the following categories.

### 6.2.1. *Filter Schemes*

These methods act as pre-filters to the classification task and involve using certain generic characteristic of the data such as correlation, information gain or entropy [155], to select and rank subsets of features. Mostly this group of techniques is used with large number of features as they have better computational complexity. They also tend to provide better generalization as compared to other techniques at comparable accuracy.

### 6.2.2. *Wrapper Schemes*

These techniques are feature selection techniques [156] that are specifically based on a classifier and use the classifiers ability to evaluate the relevance of the feature subset. This leads to feature subsets that are highly accurate for certain classifier at the cost of generalization. These are also computationally expensive so may be inappropriate for large variable sets.

## 6.3. Proposed Feature Extraction

In the feature extraction phase both geometric and appearance based features are extracted. This module takes as input the mouth ROI and extracts feature vectors which will be subsequently used in the feature selection module.

### 6.3.1. *Geometric Features*

The geometric features that we have extracted for this study include area, length of major and minor axis, eccentricity, orientation and length of the perimeter of the outer lip contour. The algorithm takes the mouth ROI and detects the outer lip contour as described in Chapter IV.5. Next the shape of the lip is characterized by several features which include the area contained inside the outer lip contour, the length of the major and minor axis and the perimeter of the lip contour. Eccentricity and orientation were also calculated from the major and minor axis to be used as a person independent feature. The output of the module for each frame of the video is organized in a feature vector for the feature selection module.

### 6.3.2. *Appearance Based Features*

We have employed several appearance based techniques to extract visual feature vectors. As these techniques result in a wide range of different type and size of features, so to provide a fairer comparison we have limited the size of all feature vectors to a standard of 300 features per frame.

*Pixel Intensity Profiles* Based on the lip contour seven vertical and one horizontal scan were defined (cf. Figure 12). Pixel intensity levels from these scan lines were then concatenated to form a feature vector of size 300.

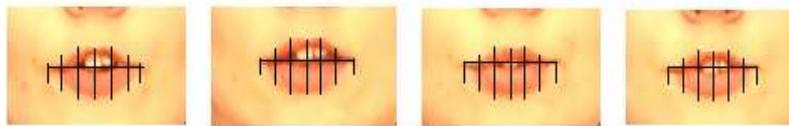


Figure 12: Pixel intensity profiles.

*Mean Removed PCA* (MR-PCA) as proposed by [157] consists of calculating the average mouth image from a video sequence and then subtracting it from each frame. This enables us to remove unwanted variation that is static and subject dependent. First a mean mouth image was calculated as

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t$$

Then new mean removed frames were calculated as

$$y^* = y_t - \bar{y}$$

These mean removed mouth frames are then used to create a PCA subspace where 300 of the maximum modes of variations are preserved.

*Optical Flow* We have calculated the dense optical flow with Lucas-Kanade method and then to reduce the feature vector to the standard size of 300 we apply an averaging filter.

*Spatio-Temporal Templates* (STT) and similar techniques [158] are generated by accumulating frame by frame image difference in a video sequence. STT captures both the location and time of motion occurrence and removes static objects. First intensity values were subtracted in consecutive frames. Then these difference images were binarized. Finally time information was added implicitly by multiplying frames with a time factor, which was a linear ramp function in our case.

*Discrete Fourier Transform* (DFT) Two dimensional fast Fourier transform was calculated for each mouth frame and 300 of the highest frequencies were preserved as the feature vector.

*Independent Component Analysis* (ICA) was used to create a subspace preserving 300 maximum variation and then feature vectors were projected into this space.

*Discrete Wavelet Transform* (DWT) A second order 2-D DWT was applied to each mouth image using Haar wavelets conversing low-low coefficients successively.

*Discrete Cosine Transform* (DCT) was applied to each mouth frame and then 300 highest coefficients were selected by a zigzag scan.

## 6.4. Feature Selection

The next module evaluates feature extracted from the previous module for relevance to person identification. We have selected one advance technique and two basic ones which are described below.

### 6.4.1. *mRMR Feature Selection*

Most of the filter based techniques are based on the concept of simple ranking, in which features are first ranked and the top most ranking features are selected. Whereas it is quite possible that the selected features are highly correlated and thus redundant. Minimal-Redundancy-Maximum-Relevance (mRMR) proposed by [159] aims to solve this problem by first selecting features set  $S$  that has maximum relevance between feature  $x_i$  and target class  $c$  by means of a similarity measure such as mutual information.

$$\max D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)$$

Then reducing redundancy by selecting features that are maximally dissimilar to each other as

$$\min R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j)$$

Finally the above two criteria are combined and optimized as

$$\max \Phi(D, R) = D - R$$

### 6.4.2. *Bhattacharya Distance (BD)*

Bhattacharyya distance measures the similarity between two discrete probability distributions. For discrete probability distributions  $p$  and  $q$  over the domain  $X$ , it is defined as:

$$D_B(p, q) = -\ln(BC(p, q))$$

Where BC is the Bhattacharyya coefficient and is defined as

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

### 6.4.3. *Mutual Information (MI)*

Mutual information of two random variables  $X$  and  $Y$  is the quantity that measures the mutual dependence of the two variables and for discrete variables can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p_1(x)p_2(y)} \right)$$

where  $p(x,y)$  is the joint probability distribution function of  $X$  and  $Y$ , and  $p_1(x)$  and  $p_2(y)$  are the marginal probability distribution functions of  $X$  and  $Y$  respectively.

## 6.5. Experiments and Results

In this section we elaborate the experimental setup and discuss the results obtained from the feature selection module. Tests were carried out on Valid Database [65] which consists of five recording sessions of 106 subjects using the third utterance. Videos were first converted to sequence of images at a frame rate of 25f/s and then feature extraction techniques (ICA, etc.) were applied to individual frames to extract feature vectors. Then feature vectors were concatenated and analyzed using feature selection methods.

Rank	mRMR	BT	MI
1	ICA	PCA	DWT
2	Intensity Profiles	DCT	DFT
3	STT	ICA	Geometric
4	DFT	DFT	PCA
5	DCT	Intensity Profiles	STT
6	DWT	STT	DCT
7	Geometric	Optical Flow	ICA
8	Optical Flow	Geometric	Optical Flow
9	PCA	DWT	Intensity Profiles

Table 5: Rank Results for mRMR, Bhattacharya Distance, Mutual Information

Rank	Rank Fusion
1	ICA
2	DFT
3	STT
4	DCT
5	Intensity Profiles
6	PCA
7	DWT
8	Geometric
9	Optical Flow

Table 6: Rank Fusion

Table 5 provide the ranking results according to the feature selection techniques described before. Just looking at these tables at first glance the results may appear to be random and uncorrelated but on a deeper analysis some conclusions can be deduced such as the Discrete Fourier Transform has average performance across all feature selection techniques. ICA always performs better than intensity profile method and that optical flow has the one of the lowest performance for all the three methods. Therefore to present a much clearer picture of the results we decided to fuse the ranking results form the three techniques. The fusion is a weighted sum of the ranks from the three techniques and is given as

$$\text{Rank Fusion} = (2 * mRMR) + BD + MI$$

The mRMR technique is given twice the weight as compared to the other techniques, the reason being that we believe the mRMR is the most advanced technique and is better able to take the mutual information and redundancy into account. From Table 6 we can clearly see a ranking; here we would like to comment about the best and the worst performing technique. As expected the best performing technique is ICA which is a supervised technique and takes the class information into account. The two worst performing techniques are the geometric features and optical flow. The reason behind the poor performance of geometric features is they contain much less information as compared to other features such as DFT i.e. size of a geometric feature is 6 while that of DFT is 300, thus it is not a fair comparison from the beginning. Whereas the poor performance of the optical flow features can be attributed to the fact that they are purely behavioural feature and lack the physical aspect of lip shape and appearance.

## 6.6. Conclusions

In this section we have presented a comparison between various feature extraction techniques and their relevance to person recognition. Geometric and appearance based features were extracted and three feature selection techniques were used to compare them. We observed weak correlation between the results from the three methods and thus decided to fuse the ranking results. A supervised technique ICA, gave the best performance, where as geometric and optical flow features performed poorly. Thus the instinctive notion is reinforced that appearance contains more information for person recognition than shape and behaviour.



## *Chapter V. Application of Lip Features*

---

### 1. Introduction

In this chapter we have elaborated some of the applications of lip features. Traditionally lip features have been used for AV speech and speaker recognition but here we have also employed them for related fields. In the first section we present our work for recognition of persons using behavioural lip features. In the next section we present a response registration system based on lip features.. In the last section we discuss a gender recognition system based on a combination of behavioural lip features with appearance.

### 2. Lip features for person recognition

#### 2.1. Introduction

In this chapter we investigate the possible contribution of lip features extracted from low quality videos for person recognition. The work is based on the intuitive notion that people have a unique speech pattern that is a characterizing behaviour of that person. Thus we have extracted two feature vectors, both based on the motion of lip and attempted to recognize people using them. The initial results reported tend to validate this original proposal, thus opening some new perspectives for design of future hybrid and efficient system.

#### 2.2. Behavioural Lip Features

Lip features as they have been described before can be divided as static and dynamic. The static features model the shape and appearance of the lip at an instant of time, while the dynamic features model that motion of the lip over time. In this study we decided to focus on the behavioural aspect of speech, thus we extracted static and dynamic features, paying special attention not to include any physical attributes of the lip shape and appearance. Once the outer lip contour has been detected behavioural features are extracted, which include normalized static features and optical flow based dynamic features.

### 2.2.1. Static Features

Geometric features such as width, height, and lip orientation have been used for some time either explicitly or implicitly in face recognition. They model the shape and thus the physical attributes of the lips which were undesirable for this study. Thus we extract geometric features and perform a normalization step to conserve only the behavioural aspect of the lip shape.

For each frame  $\Phi_t$  at time  $t$  the outer lip contour was detected as described in Chapter IV.5, and then geometric features  $Gf_t$  were extracted. Which consist of the x-y coordinates of 4 extremas points and the length of the major and minor axis of the outer lip contour.

$$P_t = [x1_t, y1_t, x2_t, y2_t, x3_t, y3_t, x4_t, y4_t, Maj_t, Min_t]$$

Normalization was then carried out which consisted of subtracting the mean value from each feature, given by the following transform,

$$x_{n,t} = p_{n,t} - \mu_n$$

for  $n = 1, \dots, 10$  and  $t = 1, \dots, T$ , where  $\mu_n$  is the mean value of the  $n$ -th feature,

$$\mu_n = \frac{1}{T} \sum_{t=1}^T p_{n,t}.$$

Finally the normalized features are concatenated in a feature vector

$$Gf_t = [x_{1,t}, \dots, x_{n,t}]$$

### 2.2.2. Dynamic Features

Dynamic features model the motion of the lips using the appearance, either in the form of color or grey level value. Several dynamic features have been studied in literature, but the most common have been the features extracted from optical flow. In this study we have also included motion features using Lucas Kanade optical flow. Dense optical flow features were calculated frame by frame. The following vector depicts the dynamic feature vector;

$$Df_t = [u_{1,1,t}, v_{1,1,t}, u_{1,2,t}, v_{1,2,t}, \dots, u_{n,m,t}, v_{n,m,t}]$$

Where  $u$  and  $v$  are the horizontal and vertical component of the motion vector calculated for the row  $n$  and column  $m$  of the mouth ROI.

## 2.3. Person recognition

Evaluation of the geometric  $Gf_t$  and dynamic  $Df_t$  behavioural features was carried out separately. The features were extracted for each person frame by frame and then concatenated in a matrix  $\mathbf{X} \in \mathfrak{R}^{D \times T}$  which at anytime consists of either the geometric  $Gf_t$  or dynamic  $Df_t$  features.

$$X = [Gf_1, Gf_2, \dots, Gf_T] \quad \Lambda \quad [Df_1, Df_2, \dots, Df_T]$$

Our person recognition module consists of GMM modelling and Bayes classifier, which are explained below.

### 2.3.1. GMM based Model Estimation

The enrolment phase consists of a probabilistic approach that estimates the distribution of feature vectors of each client in the feature space, i.e. for each individual  $k$ , we aim to represent his class conditional probability density function (PDF) of feature vectors:  $p(\mathbf{x}_n | k)$ .

Gaussian mixture models (GMMs) have been extensively used as a generic probabilistic model for approximating multivariate densities. Moreover, as GMMs are intrinsically unconstrained they are well suited to our recognition problem, in which there is no prior knowledge of user. Thus we decided to approximate each class conditional PDF by employing a Gaussian mixture models (GMMs).

A GMM is a finite mixture model of Gaussian distributions. A non-singular multivariate normal distribution of a random variable,  $\mathbf{x} \in \mathfrak{R}^D$ , is defined as:

$$\mathfrak{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{D/2} \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where  $\boldsymbol{\mu} \in \mathfrak{R}^D$  is the mean vector, and  $\boldsymbol{\Sigma} \in \mathfrak{R}^{D \times D}$  is the non-singular covariance matrix.

Then, a Gaussian mixture model probability density function (GMM-PDF) is a weighted sum of  $C$  normal distributions:

$$p(\mathbf{x} | \boldsymbol{\Theta}) \equiv \sum_{c=1}^C \alpha_c \mathfrak{N}(\mathbf{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

in which  $\boldsymbol{\Theta} = \{\alpha_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 1, \dots, C\}$  is the parameter list, and  $\alpha_c \in [0, 1]$  is the weight of the  $c$ -th Gaussian component. In addition, each  $\alpha_c$  corresponds to the a priori probability that an observation  $\mathbf{x}$  has been generated by the  $c$ -th normal source, and its value is normalised such as:  $\sum_{c=1}^C \alpha_c \equiv 1$ .

If we assume statistical independence between the  $K$  classes that correspond to the clients of our system, then the overall estimation of GMM parameters can be divided into  $K$  separate estimation problems. Hence, for each client  $k$ , his model parameters  $\Theta_k$  are obtained by solving a maximum likelihood problem. Unfortunately, the analytical approach for solving the maximum likelihood problem is intractable for GMMs with unknown and unrestricted covariance matrices and means; the solution is then to apply an optimisation strategy, such as the expectation-maximisation (EM) algorithm [214]. The EM algorithm requires an initialization step with an initial estimate of the model parameters  $\Theta^{(0)}$ . This step is important, because the choice of  $\Theta^{(0)}$  determines where the algorithm converges, or hits the boundary of the parameter space producing singular meaningless results. The common solution is using a clustering algorithm like the K-means or the fuzzy K-means [218]. The initialisation of the EM algorithm is done in two phases. Firstly, the training data is clustered into  $C$  partitions, by applying the *K-means* method or the *fuzzy K-means* one. After that, the initial parameter set,  $\Theta^{(0)}$ , is calculated by: taking the cluster means, uniform or cluster covariance matrices, and uniform or cluster weights.

In GMM modelling, the total number of Gaussian components  $C$  does not need to be guessed accurately: it is just a parameter defining the complexity of the approximating distribution. However, if  $C$  is too small, there is not an adequate amount of components to learn the feature distribution precisely. On the other hand, when  $C$  is too large the modelling is excessively complex, this may lead either to an over fitted classifier, or to singularities in the covariance matrices.

We also need to guess the minimum number of training feature vectors,  $N^{(\min)}$ , that are recommended for a reliable estimation of the GMM parameters. Firstly, we recall that the number of free parameters in a GMM, with  $C$  Gaussian components and  $D$ -dimensional real feature vectors  $\mathbf{x}_n \in \mathfrak{R}^D$ , is:

$$\eta = C * \left( \frac{1}{2} D^2 + \frac{3}{2} D \right) + C - 1$$

Then, as a rule of thumb, we empirically require a minimum number of feature vectors as in [214]:  $N^{(\min)} > 3\eta$ . It is worth noting that the number of recommended vectors increases linearly with the number of Gaussian components (the complexity of the modelling), and quadratically with the dimensionality of the feature space.

### 2.3.2. Bayesian classification

The classification task of our system is achieved by applying the probability theory and the Bayesian decision rule (also called Bayesian inference) [214], so that the classifier chooses the most probable class, or equivalently the option with the lowest risk (expected cost).

In our framework, the test vector is actually a video sequence thus, we aim to compute the video posterior probability,  $p(k | \mathbf{X})$ , which we define as the probability that all feature vectors extracted from a video  $\mathbf{X} \in \mathfrak{R}^{D \times T}$  belong to class  $k$ :

$$p(k | \mathbf{X}) \equiv p(k | \mathbf{x}_1, \dots, \mathbf{x}_T)$$

By applying the Bayes' rule, the posterior probability  $p(k | \mathbf{X})$  becomes:

$$p(k | \mathbf{X}) = \frac{p(\mathbf{X} | k)p(k)}{p(\mathbf{X})} = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_T | k)p(k)}{p(\mathbf{x}_1, \dots, \mathbf{x}_T)}$$

First of all, the divisor:

$$p(\mathbf{X}) \equiv p(\mathbf{x}_1, \dots, \mathbf{x}_T) = \sum_{k=1}^K p(\mathbf{x}_1, \dots, \mathbf{x}_T | k)p(k) = M_{\mathbf{x}}$$

is merely a scaling factor  $M_{\mathbf{x}}$ , to assure that the posterior probabilities  $p(k | \mathbf{X})$  are really probabilities (their sum is one). Hence, we can simplify the previous expression as:

$$p(k | \mathbf{X}) = \frac{p(\mathbf{X} | k)p(k)}{M_{\mathbf{x}}} = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_T | k)p(k)}{M_{\mathbf{x}}}$$

Afterwards, the a priori probability  $p(k)$  represents the probability of occurrence of each class  $k$ , and it is usually estimated from the training database. Finally, in order to calculate the video posterior probability  $p(k | \mathbf{X})$ , we have to express the joint class conditional PDF  $p(\mathbf{X} | k)$  as a function of the class conditional PDFs of feature vectors  $p(\mathbf{x}_t | k)$ , which are our user models estimated during the enrolment. This task can be problematic, unless we assume that the feature vectors  $\mathbf{x}_t$  are independent from each other; this way, the joint class conditional PDF  $p(\mathbf{X} | k)$  takes the form of:

$$p(\mathbf{X} | k) \equiv p(\mathbf{x}_1, \dots, \mathbf{x}_T | k) \cong \prod_{t=1}^T p(\mathbf{x}_t | k)$$

and the video posterior probability becomes:

$$p(k | \mathbf{X}) \cong \frac{p(k)}{M_{\mathbf{x}}} \prod_{t=1}^T p(\mathbf{x}_t | k)$$

The similarity score for the identification task  $S^{(ID)}(\mathbf{X}, \Theta_k)$ , is derived from the video posterior probability  $p(k | \mathbf{X})$  by computing the log-posterior probability, because it is analytically and numerically more practical, and the properties of the similarity function do not change thanks to the monotonicity of the logarithm. Hence,  $S^{(ID)}(\mathbf{X}, \Theta_k)$  takes the form of:

$$S^{(ID)}(\mathbf{X}, \Theta_k) = \ln p(k | \mathbf{X}) = \sum_{t=1}^T \ln p(\mathbf{x}_t | k) + \ln p(k) - \ln M_{\mathbf{X}}$$

Finally, the similarity score for the verification task  $S^{(VER)}(\mathbf{X}, \Theta_k)$ , is the log-posterior probability ratio:

$$S^{(VER)}(\mathbf{X}, \Theta_k) = \ln \left[ \frac{p(k | \mathbf{X})}{p(\bar{k} | \mathbf{X})} \right] = \sum_{t=1}^T \ln p(\mathbf{x}_t | k) - \sum_{t=1}^T \ln p(\mathbf{x}_t | \bar{k}) + 2 \ln p(k) - 1$$

where  $p(\bar{k} | \mathbf{X})$  is the posterior probability of the alternative hypothesis  $\bar{k}$ , and  $p(\mathbf{x}_t | \bar{k})$  is the impostor model (the class conditional PDF for  $\bar{k}$ ). In other words,  $p(\bar{k} | \mathbf{X})$  expresses the probability that all feature vectors extracted from a video  $\mathbf{X} \in \mathfrak{R}^{D \times T}$  do not belong to class  $k$ , and  $p(\mathbf{x}_t | \bar{k})$  represents the probability that the alternative hypothesis  $\bar{k}$  can generate  $\mathbf{x}_t$ .

Unfortunately, the estimation of the impostor model  $p(\mathbf{x}_t | \bar{k})$  is usually problematic, because it should represent the space of all possible alternatives to  $k$ , which is huge and requires a massive amount of training data. Inspired by the speaker verification domain and the work of Rosenberg et al. [219], we approximate the impostor model by using the set of other client models  $p(\mathbf{x}_t | k)$ , which are called background models or cohorts. More precisely,  $p(\mathbf{x}_t | \bar{k})$  is estimated by taking the average of the  $L$  best client models on a given test (a video in our case):

$$p(\mathbf{x}_t | \bar{k}) \cong \frac{1}{L} \sum_{l=1}^L p(\mathbf{x}_t | k^{(l)})$$

where  $k^{(l)}$  is the client model that produces the  $l$ -th highest video posterior probability  $p(k | \mathbf{X})$ .

## 2.4. Results and experiments

### 2.4.1. *Experimental set-up*

Person recognition tests were carried out on the Italian TV Database, we selected 104 video sequences for training (8 for each of the 13 individuals), and the remaining 104 were left for testing. The geometric features were extracted frame by frame and then normalized. The dynamic features were extracted and then their dimensionality was reduced using Principal component analysis (PCA) and Discrete Cosine Transform (DCT). The reason for applying dimensionality reduction is two fold, the first was to provide a fair comparison with geometric features (10 features per frame), and the other was that we wanted to use the same classification framework for both the features. Optimal number of coefficients were calculated experimentally to be 13 for both PCA and DCT, but as we wanted a fair comparison we used only 10 PCA and DCT coefficients.

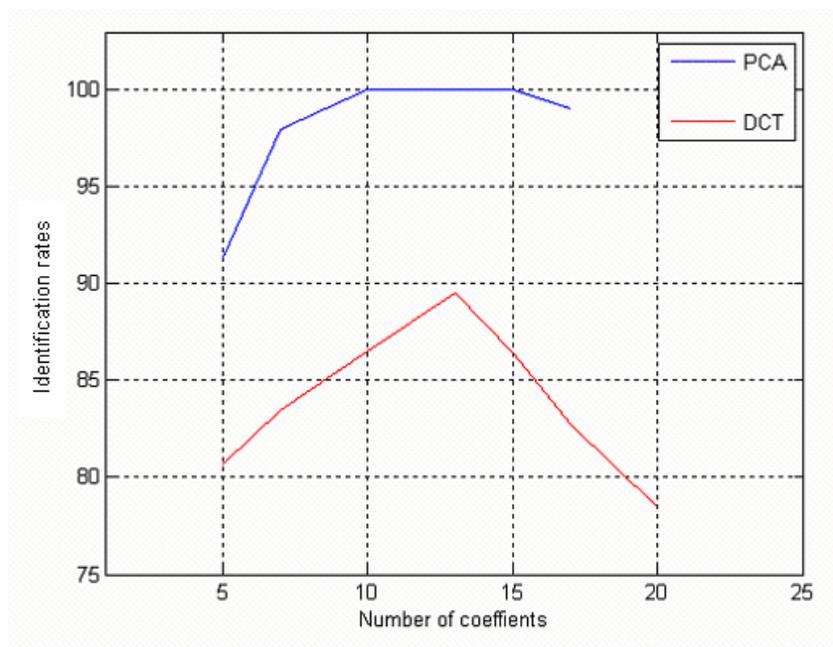


Figure 13: Identification results comparing DCT and PCA coefficients.

Client models are approximated using GMMs with 4 Gaussian components, and their parameters are estimated through the EM algorithm, which is initialised with: cluster means (computed using K-means), uniform weights and covariances. Finally, the impostor models for verification were approximated by taking the average of the best 2 background (or cohort) models.

### 2.4.2. *Comparison with EigenFace*

To give an idea of the discriminatory power of our person recognition strategy, we relate it with a recognition technique based on facial appearance: eigenface. In our implementation of the eigenface approach, we firstly pre-process all images with a histogram equalisation, colour component by component, to reduce the mismatches due to illumination variations. Next, we represent the data set by using the NTSC colour space. Once the colour components are rearranged into vectors, we apply PCA to the enrolment subset to compute a reduced face space of dimension 243, and we calculate the feature vectors by whitening the projection coefficients in the eigenspace. Then, the client models are registered into the system using their centroid vectors, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved using a nearest neighbour classifier with cosine distances in (the whitened) face space.

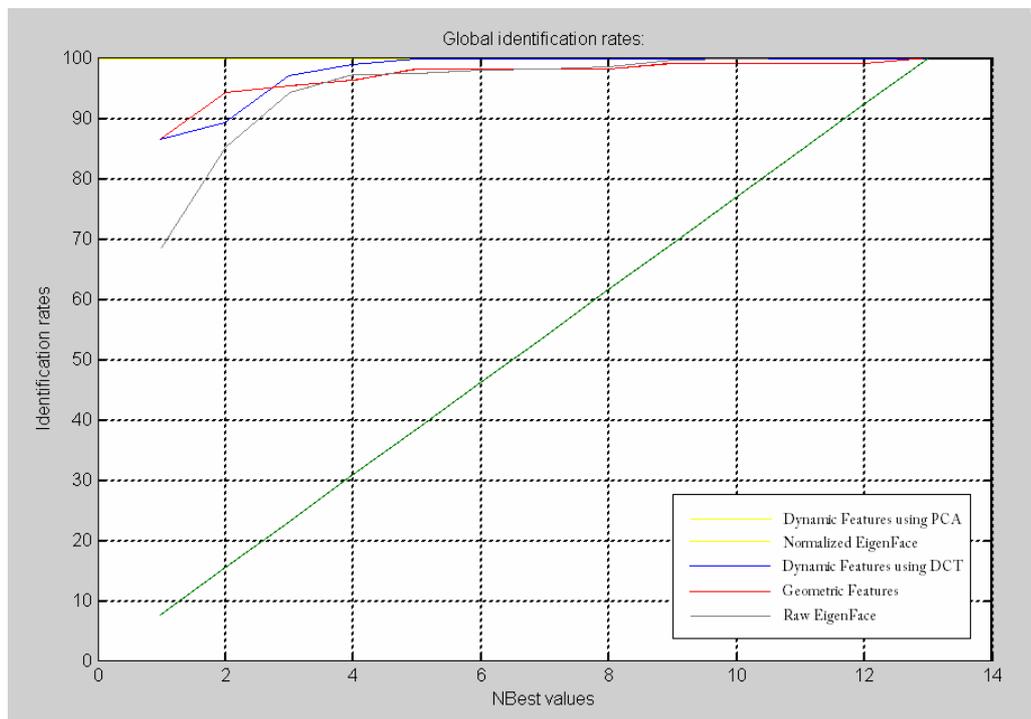
As we are aware of the sensitivity of eigenface approach to normalization, we derived two versions of the Italian TV Database, one subsampled and normalized and the other one raw. The videos are subsampled at a frame rate of 2 frames per second, and to normalise the video frames we firstly (in-plane) rotated the heads to horizontal eye position, then we cropped the face regions, and finally we aligned the images using the locations of the pupils.

### 2.4.3. *Identification and verification results*

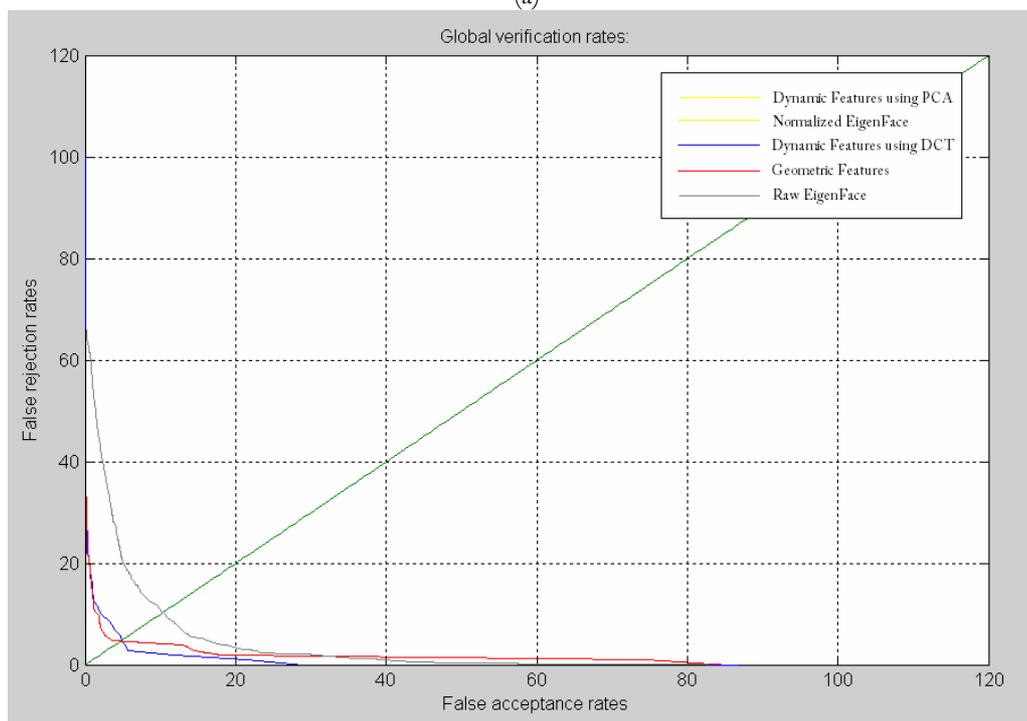
Figure 14 (a) shows the identification scores for the proposed features. The best results were obtained for Dynamic features using 10 PCA coefficients with an identification rate of 100 %. The Geometric features achieved an identification rate of 86.55% similar to Dynamic features using 10 DCT coefficients with an identification rate of 86.53 % for the best match.

Figure 14 (b) shows the Receiver Operating Characteristic (ROC) curve of our system, with False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR): the Equal Error Rate (EER) value is for geometric features is 4.69%, for Dynamic features using DCT coefficients is 4.88 % and for Dynamic features using PCA coefficients is 0 %.

The results for the normalized eigenface approach are excellent: 100.0% of CIR and 0.0% of EER, we consider this a too favourable and unrealistic situation. On the other hand for the raw version of the dataset without normalisation, the CIR decreases to 68.4% and the EER increases to 10.3%



(a)



(b)

Figure 14: Recognition Results a) Identification results b) Verification Results.

## 2.5. Conclusions

In this section we have presented a novel person recognition system based on behavioural lip features. We extracted behavioural features, which include static features, such as the normalized length of major/minor axis, coordinates of lip extrema points and dynamic features based on optical flow. Special care was taken not to include any physical attributes of the lip shape and appearance. These features were then modelled using a Gaussian Mixture Model (GMM) and finally the classification was done using a Bayesian decision rule. Experiments were carried out on a specialized database that proves that behavioural features can be used to recognize identities.

## 3. Lip features for HCI

### 3.1. Introduction

Over the past decades keyboard and mouse have been the prevalent interfaces for human computer communication, but they were designed with able-bodied individuals in mind. Unfortunately people who lack certain skills have been left out and tearing down this divide has been a challenging task. Initial efforts were made for people with fine motor disabilities by designing equipment that can substitute as pointing devices but they were specific to a certain disability and at times expensive. Next came speech based interfaces which enabled people to communicate by talking to the computer in a more natural way, but they have mostly been limited to text entry systems. A full fledged system that will allow a user with fine motor and speech disability to enter text and give commands is still far from reality.

In this section<sup>1</sup> we propose a human machine interface for the people with speech and fine motor impairment by using video input. Currently we have focused on two type of interfaces; first one for gesture recognition elaborated by a single choice question for which the user can respond by nodding of the head. The second interface for lip reading is illustrated by a multiple choice questions system where the user only articulates the lip motion of the digit of choice.

The novelty of our approach lies in proposing several image processing techniques that enable us to attain real-time (30f/s) detection of response. The yes/no detection is achieved by combining robustness of a holistic approach with the accuracy of a feature based technique. For digit recognition we have proposed a feature vector that is created by superimposing the outer lip contour of the video sequence. Using this single image as the feature vector reduces the computational cost of the classifier thus enabling us to attain results in real-time.

---

<sup>1</sup> This work was partially funded and completed during the eNTERFACE 07 workshop.

## 3.2. Head gesture recognition

Head gesture recognition systems aspire to have a better understanding of subliminal head movements that are used by humans to complement interactions and conversations. These systems vary considerably in their application from complex sign language interpretation to simple nodding of head in agreement. They also carry additional advantage for people with disabilities or young children with limited capabilities.

We focused on a simple yet fast and robust head gesture recognition system to detect the response of users to Yes/No type question. We did not wish to be limited by using specialized equipment thus we have focused our efforts in using a standard webcam for vision based head gesture recognition.

### 3.2.1. *State of Art*

Head gesture recognition methods combine various computer vision algorithms for feature extraction, segmentation, detection, tracking and classification, so categorizing them based on distinct modules would be overly complicated. We thus propose to divide the current head gesture recognition systems into the following categories.

*Holistic Approach:* This category of techniques focuses on the head as a single entity and develops algorithms to track and analyze the motion of head for gesture recognition. The positive point of these techniques is that as head detection is the main objective, they are quite robust at detecting it. The main disadvantage is the accuracy in detecting small amounts of motion.

Systems introduced in [160][161] have been based on color transforms to detect the facial skin color. In [162] the mobile contours have been first enhanced using pre-filtering and then transformed into log polar domain. [163] have build a mouse by tracking head pose using a multi-cues tracker combining, color, templates etc. in layers so if one fails the other layer can compensate for it.

*Local Feature Approach:* These algorithms detect and track local facial features such as eyes. The advantage is accuracy in motion estimation but the drawback is that local features are generally much difficult and computationally expensive to detect. [164] have proposed a “between eye” feature that is selected by a circle frequency filter. [165] have based there gesture recognition on an infra-red camera with LEDs placed under the monitor to detect accurately the location of the pupil.

*Hybrid Approach:* The aim of these algorithms is to combine holistic and local feature based techniques. Thus in reality trying to find a compromise between robustness of holistic approaches and accuracy of local feature based techniques, but most of them end up being computationally expensive as they combine various different levels of detection and tracking.

[166] have reported a head gesture based cursor system that detects a heads using a statistical model of the skin color. Then heuristics were used to detect nostrils and tracked to detect head gestures. In [167] they have combined previous work that has been done in face detection and recognition, head pose estimation and facial gesture recognition to develop a mouse controlled by facial actions.

### 3.2.2. *Proposed Method*

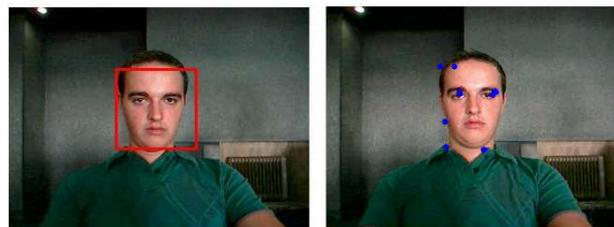
The method proposed builds upon previously developed algorithms that are well accepted. The specific requirements of our project dictate that the head gesture recognition algorithm should be robust to lighting and scale yet fast enough to maintain a frame rate of 30 f/s. On the other hand scenarios concerning occlusion and multiple heads in the scene have not been handled in the current implementation

*Face Detection:* The first module is the face detector, which is based on cascade of boosted classifiers proposed by [168]. Instead of working with direct pixel values this classifier works with a representation called “Integral Image”, created using Haar-like features. The advantage of which is that they can be computed at any scale or location in constant time. The learning algorithm is based on AdaBoost, which can efficiently select a small number of critical visual features from a larger set, thus increasing performance considerably.

The classifier has been trained with facial feature data provided along the Intel OpenCV library [169]. The face detection (cf. Figure 15) using the above classifier is robust to scale and illumination but has two disadvantages, first although it can be considered fast as compared to other face detection systems but still it attains an average performance of 15 f/s. Secondly it is not as accurate as local feature trackers. Thus head detection was only carried out in the first frame and results passed on to the next module for local feature selection and tracking.

*Feature Selection and Tracking:* The next step involves the selection of prominent features (cf. Figure 15) within the region of the image where the face has been detected. We have applied the Harris corner and edge detector [170] to find such points. The Harris operator is based on the local auto-correlation function.

Tracking of these feature points is achieved by Lucas Kanade technique [171]. It uses the spatial intensity gradient of the images to guide in search for matching location, thus requiring much less comparisons with respect to algorithms that use a predefined search pattern or exhaustive search.



a) b)  
Figure 15: a) Detected Face b) Tracking Points

*Yes/No Decision:* The final module analyzes the coordinate points provided by the tracking algorithm to take decision whether the gesture is a Yes or a No. First a centroid point is calculated from the tracked points, then the decision is taken based on the amount of horizontal or vertical motion of this centeroid. If the amount of vertical motion in the entire sequence is larger than the horizontal a yes decision is generated, similarly for No.

### 3.2.3. Experiments and Results

The development and testing was carried out on a basic 1.5 MHz laptop with 512 MB of RAM, without any specialized equipment. Video input of the frontal face was provided by a standard USB webcam with a resolution of 320X240 at 30 f/s.

Illumination and scale variability are the two main causes of errors in image processing algorithms, thus we have tried to replicate the scenarios most probable to occur in a real life situation. Although the amount of testing was limited to 10 people because of time concerns but due to the amount of variability introduced both in environmental conditions and subject characteristics (glasses/facial hair/sex), the tests are quite adequate.

*Illumination Variability:* As illumination variation is not dealt with explicitly in our algorithm, we defined three illumination scenarios (cf. Figure 16) to measure the effect of lighting change on our algorithms.



Figure 16: Light Variation

*Scale Variability:* The second important source of variability is scale, we have experimented with 3 different scale (cf. Figure 17) defined as the distance between the eyes in number of pixels. The 3 measures are S1: 15, S2: 20, S3: 30 pixels.



Figure 17: Scale Variation

*Test Questions:* The following five questions were selected due to the fact that they can be easily responded to by using head gestures of yes and no.

- Q1. Are the instructions clear?
- Q2. Are you male?
- Q3. Are you female?
- Q4. Are you a student of Computer Science?
- Q5. Do you like chocolate?

*Results:* The system was tested with 10 people who were asked 5 questions each by varying the scale and lighting, we achieved a correct recognition rate of 92 % for the Yes/No gesture. The system did have some problem with detecting the face at large distances when illuminated from the side or in direct sunlight.

### 3.3. Lip Reading

Lip-reading has been generally used to complement noisy audio signal for speech recognition but lately researchers have started to focus on using lip-reading for other applications like human computer interaction (HCI) and automatic indexing of television broadcasts.

Several situations could arise when we either do not have access to an audio capture device or the user is physically disabled, so for this work we have focused on providing a natural interface for a user to reply to a multiple choice question by using only the video signal. Currently the system can detect the motion of a users lip and recognize the first three digits of the decimal number system i.e. 1, 2, 3.

#### 3.3.1. *Proposed Method*

The algorithm proposed is a generalized lip reading system with restricted vocabulary. It proposes the use of a superimposed image of lip motion for the entire video sequence to be used as a feature vector for digit recognition. It gets the rough localization of the mouth region and then performs multiple classical image processing techniques to extract the outer lip contour and finally a support vector machine (SVM) is used to classify video as 1, 2, 3 digits from lip motion.

*Feature Vector:* Once the outer lip contour has been detected (as in Chapter IV.5) for one lip image this procedure is repeated for all the images of the video. The lip boundary for all images is then superimposed (cf. Figure 18) to obtain the final feature vector. The pixel values of this image are used directly as the feature vector for digit recognition.

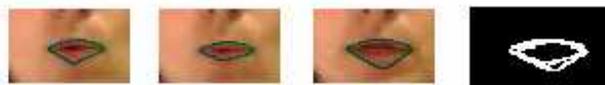


Figure 18: Lip detection sequence and superimposed image.

*Digit Recognition:* Classification of the feature vectors is performed by using a support vector machine (SVM); a supervised classification technique originally designed for a 2-class problem but now has been extended to multiple classes and can also perform regression. It first maps feature vectors into a higher-dimensional space using a kernel function, and then it builds an optimal linear discriminating function in this space. The solution is optimal because the margin between the separating hyperplanes and the nearest feature vectors is maximal.

### 3.3.2. Experiments and Results

The experiments were carried on a publicly available database [172]. Videos for 10 persons were randomly selected; each person had three repetitions of each digit (1, 2, 3). The dataset was then further divided into two sets; two third of the videos were used for training and the rest for testing.

Superimposed image of lip motion obtained from the image processing stage are now used as feature vectors for digit recognition. Classification is performed by using a 3 class SVM with RBF kernel, which can handle diverse type of data. The optimal parameters were selected using grid search method suggested by [173]. The following results were obtained when the parameters of SVM were tuned individually for each class and overall combined result.

	Class 1	Class 2	Class 3	Combined
Identification Rate	90 %	70 %	100 %	82 %

Table 7: Identification rate for digit recognition

### 3.4. Conclusions

In this section first we have introduced a real time and highly robust head gesture recognition system. It combines the robustness of a well accepted face detection algorithm with an accurate feature tracking algorithm to achieve a high level of speed, accuracy and robustness. Like all systems, our implementation does have its limitations. The first is that it cannot handle occlusion of the face and second is handling head gestures from multiple persons simultaneously in a given scene.

Secondly we have proposed an experimental digit recognition system with limited vocabulary. The novel approach in this system has been the use of a single image computed from a video sequence as a characteristic feature vector for recognition. This superimposed image greatly reduces the complexity of the feature vector and enables the use of a much simpler and efficient classifier.

## 4. Lip features for Gender

Human face contains a variety of information for adaptive social interactions amongst people. In fact, individuals are able to process a face in a variety of ways to categorize it by its identity, along with a number of other demographic characteristics, such as gender, ethnicity, and age. In particular, recognizing human gender is important since people respond differently according to gender. In addition, a successful gender classification approach can boost the performance of many other applications, including person recognition and smart human-computer interfaces.

In this section<sup>2</sup> we address the problem of automatic gender recognition by exploiting the physiological and behavioural aspects of the face at the same time. We propose a multimodal recognition approach that integrates the temporal and spatial information of the face through a probabilistic framework.

### 4.1. Related Works

There exists vast literature in social and cognitive psychology describing the impressive capabilities of humans at identifying familiar faces; though, most works deal with person recognition, and only few studies are focused on gender recognition. The automatic gender recognition from human faces has been studied since the late '80s [184], but only in the new millennium it has received significant attention from the scientific community: here we propose a short review of the latest approaches for gender recognition.

---

<sup>2</sup> This work was completed in collaboration with Dr. Federico Matta.

---

Sun et al. [185] applied principal component analysis (PCA) to represent each image as a feature vector in a low dimensional space; genetic algorithms (GA) were then employed to select a subset of features from the low dimensional representation that mostly encodes the gender information. Four different classifiers were compared in this study: the Bayesian decision making, a neural network (NN), support vector machines (SVM) and a classifier based on linear discriminant analysis (LDA). The SVM achieved the best performance in the comparative experiments. Gutta et al. [186] considered a hybrid classifier for gender determination of human faces that consisted of an ensemble of radial basis functions (RBFs) and decision trees (DTs).

Nakano et al. [187] focused on the edge information and exploited a neural network (NN) classifier for gender recognition. In particular, they computed the density histograms of the edge images, which were successively treated as input features for the NN. Lu et al. [188] exploited the range and intensity information of human faces for ethnicity and gender identification using a support vector machine (SVM). They firstly dealt with the ethnicity discrimination between Asian and non-Asian, where they exploited the relationship between the 3D shape of the human face and the ethnicity. The 3D scans were firstly registered by manually specifying six landmark points, and then they applied a grid to crop the face area and generate the feature vectors. As a similarity measure for classification they computed the posterior probabilities from the SVMs; in the end, they identified the gender of the user by comparing the probability to be a man with that of being a woman.

Moghaddam et al. [189] also proposed to classify gender from facial images (of 21x21 pixels) using support vector machines (SVMs). They tested the SVMs by implementing different kernels and they obtained the best experimental results with the Gaussian kernel, followed by the cubic polynomial kernel. Saatci and Town [190] have proposed a combined gender and expression recognition system using facial images. First the face was modelled using an Active Appearance Model (AAM), then features were extracted. Linear, polynomial and RBF based SVM were then used to classify gender and expression. Furthermore gender specific differences in appearance were exploited for expression recognition and vice versa.

Kim et al. [191] base their gender recognition system on a Gaussian Process Classifier (GPC). Facial images are first normalized to a standard dimensions and background and hair information was removed. Parameters for the GPC are learned using Expectation Maximization (EM) – Expectation Propagation (EP) algorithm. Finally GPC is used for classification. Zhiguang et al. [192] have focused on improving gender classification results by texture normalization. After scale normalization, affine fitting and Delaunay triangulation warping is employed to get a shape free texture. Lastly SVM, FLD and Adaboost are used for classification.

Rowly and Baluja [193] have proposed using adaboost using low resolution greyscale images. Several weak classifiers were build based on pixel value comparisons which lead to results just better than random chance. Then these are combined using adaboost to create a strong classifier. Tests were carried out on Feret database with an overall accuracy of 90 %. Caifeng et al. [194] have fused gait and face features for gender classification. Gait Energy Images were used for gait feature representation and normalized pixel values were used facial feature representation. These were then fused at feature level using canonical correlation analysis. Lastly support vector machines were used for classification.

## 4.2. Proposed Method

Our system is composed by two parallel complementary subsystems and a score integration step. The first recognition subsystem (identified by light yellow boxes in Figure 19) is exploiting the temporal video information and is based on unconstrained head and mouth motion. The second recognition subsystem (identified by tan boxes in Figure 19) works with the spatial information and exploits facial appearance; more precisely, it is a probabilistic extension of the original eigenface technique presented in [87] by Turk and Pentland. For a consistent integration of this heterogeneous biometric information (motion and appearance) into a unified recognition approach, both subsystems share the same probabilistic framework: a Gaussian mixture model (GMMs) approximation to represent the biometric features of each client, and Bayesian inference to calculate the similarity between tests and models. In the end, the similarity scores of the two parallel subsystems are combined in the last step (identified by a gold box in Figure 19), which operates the final identification and verification decisions after a suitable opinion fusion (or score fusion).

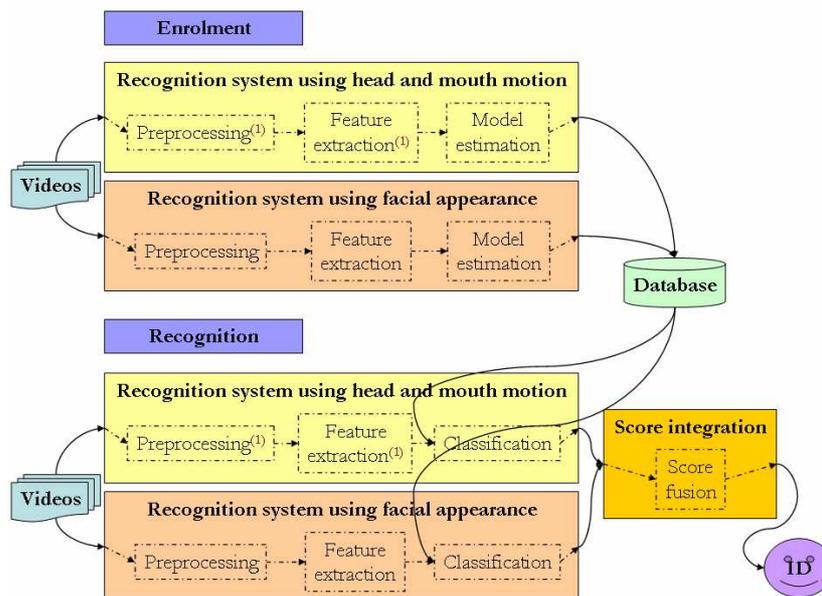


Figure 19. Architecture of the multimodal recognition system

The two subsystems and the integration step are detailed in the following three subsections.

#### 4.2.1. *Temporal recognition subsystem*

This subsystem exploits the unconstrained head and mouth motion information for gender recognition, by tracking a few facial landmarks in the image plane and by segmenting the outer lip contour in each video frame. Here we provide details of this temporal subsystem

*Pre-processing:* Each video sequence is firstly pre-processed by semi automatically detecting the face, and then by automatically tracking the facial landmarks over time using a template matching strategy in the RGB color space. Next, the outer lip contour is detected as in Chapter IV.5.

*Feature extraction:* The feature extraction step is divided in two phases: the geometrical normalization of the tracking signals, and the calculation of the feature vectors. The former part centres and scales the tracking signals; this way, after clearing the features from any dependence on absolute head location and size, the head motion information is isolated and the inter-video variation is reduced. Then the feature matrix,  $\mathbf{X} \in \mathcal{R}^{D \times N}$ , which retains the whole head and mouth discriminative information extracted from the corresponding video sequence, is generated by concatenating the following distinct features

- Head positions: the location of the head over time is included using the normalized tracking signals.
- Centred major axis of the outer lip contour.

- Centred minor axis of the outer lip contour.

*Model estimation:* The model estimation step approximates the class conditional probability density functions (PDFs) of each client, by using Gaussian mixture models (GMMs):

$$p(\mathbf{x} | \Theta) \equiv \sum_{c=1}^C \alpha_c \mathfrak{N}(\mathbf{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Where  $\mathfrak{N}(\mathbf{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$  is a singular multivariate normal distribution of a random variable,  $\mathbf{x} \in \mathfrak{R}^D$ ,  $\Theta = \{\alpha_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 1, \dots, C\}$  is the parameter list, and  $\alpha_c \in [0, 1]$  is the weight of the  $c$ -th Gaussian component. In addition, each  $\alpha_c$  corresponds to the a priori probability that an observation  $\mathbf{x}$  has been generated by the  $c$ -th normal source, and its value is normalized such as:  $\sum_{c=1}^C \alpha_c \equiv 1$ . Then, for each client  $k$ , his/her model parameters  $\Theta_k$  are obtained by solving a maximum likelihood problem through the expectation-maximization (EM) algorithm.

*Classification:* Finally, the classification step computes the similarity scores of the temporal subsystem by applying the probability theory and the Bayesian decision rule (also called Bayesian inference).

$$p(k | \mathbf{X}) = \frac{p(\mathbf{X} | k)p(k)}{M_{\mathbf{X}}} = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_N | k)p(k)}{M_{\mathbf{X}}}$$

The similarity score for the identification task is the video log-posterior probability:

$$S^{(ID)}(\mathbf{x}, \Theta_k) = \ln p(k | \mathbf{x})$$

while the similarity score for the verification mode is the video log-posterior probability ratio:

$$S^{(VER)}(\mathbf{x}, \Theta_k) = \ln \left[ \frac{p(k | \mathbf{x})}{p(\bar{k} | \mathbf{x})} \right]$$

where  $p(\bar{k} | \mathbf{x})$  is the posterior probability of the alternative hypothesis  $\bar{k}$ , and  $p(x_n | \bar{k})$  is the impostor model (the class conditional PDF for  $\bar{k}$ ).

### 4.2.2. Spatial recognition subsystem

For a consistent integration of the facial appearance information in our multimodal person recognition system, we developed a probabilistic extension of the original eigenface technique [87]. In particular, the pre-processing and feature extraction steps are kept pretty close to the standard eigenface approach, while the model estimation and classification steps are adapted to share the same probabilistic framework of the other recognition subsystem that exploits head and mouth motion.

*Pre-processing:* The pre-processing step applies image processing, the first transformation is a histogram equalization color component by color component, which is useful to reduce the impact of inter-image illumination and color variations. Then, this step converts the image signal into the most discriminative representation, which in our case is the NTSC (luminance, hue and saturation) color space. Finally, the image pixels are arranged in long vectors through a process called image vectorisation.

*Feature extraction:* The feature extraction step isolates the discriminative information that characterizes the individual and discards the irrelevant one, by applying a linear transformation from the high dimensional image space to a lower dimensional space (called the face space), which is much smaller. More precisely, each vectorised image  $\mathbf{s}_n$  is approximated with its projection in the face space  $\mathbf{v}_n \in \mathfrak{R}^D$  by the following linear transformation:

$$\mathbf{v}_n = \mathbf{W}^T (\mathbf{s}_n - \boldsymbol{\mu})$$

where  $\mathbf{W}$  is a projection matrix with orthonormal columns, and  $\boldsymbol{\mu} \in \mathfrak{R}^D$  is the mean image vector of the whole training set.

The optimal projection matrix  $\mathbf{W}$  is computed using the principal component analysis (PCA) (also called the Karhunen-Loeve transform (KLT)), which has the property of optimally representing the distribution of data in the root mean squares sense.

Once the image data set is projected into the face space, the vectors in the feature matrix,  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathfrak{R}^{D \times N}$ , are generated by computing the whitened projections in face space.

*Model estimation:* The model estimation step adopts the same probabilistic approach of the parallel subsystem using head and mouth motion for recognition. In fact, the distribution of the feature vectors of each client is modelled with a GMM, which approximates the class conditional probability density function of each user,  $k$ , in feature space:

$$p(\mathbf{x}_n | k) \cong p(\mathbf{x}_n | \Theta_k) \cong \sum_{c=1}^{C_k} \alpha_{k,c} \mathfrak{N}(\mathbf{x}_n | \boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c})$$

*Classification:* Finally, the classification step also closely resembles to the one in the temporal recognition system; in fact, it also computes the similarity scores by applying the probability theory and the Bayesian decision rule (also called Bayesian inference). In our implementation, we select only one key frame to test a given video sequence; hence, the related feature matrix contains only one feature vector,  $\mathbf{X}=\mathbf{x}\in\mathfrak{R}^D$ , and the video posterior probability is equal to the frame posterior probability.

$$p(k|\mathbf{x})=\frac{p(\mathbf{x}|k)p(k)}{p(\mathbf{x})}$$

As before, the similarity score for the identification task is the video log-posterior probability:

$$S^{(ID)}(\mathbf{x},\Theta_k)=\ln p(k|\mathbf{x})$$

while the similarity score for the verification mode is the video log-posterior probability ratio:

$$S^{(VER)}(\mathbf{x},\Theta_k)=\ln\left[\frac{p(k|\mathbf{x})}{p(\bar{k}|\mathbf{x})}\right]$$

#### 4.2.3. Score integration step

The score integration step combines the similarity scores of the two parallel subsystems by applying a suitable opinion fusion (or score fusion) strategy; after that, it takes the final identification and verification decisions using this extended measure of similarity. The score integration step calculates the multimodal similarity scores between the  $j$ -th test sequence  $\Phi_j$ , and the  $k$ -th client model by applying the weighted summation fusion (also called sum rule), which has the following general formula:

$$\xi_{j,k}^{(i)}\equiv S^{(i)}(\Phi_j,\Theta_k)=a_{j,k}\eta_{j,k}^{(i)}+b_{j,k}\rho_{j,k}^{(i)}$$

where  $a_{j,k}$ , and  $b_{j,k}$ , are the weighting values,  $\eta_{j,k}^{(i)}$  and  $\rho_{j,k}^{(i)}$  are the similarity scores of the temporal and spatial subsystems, and  $i$  specifies the identification or verification case. In particular, we selected the equal weighting of modalities, which is obtained by taking the average of the separate similarity scores, or equivalently by setting the weights as:

$$a_{j,k}=b_{j,k}=0.5$$

for  $\forall j, k$ . This choice has an interesting probabilistic interpretation; in fact, if we assume that the features related to facial motion  $\mathbf{X}_j^{(mn)}$  and those to facial appearance  $\mathbf{x}_j^{(app)}$  are statistically independent, then the multimodal similarity scores for the identification task are equal to the joint log-posterior probabilities of  $\mathbf{X}_j^{(mn)}$  and  $\mathbf{x}_j^{(app)}$ :

$$\xi_{j,k}^{(ID)} \equiv S^{(ID)}(\Phi_j, \Theta_k) = \frac{1}{2} \log p(k | \mathbf{X}_j^{(mn)}, \mathbf{x}_j^{(app)}) + \frac{1}{2} p(k)$$

except for an irrelevant translating factor, the a priori probability  $p(k)$ , which is not dependent on the test itself and it is already known before the recognition process.

### 4.3. Experiments and Results

#### 4.3.1. Database

Evaluation was carried out on the Italian TV Database, we split the whole database into an enrolment and recognition subset: 104 video sequences (8 for each of the 13 clients) are employed for the training of our system and the remaining 104 (8 for each of the 13 clients) are used for its testing.

Due to the well known high sensitivity of PCA-based recognition algorithms to facial alignment, variation in pose and scale, we derived a special version of the video database of Italian TV Database by sub sampling and manually normalizing some video frames. More precisely, for the enrolment subset we extracted 28 frames from each sequence, at a frame rate of 2 frames per second, whereas for the recognition subset we retrieved only the first key frame. After that, to normalize the video frames we firstly (in-plane) rotated the heads to horizontal eye position, then we cropped the face regions, and finally we aligned the images using the locations of the pupils.

### 4.3.2. *Experimental set-up*

*Temporal subsystem:* The head motion of each individual is represented through 8 tracking signals of 4 facial landmarks: the two eyes, the nose and the mouth. To improve the robustness of the tracking and reduce the intra-video variation, all frames are pre-processed using histogram equalization, color component by color component. During the head tracking step, the algorithm generates a template of 19 X 25 pixel for each landmark; then, the similarity scores of each color component are based on the city-block distance measure. After that, the feature extraction consists of centering the tracking signals, by applying a zero mean transformation, and using the normalized head positions and the centered major and minor axes of the outer lip contour as features for recognition; in the end, the dimensionality of the feature space is 10. Then, the client models are approximated using GMMs with 4 Gaussian components, and their parameters are estimated through the EM algorithm, which is initialized with: cluster means (computed using K-means), uniform weights and covariances. Finally, the impostor models for verification are approximated by taking the average of the best 2 background (or cohort) models.

*Spatial subsystem:* For the subsystem using facial appearance, all images are firstly pre-processed with histogram equalization, color component by color component, to reduce the mismatches due to illumination variations. Next, the data set is represented by using the NTSC color space, because it empirically provides more discriminative signals than the RGB does. Due to the well known problem of approximating high dimensional distributions with a limited amount of data, we are obliged to adopt serious restrictions on the dimension of the face space and the number of Gaussian components for a reliable GMM parameter estimation. In fact, with 228 images per person in the enrolment subset, we should use an eigenspace of dimension 10 or less for being able to reliably train 2 components, and 8 or less for 3, which is excessively constraining because too much discriminative information is lost with such a reduced space. Hence, the client models are estimated using a single Gaussian component, in a small face space of dimension 27, and the feature vectors are calculated by whitening the projection coefficients. Finally, the impostor models for verification are approximated by taking the average of the best 2 background (or cohort) models.

---

*Eigenface approach for comparison:* In our implementation of the original eigenface approach [87], we firstly pre-process all images with histogram equalization, color component by color component, to reduce the mismatches due to illumination variations. Next, we represent the data set by using the NTSC color space, because it empirically provides more discriminative signals than the RGB does. Once the color components are rearranged into large vectors, we apply the PCA to the enrolment subset to compute a reduced face space of dimension 243, and we calculate the feature vectors by whitening the projection coefficients in the eigenspace. Then, the client models are registered into the system using their centroid vectors, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved using a nearest neighbour classifier with cosine distances in (the whitened) face space.

### 4.3.3. Recognition results

The experimental results show that the integration of multiple sources of biometric information clearly improves the performance of the individual modalities. In fact, a biometric system exploiting only head motion obtains poor gender recognition scores with a CIR of 84.6% and an EER of 15.4%, we observe that the addition of mouth motion and then of facial appearance in our multi-biometric system increases the CIR to 96.2% and 99.0%, and decreases the EER to 3.8% and 1.0% respectively.

We also compared our gender recognition approach with the eigenface technique; our system performs in between the perfect recognition scores (100.0% of CIR and 0.0% of EER) of eigenfaces when applied to perfectly normalized images, which is an unrealistic and too favourable condition, and the poor results of eigenfaces when applied to not normalized frames: a CIR of 89.4% and an EER of 10.6%.

By looking at Figure 20 we can visually evaluate the benefit of integrating multiple sources of biometric information. These graphs show that, even if head motion (aqua curves, “HM”) is not such an important discriminative identifier for gender recognition applications, it can still achieve excellent results if supported by the mouth motion information (violet curves, “HM+MM”), and particularly by the facial appearance one (blue curves, “HM+MM+FA”).

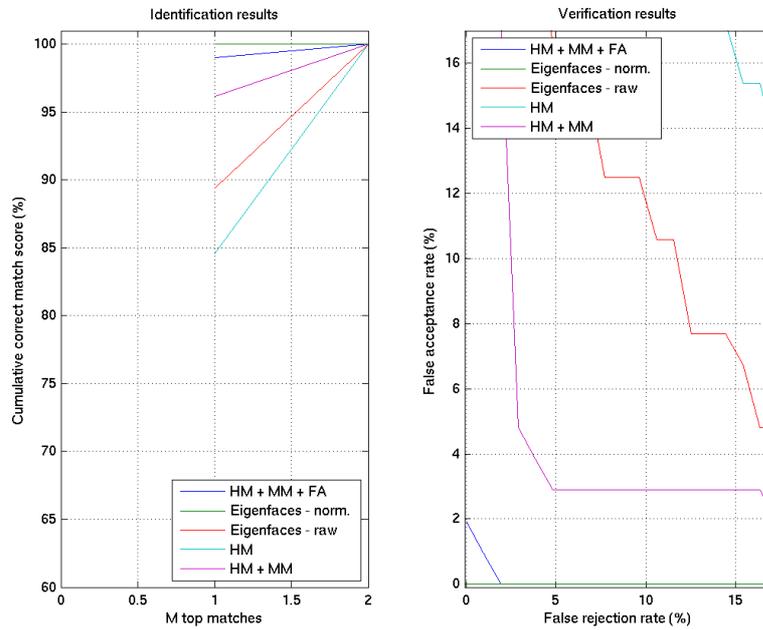


Figure 20. Gender recognition results.

In fact, these experiments are a clear demonstration of the advantage of multi-biometrics, in which complementary sources of biometric information can increase the accuracy and augment the reliability of the resulting multimodal system, by taking advantage of their redundant and richer information to compensate their individual weaknesses.

#### 4.4. Conclusion

In this section we presented a novel multimodal gender recognition system, which successfully integrates the head and mouth motion information with facial appearance by taking advantage of a unified probabilistic framework. By looking at the results of our experiments, we believe that not only facial appearance but also head and mouth motion possess a potentially relevant discriminatory power, and that the integration of different sources of biometric information from video sequences is the key strategy to develop more accurate and reliable recognition systems.



## *Chapter VI. Lip Feature Normalization*

---

### 1. Introduction

Video based face recognition has several advantages over image based techniques, the two main being, more data for pixel-based techniques, and availability of temporal information. But with these advantages there are some inconveniences also, the foremost being the augmentation of variation. In the classical image based face recognition degraded performance in face recognition has mostly been attributed to three main sources of variation in the human face, these being pose, illumination and expression. Among these, pose has been quite problematic both in its effects on the recognition results and the difficulty to compensate for it. Techniques that have been studied for handling pose in face recognition can be classified in 3 categories, first are the ones that estimates an explicit 3D model of the face [175] and then use the parameters of the model for pose compensation, second are subspace based such as eigenspace [174]. And the third type are those which build separate subspaces for each pose of the face such as view-based eigenspace [176].

Managing illumination variation in videos has been relatively less studied as compared to pose, mostly image based techniques are extended to video. The two classical image based techniques that have been extended for video with relative success are illumination cones [177] and 3D morphable models [175]. Lastly expression invariant face recognition techniques can be divided in two categories, first are based on subspace methods that model the facial deformations, such as Tsai et al. [178]. Next are techniques that use morphing techniques, like Ramachandran et al. [179], who morph a smiling into a neutral face.

In this chapter we have focused on another mode of variation that has been conveniently neglected by the research community caused by speech. The deformation caused by lip motion during speech can be considered a major cause of low recognition results, especially in videos that have been recorded in studio conditions where illumination and pose variations are minimal. In this chapter we present a novel method of handling this variation by using temporal normalization based on lip motion. The chapter is divided into two parts; the first is a synchronization method that studies lip motion and selects certain keyframes as synchronization frames. The evaluation is then carried out by comparing synchronized frames and randomly selected frames from the videos. The second part builds on the first part and uses the synchronized frame to normalize videos using lip morphing. The evaluation is carried out by comparing normalized videos with the original (non- normalized) videos.

## 2. Synchronization

In the first part we propose a temporal synchronization method that, given a group of videos for a person repeating the same phrase in all videos, studies the lip motion in one of the videos and selects synchronization frames based on a criterion of significance (optical flow). The next module then compares the motion of these synchronization frames with the rest of the videos and selects frames with similar motion as synchronization frames. For evaluation of our proposed method we use the classical eigenface algorithm to compare synchronization frames extracted from the videos and random frames to observe the improvement in a face recognition results.

The proposed synchronization method can be divided into two main parts; first is a selection method which selects frames in one of the video that are considered significant, second is a search algorithm in which the synchronization frames selected in the first video are synchronized with the remaining videos.

### 2.1. Synchronization Frame Selection

The aim of this module is to select synchronization frames from the first video of the group of videos for a specific person. Given a group of videos  $V_i$  for the person  $p$ , where  $i$  is the video index in the group, this module takes the first video  $V_1$  for each person as input and selects synchronization frames  $SF_1$ , that are considered useful for synchronization with the rest of the videos. The criterion for significance is based on amount of lip motion, hence frames that exhibit more lip motion as compared to the frames around them are considered significant. First for the video  $V_1$  the mouth region of interest (ROI)  $MI_1$  for each frame  $t$  is isolated based on tracking points provided with the database. Then frame by frame optical flow is calculated using the Lucas Kanake method (cf. Figure 21) for the entire video resulting in a matrix of horizontal and vertical motion vectors. As we are interested in a general description of the amount of lip motion in the frame we then calculate the mean of the motion vectors  $Of_t$  for each mouth ROI  $MI_t$ .

for  $t \leftarrow 1$  to  $T-1$

$$\begin{bmatrix} u_{1,1,t}, v_{1,1,t} & \cdot & \cdot & \cdot & u_{1,n,t}, v_{1,n,t} \\ \cdot & & & & \cdot \\ \cdot & & \cdot & \cdot & \cdot \\ \cdot & & & & \cdot \\ u_{m,1,t}, v_{m,1,t} & \cdot & \cdot & \cdot & u_{m,n,t}, v_{m,n,t} \end{bmatrix} = LK(MI_t, MI_{t+1})$$

$$Of_t = \sum_{m=1}^M \sum_{n=1}^N (abs(u_{m,n,t}) + abs(v_{m,n,t}))$$

end

Where  $T$  is the number of frames in the video  $V_i$ ,  $LK()$  calculates the Lucas Kanade optical flow.  $u_{m,n,t}$   $v_{m,n,t}$  are the horizontal and vertical components of the motion vectors at row  $m$  and column  $n$  of the frame  $t$ .

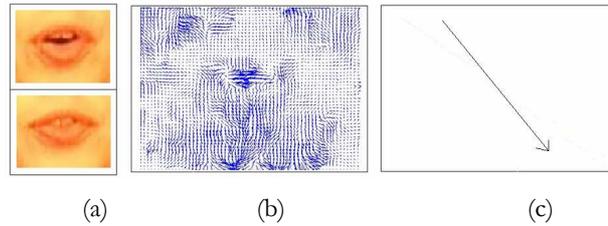


Figure 21: (a) Mouth ROI. (b) LK optical flow. (c) Mean vector.

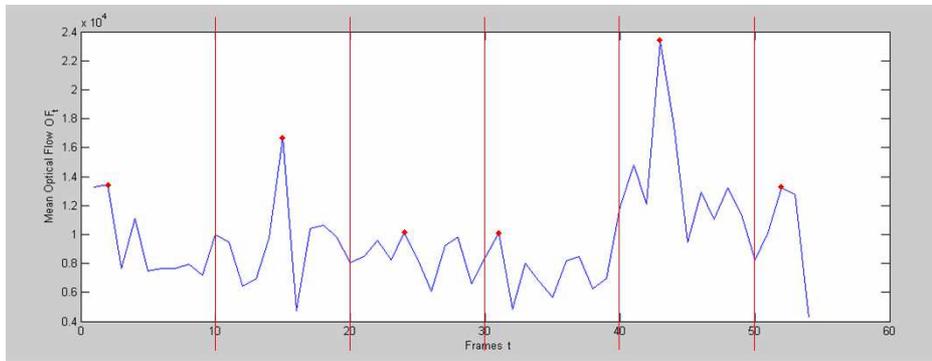


Figure 22: Mean Optical flow  $Of_i$  for Video

The next step is to select synchronization frames  $SF_t$  based on the mean optical flow  $Of_i$ , if we select frames that exhibit maximum lip motion there is a possibility that these frames might lie in close vicinity to each other. Thus we decided to divide the video into predefined segments and then select the frame with local maxima as synchronization frames.

for  $t \leftarrow 1$  to  $(T - D)$  with increments of  $D$   
 $[SF_1^1 \dots SF_1^K] = \text{Frame with value } (\max(Of_t \text{ to } Of_{t+D}))$   
 end  
 where  $D = \frac{T}{K}$

Where  $T$  is the total number of frames in the video.  $K$  is the number of synchronization frames, its value is predefined and is based on the average temporal length of the videos in the database and will be given in the experiments and results section.

## 2.2. Synchronization Frame Matching

In the previous module we have selected synchronization frames from the first video of a person and in this module we try to match these frames with the remaining videos in the group. This module can be broken down into several sub-modules, the first one is a feature extractor where we extracted two features related to lip motion. The second is an alignment algorithm that aligns the extracted lip features before matching, and the last sub-module is a search algorithm that matches the lip features using an adapted mean-square error algorithm. This results in the synchronization frame matrix  $SF_i$  for each person.

### 2.2.1. Feature Extraction

For this section we have studied the utility of two mouth features, the first one is quite simply the mouth ROI ( $MI_i$ ) as used in the previous module, the second is based on lip shape and appearance ( $LSA_i$ ) and its is based on the outer lip contour extracted in Chapter IV.5. Once the outer lip contour is detected the background is then removed and the final feature thus obtained as depicted in Figure 23. It contains the shape information in the form of lip contour and the appearance as pixel values inside the outer lip contour. Thus the feature image  $J$  may consist of either  $MI_i$  or  $LSA_i$ .



Figure 23: Lip Feature Image

### 2.2.2. Alignment.

Before the actual matching step, it is imperative that the feature images  $J$  ( $MI_i$ ,  $LSA_i$ ) are properly aligned, the reason being that some feature images maybe naturally aligned and thus have unfair advantage in matching. The alignment process is based on minimization of mean square error between feature images.

### 2.2.3. Synchronization Frame Matching.

The last module consists of a search algorithm, which tries to find frames having similar lip motion as synchronization frames selected from the first video in the rest of the videos. The algorithm is based on minimizing the mean square error, adapted for sequences of images.

Let  $J_{f(k),i,w}$  be the feature image, where  $k$  is the synchronization frame index,  $f(k)$  is the location of the synchronization frame in the video,  $i$  describes the video number and  $w$  the search window, which is fixed to  $\pm 5$  frames. Thus the search algorithm tries to find synchronization frames  $SF_i$  by matching the current feature image  $J_{f(k),1,0}$  previous feature image  $J_{f(k)-1,1,0}$  and the future feature image  $J_{f(k)+1,1,0}$  from the first video with the rest of the videos within a search window  $w$ . The search window  $w$  is created in the rest of the video centred at the location of the synchronization frame from the first video given by  $f(k)$ .

*for*  $k \leftarrow 1$  to *No of Synchronization Frames*

*for*  $i \leftarrow 2$  to *No of Videos Per Person*

*for*  $w \leftarrow f(k) - 5$  to  $f(k) + 5$

$$SF_i = \underset{j}{\operatorname{argmin}} \frac{\sum \sum (J_{f(k)-1,1,0})^2 - (J_{f(k)-1,i,w})^2 + \sum \sum (J_{f(k),1,0})^2 - (J_{f(k),i,w})^2 + \sum \sum (J_{f(k)+1,1,0})^2 - (J_{f(k)+1,i,w})^2}{(M * N)}$$

Where  $SF_i$  is the final matrix that contains the synchronization frames for all the videos  $V_i$  for one person.

### 2.3. Person Recognition

Classification was carried out using the eigenface technique [87]. The pre-processing step consists of histogram equalisation and image vectorisation (image pixels are arranged in long vectors).

We apply a linear transformation from the high dimensional image space, to a lower dimensional space (called the face space). More precisely, each vectorised image  $\mathbf{s}_n$  is approximated with its projection in the face space  $\mathbf{v}_n \in \mathfrak{R}^D$  by the following linear transformation:

$$\mathbf{v}_n = \mathbf{W}^T (\mathbf{s}_n - \boldsymbol{\mu})$$

where  $\mathbf{W}$  is a projection matrix with orthonormal columns, and  $\boldsymbol{\mu} \in \mathfrak{R}^D$  is the mean image vector of the whole training set:

$$\boldsymbol{\mu} = \frac{1}{JN} \sum_{j=1}^J \sum_{n=1}^N \mathbf{s}_{j,n}$$

in which  $J$  is the total number of sequences in the training set, and  $\mathbf{s}_{j,n}$  is the  $n$ -th vectorised image belonging to video  $\Phi_j$ . The optimal projection matrix  $\mathbf{W}$  is computed using the principal component analysis (PCA).

After the image data set is projected into the face space, the classification is carried out using a nearest neighbour classifier which compares unknown feature vectors with client models in feature space. The similarity measure adopted  $S$ , is inversely proportional to the cosine distance:

$$S(y_i, y_j) = 1 - \frac{y_i^T y_j}{\|y_i\| \|y_j\|}$$

and has the property to be bounded into the interval  $[0, 1]$ .

## 2.4. Experiments and Results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on Valid Database [65] which consists of five recording sessions of 106 subjects using the third utterance. The videos contain head and shoulder region of the subjects and the subjects are present in front of the camera from the beginning till the end.

The first video  $V_i$  was selected for the synchronization frame selection module and the rest of the 4 videos were then matched with the first video using the synchronization frame matching module. To estimate the improvement due to our synchronization process we have compared the synchronization frames  $SF_i$  and randomly selected frames using the person recognition module described before. The first video was excluded from training and testing due to its unrealistic recording conditions, 2<sup>nd</sup> and 3<sup>rd</sup> videos were used for training and 4<sup>th</sup> and 5<sup>th</sup> were used for testing both synchronization and random frames.

We apply PCA to the enrolment subset to compute a reduced face space of 243 dimensions. Then, the client models are registered into the system using their centroid vectors, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved using a nearest neighbour classifier with cosine distances.

We have created 8 datasets from our database by varying the parameters such as selection method, the type of feature image and the number of synchronization frames. The results are summarized in Table 8, the first column gives dataset number, the second column the method for selecting frames, the first 4 datasets use the proposed synchronization frame selection method and the last 4 datasets were created by selecting random frames from the videos. The third column signifies which lip features were used in the synchronization frame matching module. The fourth column is the number of synchronization frames  $K$  that were used for each video, in this study we have limited  $K$  to only 7 and 10 frames as most of the video in our database ranged from 60 to 110 frames. In case of last 4 datasets the number of synchronization frames simply signifies the number of random frames selected. The last column gives the identification rates.

Dataset	Method	Lip Feature	Number of Synchronization Frames	Identification Rates
1	Synchronization	MI	7	71.80 %
2	Synchronization	MI	10	74.18 %
3	Synchronization	LSA	7	72.28 %
4	Synchronization	LSA	10	74.02 %
5	Random	-	7	69.01 %
6	Random	-	10	69.92 %
7	Random	-	7	69.64 %
8	Random	-	10	68.85 %

Table 8: Person Recognition Results

The main result of this study is the overall improvement of identification results from synchronization frames as compared to random frames, which is evident from the Table 8. If we compare the identification results from the first 4 and last 4 datasets, it is obvious that there is an average improvement of around 4% between the 2 group of datasets. The second result that can be deduced is the improvement of recognition rates when more synchronization frames are used. The number of synchronization frames in the case of random frames simply signifies how many random frames were used and as it can be seen from the Table 8, using more random frames has no impact on the identification results. The third is insignificant change with regards to using *MI* or *LSA* as features. Here we would like to emphasize that the amount of testing for the second and third results is rather limited but this was not the main focus of this study.

## 2.5. Conclusions

In this section we have presented a temporal synchronization algorithm based on mouth motion for compensating variation caused by visual speech. From a group of videos we studied the lip motion in one of the videos and selected synchronization frames based on a criterion of significance. Next we compared the motion of these synchronization frames with the rest of the videos and selects frames with similar motion as synchronization frames. For evaluation of our proposed method we use the classical eigenface algorithm to compare synchronization frames and random frames extracted from the videos and observed an improvement of 4%.

### 3. Normalization

The second part of this chapter consists of a temporal normalization algorithm that takes the synchronization frames from the previous module and normalizes the length of the video by lip morphing. Firstly the videos are divided into segments defined by the location of the synchronization frames. Next the normalization is carried out independently for each segment of the video by first selecting an optimal number of frames for each segment and then adding and removing frames to normalize the length of the video. The evaluation is carried out by comparing normalized videos with the original videos in a person recognition scenario.

#### 3.1. Optimal Number of Frames.

Given the video  $V_i$ , it is first divided into segments  $S_q$ , where  $q$  is the number of segments and is equal to the number of synchronization frames plus one. Next the optimal number of frames  $O_q$  for each corresponding segment  $S_q$  is calculated by averaging the number of frames  $F_{i,q}$  in the corresponding segment of the videos  $V_i$ .

$$\begin{aligned} & \text{for } q \leftarrow 1 \text{ to } Q \\ & \text{for } i \leftarrow 1 \text{ to } I \\ & O_q = \frac{\sum_{i=1}^I F_{i,q}}{I} \end{aligned}$$

#### 3.2. Transcoding

The next step is to add/remove frames (commonly known as transcoding) from each segment of the video so as to make them equal to the optimal number of frames. The simplest techniques for transcoding like up/down-sampling and interpolation results in jerky and blurred videos respectively. Advanced technique such as motion compensated frame rate conversion [180], use block matching to estimate and compensate for motion but are imperfect as they lack information about the type of motion and thus frequently consider a uniform linear model of motion. As for this study we already have an estimation of lip motion from previous modules, we decided to use image morphing instead of block matching/compensation which results in visually superior results.

Morphing is the process of creating intermediate or missing frames from existing frames. Mesh morphing [181], one of the well studied techniques consists of creating a morphed frame  $I_m$  from source frame  $I_s$  and target frame  $I_t$  by selecting corresponding feature points in  $I_s$  and  $I_t$ , creating a mesh based on these feature points, warping  $I_s$  and  $I_t$  and finally interpolating warped frames to obtain the morphed frame  $I_m$ . In our study morphing was carried out only on the lip ROI as this region exhibits the most significant motion in the video. Lip ROI was first isolated and outer lip contour detected as in Chapter IV.5. These Lip ROI formed the  $I_s$  and  $I_t$  frames, feature points consisted of the 4 extremas of the outer lip contour (top, bottom, left, right). Mesh morphing was then carried out. Finally the morphed Lip ROI was superimposed on the original image to obtain the morphed frame.

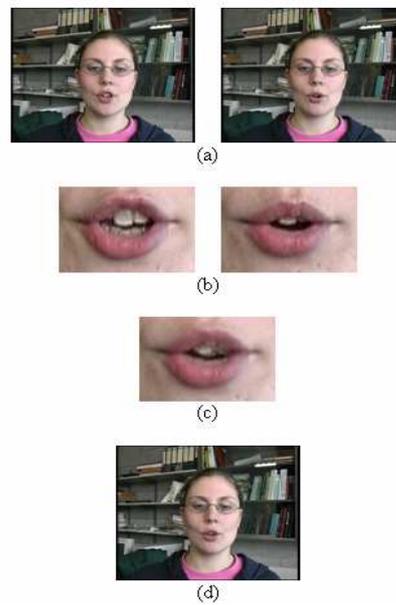


Figure 24:(a) Existing Frames (b) Lip ROI (c) Morphed Lip ROI (d) Morphed Frame

Decision regarding the number of frames to be added/removed is taken by comparing the number of frames in each segment  $S_q$  to the optimal number of frames; the frames are then added/removed at regularly spaced intervals of the segment. Addition of a frame consists of creating a morphed frame  $I_i$  from previously existing frames,  $I_{i-1}$  and  $I_{i+1}$ . Similarly frame  $I_i$  is removed by morphing frames  $I_{i-1}$  and  $I_i$  and replacing  $I_{i-1}$  with the morphed frame, and replacing frame  $I_{i+1}$  with the morphed frame from  $I_i$  and  $I_{i+1}$ . Finally deleting the frame  $I_i$ . Thus

*Frame Addition*

$$I_i \leftarrow \text{Morph}(I_{i-1}, I_{i+1})$$

*Frame Deletion*

$$I_{i-1} \leftarrow \text{Morph}(I_{i-1}, I_i)$$

$$I_{i+1} \leftarrow \text{Morph}(I_{i+1}, I_i)$$

$$\text{Delete}(I_i)$$

### 3.3. Person recognition

For testing our normalization algorithm we used a spatio-temporal method proposed by [174]. It consists of two modules: Feature Extraction, which transforms input videos into “X-ray images” and extracts low dimensional feature vectors, and Person Recognition, which generates user models for the client database (enrolment phase) and matches unknown feature vectors with stored models (recognition phase).

#### 3.3.1. Feature Extraction

Inspired by the application of discrete video tomography [182] for camera motion estimation, we compute the temporal X-ray transformation of a video sequence, to summarize the facial motion information of a person into a single X-ray image. It is important to notice that we restrict our framework to a fixed camera; hence, the video X-ray images represent the motion of the facial features and some appearance information, which is the information that we use to discriminate identities.



Figure 25: Original Frames and Temporal X-ray Image.

Given an input video of length  $T_i$ ,  $V_i \equiv \{I_{i,1}, \dots, I_{i,T_i}\}$ , the Feature Extractor module first calculates the edge image sequence  $E_i$ , obtained by applying the Canny edge-finding method [183] frame by frame:

$$E \equiv \{J_{i,1}, \dots, J_{i,T_i}\} = f_{EF}(V_i)$$

Then, the resulting binary frames,  $J_{i,t}$ , are temporally added up to generate the X-ray image of the sequence:

$$X_i = C \sum_{t=1}^{T_i} J_{i,t}$$

where  $C$  is a scaling factor to adjust the upper range value of the X-ray image.

After that, the Feature Extractor reduces the X-ray image space to a low dimensional feature space, by applying the principal component analysis (PCA) (also called the Karhunen-Loeve transform (KLT)): PCA computes a set of orthonormal vectors, which optimally represent the distribution of the training data in the root mean squares sense. In the end, the optimal projection matrix,  $\mathbf{P}$ , is obtained by retaining the eigenvectors corresponding to the  $M$  largest eigenvalues, and the X-ray image is approximated by its feature vector,  $y_i \in \mathfrak{R}^M$  calculated using the following linear projection:

$$y_i = \mathbf{P}^T (x_i - \mu)$$

where  $\mathbf{x}_i$  is the X-ray image in a vectorial form and  $\mu$  is the mean value.

### 3.3.2. Person Recognition

During the enrolment phase, the Person Recognizer module generates the client models and stores them into the system. These representative models of the users are the cluster centres in feature space that are obtained using the enrolment data set.

For the recognition phase, the system implements a nearest neighbour classifier which compares unknown feature vectors with client models in feature space. The similarity measure adopted  $\mathcal{S}$ , is inversely proportional to the cosine distance:

$$\mathcal{S}(y_i, y_j) = 1 - \frac{y_i^T y_j}{\|y_i\| \|y_j\|}$$

and has the property to be bounded into the interval  $[0, 1]$ .

### 3.4. Experiments and results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on Valid Database [65] which consists of five recording sessions of 106 subjects using the third utterance. The first video was selected for the synchronization frame selection module and the rest of the 4 videos were then synchronized with the first video using the synchronization frame matching module. Finally all videos were temporally normalized.

To estimate the improvement due to our normalization process we have compared the normalized videos generated by our algorithm to original non-normalized videos using the person recognition module described above. First 3 videos were used for training and the rest 2 were used for testing. The number of synchronization frames in this study have been set to 7, as the average number of frames per video in our database was approximately 70. The recognition system has been tested using a feature space of size 190, constructed with the enrolment data set. The video frames are also pre-processed using histogram equalization, in order to reduce the illumination variations between different sequences.

Method	CIR % (1 <sup>st</sup> )	CIR % (5 <sup>th</sup> )	CIR % (10 <sup>th</sup> )	EER %
Normalized Video	69.02 %	82.60 %	89.13 %	10.1 %
Original Video	65.21 %	81.52 %	85.86 %	11.9 %

Table 9: Person Recognition Results

The identification and verification results are summarized in Table 9; its columns report the correct identification rates (CIR), computed using the best, 5-best and 10-best matches, and the equal error rates (EER) for the verification mode. We notice that the recognition system using normalized videos performs better than the analogous one working with non-normalized videos.

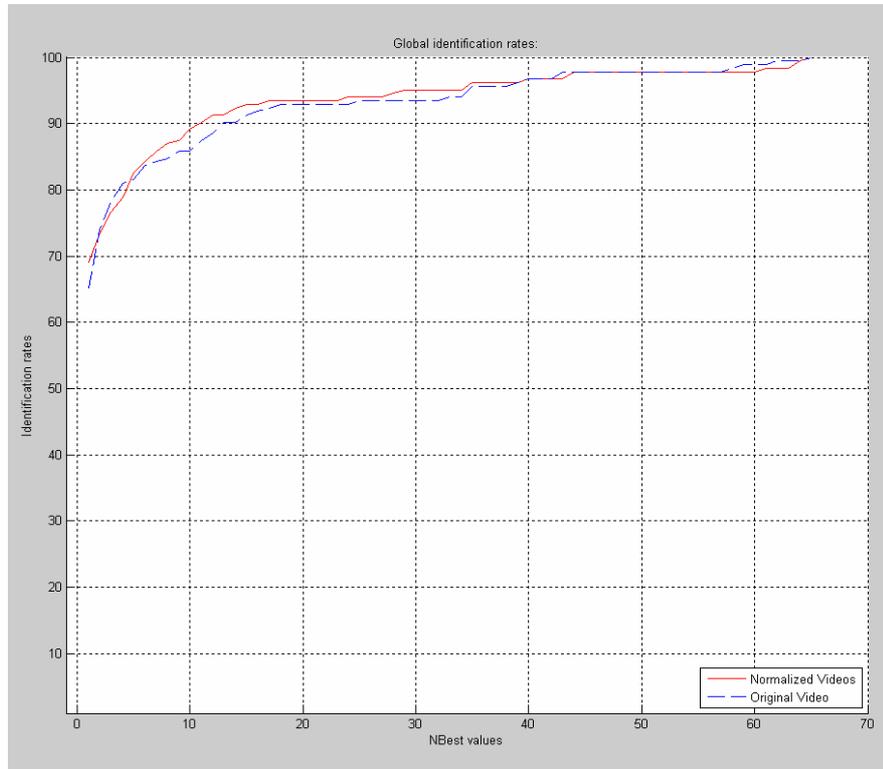


Figure 26 : Correct Identification Rates (CIR)

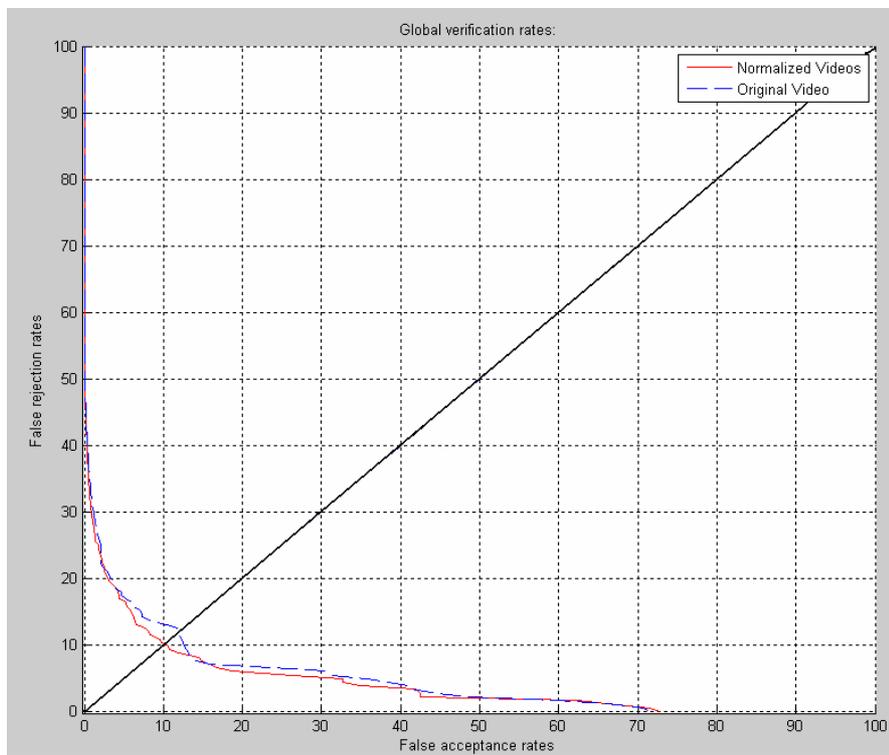


Figure 27 : Verification Rates (EER)

### 3.5. Conclusions

In this section we have presented a temporal normalization algorithm based on mouth motion for compensating variation caused by visual speech. Using the synchronization frames from the previous module we normalized the length of the video. Firstly the videos were divided into segments defined by the location of the synchronization frames. Next normalization was carried out independently for each segment of the video by first selecting an optimal number of frames and then adding/removing frames to normalize the length of the video. The evaluation was carried out by using a spatio-temporal person recognition algorithm to compare our normalized videos with non-normalized original videos, an improvement of around 4% was observed.



## *Chapter VII. Conclusions*

---

### 1. Concluding Summary

In this thesis we have presented a detailed work on various aspects of using lip features as a biometric identifier. In Chapter II we provide an introduction to the fundamental notions of biometrics and in Chapter III we reviewed the literature on audio-visual person recognition using lip information. We presented the various steps involved in audio-visual person recognition from pre-processing to classification, and several examples of existing systems. We concluded that majority of the systems use visual lip features as an enhancement of the audio based systems. The systems based only on visual lip features tend to focus on the physical aspect and are generally tested on text dependent databases containing short duration sentences.

In Chapter IV we present a novel lip detection algorithm based on fusion of well established image processing techniques, with an extensive evaluation of the detection process using a database that represented real world conditions. The numerical results presented were deemed satisfactory and this fact was confirmed on visual inspection of the detected contours. An evaluation of various lip features for person recognition was also carried out, we extracted various geometric and appearance based lip features and compared them using three feature selection measures. We observed that ICA, which is a supervised method performed the best, while the two worst performing techniques were the geometric features and optical flow. Probably the main reason behind this poor performance is that the geometric features contain much less information as compared to other physical features such as DFT. On the other hand the poor performance of optical flow could be based on the fact that it is a purely behavioral feature and lacks the discriminant power of lip shape and appearance.

In Chapter V we present a person recognition system based on behavioural lip features. We took special care not to include any physical traits of the lip motion during speech. We modelled the static and dynamic features extracted from lip contour using a Gaussian Mixture Model (GMM) and the classification was done using a bayesian decision rule. Although the two features studied i.e. normalised geometric features and optical flow based dynamic features were the worst performing features in the feature selection evaluation, they performed quite well on a limited database. These initial results validate that a behavioural speech pattern can be learned from long duration video and used to verify identities.

In Chapter VI we focused on another aspect of visual speech i.e. as a source of variation in face recognition systems. We proposed a temporal normalization method for visual speech and investigate its effects on person recognition. We first synchronized the videos and compared these synchronized frames to randomly selected frames using the eigenface approach. We observed a marked improvement which convinced us to further evaluate this topic. We then normalized the frames using lip morphing and compared the normalized videos obtained from our method to non-normalized original videos using a spatio-temporal face recognition technique. Although we only achieve 4%, which is almost the same as we obtained using only synchronization, we would like to point out that the person recognition modules were completely different.

In Chapter VIII we present some initial results of our work on global face and local eye motion for person recognition.

## 2. Future Works

We have presented lip analysis techniques and their consequences on person recognition. Although the results are quite promising, there are some limitations which warrant further work, we will discuss them in this section.

In Chapter IV we presented a lip detection method based on fusion of two independent methods, there is still some room for improvements. Currently we compensate for errors by fusion, we would like to automatically evaluate the results from the independent methods and detect failure, then propose an appropriate fusion approach. Also we only tested two fusion approaches; it would be interesting to study others also.

We then evaluated the lip features for their relevance to person recognition. Further improvements to the proposed study could be in the form of more feature extraction techniques, even though we have tried to include all state of the art techniques in our study but still there exist several techniques that differ in a minor way from the generalized techniques presented here. The second improvement that we envisage is using a wrapper type feature selection.

In Chapter V we presented a person recognition method based on behavioural lip features. Several improvements can be made to our system, the most essential would be to test our methods on a much larger database but unfortunately such a database is not currently available. Another aspect that we would like to explore is to test our methods on a natural conversational database, as the current database consists of TV news reporters.

A major improvement could be to focus on the analysis of various other behavioural aspects of the human face such as blinking of eyes or motion of the pupils. This approach may show more important discriminating power, capturing the details of personal movement. In our person recognition approach we have assumed the independence of features, which is not actually true, thus it would be interesting to study methods that exploit this interdependence such as bayssienne networks. Another possibility is to use our biometric system, based on mouth motion, and integrate it in a multimodal one; for this purpose it could be possible to couple it with a physical modality such as appearance.

We then presented an application of lip features for HCI. In the future we would like to develop a real time head gesture and lip reading capable of understanding several complex head gestures with an enhanced vocabulary for lip reading. Other important contribution could be enhancing the interface by gaze estimation and personalizing modalities for specific user.

Next we described a multimodal gender recognition system. There are different ways to improve our present multimodal approach. First of all, the temporal subsystem can be improved by increasing the accuracy of the tracking signals, for example by implementing a more robust tracker. Afterwards, the static subsystem can apply a more discriminating space reduction technique, like LDA, CCA, etc., or a more performing strategy if compatible with our probabilistic framework. Finally, more elaborate fusion techniques, like a post-classifier, might improve the integration of the discriminating information of our system, but the amount of required data could be a problem.

In Chapter VI we elaborated on a temporal normalization method using lip motion. One important item of our normalization method are the number of synchronization frames used, as it can be observed from Table 8 that the number of synchronization frames has an impact on results and it requires further evaluation. Currently for the first part which is based on synchronization frame for person recognition uses an image based classifier i.e. eigenface, whereas the second part based on normalized videos for person recognition using a spatio-temporal approach. We would like to employ a unified classifier so that the results from the two methods can be compared meaningfully.

The database used in these experiments consisted of short sentences; it would be interesting to see results of normalization on other databases. Another specificity of the database used was that although it did not contain any pose variation, strong illumination variation was present, which has negatively affected recognition results. Further improvements to the proposed work could be inclusion of spatial normalization, such as pose and illumination.

### 3. Publications

1. Usman Saeed and Jean-Luc Dugelay  
Temporal normalization of videos using visual speech  
MiFor'09 : 1st ACM Workshop on Multimedia in Forensics, October 19-24, 2009, Beijing, China .
2. Abdelaali Benaiss, Usman Saeed, Jean-Luc Dugelay and Mohamed Jedra  
Impostor detection using facial stereoscopic images  
Eusipco 2009, 17th European Signal Processing Conference, August 24-28, 2009, Glasgow, Scotland
3. Federico Matta, Usman Saeed, Caroline Mallauran and Jean-Luc Dugelay  
Facial gender recognition using multiple sources of visual information  
MMSP 2008, 10th IEEE International Workshop on MultiMedia Signal Processing, October 8-10, 2008, Cairns, Queensland, Australia.
4. Usman Saeed and Jean-Luc Dugelay  
Facial video based response registration system  
Eusipco 2008, 16th European Signal Processing Conference, August 25-29, 2008, Lausanne, Switzerland
5. Gopal Ananthkrishnan, Hamdi Dibeklioglu, Martin Lojka, Adolfo Lopez, Serafeim Perdakis, Usman Saeed, Albert Ali Salah, Dimitrios Tzovaras and Athanasios Vogianou  
Activity-related biometric authentication  
eNTERFACE 2008, 4th Summer Workshop on Multimodal Interfaces, August 4th-29th, 2008, Paris, France
6. Jerome Allasia, Ana Cristina Andres del Valle, Dragos Catalin Barbu, Ionut Petre, Usman Saeed and Jerome Urbain  
Multimodal services for remote communications  
eNTERFACE 2007, 3rd SIMILAR Workshop on Multimodal Interfaces, July 16-August 10th, 2007, Istanbul, Turkey
7. Usman Saeed and Jean-Luc Dugelay  
Person recognition from video using facial mimics  
ICASSP 2007, 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing, April 15 - 20, 2007, Honolulu, USA
8. Usman Saeed, Federico Matta and Jean-Luc Dugelay  
Person recognition based on head and mouth dynamics  
MMSP 2006, IEEE International Workshop on Multimedia Signal Processing, October 3-6, 2006, Victoria, Canada



## Chapter VIII. Appendices

### Face and Eye Features for Person Recognition

In this chapter, we present some initial results for using global face dynamics and local eye motion for person recognition. Facial movements are analyzed by calculating the degree of face symmetry and angle of the face in each video frame. Eye motion is composed of eyelid and pupil motion in the video. Statistical features are then computed from these signals, and used for discriminating identities. The classification task is done using a Gaussian Mixture Model (GMM) approximation and Bayesian classifier.

#### 1. Facial feature extraction

This module is responsible for extraction of parameters relating to the rigid face motion. The parameters extracted are as under.

##### 1.1. Face Angle

Face angle is defined as the angle the face makes with the horizontal axis of the image plane. Using the tracking points provided along the database, the slope between the nose point  $P_n(x_n, y_n)$  and the mouth point  $P_m(x_m, y_m)$  is calculated as:

$$slope = \frac{(y_m - y_n)}{x_m - x_n}$$



Figure 28: Facial feature points with face angle.

The angle of the face with the horizontal image axis is calculated by taking the inverse tangent of the slope.

$$Angle = \tan^{-1}(slope)$$

## 1.2. Face Symmetry

The second parameter extracted is the face symmetry, defined as the degree of difference between the left and right sides of the face. First we remove the background using an approximation of an ellipse based on the nose point as its center.

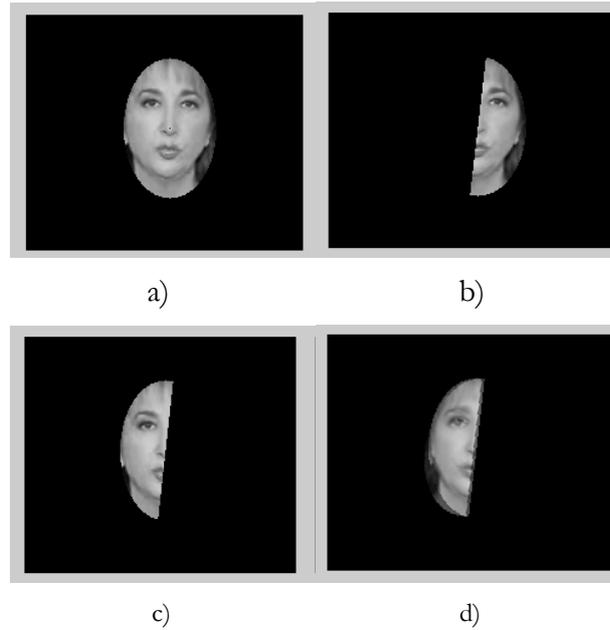


Figure 29: a) Background removed facial image b) Right facial Image c) Left facial Image d) Left-Right Overlaid Image.

Using the nose point  $P_n(x_n, y_n)$ , mouth point  $P_m(x_m, y_m)$  and eye points  $P_{e1}(x_{e1}, y_{e1})$ ,  $P_{e2}(x_{e2}, y_{e2})$  the major and minor axis of the ellipse are defined as:

$$MajorAxis = \left( y_m + \frac{(y_m - y_n)}{2} \right) - \left( y_{e1} + \frac{(y_m - y_n)}{2} \right)$$

$$MajorAxis = \left( x_{e2} + \frac{(x_{e2} - x_{e1})}{8} \right) - \left( x_{e1} + \frac{(x_{e2} - x_{e1})}{8} \right)$$

All image pixels outside the ellipse as defined above are set to zero and then from the slope, the equation of line is calculated as:

$$y - y_m = slope(x - x_m)$$

This line is used as the boundary of the left and right side of the face. Depending on the fact the slope is positive or negative, one side of the face is selected. The selected side is flipped and the aligned with other side. The alignment consists of rigid rotation and translation. Once both sides are aligned the normalized MSE between the two corresponding sides is calculated.

$$MSE = \frac{1}{NM} \sum_{x=0}^N \sum_{y=0}^M (Right(x, y) - Left(x, y))^2$$

## 2. Eye Dynamics

This module extracts parameters related to the motion of the eye which includes estimation of pupil motion and blinking. The blinking algorithm is based on a combination of image subtraction and optical flow calculation. First based on the tracking points a region of interest is selected around each of the eye. Using the assumption that global change is minimal between two consecutive ROIs, consecutive frames are subtracted to detect change. A threshold equal to twice the mean error value in the entire sequence is fixed to detect significant change, but this change can also be due to other phenomena such as global change or error in tracking, thus optical flow was also calculated using the Lucas Kanade method [171].

A mean motion vector was then calculated to have an estimate of the overall motion in the ROI. Finally the entire sequence was searched for downward, stationary and upward motion, stationary stage being when the eye is closed during a blink. The sequences thus selected are considered as blinks.

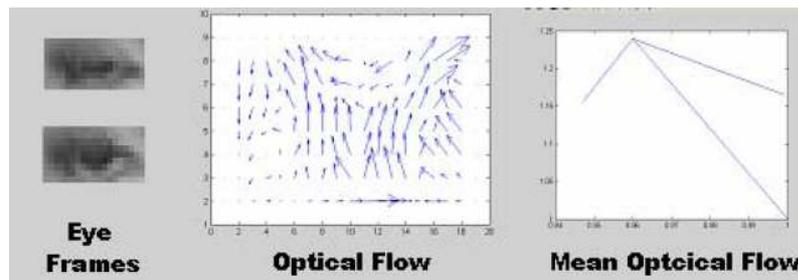


Figure 30: Optical flow of eye motion.

Then we estimate the location of the pupil. The main principle of the algorithm is that the pupil is the darkest region in the eye ROI. The algorithm is defined as:

- Detect the global minima in the eye ROI.
- Select points with value close to global minima as possible pupil candidates.

- Convolve candidates with a circular mask of 5 pixels.
- Select pixel with max response as pupil center.

Circular mask of 5 pixels corresponds to the average size of pupil in our database.

### 3. Person recognizer module

The recognition module is similar to the one described in Chapter V.2.3, it exploits the individual feature vectors extracted from video sequences for classification purposes. The local and global features are firstly concatenated in various combinational vectors, which are subsequently used for training a Gaussian Mixture Model (GMM) for each person in the database. Classification is then performed by calculating the class conditional probability density functions in a Bayesian classifier.

### 4. Experimental Results And Discussions

Test were carried out on the Italian TV Database, we selected 104 video sequences for training (8 for each of the 13 individuals), and the remaining 104 were left for testing. Several combination of the feature vectors were tested to ascertain which features played a dominating role in recognition. The Table 10 describes the combination of vectors used and the best identification results obtain and the number of GMM components used to model the features.

Feature Vectors	Identification rates %	GMM components
Face Angle	23.0	2
Face Symmetry	36.5	4
Combined Face	34.5	3
Pupil Location	60.5	1
Blinking	38.5	3
Combined Eye	59.5	1
All Combined	65.5	2

Table 10: Results for feature vectors in identification rates.

As it is evident from the table, the best performance (65.5% identification) is obtained by using all the feature vectors. Using face angle exhibited the worst results; a possible explanation could be lack of data as each person mostly moved his head by a few degrees. On the other hand the pupil location provides quite good results i.e. 60.5% identification rate.

## 5. Conclusions And Future Works

In this chapter we have presented a nascent study that explores the possibility of using facial dynamics for person recognition. It is based on global face and local eye motion. The preliminary results provide insight into the comparative potential of each feature, but the environment and nature of application must be kept in mind. In this study the highest identification rate was achieved when using the pupil location. Several improvements can be made to our system, one major improvement could be to refine the signal extraction process e.g. using snakes or deformable templates could be used for local features.



## References

- [1] F. Matta, "Video person recognition strategies using head motion and facial appearance," University of Nice Sophia-Antipolis, 2008.
- [2] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi and M. Bone, "Facial recognition vendor test 2002: evaluation report," <http://www.frvt.org/FRVT2002/>, March 2003.
- [3] J. G. Wilpon, L. R. Rabiner, and T. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *Tech. J. Bell Labs*, vol. 63, no. 3, pp. 479–498, 1984.
- [4] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Tech. J. Bell Labs*, vol. 54, no. 2, pp. 297–315, 1975.
- [5] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 432–435, Hong Kong, 2003.
- [6] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.
- [7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [8] J. Ajmera, "Robust audio segmentation," Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2004.
- [9] J. A. Haigh, "Voice activity detection for conversational analysis," M.Phil. thesis, Univ. of Wales, Swansea, U.K., 1994.
- [10] J. M. Naik and D. M. Lubensky, "A hybrid HMM-MLP speaker verification algorithm for telephone speech," In *Proc. IEEE ICASSP*, vol. 1, pp. 153–156, 1994.
- [11] R.W. Schafer and L. R. Rabiner, "Digital representations of speech signals" *Proc. of IEEE*, vol. 63, no. 4, pp. 662–677, 1975.
- [12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.
- [13] F. Itakura, "Line spectrum representation of linear predictive coefficients," *Trans. Committee Speech Research, Acoustical Soc.*, vol. S75, p. 34, Japan, 1975.
- [14] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.
- [15] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.
- [16] S. Furui, "An overview of speaker recognition technology," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, pp.1–9, 1994.

- 
- [17] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition," *AT&T Technical Journal*, Vol. 66, No. 2, pp. 14-26, 1987.
- [18] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, no. 1, pp. 43-49, 1978.
- [19] A. Higgins, "YOHO Speaker Verification," *Speech Research Symposium*, Baltimore, 1990.
- [20] A. Higgins, L. Bahler, and J. Porter, "Voice Identification Using Nearest Neighbor Distance Measure," In *International Conference on Acoustics, Speech, and Signal Processing in Minneapolis*, pp. 75-378, 1993.
- [21] J. Oglesby, "Neural models for speaker recognition," Ph.D. dissertation, Univ. College of Swansea, Swansea, U.K., 1991.
- [22] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech Audio Processing*, pt. II, vol. 2, no. 1, pp. 194-205, 1994.
- [23] M. M. Homayounpour and G. Chollet, "Neural net approaches to speaker verification: Comparison with second-order statistic measures," *Proc. IEEE ICASSP*, vol. 1, pp. 353-356, 1995.
- [24] Y. Bennani, F. F. Soulie, and P. Gallinari, "A connectionist approach for automatic speaker identification," in *Proc. IEEE ICASSP*, vol. 1, pp. 265-268, 1990.
- [25] Y. Bennani, "Probabilistic cooperation of connectionist expert modules: Validation on a speaker identification task," in *Proc. IEEE ICASSP*, vol.1, pp. 541-544, 1993.
- [26] L. Rudasi and S. A. Zahorian, "Text-independent talker identification using neural networks," *J. Acoust. Soc. Amer.*, pt. Suppl. 1, vol. 87, no. S104, 1990.
- [27] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing*, vol. 11, no. 4, pp. 18-32, 1994.
- [28] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.
- [29] S. Dupont and J. Luetttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141-151, Sept. 2000.
- [30] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.
- [31] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," *Signal Processing*, editor A. Oppenheim. Englewood Cliffs: Prentice-Hall, 1993.
- [32] D. L. Hall, "Mathematical Techniques in Multisensor Data Fusion," Norwood, MA: Artech House, 1992.
- [33] D.L. Hall, J. Llinas, "Multisensor data fusion," *Handbook of Multisensor Data Fusion*, D. L. Hall and J.Llinas (Eds.), CRC Press, pp. 1-10, USA, 2001.
- [34] S.S. Iyengar, L. Prasad, H. Min, "Advances in Distributed Sensor Technology," Prentice Hall PTR, New Jersey, 1995.
- [35] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. IEEE ICASSP*, vol. 2, pp. 833-836, 1996.
- [36] C. C. Chibelushi, J. S. D. Mason, and F. Deravi, "Feature-level data fusion for bimodal person recognition," *Proc. 6th IEE Int. Conf. Image Processing and its Applications*, pp. 399-403, 1997.

- [37] S. Bengio, "Multimodal authentication using asynchronous HMMs," Proc. 4th International Conf. Audio- and Video-based Biometric Person Authentication, Guildford, pp. 770-777, 2003.
- [38] K. S. Lawrence and I. J. Michael, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones," Mach. Learn., vol. 37, no. 1, pp. 75-87, 1999.
- [39] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in Proc. IEEE CCVPR, pp. 994-999, 1997.
- [40] V. Radova, J. Psutka, "An approach to speaker identification using multiple classifiers," Proc. IEEE Conf. Acoustics, Speech and Signal Processing, vol. 2, pp. 1135-1138, 1997.
- [41] R.C. Luo, M.G. Kay, "Introduction", Multisensor Integration and Fusion for Intelligent Machines and Systems, R.C. Luo and M.G. Kay (eds.) Ablex Publishing Corporation, pp. 1-26, Norwood, NJ, 1995.
- [42] L. A. Alexandre, A. C. Campilho, M. Kamel, "On combining classifiers using sum and product rules," Pattern Recognition Letters vol. 22, pp. 1283-1289, 2001.
- [43] V. Chatzis, A. G. Bors, and I. Pitas, "Multimodal decision-level fusion for person authentication," IEEE Trans. Syst., Man, Cybern. A, vol. 29, pp. 674-680, 1999.
- [44] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," IEEE Trans. Neural Networks, vol. 10, pp. 1065-1074, 1999.
- [45] L.L. Mok, W.H. Lau, S.H. Leung, S.L. Wang, H. Yan, "Person authentication using ASM based lip shape and intensity information," International Conference on Image Processing, vol.1, no., pp. 561-564, 2004.
- [46] T. Wark, S. Sridharan, V. Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," Proceedings of Fourteenth International Conference on Pattern Recognition, vol.1, pp.123-125, Aug 1998.
- [47] A.G. de la Cuesta, Z. Jianguo, P. Miller, "Biometric Identification Using Motion History Images of a Speaker's Lip Movements," Machine Vision and Image Processing Conference, pp.83-88, 2008.
- [48] J. Luetttin, N.A. Thacker, S.W. Beet, "Speaker identification by lipreading," Proceedings of Fourth International Conference on Spoken Language, vol.1, pp. 62-65, 1996.
- [49] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in International Conference of Audio- and Video-Based Person Authentication, pp. 260-267, U.K., 2003.
- [50] M.I. Faraj, J. Bigun, "Motion Features from Lip Movement for Person Authentication," 18<sup>th</sup> International Conference on Pattern Recognition, vol. 3, pp.1059-1062, 2006.
- [51] H.E. Cetingul, Y. Yemez, E. Engin, A.M. Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," IEEE Transactions on Image Processing, vol.15, no.10, pp.2879-2891, 2006.
- [52] T. Wagner, U. Dieckmann, "Multi-sensorial inputs for the identification of persons with synergetic computers," IEEE International Conference of Image Processing, vol.2, no., pp.287-291, 1994.
- [53] H. Pan, L. Zhi-Pei, T.S. Huang, "A new approach to integrate audio and visual features of speech," International Conference on Multimedia and Expo, vol.2, pp.1093-1096, 2000.

- 
- [54] C.C. Broun, X. Zhang, R.M. Mersereau, M. Clements, "Automatic speechreading with application to speaker verification," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.1, pp. I-685-I-688, 2002.
- [55] P. Jourlin, J. Luettin, D. Genoud and H. Wassner, "Acoustic-labial Speaker Verification," In Proceedings of the First international Conference on Audio- and Video-Based Biometric Person Authentication, 1997.
- [56] N. Fox, R. B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," Proc. 4th International Conference on Audio and Video Based Biometric Person Authentication, 2003.
- [57] T. Wark, S. Sridharan, V. Chandran, "The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs," Proceedings of International Conference on Acoustics, Speech, and Signal Processing, vol. 6, pp.2389-2392, 2000.
- [58] A. Kanak, E. Erzin, Y. Yemez, A.M. Tekalp, "Joint audio-video processing for biometric speaker identification," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp. II-377-80 vol.2, 2003.
- [59] M. Ichino, H. Sakano, N. Komatsu, "Multimodal Biometrics of Lip Movements and Voice using Kernel Fisher Discriminant Analysis, 9<sup>th</sup> International Conference on Control, Automation, Robotics and Vision, pp.1-6, 2006.
- [60] M. I. Faraj, J. Bigun, "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition," IEEE Transactions on Computers, vol.56, no.9, pp.1169-1175, 2007.
- [61] T. Chen, "Audiovisual speech processing," IEEE Signal Processing Mag., 2001.
- [62] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, "AVICAR: Audio-visual speech corpus in a car environment," Conf. Spoken Language, 2004.
- [63] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," 1<sup>st</sup> Int. Conf. Audio- and Video-Based Biometric Person Authentication, 1997.
- [64] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," 2<sup>nd</sup> Int. Conf. Audio- and Video-Based Biometric Person Authentication, 1999.
- [65] N. A. Fox, B. O'Mullane, and R.B. Reilly, "The realistic multi-modal VALID database and visual speaker identification comparison experiments," 5<sup>th</sup> International Conference on Audio- and Video-Based Biometric Person Authentication, 2005.
- [66] V. Popovici, J. Thiran, E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, "The BANCA Database and Evaluation Protocol," 4<sup>th</sup> International Conference on Audio- and Video-Based Biometric Person Authentication, 2003.
- [67] A.J. O'Toole, J. Harms, S.L. Snow, D.R. Hurst, M.R. Pappas, J.H. Ayyad, H. Abdi, "A video database of moving faces and people," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no.5, pp.812-816, 2005.
- [68] B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquo, R. Ingold, D. Von Rotz, "MyIdea - Multimodal Biometrics Database, Description of Acquisition Protocols," In proc. of Third COST 275 Workshop, pp 59-62, U.K., 2005.
- [69] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. L. les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacrétaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities," in Audio- and Video-Based Biometric Person Authentication, pp. 1056, 2003.

- 
- [70] T. J. Hazen, K. Saenko, C.-H. La, and J. Glass, "A segment-based audio-visual speech recognizer: Data collection, development and initial experiments," International Conference on Multimodal Interfaces, 2004.
- [71] C. Sanderson, and K.K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," Pattern Recognition, vol.2, pp. 293-302, 2003.
- [72] T. Sakai, M. Nagao, and T. Kanade, "Computer analysis and classification of photographs of human faces," in Proc. First USA—Japan Computer Conference, pp. 2-7, 1972.
- [73] J. Choi, S. Kim and P. Rhee, "Facial components segmentation for extracting facial feature," in Proceedings Second International Conference on Audio- and Video-based Biometric Person Authentication, 1999.
- [74] S.A. Sirohey, "Human Face Segmentation and Identification," Technical Report CS-TR-3176, Univ. of Maryland, 1993.
- [75] J. Huang, S. Gutta, and H. Wechsler, "Detection of human faces using decision trees," in IEEE Proc. of 2<sup>nd</sup> Int. Conf. on Automatic Face and Gesture Recognition, Vermont, 1996.
- [76] R. Herpers, M. Michaelis, K.-H. Lichtenauer, and G. Sommer, "Edge and keypoint detection in facial regions," in IEEE Proc. of 2<sup>nd</sup> Int. Conf. on Automatic Face and Gesture Recognition, pp. 212–217, Vermont, 1996.
- [77] S. McKenna, S. Gong, and J. J. Collins, "Face tracking and pose representation," in British Machine Vision Conference, Scotland, 1996.
- [78] J. Yang and A. Waibel, "A real-time face tracker," in IEEE Proc. of the 3<sup>rd</sup> Workshop on Applications of Computer Vision, Florida, 1996.
- [79] J. L. Crowley and F. Berard, "Multi-model tracking of faces for video communications," in IEEE Proc. of Int. Conf. on Computer Vision and Pattern Recognition, Puerto Rico, 1997.
- [80] H. P. Graf, E. Cosatto, D. Gibson, E. Petajan, and M. Kocheisen, "Multi-modal system for locating heads and faces," in IEEE Proc. of 2<sup>nd</sup> Int. Conf. on Automatic Face and Gesture Recognition, pp. 277–282, Vermont, 1996.
- [81] C. H. Lee, J. S. Kim and K. H. Park, "Automatic human face location in a complex background," Pattern Recog. vol. 29, pp. 1877–1889, 1996.
- [82] Y. Dai and Y. Nakano, "Face-texture model based on sgld and its application," Pattern Recog. vol. 29, pp. 1007–1017, 1996.
- [83] F. Luthon and M. Lievin, "Lip motion automatic detection," in Scandinavian Conference on Image Analysis, Lappeenranta, Finland, 1997.
- [84] M. K. Hu, "Visual pattern recognition by moment invariants," IRE Transactions on Information Theory, vol. 8, pp. 179–187, 1962.
- [85] S. McKenna, Y. Raja, and S. Gong, "Tracking Colour Objects Using Adaptive Mixture Models," Image and Vision Computing, vol. 17, nos. 3/4, pp. 223-229, 1998.
- [86] P. J. L. Van Beek, M. J. T. Reinders, B. Sankur, and J. C. A. Van Der Lubbe, "Semantic segmentation of videophone image sequences," in Proc. of SPIE Int. Conf. on Visual Communications and Image Processing, pp. 1182–1193, 1992.
- [87] M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cog. Neurosci. vol. 3, pp. 71–86, 1991.
- [88] S. McKenna, S. Gong, and H. Liddell, "Real-time tracking for an integrated face recognition system," in 2<sup>nd</sup> Workshop on Parallel Modelling of Neural Operators, Faro, Portugal, 1995.

- 
- [89] M. U. Ramos Sanchez, J. Matas, and J. Kittler, "Statistical chromaticity models for lip tracking with b-splines," in *Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, Switzerland, 1997.
- [90] L. C. De Silva, K. Aizawa, and M. Hatori, "Detection and tracking of facial features by using a facial feature model and deformable circular template," *IEICE Trans. Inform. Systems*, pp. 1195–1207, 1995.
- [91] S. H. Jeng, H. Y. M. Liao, C. C. Han, M. Y. Chern, and Y. T. Liu, "Facial feature detection using geometrical face model: An efficient approach," *Pattern Recog.*, 1998.
- [92] F. Smeraldi, O. Carmona, and J. Bigun, "Saccadic search with Gabor features applied to eye detection and real-time head tracking," *Image Vision Comput.*, vol. 18, pp. 323–329.
- [93] M. C. Burl, T. K. Leung, and P. Perona, "Face localization via shape statistics," in *Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995.
- [94] W. Huang, Q. Sun, C. P. Lam, and J. K. Wu, "A robust approach to face and eyes detection from images with cluttered background," in *Proc. of International Conference on Pattern Recognition*, 1998.
- [95] D. Maio and D. Maltoni, "Real-time face location on gray-scale static images," *Pattern Recog.* vol. 33, pp. 1525–1539, 2000.
- [96] A. R. Mirhosseini, H. Yan, K.-M. Lam, and T. Pham, "Human face image recognition: An evidence aggregation approach," *Computer Vision and Image Understanding*, 1998.
- [97] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," in *Proc. of 1<sup>st</sup> Int. Conf. on Computer Vision*, London, 1987.
- [98] S. R. Gunn and M. S. Nixon, "A dual active contour for head and boundary extraction," in *IEE Colloquium on Image Processing for Biometric Measurement*, London, pp. 6/1, 1994.
- [99] C. L. Huang and C.W. Chen, "Human facial feature extraction for face interpretation and recognition," *Pattern Recog.* vol. 25, pp. 1435–1444, 1992.
- [100] T. Yokoyama, Y. Yagi, and M. Yachida, "Facial contour extraction model," in *IEEE Proc. of 3<sup>rd</sup> Int. Conf. on Automatic Face and Gesture Recognition*, 1998.
- [101] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vision*, vol. 8, pp. 99–111, 1992.
- [102] G. Chow and X. Li, "Towards a system for automatic facial feature detection," *Pattern Recog.*, vol. 26, 1739–1755, 1993.
- [103] J. Huang and H. Wechsler, "Eye detection using optimal wavelet packets and radial basis functions," *Int. J. Pattern Recog. Artificial Intell.*, vol. 13, 1999.
- [104] A. Shackleton and W. J. Welsh, "Classification of facial features for recognition," in *IEEE Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 573–579, Hawaii, 1991.
- [105] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic tracking, coding and reconstruction of human faces, using flexible appearance models," *IEEE Electron. Lett.*, vol. 30, pp. 578–579, 1994.
- [106] A. Lanitis, A. Hill, T. Cootes, and C. Taylor, "Locating facial features using genetics algorithms," in *Proc. of Int. Conf. on Digital Signal Processing*, pp. 520–525, Cyrus, 1995.
- [107] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer.* Vol. 4, 519–524, 1987.

- [108] L. Meng and T. Nguyen, "Two subspace methods to discriminate faces and clutters," in Proceedings of the 2000 International Conference on Image Processing, 2000.
- [109] M.-H. Yang, N. Ahuja, and D. Kriegman, "Face detection using mixtures of linear subspaces," in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000.
- [110] Q. Song and J. Robinson, "A feature space for face image processing," in Proceedings of the 15th International Conference on Pattern Recognition, vol II, 2000.
- [111] M. Tanaka, K. Hotta, T. Kurita, and T. Mishima, "Dynamic attention map by using model for human face detection," in Proc. of International Conference on Pattern Recognition, 1998.
- [112] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 45-51, 1998.
- [113] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 746-751, 2000.
- [114] T. Rikert, M. Jones, and P. Viola, "A Cluster-Based Statistical Model for Object Detection," Proc. Seventh IEEE Int'l Conf. Computer Vision, vol. 2, pp. 1046-1053, 1999.
- [115] R.J. Qian and T.S. Huang, "Object Detection Using Hierarchical MRF and MAP Estimation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 186-192, 1997.
- [116] M. Propp and A. Samal, "Artificial neural network architecture for human face detection," Intell. Eng. Systems Artificial Neural Networks, vol. 2, pp. 535-540, 1992.
- [117] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, pp. 23-38, 1998.
- [118] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition, pp. 38-44, 1998.
- [119] R. Vaillant, C. Monrocq, and Y. Le Cun, "An Original Approach for the Localisation of Objects in Images," IEE Proc. Vision, Image and Signal Processing, vol. 141, pp. 245-250, 1994.
- [120] R. Feraud, O. Bernier, and D. Collobert, "A constrained generative model applied to face detection," Neural Process. Lett. vol. 5, pp. 73-81, 1997.
- [121] S.-H. Lin, S.-Y. Kung, and L.-J. Lin, "Face recognition/detection by probabilistic decision-based neural network," IEEE Trans. Neural Networks, vol. 8, pp. 114-132, 1997.
- [122] D. Roth, "The SNoW Learning Architecture," Technical Report UIUCDCS-R-99-2102, UIUC Computer Science Department, 1999.
- [123] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 130-136, 1997.
- [124] C. Papageorgiou and T. Poggio, "A Trainable System for Object Recognition," Int'l J. Computer Vision, vol. 38, no. 1, pp. 15-33, 2000.
- [125] F. Samaria and S. Young, "HMM Based Architecture for Face Identification," Image and Vision Computing, vol. 12, pp. 537-583, 1994.

- 
- [126] F.S. Samaria, "Face Recognition Using Hidden Markov Models," PhD thesis, Univ. of Cambridge, 1994.
- [127] A.V. Nefian and M. H. H III, "Face Detection and Recognition Using Hidden Markov Models," Proc. IEEE Int'l Conf. Image Processing, vol. 1, pp. 141-145, 1998.
- [128] A. Hulbert, and T. Poggio, "Synthesizing a Color Algorithm From Examples," in Science. vol. 239, pp. 482-485, 1998.
- [129] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition," in Proceedings of IAPR Workshop. pp. 318-321, 2002.
- [130] S.-H. Leung, S.-L. Wang, W.-H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," IEEE Transactions on Image Processing, vol.13, no.1, pp.51-62, 2004.
- [131] S. Lucey, S. Sridharan and V. Chandran, "Adaptive mouth segmentation using chromatic features," Pattern Recogn. Lett. vol. 23, pp. 1293-1302, 2002.
- [132] X. Zhang and R. M. Mersereau, "Lip feature extraction toward an automatic speechreading system," in Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. 226-229, Canada, 2000.
- [133] S. Lucey, S. Sridharan, V. Chandran, "Initialised eigenlip estimator for fast lip tracking using linear regression," Proceedings. 15th International Conference on Pattern Recognition, vol.3, pp.178-181, 2000.
- [134] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A couple HMM for audio-visual speech recognition," in Proc. ICASSP, pp. 2013-2016, 2002.
- [135] C. Bregler, and S.M. Omohundro, "Nonlinear Manifold Learning for Visual Speech Recognition," in Proceedings of International Conference of Computer Vision. pp 494-499, 1995.
- [136] Y.-P. Guan, "Automatic extraction of lips based on multi-scale wavelet edge detection," IET Computer Vision , vol.2, no.1, pp.23-33, March 2008.
- [137] M. Sadeghi, J. Kittler, K. Messer, "Modelling and segmentation of lip area in face images," IEE Proceedings Vision, Image and Signal Processing, vol.149, no.3, pp. 179-184, 2002.
- [138] M. Lievin, F. Luthon, "Unsupervised lip segmentation under natural conditions," Proceedings of International Conference on Acoustics, Speech, and Signal Processing, vol.6, pp. 3065-3068 ,1999.
- [139] Y. Tian, T. Kanade, and J. Cohn, "Robust lip tracking by combining shape, color and motion," in Proc. ACCV, pp. 1040-1045, 2000.
- [140] R. Kaucic, B. Dalton, and A. Blake, "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications," In Proceedings of the 4th European Conference on Computer Vision, vol. II, 1996.
- [141] T. Coianiz, L. Torresani, and B. Caprile, "2D deformable models for visual speech analysis," NATO Advanced Study Institute: Speech reading by Man and Machine, pp. 391-398, 1995.
- [142] M. E. Hennecke, K. V. Prasad, and D. G. Stork, "Using deformable templates to infer visual speech dynamics," in Proc. 28th Annu. Asilomar Conf. Signals, Systems, and Computers, pp. 578-582, 1994.
- [143] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audiovisual speech recognition using MPEG-4 compliant visual features," EURASIP J. Appl. Signal Processing, pp. 1213-1227, 2002.

- 
- [144] N. Eveno, A. Caplier, P. Coulon, "Accurate and quasi-automatic lip tracking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 706 – 715, 2004.
- [145] T. F. Cootes, "Statistical Models of Appearance for Computer Vision," Technical report, University of *Manchester*, 2004.
- [146] L. Zhaorong, A. Haizhou, "Texture-Constrained Shape Prediction for Mouth Contour Extraction and its State Estimation," 18<sup>th</sup> International Conference on Pattern Recognition, vol.2, pp.88-91, 2006.
- [147] C.L. Huang, and Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," in Journal of Visual Communication and Image Representation. vol. 8, pp. 278-290, 1997.
- [148] S. Werda, W. Mahdi, A. Ben-Hamadou, "Colour and Geometric based Model for Lip Localisation: Application for Lip-reading System," 14th International Conference on Image Analysis and Processing, pp.9-14, 2007.
- [149] L.L. Mok, W.H. Lau, S.H. Leung, S.L.Wang, H. Yan, "Person authentication using ASM based lip shape and intensity information," International Conference on Image Processing, vol.1, pp. 561-564, 2004.
- [150] M. Chan, "Automatic lip model extraction for constrained contour-based tracking," in Proc. IEEE International Conference on Image Processing, vol. 2, pp. 848-851, Kobe, Japan, October 1999.
- [151] C. Bouvier, P.-Y. Coulon, X. Maldague, "Unsupervised Lips Segmentation Based on ROI Optimisation and Parametric Model," IEEE International Conference on Image Processing, vol.4, pp. 301-304, 2007.
- [152] K. Michael, W. Andrew, and T. Demetri, "Snakes: active Contour models," In Proc. International Journal of Computer Vision, vol. 1, pp. 259-268. 1987.
- [153] N. S. Thejaswi and S. Sengupta, "Lip Localization and Viseme Recognition from Video Sequences," Fourteenth National Conference on Communications, India, 2008.
- [154] F. Bourel, C. C. Chibelushi and A. A. Low, "Robust Facial Feature Tracking", in Proceedings of the 11th British Machine Vision Conference, vol. 1, pp. 232–241. UK, 2000.
- [155] M.A. Hall, and L.A. Smith, "Practical feature subset selection for machine learning," In Proceedings of the 21<sup>st</sup> Australian Computer Science Conference, pp. 181–191, Australia, 1998.
- [156] R. Kohavi, and G. John, "Wrapper for Feature Subset Selection," Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.
- [157] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," In Proc. of International Conference on Image Processing, vol. 3, pp. 173–177, U.S.A. 1998.
- [158] W.C. Yau, D.K. Kumar, and H. Weghorn, "Visual Speech Recognition Using Motion Features and Hidden Markov Models," In Proc. of International Conference on Computer Analysis of Images and Patterns, pp. 832-839, Austria, 2007.
- [159] C. Ding, and H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," In Proc. of 2<sup>nd</sup> IEEE conference on Computational Systems Bioinformatics, pp. 523-528, U.S.A., 2003.
- [160] P. Lu, X. Huang, X. Zhu and Y. Wang, "Head Gesture Recognition Based on Bayesian Network," in Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, 2005, pp. 492.

- 
- [161] Pei Chi Ng and L.C. De Silva, "Head gestures recognition," in Proceedings of International Conference on Image Processing, vol.3, pp.266-269, 2001.
- [162] A. Benoit, and A. Caplier, "Head nods analysis: interpretation of non verbal communication gestures," in Proceedings of International Conference on Image Processing, vol.3, pp. 425-8, 2005.
- [163] K. Toyama, "Look, Ma--No Hands! Hands free cursor control with real-time 3D face tracking," in Proceedings of Workshop on Perceptual User Interface, 1998.
- [164] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes," in Proceedings of 4th International Conference on Automatic Face and Gesture Recognition, pp.40-45, 2000.
- [165] A. Kapoor and R. Picard, "A real-time head nod and shake detector," in Proceedings of Workshop on Perspective User Interfaces, 2001.
- [166] V. Chauhan and T. Morris, "Face and feature tracking for cursor control," in Proceedings of 12th Scandinavian Conference on Image Analysis, 2001.
- [167] P. Hong and T. Huang, "Natural Mouse-a novel human computer interface," in Proceedings of International Conference on Image Processing, vol.1, pp.653-656, 1999.
- [168] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, vol.1, pp. 511-518, 2001.
- [169] <http://www.intel.com/technology/computing/opency/>
- [170] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in Proceedings of 4<sup>th</sup> Alvey Vision Conference, pp.147-151, 1988.
- [171] B.Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proceedings of DARPA Image Understanding Workshop, pp. 121-130, 1981.
- [172] [ww.cmpe.boun.edu.tr/enterface07/outputs/final/p12docs.zip](http://ww.cmpe.boun.edu.tr/enterface07/outputs/final/p12docs.zip)
- [173] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," in Research Report of National Taiwan University, 2003.
- [174] F. Matta, J-L. Dugelay, "Tomofaces: eigenfaces extended to videos of speakers," In Proc. of International Conference on Acoustics, Speech, and Signal Processing, pp.1793-1796, 2008.
- [175] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," In PAMI, vol. 9, pp. 1063-1074, 2003.
- [176] K. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," In Proc of CVPR, pp. 852-859, 2005.
- [177] A. S. Georghiades, D. J. Kriegman, and P. N. Belhumeur, "Illumination cones for recognition under variable lighting: Faces, In Proc of CVPR, pp. 52-59, 1998.
- [178] P. Tsai, T. Jan, T. Hintz, "Kernel-based Subspace Analysis for Face Recognition," In Proc of International Joint Conference on Neural Networks, pp.1127-1132, 2007.
- [179] M. Ramachandran, S.K. Zhou, D. Jhalani, R. Chellappa, "A method for converting a smiling face to a neutral face with applications to face recognition," In Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp. 977-980, 2005.
- [180] K. Ugiyama, T. Aoki, S. Hangai, "Motion compensated frame rate conversion using normalized motion estimation," In Proc. IEEE Workshop on Signal Processing Systems Design and Implementation, pp. 663-668, 2005.

- 
- [181] G. Wolberg, "Recent Advances in Image Morphing," In Proceedings of the Conference on Computer Graphics international, USA, 1996.
- [182] A. Akutsu, and Y. Tonomura, "Video tomography: an efficient method for camerawork extraction and motion analysis," In Proceedings of the Second ACM international Conference on Multimedia, pp. 349-356, USA, 1994.
- [183] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 8, pp. 679-698, 1986.
- [184] B.A. Golomb, D.T. Lawrence and T.J. Sejnowski, "Sex-Net: a neural network identifies sex from human faces," in Proceedings of Advances in neural information processing systems, pp. 572-577, 1990.
- [185] Z. Sun, G. Bebis, X. Yuan and S.J. Louis, "Genetic feature subset selection for gender classification: a comparison study", in IEEE Proceedings on Applications of Computer Vision, pp. 165-170, 2002.
- [186] S. Gutta, J.R.J. Huang, P. Jonathon and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces", in IEEE Transactions on Neural Networks, vol. 11, no. 4, pp. 948-960, 2000.
- [187] M. Nakano, F. Yasukata and M. Fukumi, "Age and gender classification from face images using neural networks," in Signal and Image Processing, 2004.
- [188] X. Lu, H.Chen and A.K. Jain, "Multimodal facial gender and ethnicity identification", in Advances in Biometrics, vol. 3832/2005, pag. 554-561, 2005.
- [189] B. Moghaddam and M.-H. Yang, "Gender classification with support vector machines", in IEEE Proceedings on Automatic Face and Gesture Recognition, pp. 306-311, 2000.
- [190] Y. Saatci and C. Town, "Cascaded classification of gender and facial expression using active appearance models," in Automatic Face and Gesture Recognition, pp. 393-398, 2006.
- [191] H.-C. Kim, D. Kim, Z. Ghahramani and S.Y. Bang, "Appearancebased gender classification with Gaussian processes," in Pattern Recognition Letters, vol. 27, no. 6, pp. 618-626, 2006.
- [192] Y. Zhiguang, L. Ming and A. Haizhou, "An Experimental Study on Automatic Face Gender Classification," in IEEE Proceedings on Pattern Recognition, pp. 1099-1102, 2006.
- [193] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," in International Journal of Computer Vision, vol. 71, no. 1, pp. 111- 119, 2007.
- [194] S. Caifeng, G. Shaogang, and P.W. McOwan, "Learning gender from human gaits and faces," in IEEE Proceedings on Advanced Video and Signal Based Surveillance, pp.505-510, 2007.
- [195] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improved connected letter recognition by lipreading," in Proc. IEEE ICASSP, vol. 1, pp. 557-560, 1993.
- [196] P. Duchnowski, U. Meier, and A.Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in Proc. ICSLP, vol. 2, pp. 547-550, 1994.
- [197] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for broadcast news: Some fusion techniques," in Proc. Works. Multimedia Signal Processing, pp. 161-167, Denmark, 1999.
- [198] T. Wark, S. Sridharan, and V. Chandran, "Robust speaker verification via fusion of speech and lip modalities," in Proc. Int. Conf. Acoustics, Speech Signal Processing, pp. 3061-3064, USA, 1999.

- 
- [199] P. S. Aleksic, G. Potamianos, and A. K. Katsaggelos, "Exploiting visual information in automatic speech processing," in *Handbook of Image and Video Processing*, pp. 1263–1289, A. Bovik, Ed. New York: Academic, Jun. 2005.
- [200] E. D. Petajan, N. M. Brooke, B. J. Bischoff, and D. A. Boddoff, "Experiments in automatic visual speech recognition," in *Proc. 7th FASE Symp.*, pp. 1163–1170, Book 4, 1988.
- [201] S. Nishida, "Speech recognition enhancement by lip information," in *Proc. CHI*, pp. 198–204, 1986.
- [202] I. Matthews, J. A. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition," in *Proc. 4th ICSLP*, vol. 1, pp. 38–41, 1996.
- [203] I. Matthews, G. Potamianos, C. Neti, and J. Luetin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. Int. Conf. Multimedia Expo*, pp. 22–25, 2001.
- [204] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Processing*, vol. 2002, no. 11, pp. 1228–1247, 2002.
- [205] C. Benoit, "On the production and the perception of audio-visual speech by man and machine," in *Proc. Symp. Multimedia Communications and Video Coding*, pp. 277–284, 1995.
- [206] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke, "Lipreading by neural networks: Visual preprocessing, learning, and sensory integration," in *Advances in Neural Information Processing Systems*, pp. 1027–1034, J. Cowan, G. Tesauero, and J. Alspector, Eds. San Mateo, CA: Morgan Kaufmann, 1994.
- [207] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Eur. Conf. Computer Vision*, pp. 484–498, Germany, 1998.
- [208] S. W. Foo, Y. Lian, and L. Dong, "Recognition of visual speech elements using adaptively boosted hidden Markov models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 693–705, 2004.
- [209] J. F. G. Perez, A. F. Frangi, E. L. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. I, pp. 473–476, 2005.
- [210] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.
- [211] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audiovisual speech recognition using MPEG-4 compliant visual features," *EURASIP J. Appl. Signal Process.*, pp. 1213–1227, 2002.
- [212] A.G. de la Cuesta, Z. Jianguo, P. Miller, "Biometric Identification Using Motion History Images of a Speaker's Lip Movements," *International Machine Vision and Image Processing Conference*, vol., pp. 83-88, 2008.
- [213] J.S.D. Mason, J. Brand, R. Auckenthaler, F. Deravi, C. Chibelushi, "Lip signatures for automatic person recognition," *IEEE 3<sup>rd</sup> Workshop on Multimedia Signal Processing*, pp.457-462, 1999.
- [214] P. Paalanen, J.-K. Kamarainen, J. Ilonen and H. Kalviainen, "Feature representation and discrimination based on Gaussian model probability densities," *Practices and algorithms*, Research report of the Lappeenranta University of Technology, no. 95, 2005.

- [215] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879-2891, 2006.
- [216] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition," in *Proceedings of the IAPR Workshop on Machine Vision Application*, pp. 318-321, Japan, 2002.
- [217] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*, Cambridge, MA: MIT Press, 2004.
- [218] J.C. Bezdec, "Pattern recognition with fuzzy objective function algorithms," Plenum Press, 1981.
- [219] A.E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang and F.K. Soong, "The use of cohort normalized scores for speaker verification," *Proceedings of Spoken Language Processing*, pp. 599-602, 1992.
- [220] A.W.-C. Liew, L. Shu Hung, L. Wing Hong, "Segmentation of color lip images by spatial fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol.11, no.4, pp. 542-549, 2003.
- [221] Y.-P. Guan, "Automatic extraction of lips based on multi-scale wavelet edge detection," *IET Computer Vision*, vol.2, no.1, pp.23-33, March 2008.
- [222] T.F. Chan, L.A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol.10, no.2, pp.266-277, 2001.
- [223] C. Garcia, M. Delakis. "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(11), pp.1408-1423, 2004.



*Reconnaissance des Personnes par  
l'Analyse des Lèvres*

*Usman Saeed*

## 1. Introduction

La sécurité est devenue un thème récurrent au 21<sup>ème</sup> siècle entraînant un attrait important pour la biométrie et la vidéo-surveillance. Parmi les différents systèmes d'identification biométriques utilisés de nos jours, tels que les empreintes digitales ou encore l'ADN, l'identification basée sur le visage comporte plusieurs atouts, notamment elle est non intrusif, facile à acquérir et bien acceptée par le public en général.

Le terme biométrie, d'origine grecque, vient des mots *Bios* (vie) et *Metron* (mesure) et signifie "mesure de la vie". La biométrie se définit comme

*"La mesure et l'analyse des caractéristiques physiques ou comportementales pour vérifier l'identité des personnes"*

Les identificateurs biométriques sont généralement classés en deux catégories : physiques ou comportementaux, toutefois, les identificateurs qui contiennent ces deux caractéristiques sont classés comme hybrides.

- Physiques: Empreintes digitales, Iris, ADN, Retina.
- Behavioural: Démarche
- Hybrid: Voix, Visage, Signature

Un système typique de reconnaissance est composé de deux modules, l'enrôlement et la reconnaissance. L'objectif de la phase enrôlement consiste à enregistrer de nouveaux utilisateurs dans le système de reconnaissance et le module de reconnaissance à les identifier.

Les modes de fonctionnement décrivent l'ensemble des conditions dans lesquelles un système de reconnaissance biométrique marche. Un système de reconnaissance biométrique a deux principaux modes de fonctionnement, la vérification et l'identification. Dans un cas de vérification, un utilisateur présente ses identificateurs biométriques et déclare une identité, le système de reconnaissance vérifie ensuite sa demande et décide de l'accepter ou la rejeter. Dans un cas d'identification, un utilisateur présente ses identificateurs biométriques, mais ne fait aucune déclaration sur son identité, le système compare ensuite ce modèle à tous les modèles dans la base de données pour trouver l'identité la plus probable.

## 2. Etat de l'Art

Cette thèse aborde divers sujets tels que la reconnaissance de la personne, la reconnaissance du sexe ou encore la normalization de video. Il est donc impossible de présenter un bilan détaillé sur tous les sujets. Par conséquent nous présentons un bilan détaillé sur l'identification de la personne audio-visuel (AV) en utilisant le mouvement des lèvres. La Figure 4 décrit les étapes impliquées dans un système audio-visuel pour la reconnaissance des personnes.

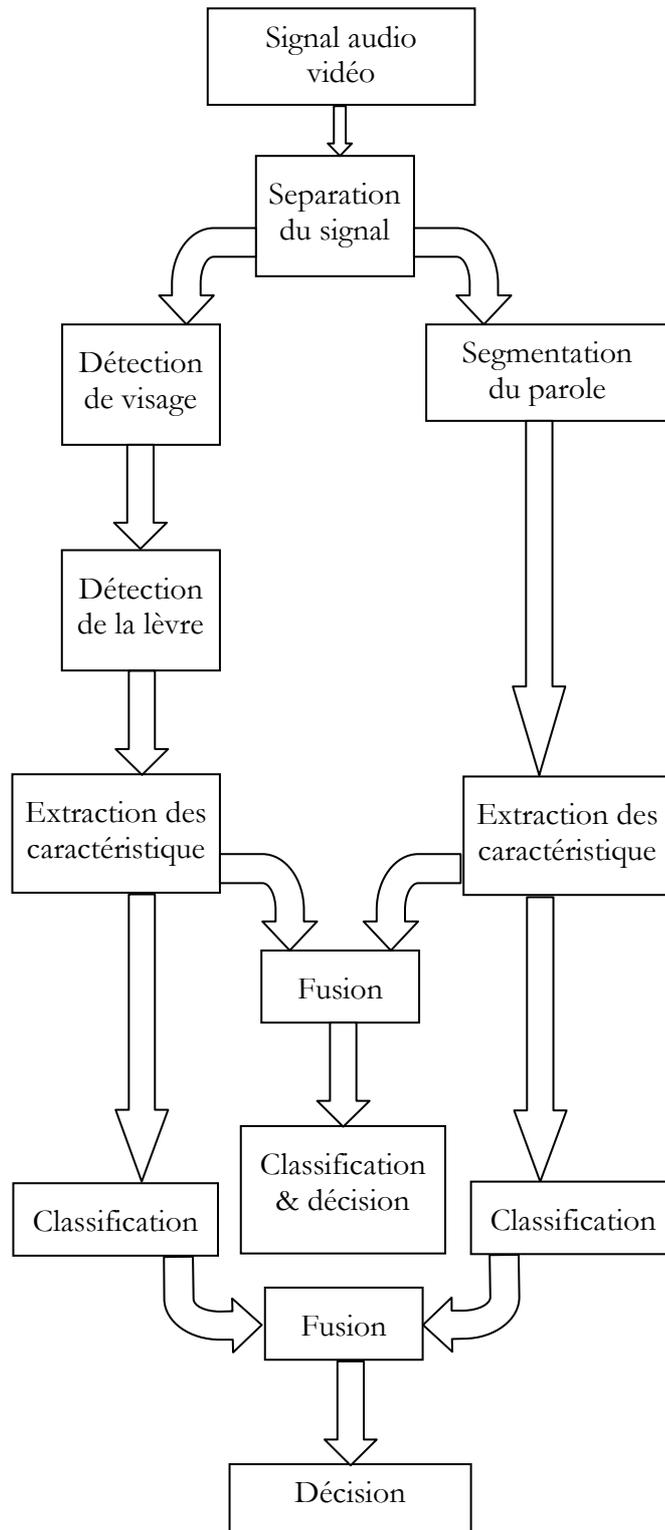


Figure 31: Système des reconnaissance audio-visuel des personnes

La Table 11 présente un résumé des systèmes qui utilise les caractéristique visuelles des lèvres pour l'identification des personnes.

Systeme	Caractéristique	Classification	Base de données	Résultats
Wark et al. 1998 [3]	ACP/ADL sur vecteurs de couleur	GMM	M2VTS 37 personnes contenu: 0-9	90% CIR
Mok et al. 2004 [2]	Paramètres des ASM ACP sur vecteurs d'intensité	HMM/GMM	Privé 40 personnes contenu: 3725	98% CIR
Cuesta et al. 2008 [4]	Image de l'histoire du mouvement	Classification bayésienne	Privé 9 personnes contenu: 0-9	100% CIR
Luetttin et al. 1996 [5]	Paramètres des ASM ACP sur vecteurs d'intensité	HMM/GMM	Tulip 12 personnes contenu:4 Chiffres	97.9 % CIR
Lucey et al. 2003 [49]	ACP/ADL sur la région de la bouche	HMM	M2VTS 36 personnes contenu: 0-9	19.71 % ER
Faraj et al. 2006 [50]	Les vecteurs de mouvement	GMM/HMM	XM2VTS 200 personnes Contenu: 3 Phrase	78% CVR
Cetingul et al. 2006 [8]	TCD des vecteurs de mouvement Distance entre les point sur les lèvres	HMM	MVGL-AVD 50 personnes contenu: Nom, 348-572	5.2 % EER

Table 11: Résumé des systèmes qui utilise les traits visual des lèvres

### 3. Bases de Données

Les bases de données utilisées doivent répondre à deux principales exigences afin de pouvoir évaluer correctement les performances des systèmes, notamment elles doivent comporter :

- Suffisamment de données par utilisateur pour permettre l'apprentissage et la reconnaissance en utilisant des données relatives à la biométrie comportementale.
- Plusieurs sessions d'individus répétant la même phrase.

Malheureusement, il n'existe aucune base de données qui répond à nos exigences, plus particulièrement, la première exigence est la plus difficile à réaliser. Donc nous avons décidé d'utiliser deux bases de données pour nos expériences, Valid et Italian TV Database, qui sont présentées ci-dessous.

La base de données Valid [65] se décompose en cinq sessions d'enregistrement de 106 personnes (77 hommes, 29 femmes) sur une période d'un mois. Le contenu de la base de données se compose de trois phrases par session (en anglais),

1: “<Nom complet du person>”

2: “5 0 6 9 2 8 1 3 7 4”

3: “Joe took father’s green shoe bench out”

Les cinq sessions ont été enregistrées sur une période d'un mois, la première session a été enregistrée sous des conditions acoustiques et d'éclairage idéales. Les quatre autres sessions ont des cas de la vie courante ou aucune spécificité n'a été imposée quant aux conditions d'éclairage ou d'enregistrement.

La base de données “Italian TV Database” compilées par [1], a été enregistrée par la chaîne italienne RAI 1, sur une période de 21 mois. Elle se compose de 208 vidéos de 13 personnes (8 hommes et 5 femmes) et chaque vidéo comporte 13 secondes. La Figure 6 illustre les 7 premières images de quelques uns des locuteurs.



Figure 32: Premières 7 images de certains des orateurs

		<b>Complet</b>	<b>Enrôlement</b>	<b>Reconnaissance</b>
<b>Detail Global</b>	Nombre d'Individus	13	13	13
	Nombre d'Hommes	8	8	8
	Nombre de Femmes	5	5	5
	Nombre de Vidéos	208	104	104
	Nombre d'Image	68640	34320	34320
	Longueur de Vidéos	45min. 49sec.	22 min. 54sec.	22 min. 54sec.
<b>Par Personne</b>	Nombre de Vidéos	16	8	8
	Nombre d'Image	5280	2640	2640
	Longueur de Vidéos	3min. 31sec.	1min. 46sec.	1min. 46sec.
<b>Par Vidéo</b>	Nombre d'Image	330	330	330

	Longueur	13 sec.	13 sec.	13 sec.
<b>Resolution</b>	Spatial	Hauteur 192 pixel	Hauteur 192 pixel	Hauteur 192 pixel
		Largeur 224 pixel	Largeur 224 pixel	Largeur 224 pixel
	Temporal	24.97 f/s	24.97 f/s	24.97 f/s
<b>Compression</b>	Taux de Compression	118 Kb/s	118 Kb/s	118 Kb/s
	Format	Windows Media Video 9	Windows Media Video 9	Windows Media Video 9

Table 12: Détails techniques de "Italian TV database". [1]

#### 4. Détection des lèvres

Une première contribution majeure est la mise au point d'un algorithme de détection des lèvres en fusionnant deux méthodes indépendantes. La première méthode est basée sur la détection de contours alors que la seconde est orientée sur la segmentation. Chacune ayant des caractéristiques distinctes et donc présente différentes forces et faiblesses. On exploite leurs points forts en combinant les deux méthodes par fusion. L'architecture du système est donnée par Figure 7.

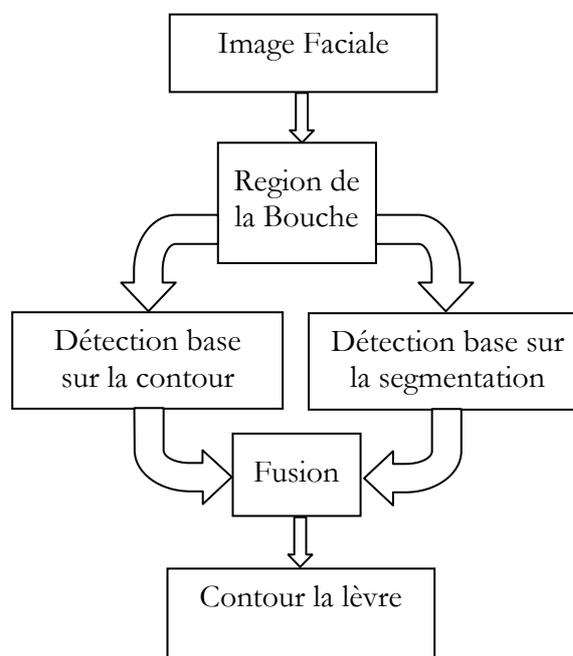


Figure 33: L'architecture du système

Le premier algorithme est basé sur une méthode de détection des contours, Il comporte deux étapes, La première est une transformation pour améliorer la couleur des lèvres proposé par [216]

$$I = \frac{2G - R - 0.5B}{4}$$

La prochaine étape est l'extraction du contour des lèvres. Pour cela, nous avons utilisé les contours actifs [152].

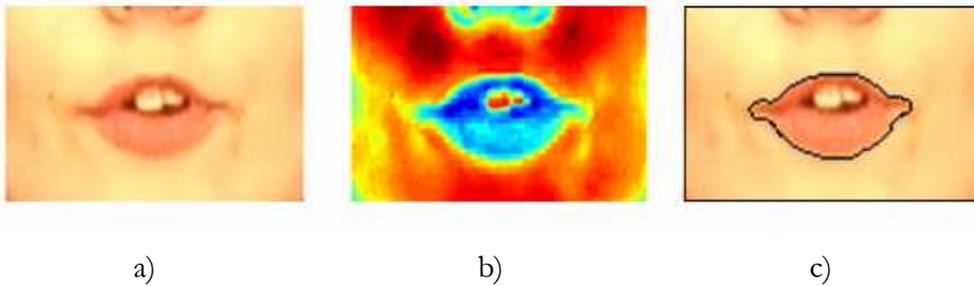


Figure 34: a) Region de la bouche b) Transformation des couleurs c) Détection de contour.

La deuxième approche est basée sur la segmentation, après une transformation des couleurs vers le domaine YIQ. [153] ont présenté que la séparation entre la peau et les lèvres peut être réalisée avec succès dans le domaine YIQ. Donc, nous avons transformé les images de RGB vers l'espace couleur YIQ en utilisant l'équation ci-dessous et en conservant le canal Q.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.31135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Après nous avons appliqué une technique appelée “active contours without edges” [222]. Cette dernière modélise les intensités en différente région de l'image et utilise comme le même critère d'arrêt que pour les contours actifs.

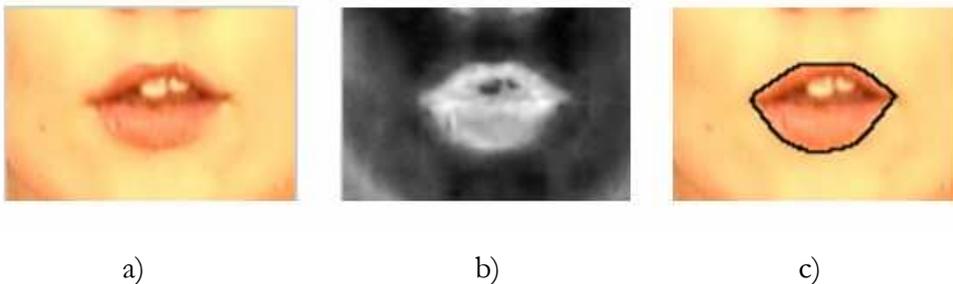


Figure 35: a) Region de la bouche b) Transformation des couleurs c) Détection de la région

Nous sommes confrontés à deux types d'erreurs aussi nous avons proposé des méthodes pour la détection et la correction des erreurs. Le premier type d'erreur est lorsque la lèvre a été totalement manquée et un autre objet a été détecté. Cette erreur peut être facilement détectée en appliquant des contraintes sur les valeurs relatives à la localité. Si cette erreur a été observée, les résultats de détection sont rejetés. Le second type d'erreur est rencontré quand la lèvre n'est pas détectée dans son intégralité, ces erreurs sont difficiles à détecter c'est pourquoi nous avons proposé d'utiliser la fusion comme mesure corrective, sous condition que les deux techniques de détection ne failliront pas simultanément.

Les résultats de détection obtenus par ces méthodes indépendantes ont été ensuite fusionnés en utilisant les opérateurs logiques ET et OU. Les tests ont été effectués sur la base de données Valid [65] qui comporte cinq sessions d'enregistrements de 106 personnes. Une image a été extraite pour chacune des cinq vidéos afin de créer une base de données de 530 images faciales.

Pour évaluer l'algorithme de détection de la lèvre nous avons utilisé les deux mesures suivantes, proposées par [221]. La première mesure détermine le pourcentage de Chevauchement (OL), entre la région de la lèvre segmentée  $A$  et la vérité du terrain  $A_G$ . Elle est définie par

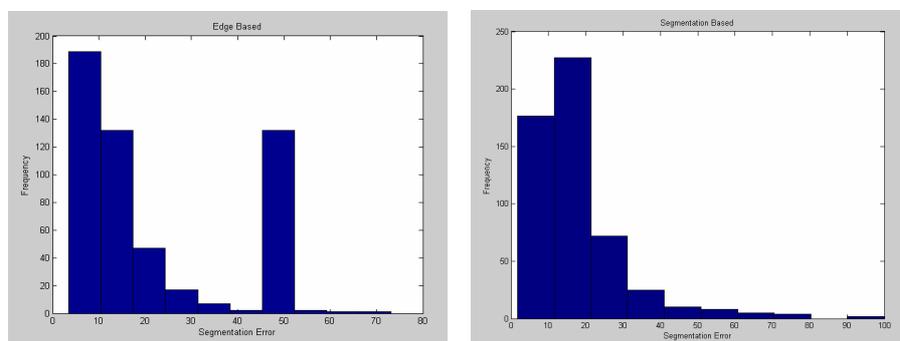
$$OL = \frac{2(A \cap A_G)}{A + A_G} * 100$$

La seconde mesure est l'erreur de segmentation ( $SE$ ), définie comme

$$SE = \frac{OLE + ILE}{2 * TL} * 100$$

OLE (erreur lèvre externe) est le nombre de pixels n'appartenant pas à la lèvre étant classés comme pixels des lèvres. ILE (erreur lèvre interne) est le nombre de pixels relatifs aux lèvres classés comme des pixels n'y appartenant pas. TL désigne le nombre de pixels se rapportant aux lèvres dans la vérité du terrain.

Figure 36 représente l'erreur de segmentation sous la forme d'histogrammes, et les erreurs moyennes sont données dans Table 13.



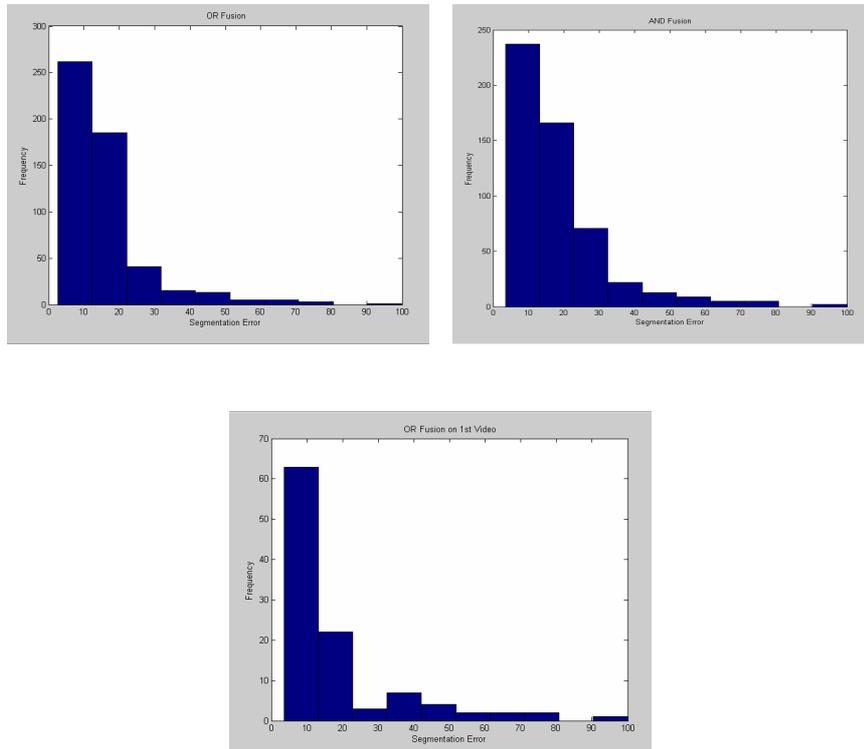


Figure 36: Histogrammes des erreurs de segmentation

Méthode de détection des lèvres	Erreur de segmentation moyenne (SE) %	Chevauchement moyenne (OL) %
Segmentation	17.8225	83.6419
Contour	22.3665	65.6430
OU Fusion	15.6524	83.9321
ET Fusion	18.4067	84.2452
Ou Fusion sur 1 <sup>er</sup> Vidéo	13.9964	87.1492

Table 13: Résultats de détection pour les lèvres



Figure 37: Exemple d'images avec 15% d'erreur de segmentation

Les techniques de fusion ont été appliquées et de meilleurs résultats ont été observés pour la fusion OU. L'erreur de segmentation minimale obtenue est d'environ 15%, ce qui peut paraître assez grand et en se référant à la Figure 11, il est évident que le fait d'inclure même une faible partie de la peau peut causer ce pourcentage d'erreur.

## 5. L'évaluation des caractéristiques des lèvres

Nous allons maintenant étudier l'influence des caractéristiques visuelles du mouvement des lèvres sur la reconnaissance des personnes. Pour cela, on extrait diverses caractéristiques géométriques et d'autres basées sur l'apparence des lèvres. On les compare en utilisant trois mesures de sélection: mRMR, Distance Bhattacharya et l'information mutuelle.

Les caractéristiques géométriques qu'on a extraites pour cette étude comprennent la superficie, longueur de l'axe majeur et mineur, l'excentricité, l'orientation et la longueur du périmètre du contour des lèvres.

Les caractéristiques basées sur l'apparence sont

- Profil d'intensité des pixels (PI)
- Moyenne enlevé ?? ACP [157]
- Flux optique (FO)
- Spatio-Temporal Templates (STT) [158]
- Transformée de Fourier discrète (TFD)
- Analyse en composantes indépendantes (ACI)
- Transformée en ondelettes discrète (TOD)
- Transformée en cosinus discrète (TCD)

L'évaluation a été effectuée en utilisant les trois techniques suivantes :

*Minimal-Redundancy-Maximum-Relevance* (mRMR) proposé par [159] sélectionne d'abord les caractéristiques  $S$  qui a un maximum de pertinence entre caractéristiques  $x$  et la classe  $c$ , par une mesure de similarité telles que l'information mutuelle.

$$\max D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)$$

Après la redondance est réduite en sélectionnant les caractéristiques qui sont les plus dissemblables

$$\min R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j)$$

Enfin, les deux critères sont combinés et optimisés

$$\max \Phi(D, R) = D - R$$

La distance de Bhattacharyya mesure la similarité entre deux distributions de probabilité discrètes. Pour les distributions de probabilité discrètes  $p$  et  $q$ , elle est définie par:

$$D_B(p, q) = -\ln(BC(p, q))$$

BC est le coefficient de Bhattacharyya

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

L'information mutuelle de deux variables aléatoires  $X$  et  $Y$  mesure la dépendance mutuelle des deux variables et pour les variables discrètes peuvent être définies comme:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p_1(x)p_2(y)} \right)$$

Les tests ont été effectués sur base de données Valid [65] qui se compose de cinq sessions d'enregistrements de 106 personnes.

Rang	mRMR	BT	MI
1	ACI	ACP	DWT
2	PI	TCD	TFD
3	STT	ACI	Géométrique
4	TFD	TFD	ACP
5	TCD	PI	STT
6	DWT	STT	TCD
7	Géométrique	FO	ACI
8	FO	Géométrique	FO
9	ACP	DWT	PI

Table 14: Résultats pour mRMR, Distance de Bhattacharyya, Information Mutuelle

---

<b>Rang</b>	<b>Rang Fusion</b>
1	ACI
2	TFD
3	STT
4	TCD
5	PI
6	ACP
7	TOD
8	Géométrie
9	FO

Table 15: Rang fusion

Nous avons observé une corrélation faible entre les résultats des trois méthodes et a donc décidé de fusionner les résultats du classement donné par l'équation suivante.

$$\text{Rang Fusion} = (2 * mRMR) + BD + MI$$

La méthode ICA a donné les meilleurs résultats nettement supérieurs à ceux donnés par les caractéristiques géométriques ou le flux optique. Donc notre hypothèse que l'apparence contient plus d'informations pour la reconnaissance des personnes que la forme et le comportement est validée.

## 6. Reconnaissance des personnes grace aux caractéristiques des lèvres.

Pour la deuxième contribution majeure on extrait les caractéristiques qui modélisent l'aspect comportemental du mouvement des lèvres lorsque la personne parle afin de les exploiter pour la reconnaissance des personnes. Les caractéristiques du comportement incluent des caractéristiques statiques, telles que la longueur normalisée de l'axe majeur / mineur, les coordonnées des points relatifs aux extrema des lèvres et des caractéristiques dynamiques en fonction du flux optique.

Les caractéristiques géométriques se composent des longueurs de l'axe majeur / mineur et les coordonnées de points extrema des lèvres.

$$P_t = [x1_t, y1_t, x2_t, y2_t, x3_t, y3_t, x4_t, y4_t, Maj_t, Min_t]$$

La normalisation a ensuite été effectuée en soustrayant la valeur moyenne de chaque caractéristique

$$x_{n,t} = p_{n,t} - \mu_n$$

pour  $n = 1, \dots, 10$  et  $t = 1, \dots, T$ ,  $\mu_n$  est la valeur moyenne pour la dimension  $n$ ,

$$\mu_n = \frac{1}{T} \sum_{t=1}^T p_{n,t}$$

Les caractéristiques dynamiques sont composées de flux optique qui ont été calculées image par image.

$$Df_t = [u_{1,1,t}, v_{1,1,t}, u_{1,2,t}, v_{1,2,t}, \dots, x_{n,m,t}, y_{n,m,t}]$$

$u$  et  $v$  sont les composantes horizontale et verticale du vecteur de mouvement, calculé pour la ligne  $n$  et la colonne  $m$  de l'image de bouche.

Ces caractéristiques ont ensuite été modélisées en utilisant un mélange de modèles gaussiens

$$p(\mathbf{x} | \Theta) \equiv \sum_{c=1}^C \alpha_c \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

$\Theta = \{\alpha_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 1, \dots, C\}$  est la liste de paramètres, et  $\alpha_c \in [0,1]$  est le poids de la  $c$ -ième composante de la gaussienne.

Enfin le classement a été fait en utilisant une règle de bayésienne. La probabilité a posteriori pour la vidéo  $p(k | \mathbf{X})$  est:

$$p(k | \mathbf{X}) = \frac{p(\mathbf{X} | k)p(k)}{p(\mathbf{X})} = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_T | k)p(k)}{p(\mathbf{x}_1, \dots, \mathbf{x}_T)}$$

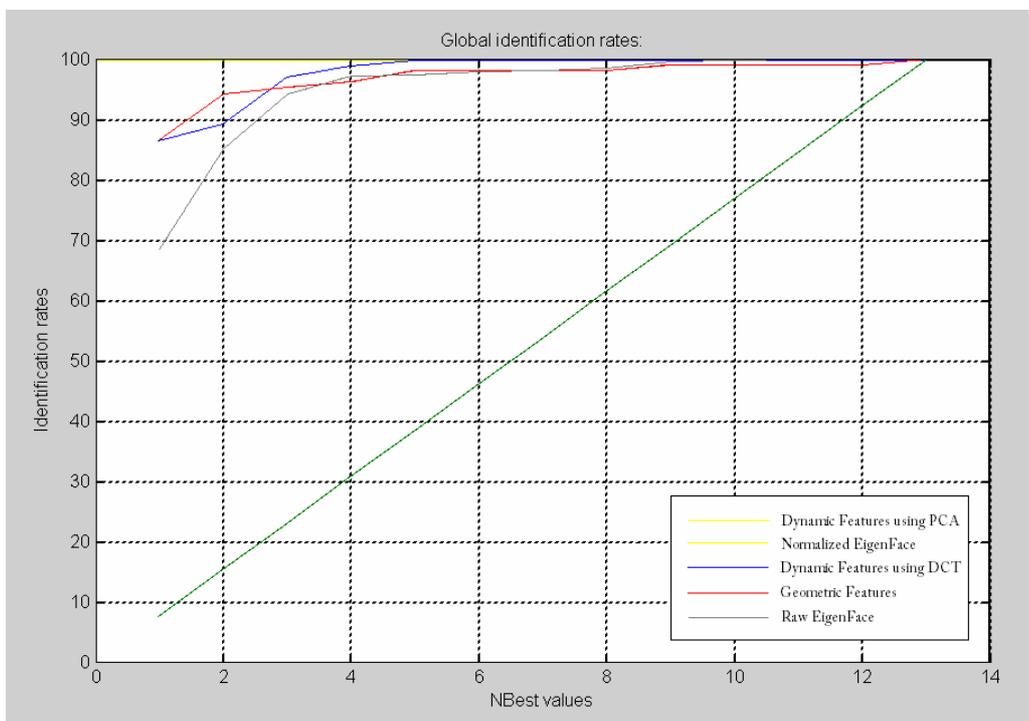
Supposons que les caractéristiques sont indépendantes; la densité de probabilité  $p(\mathbf{X} | k)$  de la classe devient:

$$p(\mathbf{X} | k) \equiv p(\mathbf{x}_1, \dots, \mathbf{x}_T | k) \cong \prod_{t=1}^T p(\mathbf{x}_t | k)$$

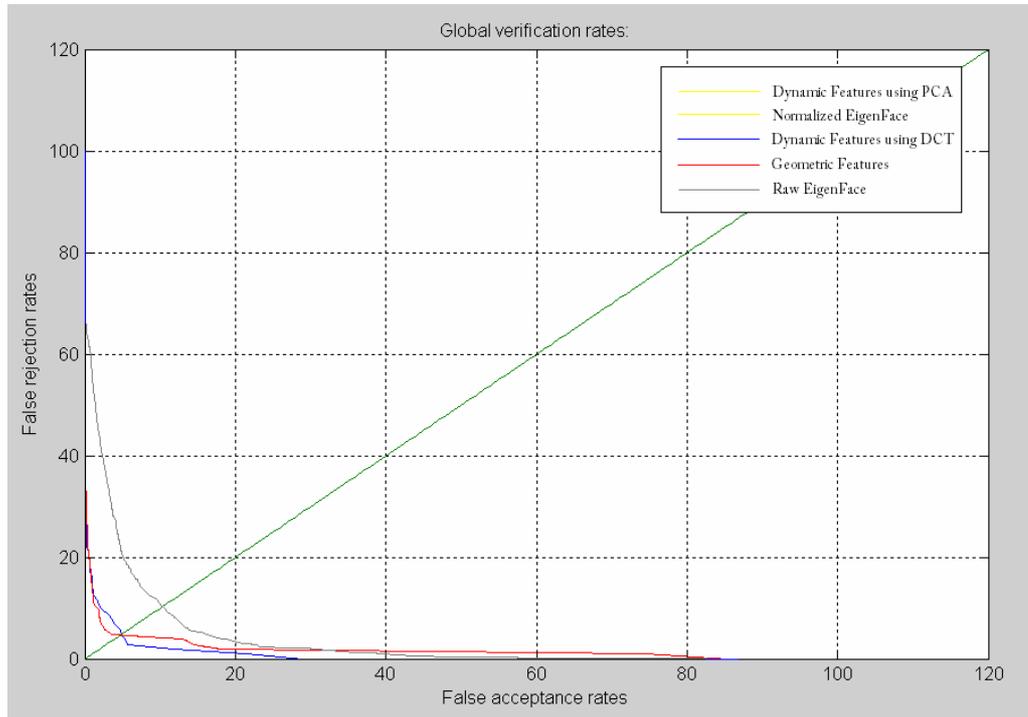
Et la probabilité a posteriori pour la video devient:

$$p(k | \mathbf{X}) \cong \frac{p(k)}{M_{\mathbf{X}}} \prod_{t=1}^T p(\mathbf{x}_t | k)$$

Les tests ont été effectués sur la base de données Italian TV Database. 104 séquences vidéo ont été sélectionnées pour l'apprentissage (8 pour chacun des 13 individus), alors que les 104 autres ont été gardées pour l'évaluation. Les caractéristiques dynamiques ont été extraites et leurs dimensions ont été réduites à l'aide d'un ACP et TCD.



(a)



(b)

Figure 38: Résultats de la reconnaissance a) Identification b) Vérification

La Figure 14 représente les résultats obtenus pour les caractéristiques dynamiques tout d'abord réduites grâce à un ACP, performance les plus intéressantes, ou encore avec TCD ou enfin avec les caractéristiques géométriques normalisées.

## 7. La normalisation des caractéristiques de lèvres

La troisième contribution majeure de cette thèse correspond à introduire une nouvelle méthode de normalisation temporelle pour la variation causée par le mouvement des lèvres pendant le discours. Ainsi, pour une série de plusieurs vidéos dans lesquelles une personne répète la même phrase plusieurs fois, nous avons étudiés le mouvement général des lèvres et sélectionné certaines images clés comme images de synchronisation.

Au début le flux optique est calculé par la méthode de Lucas Kanade frame par frame, (cf. Figure 21) pour toute la vidéo. Après nous calculons la moyenne des vecteurs de mouvement pour chaque frame.

```

for t ← 1 to N - 1
    [um,n,t vm,n,t] = LK(MIt, MIt+1)
    Oft = ∑m=1M ∑n=1N (abs(um,n,t) + abs(vm,n,t))
end

```

$N$  est le nombre de frames dans la vidéo,  $LK()$  calcule le flux optique par la method Lucas Kanade.

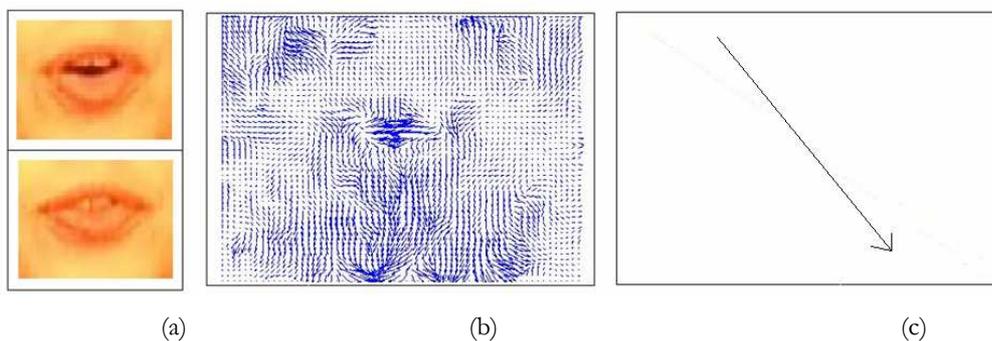


Figure 39: (a) Région de la bouche. (b) LK flux optique (c) vector moyenne

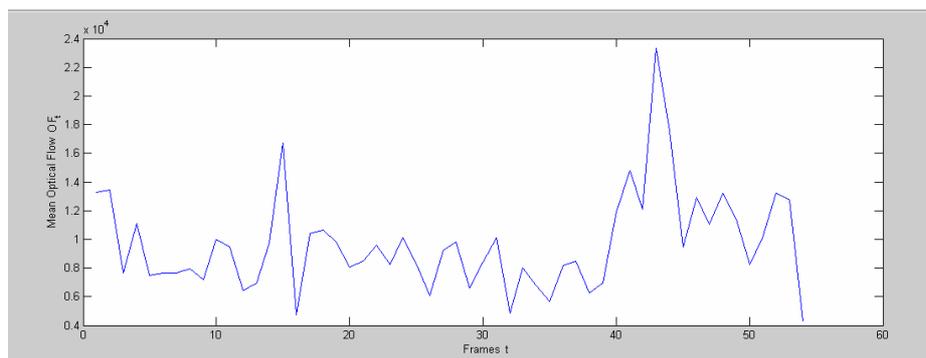


Figure 40: Flux optique moyenne  $Of_i$  pour la Vidéo

L'étape suivante consiste à sélectionner des frames de synchronisation basé sur les flux optique moyenne, si nous sélectionnons les frames qui présentent le mouvement de lèvres maximale, il y a une possibilité que ces frames puissent être proches les unes des autres. Donc nous avons décidé de diviser la vidéo en segments prédéfinis et par la suite sélectionner les frames se rapportant à des maxima locaux telles que les frames de synchronisation.

---

*for*  $t \leftarrow 1$  *to*  $(N - D)$  *with increments of*  $D$   
 $SF_1 = \text{Frame with value } (\max(Of_t \text{ to } Of_{t+D}))$   
*end*  
*where*  $D = \frac{N}{K}$

$N$  est le nombre total de frames dans la vidéo.  $K$  est le nombre de frames de synchronisation.

Après, le reste des vidéo est synchronisé par rapport aux frames de synchronisation de la première vidéo. Un algorithme de recherche basé sur l'erreur quadratique moyenne comme défini ci-dessous est utilisé

*for*  $k \leftarrow 1$  *to* *No of Synchronization Frames*  
*for*  $i \leftarrow 2$  *to* *No of Videos Per Person*  
*for*  $w \leftarrow f(k) - 5$  *to*  $f(k) + 5$   
 $SF_i = \operatorname{argmin} \frac{\sum \sum ((J_{f(k)-1,1})^2 - (J_{f(k)-1,i,w})^2) + \sum \sum ((J_{f(k),1})^2 - (J_{f(k),i,w})^2) + \sum \sum ((J_{f(k)+1,1})^2 - (J_{f(k)+1,i,w})^2)}{(M * N)}$

$SF_i$  est la matrice de sortie ? qui contient les frames de synchronisation pour toutes les vidéos pour une personne donnée.

Enfin toutes les vidéos sont normalisées temporellement basée sur le morphing ? des lèvres. La région de la bouche de « frames consécutifs » [I don't know how to translate that] est d'abord isolée puis transformé en utilisant en le « Mesh morphing » [181].

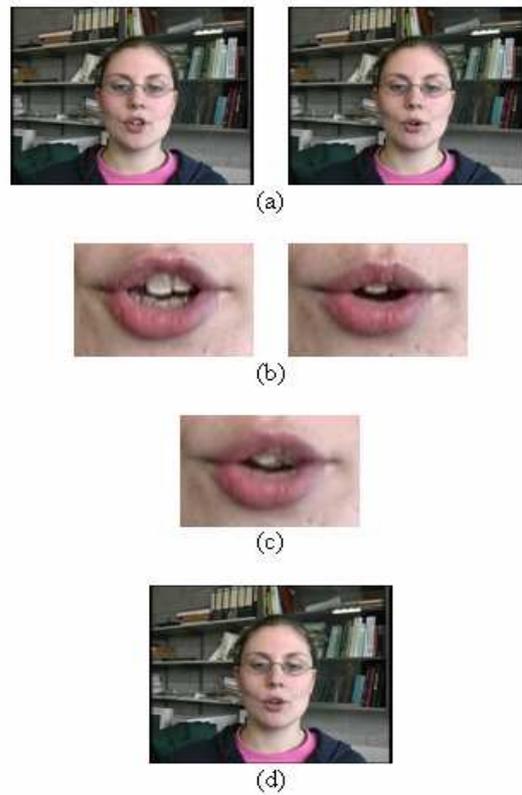


Figure 41:(a) Frames existantes (b) Région de la bouche (c) Région de la bouche apres morphing (d) Frame apres morphing

Les frames sont ajoutés et supprimés en utilisant l'algorithme ci-dessous

#### *Frame Addition*

$$I_i \leftarrow \text{Morph}(I_{i-1}, I_{i+1})$$

#### *Frame Removal*

$$I_{i-1} \leftarrow \text{Morph}(I_{i-1}, I_i)$$

$$I_{i+1} \leftarrow \text{Morph}(I_{i+1}, I_i)$$

$$\text{Delete}(I_i)$$

L'évaluation de notre algorithme de normalisation a été effectuée en utilisant un algorithme spatio-temporel pour la reconnaissance des personnes, proposé par [174]

Il détecte d'abord les contours frames par frames par l'application de la méthode Canny [183]. Puis, les images binaires obtenues sont additionnées temporellement.

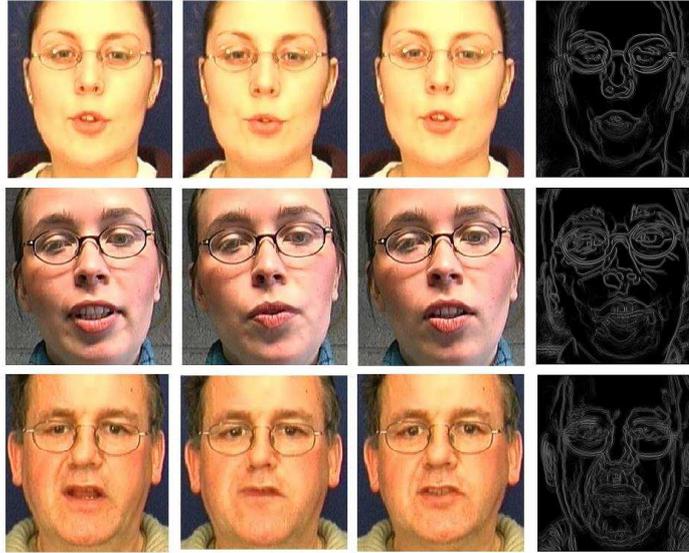


Figure 42: Frames original et and image X-ray.

Pendant la phase d'apprentissage, le système génère des modèles clients basés sur les centres des clusters. Pour la phase de reconnaissance, le système utilise un classificateur des plus proches voisins.

Les tests ont été effectués sur la base de données Valid [65] qui comporte cinq sessions d'enregistrement de 106 personnes. Pour estimer l'amélioration due à notre processus de normalisation, nous avons comparé les vidéos normalisées générées par notre algorithme avec vidéo non-normalisée. Les résultats sont présentés dans le tableau 5.

Méthode	CIR % (1 <sup>er</sup> )	CIR % (5 <sup>ème</sup> )	CIR % (10 <sup>ème</sup> )	EER %
Vidéo normalisées	69.02 %	82.60 %	89.13 %	10.1 %
Vidéos originales	65.21 %	81.52 %	85.86 %	11.9 %

Table 16: Résultats pour la reconnaissance des personnes

## 8. Conclusions

Dans cette thèse nous avons présenté un travail approfondi sur les différents aspects de l'utilisation des caractéristiques des lèvres. Nous avons tout d'abord introduit les notions fondamentales de la biométrie et nous avons par la suite dressé un état de l'art sur l'identification de la personne en utilisant les informations audio-visuelle des lèvres. Après, nous avons présenté un algorithme de détection des lèvres basée sur la fusion de deux techniques de traitement d'image . Ensuite, nous avons présenté un système de reconnaissance de personne basé sur les caractéristiques comportementales des lèvres. Enfin nous avons proposé une méthode de normalisation temporelle pour le discours visuel et étudié ses effets sur la reconnaissance de personne.

---

## References

- [1] F. Matta, "Video person recognition strategies using head motion and facial appearance," University of Nice Sophia-Antipolis, 2008.
- [2] L.L. Mok, W.H. Lau, S.H. Leung, S.L. Wang, H. Yan, "Person authentication using ASM based lip shape and intensity information," International Conference on Image Processing, vol.1, no., pp. 561-564, 2004.
- [3] T. Wark, S. Sridharan, V. Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," Proceedings of Fourteenth International Conference on Pattern Recognition, vol.1, pp.123-125, Aug 1998.
- [4] A.G. de la Cuesta, Z. Jianguo, P. Miller, "Biometric Identification Using Motion History Images of a Speaker's Lip Movements," Machine Vision and Image Processing Conference, pp.83-88, 2008.
- [5] J. Luetttin, N.A. Thacker, S.W. Beet, "Speaker identification by lipreading," Proceedings of Fourth International Conference on Spoken Language, vol.1, pp. 62-65, 1996.
- [6] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in International Conference of Audio- and Video-Based Person Authentication, pp. 260–267, U.K., 2003.
- [7] M.I. Faraj, J. Bigun, "Motion Features from Lip Movement for Person Authentication," 18<sup>th</sup> International Conference on Pattern Recognition, vol. 3, pp.1059-1062, 2006.
- [8] H.E. Cetingul, Y. Yemez, E. Engin, A.M. Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," IEEE Transactions on Image Processing, vol.15, no.10, pp.2879-2891, 2006.
- [9] N. A. Fox, B. O'Mullane, and R.B. Reilly, "The realistic multi-modal VALID database and visual speaker identification comparison experiments," 5<sup>th</sup> International Conference on Audio- and Video-Based Biometric Person Authentication, 2005.
- [10] K. Michael, W. Andrew, and T. Demetri, "Snakes: active Contour models," In Proc. International Journal of Computer Vision, vol. 1, pp. 259-268. 1987.
- [11] N. S. Thejaswi and S. Sengupta, "Lip Localization and Viseme Recognition from Video Sequences," Fourteenth National Conference on Communications, India, 2008.
- [12] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," In Proc. of International Conference on Image Processing, vol. 3, pp. 173–177, U.S.A. 1998.
- [13] W.C. Yau, D.K. Kumar, and H. Weghorn, "Visual Speech Recognition Using Motion Features and Hidden Markov Models," In Proc. of International Conference on Computer Analysis of Images and Patterns, pp. 832-839, Austria, 2007.
- [14] C. Ding, and H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," In Proc. of 2<sup>nd</sup> IEEE conference on Computational Systems Bioinformatics, pp. 523-528, U.S.A., 2003.
- [15] F. Matta, J-L. Dugelay, "Tomofaces: eigenfaces extended to videos of speakers," In Proc. of International Conference on Acoustics, Speech, and Signal Processing, pp.1793-1796, 2008.

- [16] G. Wolberg, "Recent Advances in Image Morphing," In Proceedings of the Conference on Computer Graphics international, USA, 1996.
- [17] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 8, pp. 679-698, 1986.
- [18] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition," in Proceedings of the IAPR Workshop on Machine Vision Application, pp. 318-321, Japan, 2002.
- [19] Y.-P. Guan, "Automatic extraction of lips based on multi-scale wavelet edge detection," *IET Computer Vision*, vol.2, no.1, pp.23-33, March 2008.
- [20] T.F. Chan, L.A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol.10, no.2, pp.266-277, 2001.