

Robust Visual Speech Recognition Using Optical Flow Analysis and Rotation Invariant Features

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Ayaz Ahmed Shaikh

B.Eng. (Electronic), M.Phil (Communication Systems & Networks)

Mehran University of Engineering and Technology

School of Electrical and Computer Engineering
Science, Engineering and Technology Portfolio

RMIT University

August 2011

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Ayaz Ahmed Shaikh

Dated:

Acknowledgement

I particularly wish to thank the following people:

- Most especially Professor Mike Austin and Dr. Paul Beckett. Professor Austin for his mentoring and encouragement to submit this thesis, his generous support and his availability being invaluable for me, as was the support and encouragement from Dr. Beckett throughout my PhD candidacy.
- Dr. Jayavardhana Gubbi, of the University of Melbourne, who generously devoted considerable time for guidance, and was always available for ongoing discussions and moral support.
- Associate Prof. Dr. Dinesh K Kumar for providing the ideas, inspiration, motivation and the necessary support to commence the research in the field of visual speech recognition.
- A special thank you to Dr. Wai Chee Yau for providing the dataset.
- My colleagues Tayab, Arun, Fawaz, Khalifa, Premith, Noura and Kashfia for advice and fruitful exchange of ideas.
- Additionally, I thank Bryan Glynn for carefully reading the manuscript and for giving helpful hints and suggestions in writing this thesis.

I acknowledge Mehran University of Engineering and Technology, Pakistan, for their financial support, making it possible for me to commence this research and to proceed to its eventual completion.

Finally, I want to thank my family. I am deeply grateful to my wife Shazia, whose patience and support encouraged me daily. In addition, I like to thank my mother and my parents-in-law for their continued prayers and encouragement.

Ayaz Shaikh

Keywords

Lip-reading, visual speech recognition, Speech reading, visual feature extraction, temporal speech segmentation, classification of visual features, Support Vector Machines

Table of Contents

Declaration.....	ii
Acknowledgement	iii
Keywords.....	iv
Table of Contents.....	v
Abstract.....	ix
List of Tables	xii
List of Figures	xiii
Abbreviations and Acronyms	xv
Chapter 1	1
Introduction.....	1
1.1 Pattern Recognition	3
1.1.1 Feature Extraction.....	4
1.1.2 Classification	4
1.2 Motivation.....	5
1.3 Scope of Thesis.....	5
1.4 Aims and Objectives of the Research	6
1.5 Outline of Thesis	8
1.6 Publications Resulting from this Research.....	9
Chapter 2.....	11
An Overview of AVSR Systems.....	11
2.1 Introduction	11
2.2 Non-Audio Speech Modalities	12
2.3 The Review of AVSR and VSR.....	17
2.4 Anatomy of the Human Speech Production System.....	20
2.5 Linguistics of Visual Speech	22
2.6 Visual Speech Perception by Humans.....	23
2.7 Speech Reading Proficiency	24
2.8 Significance of Facial Parts in Lip-reading.....	24
2.9 Basic Components of Visual Speech Recognition System.....	25

2.10	Brief Description of Proposed VSR System	27
2.11	Summary	30
Chapter 3		31
Image and Video Preprocessing.....		31
3.1	Visual Front End	31
3.1.1	Face Detection	33
3.1.1	Spatial Segmentation	34
3.2	Choice of Utterances.....	36
3.2.1	Dataset Exploited for this Study	36
3.3	Temporal Segmentation	39
3.4	Image Noise Reduction	46
3.5	Issues that Need to be Considered for the Development of Visual Speech Recognition.....	46
3.6	Summary	48
Chapter 4		49
Mouth Movement Representation Using Optical Flow Based Motion Template		49
4.1	Introduction	51
4.2	Development of Optical Flow Based Motion Templates	52
4.2.1	Optical Flow Motion Estimation	53
4.2.2	Gradient Based Approach	53
4.2.3	Global Methods.....	54
4.2.4	Local Methods.....	55
4.2.5	Optical Flow Computation Used in this Research.....	57
4.2.6	Non Overlapping Block Based Approach	59
4.2.6.1	Block Optimization	62
4.2.7	Normalization of speed of speech	63
4.3	Summary	64
Chapter 5		65
Mouth Movement Representation Using Directional Motion History Images.....		65
5.1	Motion Representation Theory	66
5.1.1	Basics of Motion History Image (MHI)	66
5.1.2	Development of Directional Motion History Images (DMHIs).....	69

5.2	Advantages and Disadvantages of Directional Motion History Images.....	74
5.3	Summary	77
Chapter 6		78
Visual Speech Feature Extraction and Classification		78
6.1	Classification of Visual Feature Extraction.....	79
6.1.1	Shape Based Feature Extraction	79
6.1.2	Appearance Based Feature Extraction.....	81
6.1.3	Hybrid Features.....	82
6.1.4	Motion Based Feature Representation.....	82
6.2	Visual Speech Feature Extraction	84
6.2.1	Zernike Moments.....	86
6.2.1.1	Square-to-Circular Image Coordinate Transformation	86
6.2.1.2	Computation of ZM.....	88
6.2.1.3	Rotation Invariance of ZM	89
6.2.2	Hu Moments	91
6.3	Classifier for Lip-reading	94
6.4	Support Vector Machine.....	95
6.4.1	Optimal Separating Hyper-plane	96
6.4.2	The Non-Separable Data: Soft Margin Hyper-plane	99
6.4.3	Kernel Trick	100
6.4.4	Multiclass Kernel Machines	103
6.5	Summary	104
Chapter 7		106
Experimental Results		106
7.1	Experimental Setup 1: Performance Evaluation of an Optical Flow Based Motion Template 107	
7.1.1	Methodology for Classification	107
7.1.2	Viseme Classification Results	109
7.2	Experimental Setup 2: Performance Evaluation of DMHIs.....	116
7.2.1	Comparing the Performance of DMHI vs MHI	118
7.3	Summary	119
Chapter 8.....		121

Conclusions and Future Directions	121
8.1 Conclusions	121
8.2 Future Work	123
References	126

Abstract

The focus of this thesis is to develop computer vision algorithms for visual speech recognition. Potential applications of such a system include the lip-reading mobile phones, human computer interface (HCI) for mobility-impaired users, in-vehicle systems, robotics, surveillance, improvement of speech based computer control in a noisy environment and for the rehabilitation of the persons who have undergone a laryngectomy surgery, or older citizens who require extensive effort to speak, but can move mouth easily rather than actually pronouncing.

In the literature, there are several models and algorithms available for visual feature extraction to enhance the performance of existing acoustic speech recognition systems. These features are extracted from static mouth images and characterized as appearance and shape based features. However, these underlying methods rarely incorporate the time dependent information of mouth motion and dynamics. Nevertheless, there are no commercially available visual speech recognition systems. This reflects the need to further address the research challenges.

This dissertation presents two optical flow based approaches of visual feature extraction, which capture the dynamics of mouth motions in a video during speaking. These dynamics of mouth motion are used to classify and identify the visemes in a video. The motivation for using motion features is, because the human perception of lip-reading is concerned with the temporal dynamics of mouth motion. The majority of existing speech recognition systems is based on audio-visual signals and has been developed for speech enhancement and is prone to acoustic noise. Considering this problem, the main focus of this research is to investigate and develop a visual only speech recognition system which should be suitable for noisy environments.

The first approach is based on extraction of features from the optical flow vertical component. Preliminary experiments have revealed that the salient speech features are available in the vertical component of optical flow, whereas the horizontal features have a smaller contribution in normal speech. The optical flow vertical component is

decomposed into multiple non-overlapping fixed scale blocks and statistical features of each block are computed for successive video frames of an utterance. A common problem with existing systems is their high error rates due to the variation in speed of speech and the style of people to speak, especially across national and cultural boundaries, making them very subject dependent. To overcome this issue, the proposed system is made robust to the speed of speech by normalizing each utterance to a fixed number of features.

In the second approach, four directional motion templates based on optical flow are developed, each representing the consolidated motion information in an utterance in four directions (*i.e.*, up, down, left and right). This approach is an evolution of a view based approach known as motion history image (MHI) which implicitly encodes the spatio-temporal components of an image sequence into a grey scale scalar valued MHI. One of the main issues with the MHI method is its motion overwriting problem because of self-occlusion which happens when the motion is repeated in the same location at different times within the utterance. DMHIs seem to solve this issue of overwriting, that technique consists of four directional motion history images, rather than a single image as used in the MHI technique. Two types of image descriptors, Zernike moments and Hu moments are used to represent each image of DMHIs. A support vector machine (SVM) was used to classify the features obtained from the optical flow vertical component, Zernike and Hu moments separately. For identification of visemes, a multiclass SVM approach was employed.

A video speech corpus of seven subjects was used for evaluating the efficiency of the proposed methods for lip-reading. The experimental results demonstrate the promising performance of the optical flow based mouth movement representations. The advantages and limitations of both the techniques for visual speech recognition were identified and validated through experiments. Performance comparison between DMHI and MHI based on Zernike moments, shows that the DMHI technique outperforms the MHI technique.

A video based *ad hoc* temporal segmentation method is proposed in the thesis for isolated utterances. It has been used to detect the start and the end frame of an utterance from an

image sequence. The technique is based on a pair-wise pixel comparison method. The efficiency of the proposed technique was tested on the available data set with short pauses between each utterance.

List of Tables

Table 3.1: Fourteen <i>visemes</i> defined in MPEG-4 and the average number of frames for each viseme of the used dataset.	37
Table 3.2: Results of temporal segmentation for 14 visemes (three epochs)	41
Table 3.3: Average Frame Error Rate of 10 Epochs at Start of an Utterance	44
Table 3.4: Average Frame Error Rate of 10 Epochs at End of an Utterance.	45
Table 4.1: Average vertical and horizontal movement for an utterance.	60
Table 4.2: Average classification results of seven different block sizes.	62
Table 6.1: Average classification results of three different numbers of ZMs.	90
Table 6.2: Zernike Moments from 0 th to 14 th order.	91
Table 7.1: Average Classification Results of 14 Visemes in Terms of Specificity, Sensitivity and Accuracy, block size 240×40. (All values in %).....	110
Table 7.2: Average classification results of individual one-vs-rest binary class SVM for 14 visemes (All values in %)	112
Table 7.3: Confusion Matrix for 14 visemes using hierarchical multi class SVM	115
Table 7.4: Classification results of individual one-vs-rest class SVM using ZM (All values in %)	117
Table 7.5: Classification results of individual one-vs-rest class SVM using Hu moments (All values in %)	117
Table 7.6: Comparison of DMHI vs MHI (All values in %).	119

List of Figures

Figure 1.1: A general pattern recognition procedure	4
Figure 2.1: Ultrasound and video camera based speech recognition system	13
Figure 2.2: Placement of NAM microphone for silent communication. Vibration of the vocal-tract resonance is captured from the tissues, generated by airflow noise in the constricted laryngeal airflow.	14
Figure 2.3: Human speech production system	21
Figure 2.4: Different visemes but same phonemes. Image (a) shows viseme /m/ and image (b) shows /n/ while these are acoustically similar phonemes.	22
Figure 2.5: Basic Components of a VSR system.	26
Figure 2.6: A block diagram of the proposed Visual Speech Recognition system.....	28
Figure 3.1: Example images of seven subjects with different skin tones.	38
Figure 3.2: Results of Temporal Segmentation (a) Squared mean difference of accumulative frames intensities (b) Result of smoothing data by moving average window (c) Result of further smoothing by Gaussian filtering (d)Segmented data (blue blocks indicate starting and ending points)	42
Figure 3.3: Results of Temporal Segmentation of all 14 visemes for a single user using the proposed method.	43
Figure 4.1: Example of optical flow computation using two consecutive images.	59
Figure 4.2: System flow diagram of non-overlapping block based approach.	61
Figure 4.3: Vertical block based approach.....	62
Figure 4.4: Direction of movement of each muscle and the block arrangement used for optical flow feature extraction.	63
Figure 5.1: Development of MHI images for two utterances /a/ and /m/.....	68

Figure 5.2: Conceptual framework of optical flow vector separation into four directions.	71
Figure 5.3: (a) Four directional motion history images (DMHIs) of first three frames of an utterance /a/, (b) Complete DMHIs of an image sequence of an utterance /a/.	73
Figure 6.1: Shape based features represent the physical shape of the mouth such as height and width, given in (a). Appearance based features consider the complete ROI, shown in (b).	80
Figure 6.2: The square-to-circular image coordinates transformation of a motion template before Zernike moments computation.	87
Figure 6.3: Representation of linearly separable data of two classes, class 1 is represented by grey dots and class 2 is represented by pink dots. An optimal hyper plane separates the two classes. The four support vectors are shown as black and dark pink dots.	97
Figure 7.1: Development of optical flow vertical component based motion template (block size is 240×40).	107
Figure 7.2: Representation of vertical component division in rectangular blocks.	111
Figure 7.3: Development of optical flow vertical component based motion template (block size is 48×40).	111
Figure 7.4: Cross validation Results (a) Accuracy, (b) Specificity and (c) Sensitivity.	114

Abbreviations and Acronyms

AAM	Active Appearance Model
ANN	Artificial Neural Networks
ASM	Active Shape Model
ASR	Automatic Speech Recognition
AVSR	Audio Visual Speech Recognition
DCT	Discrete Cosine Transform
DMHI	Directional Motion History Image
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
EGG	Electroglottograph
EMA	Electromagnetic Articulography
EMG	Electromyography
FFT	Fast Fourier Transform
GPU	Graphics Processing Units
HM	Hu Moment
HMHH	Hierarchical Motion History Histogram
HMM	Hidden Markov Model
HSV	Hue Saturation Value
ICA	Independent Component Analysis
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
LGE	Lip Geometry Estimation
MHI	Motion History Image

MLP	Multilayer Perceptron
MMHI	Multilevel Motion History Image
MRF	Markov Random Field
MSA	Multi-scale Spatial Analysis
NAM	Non-Audible Murmur
PCA	Principal Component Analysis
RBF	Radial Basis Function
RGB	Red Green Blue
ROI	Region of Interest
SVM	Support Vector Machine
VSR	Visual Speech Recognition
WER	Word Error Rate
ZM	Zernike Moment

Chapter 1

Introduction

Researchers have always aimed to develop a sophisticated machine that can perform tasks like human beings. The motivation for this effort is from the practical need of intellectual tasks to be accomplished in an efficient way. These intellectual tasks are based on realization, evaluation and interpretation of information from sensors. This can be resembled to perception. Perception is the process in human beings of attaining knowledge about the environment, processing it and reacting to it accordingly [1]. It depends on complex functions of the nervous system, but is performed effortlessly. It is nearly impossible to know the exact intrinsic mechanism of perception. However, scientists have been attempting to develop computer algorithms that replicate human intelligence and the field is commonly known as artificial intelligence. Pattern recognition is the core area of artificial intelligence that enables machines to accept external inputs and react accordingly. Due to the unknowns behind such understanding, the pattern recognition field provides the facility to exploit the abstract mathematical model of perception.

A human's ability to lip-read by perceiving the lips, teeth and tongue is a hook of the chain of perception. It provides useful visual information about speech. Around four decades ago the studies by Hazard [2] and Jeffers and Barley [3, 4] demonstrated that a human listener always gets advantages from visual cues, such as lip and tongue movements and also from facial expressions and hand gestures to increase the level of speech intelligibility in noisy conditions. The process of using visual only modality is often referred to as lip-reading or speech reading, that is to perceive what someone is saying by watching his lip movements. Motivated by this ability in humans, researchers have aimed to develop an audio-visual speech recognition (AVSR) system.

The most natural and easy way of communication between humans is speech. Due to its naturalness, much research has been conducted to develop speech based human computer

interfaces (HCIs) [5-7], where speech is the basic mode of interaction with machines. While these systems have shown promising success in well defined applications such as call centres, dictations or call routing in reasonably noiseless environments, they are yet to attain the target where these systems can be deployed anywhere and everywhere. The major reason behind this is the susceptibility to noise, which degrades the performance of audio speech recognition systems. Noise robust algorithms such as, feature compensation [8], implementations of microphone arrays [9], nonlocal means denoising method [10], variable frame rate analysis [11] or noise adaptation algorithms [9, 12, 13] have presented significant improvement in speech recognition under noisy environment, however, such algorithms are not exactly prone to noise due to the difficulty in modelling the random characteristic of non-stationary noise.

In recent years, an alternative method to overcome the limitations of audio speech recognition has gained more interest. This is by the use of a multimodal conversational system. Besides speech input, multimodal conversational systems also support inputs from other modalities such as visual [14] and facial muscle activity [15] to identify the utterances. Compared to the conventional speech-only interfaces in spoken dialog systems, multimodal conversational interfaces provide better interpretation of user input due to mutual disambiguation among complementary modalities [16]. However, sensor based systems have an obvious disadvantage in that they require the user to place the sensors on to the face. One of the limitations of the muscle monitoring approaches is its low reliability low reliability. Visual modality based systems are non intrusive and users are not require to place the sensors on face and hence have emerged as more desirable options.

Visual speech information is being used to increase the robustness of speech recognition. Much research has been conducted where visual signals are fused with existing audio-only speech recognition systems (ASR) to augment the audio recognition [14, 17-22]. The McGurk effect [23] demonstrates that inconsistency between audio and visual information can result in perceptual confusion. Visual information plays an important role especially in noisy environments. AVSR systems are useful for many applications. However, such systems are not suitable for people with hearing and speech impairments.

Contribution of audio features in speech recognition systems still plays an important role than visual features. However, in some cases, it is hard to extract pertinent information from the acoustic signal. There are many applications in which it is essential to identify the desired speech signal under extremely adverse acoustic environments such as detecting a person's speech through a glass window, from a distance or a person speaking in a very noisy crowd or in a factory. In addition, if there is no assisting sign language for a TV broadcast or speeches, visual information is the only source of information for hearing impaired people. In such applications, the performance of traditional speech recognition is very limited.

There are a few works focusing on lip movement representations for speech recognition solely with visual information [24-27]. This is known as visual speech recognition (VSR). This research is an incremental effort in the field of VSR. Generally, VSR systems are comprised of several pattern recognition stages; among them the most important step deals with the computation of the visual features that are extracted in order to produce a compact representation that describes either the visual appearance or the shape of the lips in each image. The result of the feature extraction is used to generate feasible visual speech models that represent the lip motions during the speech process. Hence the main focus of this research is visual feature extraction.

1.1 Pattern Recognition

Pattern recognition deals with the assignment of some label to a given input value through their observable information, such as intensity values of an image, frequency components available in an EMG signal. Generally pattern recognition system is divided into three main components, as shown in Figure 1.1. A sensor (camera in this research) transforms the observable information such as images or sounds into signals that can be analysed by a computer system. A feature extraction component computes the signal properties/ features suitable for classification. Finally the extracted features are assigned to a classifier to sense the class of the signal based on certain types of measures, such as distance, likelihood and Bayesian.

1.1.1 Feature Extraction

The main objective of the feature extractor is to transform the input signals into a set of properties which are very similar to the signals of same category and are very distinctive to the signals of different category. This leads to the idea of obtaining the robust features that are invariant to irrelevant transformation such as rotation and scaling of the input signal. In this research, the rotation and scale invariant features are obtained to compensate the varying view angle orientation and distance of camera from the subject.

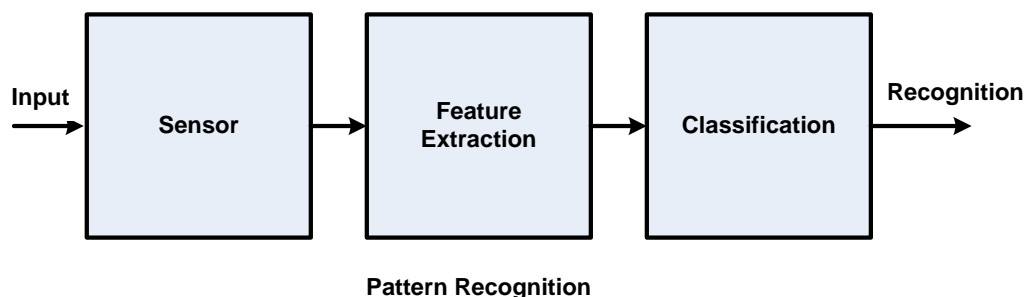


Figure 1.1: A general pattern recognition procedure

1.1.2 Classification

The task of a classifier is to assign an input feature vector to one of the existing classes, based on specific classification measures. Conventional classification measures include distance (Mahalanobis or Euclidean distance), likelihood and Bayesian *a posterior* probability. The decision boundaries generated by these measures are linear and hence are known as linear classifier. However, such type of classifiers are limited to solve the linearly separable data and are unable to handle complex non-linear decision boundaries and have little computational flexibility. Support vector machine (SVM) is a classifier with non-linear formulation, developed with the idea of classification which afford dot-products can be computed efficiently in higher dimensional feature space [28]. Classes which are non-linear and inseparable in the original space can be linearly separated by mapping them to a new higher dimensional feature spaces. SVM can handle the classes with complex non-linear decision boundaries because of this advantage SVM is the choice of this research as a visual speech classifier.

1.2 Motivation

Studies of human speech production and perception have shown that the visual movement of the speaker's lips and face are an important factor in human communication. Hearing impaired persons extensively use the visual cues and some individuals perform lip-reading to the level which enables almost perfect speech perception [29]. Normal hearing subjects also get the benefit of visual information to improve their speech perception, especially in noisy environments [30]. This is the motivation behind visual speech recognition. Reisberg et al. [31] have shown that normal-hearing subjects who see the talker's face perceive speech more accurately, even in noise free environments. Motivated by these psychological studies, several researchers [14, 18, 22, 32-34] have developed speech reading systems, mainly to demonstrate the potential use of visual information to improve the robustness of acoustic speech recognition systems in noise. While these systems have validated the benefit of visual speech information, there is still much discussion require about determining which visual features are important for visual only speech reading, how to extract them and how to represent them automatically in a robust manner.

1.3 Scope of Thesis

Obtaining robust visual speech features is a difficult task due to the varying appearance of different persons and due to appearance variability during speech production. Varying illumination and orientation of the face cause further difficulties in feature extraction. For real world applications, whether it is an office, factory or a railway station the environmental noise creates additional problems. Whereas the extraction of acoustic speech features is fairly established, important visual speech features for lip reading are relatively unknown and the investigation of different feature extraction methods is still subject to ongoing research. In an attempt to solve this problem, the work in this thesis has solely concentrated on a visual only signal and to avoid environmental noise, the acoustic signal is not used. The work has focused on researching and developing methods of robust visual feature extraction.

Considering the above facts, the scope of this thesis was to investigate the important visual features of speech, resembling to human perception. The motion based optical flow estimation method was adopted, which gives the motion estimation between consecutive images, analogous to human perception. The emphasis of the method is to provide robust features which perform well for different subjects without the use of artificial markers and sensors on the face of a subject.

1.4 Aims and Objectives of the Research

Inspired by the human perception of lip-reading, the general objective of this research can be defined as:

“To design a visual only speech recognition system based on robust motion features that can recognize a limited vocabulary dataset at viseme level”

One common difficulty with AVSR systems is their high error rates due to the large variation in the way people speak, especially across national and cultural boundaries [35], making them very subject dependent and therefore unsuitable for multiple subject applications. There is also the difference between the speeds of speaking of a subject when repeating an utterance. To overcome this issue, the proposed system is made robust to the speed of speech by normalizing each utterance to a fixed number of features. The other concern with these systems is their lack of robustness against variations in lighting conditions plus the angle and distance of the camera along with their sensitivity to variations in the colour and texture of the speaker’s skin. Both of these problems will prevent the wide deployment of these systems. Techniques previously reported in the literature exhibit both of these problems [26]. Further, the majority of video based techniques reported to date are based on multimodal configuration and have been developed for speech enhancement, not as stand-alone visual speech recognition systems [34, 36, 37].

In this research, two novel methods of feature extraction are proposed that overcome the above shortcomings. These methods depend solely on visual features so that acoustic noise will not affect the system. The proposed techniques use optical flow estimation that

measure the movement of a mouth in consecutive images and which are insensitive to background and lighting conditions. In the first approach it is concluded that the optical flow vertical component retains most of the speech information compared to the horizontal component. Hence, the optical flow vertical component has been divided into blocks that represent different sections of the mouth. The average directionality of each block of optical flow has been used as the feature to represent the movement.

In the second approach, optical flow based four directional templates are developed, each representing the consolidated motion information of an utterance in a particular direction (up, down, left and right). This approach is an enhanced technique of a view based approach known as motion history image (MHI) which implicitly encodes the temporal component of an image sequence into a scalar valued motion template. One of the key constraints of the MHI method is its motion overwriting problem due to self-occlusion which happens when the motion is repeated in the same location at different times within the utterance. DMHIs seem to solve this issue of overwriting, that technique consists of four directional motion history images, rather than a single image as used in the MHI technique. Two types of image descriptors, Zernike moments and Hu moments, are used to represent each image of DMHIs. The selection of these features is based on their robustness to the rotation and scaling invariance which helps in varying view angle and distance of camera from subject.

Support Vector Machine (SVM), a state-of-the-art classifier, has been used to classify the features of visemes obtained from the optical flow vertical component, Zernike and Hu moments separately and for identification a multi-class SVM approach has been employed.

Another shortcoming with automatic analysis of video data is the need for manual intervention due to the need for segmentation of the video. While audio assisted video speech analysis uses the audio cues for segmentation, this is not possible in video only speech recognition. One achievement of this work is that automatic segmentation of the visual speech data has been achieved solely based on video signals. The automatic

segmentation of the video data is based on a pair-wise pixel comparison method [38] to identify the start and end frame of each utterance.

1.5 Outline of Thesis

This thesis has mainly focused on feature extraction and temporal segmentation algorithms and then classification of the extracted features has performed. The outline of the thesis is as follows:

Chapter 2 reviews the literature of speech recognition, including non-audio speech modalities and audio-visual speech recognition, detailing its history and progression. This chapter also contains the physiological and linguistic aspects of speech production and visual speech perception of humans. This chapter describes the basic components of a visual speech recognition system and finally a brief description of the proposed VSR system is presented.

Chapter 3 is concerned with image and video pre-processing to make the videos suitable for visual feature extraction. It gives a brief introduction of visual front end processing including a literature review of face and region of interest (ROI) extraction, and provides a description of the dataset used in this study. It also describes a new *ad hoc* method for detection of the start and end frame of an utterance for the used dataset.

Chapter 4 presents one of the main contributions of this research. Generally, the visual features for lip-reading are extracted from static mouth images and characterized as appearance and shape based features. As opposed to shape and appearance based features that describe the underlying static shapes of mouth. Time-based motion features directly represent the dynamics of mouth movement across the video frames which are analogous to human perception. The optical flow based motion features are extracted, based on these features a novel block based approach to represent the mouth motion is presented. Variations in speed of speech between inter and intra subject can give rise to inexact viseme recognition, this variation is normalized by two phase approach.

Chapter 5 presents a further contribution of this research and describes the basic theory of motion history image. This leads to the development of optical flow based directional motion history images (DMHIs) which is a novel approach to represent the mouth movement by four directional motion templates. Advantages and disadvantages of the proposed technique are then discussed.

Chapter 6 provides a review of the visual feature extraction techniques for VSR, it describes two important rotation and scale invariant image descriptors, Zernike and Hu moments, which were used to extract the important features of DMHIs. Finally the classification of visual speech is broached, with a detailed description and justification of Support vector machine (SVM) classifier as a choice.

Chapter 7 reports on the experimental setup and methodology for classification. This chapter also evaluates the performance of the proposed motion templates computed by the optical flow vertical component and the directional motion history images. In addition, the results of proposed DMHI technique are compared with the underlying MHI technique, the results indicated that the performance of DMHI was better than MHI.

Chapter 8 concludes the thesis and provides the future directions in this research topic.

1.6 Publications Resulting from this Research

Journals

1. **A. A. Shaikh**, D. K. Kumar, & J. Gubbi, “Visual Speech Recognition Using Optical Flow and Support Vector Machines”. *International Journal of Computational Intelligence and Applications*, 10, **pp. 167-187**. 2011 (ERA-A).
2. **Ayaz A. Shaikh**, D. K. Kumar, & J. Gubbi “Automatic Visual Speech Segmentation and Recognition Using Directional Motion History Images and Zernike Moments”, Submitted to The Visual Computer, (ERA-B).

International Conferences

3. **Ayaz A. Shaikh**, Dinesh K. Kumar, Premith Unnikrishnan, Jayavardhana Gubbi, “Visual Speech Recognition using Directional Motion History Images”, published in Proceedings of the International Conference on Intelligence and Information Technology (ICIIT’10), Lahore, Pakistan. 28-30 October, 2010. ISBN: 978-1-4244-8138-5.
4. **Ayaz A. Shaikh**, Dinesh K. Kumar, Wai C. Yau, M. Z. Che Azemin, Jayavardhana Gubbi, “Lip Reading using Optical Flow and Support Vector Machines”, published in Proceedings of The 3rd International Conference on Image and Signal Processing (CISP’10), Yantai, China, 16-18 October, 2010. ISBN: 978-1-4244-6514-9.

Workshop Posters

5. **Ayaz A. Shaikh**, Jayavardhana Gubbi, Dinesh K. Kumar, Marimuthu Palaniswami, “Robust motion features for visual speech recognition”, Poster Presentation in the 5th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, 7-10 December, 2009.
6. **Ayaz A. Shaikh**, Dinesh K. Kumar, Jayavardhana Gubbi, “Visual Speech Recognition Using Optical Flow”, HCSNet Workshop on Movement and Motion Capture, Macquarie University, Sydney, Australia 25-26 September, 2009.

Chapter 2

An Overview of AVSR Systems

Audio visual speech recognition (AVSR) is a combination of diverse research fields such as linguistic, physiology of human speech production, and psychology of human perception. In addition to this, pattern recognition, video processing and computer vision fields are also incorporated in the development of AVSR systems. It is important for researchers in the field of AVSR to have some basic knowledge of linguistic, human speech production and of human perception. They should have a sound knowledge about the key elements of the other fields mentioned is a necessity, so that computer based visual speech recognition can be optimized up to maximum possible level of human visual speech perception.

This chapter provides a complete overview of the AVSR and VSR systems, before discussing the AVSR brief overview of speech recognition by non speech modalities is presented. The chapter then focuses on human speech production and linguistics of visual speech, followed by the human visual speech perception. Pertinent regions of the face that contain most important visual cues are also discussed. Finally the chapter highlights the basic components of the general visual speech recognition system and it is followed by a brief review of the proposed visual speech recognition system.

2.1 Introduction

Human speech perception is greatly improved by observing a speaker's lip movements as distinct from listening to the voice [39]. Mainstream automatic speech recognition (ASR) systems have focused exclusively on the latter: the acoustic signal. Recent advances have led to purely acoustic-based ASR systems yielding excellent results in quiet or noiseless environments [40]. As a result, these systems have been used for a variety of applications, such as in call centres, car navigation systems, audio based phone dialling,

dictation and translation assistance, and including applications from speech and language technologies to applications in intelligence and military services. In the real world however, their performance drops dramatically due to the presence of noise such as:

- in a typical office environment: human conversations;
- in industry: the noise of machinery;
- on the road: the noise of traffic;
- at a railway station: the noise of trains as well as the announcements of train arrivals and departures.

Noise robust algorithms such as, feature compensation[8], feature-normalization algorithms [13], variable frame rate analysis [11], microphone arrays [9] and other approaches [12] have demonstrated limited success in this regard.

To overcome this limitation, non-audio speech modalities have been considered, such as visual information [41], facial plethysmograms measuring intra-oral pressure and surface electromyography (sEMG) signals of a speaker's facial muscles[42, 43], to augment acoustic information [14]. These systems require sensors to be placed on the face of the speaker and are thus intrusive and impractical in most situations, whereas audio-visual systems are not suitable for people who are unable to produce sound due to speech impairments. In such situations visual only speech recognition systems are the solution. Visual speech recognition is the core focus of this dissertation. In past decades a lot of research effort has been applied to the development of visual based speech recognition systems. Systems that recognize speech from the shape and movement of the speech articulators such as the lips, tongue and teeth of the speaker have been considered, to overcome the shortcomings of other speech recognition modalities [26].

2.2 Non-Audio Speech Modalities

A number of options with non-audio speech modalities have been proposed to overcome the shortcomings of speech recognition systems and represented as the silent speech interface (SSI) [44]. A brief review of these SSIs is given below:

- Electromagnetic Articulography (EMA) is a sensor based movement capturing technique which captures the movement of fixed points on the articulators. The shape of the vocal tract is an important part of speech production. Various researchers have considered monitoring the movement of a set of fixed points within the vocal tract by implanting the coil sensors. Fagan et al. [45] investigated a system in which the magnetic sensors were pasted inside the mouth of a subject on the tongue, lips and teeth, and a set of six dual axis magnetic sensors were fixed on a pair of spectacles. In order to measure the performance of the system, the subject was asked to repeat a set of 9 words and 13 phonemes, each ten times. A 90% recognition rate was achieved under laboratory conditions.
- The most important articulator of the speech production system is the tongue, which is almost invisible while speaking. In the Ouisper project [46], an ultrasound imaging technique was used to capture the movement of the tongue during speaking. Ultrasound imagery is a non invasive and clinically safe procedure. An ultrasonic probe placed under the chin can provide a partial view of the tongue surface in the mid-sagittal plane. In their work the ultrasound imaging system combined with a standard video camera focused on the speaker's lips is used. Visual features from these two modalities are used to drive a speech synthesizer, known as a “silent vocoder”, as illustrated in Figure 2.1.

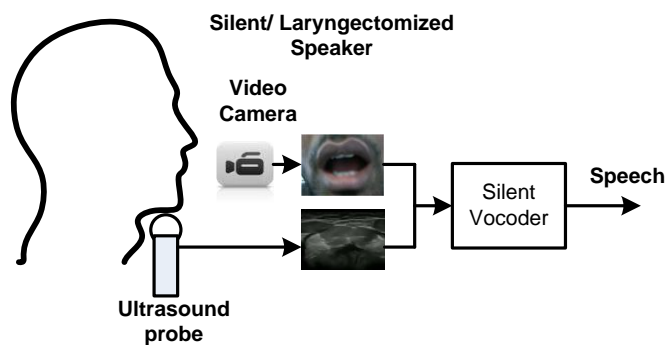


Figure 2.1: Ultrasound and video camera based speech recognition system

- NAM (Non-Audible Murmur) systems are based on a special acoustic sensor (microphone) attached to the speaker's skin, just behind and below the ear (see

Figure 2.2) over the soft tissue in the orofacial region. It senses the low amplitude sounds produced by laryngeal airflow noise and its resonance in the vocal tract [47, 48]. The sensor is capable of sensing the low amplitude sounds that can hardly be perceived by nearby listeners, and is insensitive to environmental noise. Insensitivity to noise has motivated researchers in the speech recognition area to use NAM sensors as one of the solutions for robust speech recognition in noisy environments [49-52].

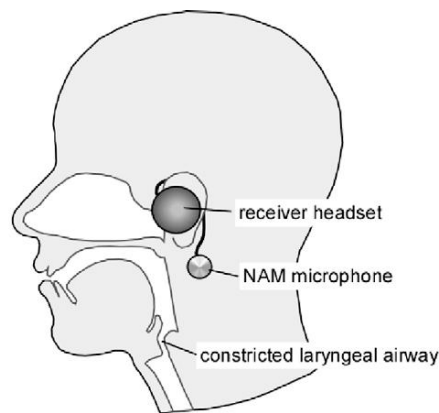


Figure 2.2: Placement of NAM microphone for silent communication. Vibration of the vocal-tract resonance is captured from the tissues, generated by airflow noise in the constricted laryngeal airflow.

- Use of a glottal activity sensor is another approach to solve the problem of de-noising speech signals corrupted by background noise. Glottal waveforms, usually obtained from the throat, forehead or crown of the head and ear can be used in conjunction with the acoustic signal received from a standard close talk microphone to augment the acoustic speech signal and improve speech quality [53]. A variety of vibration and electromagnetic sensors have been introduced such as
 - Physiological Microphone
 - Bone Microphone
 - Throat Microphone
 - In ear Microphone

The electroglottograph (EGG) [54-56], is a non-invasive measurement device designed to detect changes in electrical impedance during voice production. It is based on two gold-plated electrodes being applied on the surface of the neck on either side of the larynx. Typically, changes in electrical impedance are produced by vibrating vocal folds. When the vocal folds are closed, the electric impedance decreases, however larger impedance is produced when they are open. Glottal vibration in this way induces a signal of some 1V RMS on a 2–3MHz carrier, which is quite readily detectable. A drawback of the technique is its sensitivity to the exact positioning of the electrodes.

- Electromyography (EMG) is the process of recording electrical muscle activity using surface electrodes. When a muscle fibre is activated by the central nervous system, small electrical currents in the form of ion flows are generated. These electrical currents move through the body tissue, encountering a resistance which creates an electrical field. The resulting potential difference can be measured between certain regions on the body surface at the skin. The amplified electrical signal obtained from measuring these voltages over time can be fed directly into electronic devices for further processing. EMG signals have been used for many clinical applications including identifying neuromuscular diseases, diagnosis of low back pain and measuring motor control disorders, to control of prosthetic devices such as hands and arms. Whereas a lot of research has been conducted since 1985 in the area of speech recognition, surface EMG based methods have received attention more recently. Surface EMG signals associated with the speech muscles record the activity of the human articulation and thus allow one to trace back a speech signal even if it is unspoken [42, 43]. Recently, researchers have focused on overcoming the limitations of sEMG based speech recognition systems. Schultz and Wand [15] used alternative articulatory phonetic features to improve the classification results, whilst Walliczek *et al.* [57] and Schultz and Wand [15] used smaller acoustic units than words, enabling large vocabulary recognition by concatenating these smaller units.

- Electroencephalography (EEG) is the measure of potential difference (voltage) corresponding to different neuronal activities in the brain, recorded by attaching multiple electrodes on the scalp. Besides tremendous clinical applications, researchers have proved that EEG signals are useful for voiceless communication. Brain Computer Interface (BCI) based on EEG signals is current area of research in the biomedical and pattern recognition field. Suppes *et al.* [58] developed the first EEG and MEG (magneto encephalography) based isolated word recognition system. Wester and Schultz [59] investigated a new approach which directly recognizes “unspoken speech” in brain activity measured by EEG signals. Unspoken speech refers to the process in which a user imagines speaking a given word without actually producing any sound, indeed without performing any movement of the articulatory muscles.

All of the technologies presented above have shown encouraging results in the form of non acoustic modalities and to certain extent have resolved the issue of background noise. However, the sensor-based approaches are intrusive, impractical in most scenarios and have common challenges. Users need to fix electrodes around the face, neck and even inside the mouth. The sensors used must be very precisely positioned to obtain the optimal results.

Speech recognition based on a visual speech signal is the least intrusive [14], does not require sensors to be attached to the head and scalp, and is non-constraining and noise robust. Other benefits of VSR system are in cases, where persons have had Laryngectomy surgery and for weak or aged persons who can speak only with much effort. These people can move their lips easily rather than speaking. At the same time the attendance of an important call at any location can be a very useful service. A non-invasive visual speech recognition system built into a mobile phone can resolve these issues by communicating speechlessly. The hardware for such a system can be as simple as a webcam or a camera built into a mobile phone.

How do humans lip-read exactly and perceive audio visual speech? This is still a question to be answered. However, it is essential to look at the physiological and psychological

aspects of human speech production and perception in order to acquire the necessary information for replication of speech recognition by machines. In the following sections a review of AVSR is given. This is followed by a brief review of human speech production and perception in the context of physiological processes.

2.3 The Review of AVSR and VSR

This section presents a review of some related research on AVSR and VSR systems. AVSR involves the process of interpreting the audio-visual information contained in a video in order to extract the information necessary to establish the communication at perceptual level between humans and computers, whereas VSR systems are solely based on visual speech processing. In a noisy environment, although voice signal is not understandable, the associated mouth movements create sufficient visual cues that are able to be exploited, to develop VSR systems. Engineers have continued research into recognizing speech in a noisy environment since the 1890s [60]. This interest of researchers increased during the war years of the 1940s and 1950s, when engineers working in this field were attempting to develop a system that could establish a communication between pilots and air traffic controllers [61]. The first known work on audio-visual speech recognition was published by Sumby and Pollack in 1954 [30]. After three decades of Sumby and Pollack publication, in 1984 Petajan [41] presented the first AVSR system. In this system Petajan extracted the shape based features such as mouth height, width perimeter and area from the black and white images of the speaker's mouth. The next major advance in AVSR research was a decade later when Bregler and Koing [62] published their work using eigenlips. In the same year Duchnowski *et al.* [63] extended the technique of eigenlips using linear discriminant analysis (LDA) for the visual speech features. In 1997 Adjoundari and Benoit [64] focused on the problem of audio-visual feature fusion.

AVSR is a technology generally based on pre-recorded audio-visual dataset. IBM has recorded a high quality audio-visual dataset, because of that the major progress in the field is based on the work conducted by IBM. That work was led by Gerasimos Potamianos and his fellows, but the dataset however, was not publically available. In

addition to large vocabulary experiments, Potamianos *et al.* in 2003 conducted AVSR experiments in challenging environments, where data was captured in office and in car scenarios [65]. They demonstrated that the performance measure is considerably degraded in both modalities in the challenging environments. In 2004, IBM developed an AVSR system, in which they used a special wearable headset containing infra-red based audio visual system [66] which constantly focus on the speaker's mouth. The motivation behind this work was to develop a real time system, which reduces the computational burden of pre-processing such as face and lip localization and also to reduce the effects due to visual variability such as face orientation, lighting effects and background. In this work they found that their approach achieved results comparable to normal AVSR systems. This is the motivation behind this work, the proposed system is based on the dataset recorded by a fixed camera focusing on the user's mouth in an office environment.

Potamianos *et al.* [14, 67] have presented a detailed analysis of Audio-visual speech recognition approaches, their progresses and challenges. Generally, the systems reported in the literature are concerned with advancing theoretical solutions to various subtasks associated with the development of AVSR systems. There are very few papers that have considered the complete system. Generally the development of AVSR can be divided into the following categories: audio feature extraction, visual feature extraction, temporal segmentation, audiovisual feature fusion and classification of the features to identify the utterance. The proposed visual only system does not have any audio data and hence audio feature extraction and its' fusion with visual features is not related to this work.

The visual feature extraction techniques that have been applied in the development of VSR systems can be categorized into the following: shape-based (geometric), intensity/image-based and a combination of both. The description of these categories is detailed in Chapter 6. Sagheer *et al* [68] compared their appearance based Hyper Column Model (HCM) with the image transform based Fast Discrete Cosine Transform (FDCT) on two different datasets comprising Arabic and Japanese language elements. They demonstrated that the appearance based HCM technique outperformed the image transform technique in visual speech recognition by 6.25% overall in each of the datasets.

Recently Zhao *et al.* (2009) [69] introduced local spatio-temporal descriptors, instead of global parameters, to recognize isolated spoken phrases based solely on visual input, obtaining a speaker independent recognition rate of approximately 62% and a speaker dependent result of around 70%.

The appearance based approaches have proved to be inappropriate when the data set contains non constant illumination conditions. To cope with this problem, the statistical model of both shape and appearance are considered in literature. Active Shape Model (ASM), Active Appearance Model (AAM) and Multiscale Spatial Analysis (MSA) have found wide applications in visual speech recognition. These approaches were all proposed by Matthews *et al* [24] to extract visual features. Continuing their work on visual feature extraction, Papandreou *et al* [17] focused on multimodal fusion scenarios, using audiovisual speech recognition as an example. They demonstrated that their *visemic* AAM (based on digits 0-9) with six texture coefficients outperforms their PCA-based technique with 18 texture coefficients, achieving a word accuracy rates of 83% and 71% respectively. However, these techniques are computationally expensive and require manual intervention for complex sets of land-marking on face to define the shape of an object or the face. In general the performance of the intensity based methods is better than that achieved by the shape-based VSR techniques [35]. In addition to this intensity-based approaches do not require *a priori* statistical lips models and this fact allows the development of computationally efficient VSR systems [24].

Feature extraction methods have utilized motion analysis of image sequences representing dynamics of the lips while uttering speech. In comparison to shape based features, the global intensity and motion based features are independent from the speaker's mouth shape. Yau *et al* [26] have described a lip reading system based on dynamic visual speech features using an approach called motion history image (MHI) or Spatio-Temporal Templates (STT). MHI is a grey scale image which shows the temporal and spatial location of the movements of speech articulators occurring in the image sequence. The advantage of spatio-temporal template based methods is that the continuous video frames are compressed into a single grey scale image such that dominant motion information is retained. The MHI method is less expensive to compute,

by keeping a history of temporal changes at each pixel location [70]. However, MHI has a problem of limited memory due to which the older motion history is soon over-written by new data [71], resulting in inaccurate lip motion description and therefore leading to low accuracies in viseme recognition. In order to cope with this problem, this research proposes an enhanced version of MHI. In this technique, rather than a single grey scale image four directional motion history images (DMHIs) are proposed to overcome the issue of motion overwriting and shown to have outperformed the MHI. A detailed description of the proposed technique is presented in chapter 5. Iwano *et al* [34] and Mase *et al* [72] reported their lip-reading systems for recognizing connected English digits using an optical flow (OF) analysis. Iwano *et al* [34] adopted a hybrid approach in which the shape and dynamic visual speech features are considered together, lip contour geometric features (LCGFs) and lip motion velocity features (LMVFs) of the side-face images are calculated. Optical flow based, two LMVFs (the horizontal and the vertical variances of flow-vector components) are calculated for each frame and normalized to the maximum values in each utterance. The technique achieved digit recognition errors of 24% using a visual-only method with LCGF, 21.9% with LMVF and 26% with LCGF and LMVF combined.

2.4 Anatomy of the Human Speech Production System

In human speech production, the mouth is analogous to an audio filter where a variation in the shape of the mouth cavity and lip movement produces different sounds. The shape of the vocal tract is an important physiological aspect of the human speech production system. The main speech producing organs are the lungs, laryngeal pharynx (beneath the epiglottis), oral pharynx (behind the tongue, between the epiglottis and velum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above the velum, rear end of nasal cavity), and the nasal cavity (above the palate and extending from the pharynx to the nostrils).

An acoustic wave is produced when inhaled air is exhaled from the lungs and passes by the bronchi and trachea through the vocal folds. This source of excitation can be characterized as voiced and unvoiced which is based on the tightened or relaxed

conditions of the vocal folds. If the vocal cord muscles are tightened then airflow is modulated by the vocal folds and causes vibration, phonated excitation occurring. Correspondingly, if the vocal folds are relaxed, air pressure does not produce vibration and whispered excitation is produced. Speech produced by phonated excitation is called voiced, and that produced by whispered excitation is called unvoiced. The main articulators of speech production are the vocal cords, tongue, teeth, velum, jaw and lips. Figure 2.3 shows the human speech production system. The lips are the only articulator that is fully visible whereas the tongue and teeth are partially visible from the frontal view of the face. The movement of other speech articulators such as the velum and glottis is invisible. The visual information that can be extracted from a sequence of images encompasses lips, tongue and teeth. Important information about the invisible articulators such as complete movement of tongue and vocal cord vibration can be extracted by the sensor based methods briefly explained in Section 2.2.

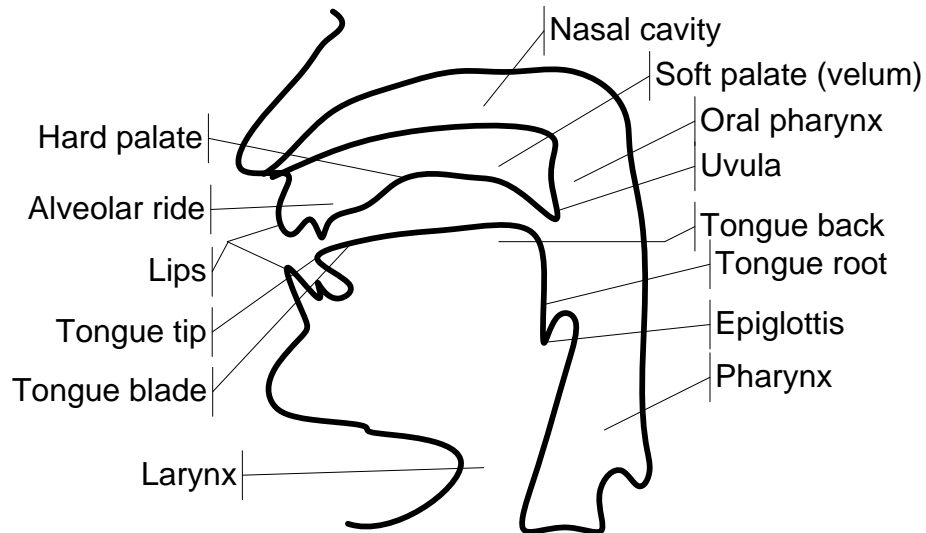


Figure 2.3: Human speech production system

The information contained in visual features is not sufficient for complete classification because phonemes such as dental, palatal and fricative consonants involve the tongue, palate and teeth which are not fully visible, and cause error in classification. One way to improve recognition rates is by incorporating contextual information as a post processing

step. However, as humans have clearly adapted to make use of this visual information, as will be shown later in this chapter, the study of how visible speech is related to acoustic speech is important.

2.5 Linguistics of Visual Speech

A *phoneme* is the smallest basic unit of audible speech that distinguishes one word sound from another. For example the word ‘cat’ and ‘rat’, the phone /c/ and /r/ have distinct sounds. Just like a phoneme, a *viseme* is the basic visual unit of speech. All of the phonemes are not completely represented by visible articulators. Commonly each viseme represents many linguistically distinct phonemes because visual speech is only partially visible [73]. For example the phonemes /p/, /b/ and /m/ are phonetically different but are represented as the same viseme class. Correspondingly, there are some phonemes that are acoustically indistinct but distinct in the visual domain [74] such as /n/ and /m/, as depicted in Figure 2.4. The typical number of phonemes defined in audio speech recognition is from 40 to 50 [75]. The set of English visemes can be determined by applying statistical methods to cluster the phonemes to represent a viseme [14, 76, 77].



Figure 2.4: Different visemes but same phonemes. Image (a) shows viseme /m/ and image (b) shows /n/ while these are acoustically similar phonemes.

Generally, the range of visemes used in visual speech recognition is in the range of 12 to 20 [36, 75] as compared to 50 phonemes in English. The number of English visemes varies mainly with respect to the accent guided by the geographical location, education

and age of the speaker. It is difficult to define a standard and a universal set of visemes suitable for all speakers [78] from different geographical locations whereas it is required to define a standard and universal set of optimized visemes for English language at least.

2.6 Visual Speech Perception by Humans

The first widely recognised work on visual contribution to speech recognition in noisy environments was published half a century ago by Sumby and Pollack [30] in (1954). In their experiments they observed the effects of noise on human perception, considering the visual information available to the listener. From their experiments it has been conclusively proved that the audio signal is the basic modality for speech communication. Listeners utilize their ability of perception in noisy environments, however by using visual information in the recognition and comprehension of speech. In 1987, Reisberg *et al.* [31] showed that listeners with normal hearing ability also take advantage of visual cues in speech recognition even when clear audible speech is spoken. This makes it obvious that visual speech signals contain a significant amount of information. The significance of visual cues is also obvious in the lip-reading ability of hearing impaired people to lip read. Human speech perception is bimodal. That was proved by McGurk and Macdonald in 1976 [23]. They demonstrated that when a person is presented with a video with a different audio recorded over it, a third sound can be perceived rather than either of the two actually presented to the person in either modality. This is known as McGurk effect [23]. The most commonly presented example in this literature is of a person watching a video of a speaker's face pronouncing /ga/ but hearing /ba/ simultaneously. The person perceives the hearing sound /da/, differently from either of the two actually presented to him. The reason for this is the visual /ga/ is more alike to visual /da/ than to visual /ba/. Correspondingly, hearing /ba/ is more alike to hearing /da/ in comparison to hearing /ga/. Thus the human audio visual sensory information is naturally synchronized.

Summerfield [60] demonstrated the advantages of human speech perception in lip-reading. Once the listener receives the acoustic signal that helps in localizing the speaker while the speaker continues to talk, the listener can further localize the articulators of

speech and can get the complimentary visual information to augment the acoustic signal. Summerfield [60] also demonstrated that many of the indistinct phonemes have distinct visemes that can be easily recognized by viewing the speakers lips. Contrary to this he also showed that the phonemes corresponding to a particular viseme can be acoustically quite different but visually the same, such as the phones /t/ and /d/ having completely different sounds whereas they have the same viseme.

Moreover, interpretation of a speech signal can be amplified by the perception of facial gestures. Not only speech, but paralinguistic information of the speaker's emotional state can be obtained from facial gestures [79]. Especially where the acoustic signal is unclear, because of background noise, watching the facial gestures can enhance speech intelligibility.

2.7 Speech Reading Proficiency

Recent research has shown motivation towards human speech recognition that can be more accurate. The most proficient lip-readers observed to date have been hearing impaired persons, who are only able to rely on speech reading (not sign language) for communication. For these people, the accuracy level for word perception on a set of isolated pre-recorded sentences was approximately 65 to 85% [80]. Although adults with normal hearing are generally less accurate lip-readers, their accuracy levels can also be moderately high [80-82].

2.8 Significance of Facial Parts in Lip-reading

Various researchers have focused on identifying the pertinent regions of the face that contain the most important features for speech perception. It has been accepted that the most informative region of the face is around the lips [39]. Benoit et al [83] demonstrated that the lips contain on average two thirds of the speech information. The rest of the information is spread over the speaker's face. In their research it was concluded that the addition of the region of the jaw along with the lips increases the human perception of visual information. It was also conclude that the frontal view of a speaker provides more

information as compared to the side view. McGrath *et al.* [84] explored that the lips contain more than half the visual information revealed on the face of an English speaker. This validates the focus on the area of the lips as the Region of Interest (ROI) in all audiovisual speech recognition systems. In the study of Brooke and Summerfield [85] it was demonstrated that for the efficient recognition of vowels the visible articulators such as the tongue and teeth play an important role. Finn [86] found that for recognition of consonants, size and shape of the lips are the pertinent features. Lip rounding and the areas around the lips hold significant features for vowel recognition Montgomery and Jackson [87]. They also verified that there is significant variability in style of speaking between speakers, and their way of moving lips and tongue in speaking.

2.9 Basic Components of Visual Speech Recognition System

This section describes the basic components of a visual speech recognition system. In general, a VSR system consists of the following parts:

- Image capturing devices
- Pre-processing
 - Noise filtering
 - Face identification
 - Lip localization
 - Temporal segmentation
- Feature extraction
- Dimensionality reduction of features
- Classification and recognition of speech

The general block diagram of a VSR system is shown in Figure 2.5. The image capturing devices are the video cameras. Most of the research in the area of AVSR and VSR is based on pre recorded datasets, with the intention to develop the real time system. After getting the desired dataset, the next step is noise filtering; it is to make the dataset suitable for further processing. In the pre-processing section face detection is performed followed by lip localization. In lip localization the area around the lips commonly known

as region of interest (ROI) is extracted from each frame of a video sequence. In this research pre-segmented dataset, which contains only ROI is used. An important step in VSR system is the automatic temporal segmentation, it is to locate the start and end frame of an utterance from a sequence of images containing multiple utterances.

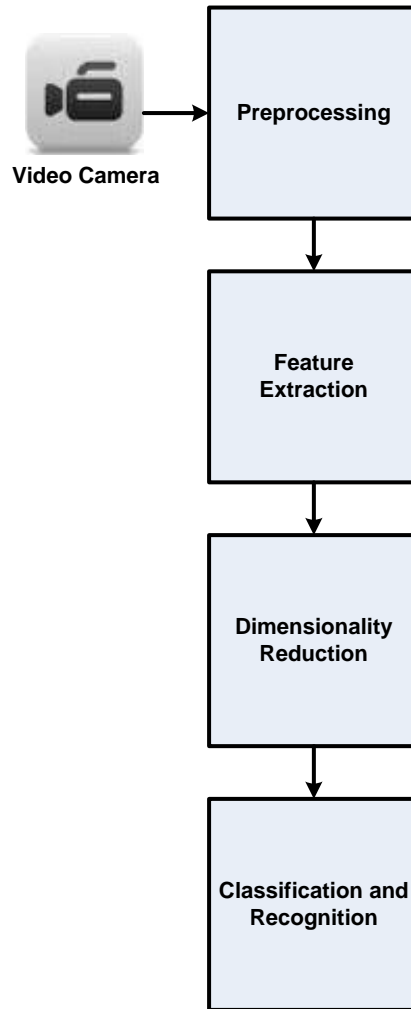


Figure 2.5: Basic Components of a VSR system.

Once the temporal window of an utterance is identified, the next step is to find the significant visual features that represent the dynamics of the mouth movements efficiently either by the visual appearance or by the shape of the lips in each frame of a video. To produce a compact representation of extracted features suitable for the statistical classifiers, generally some feature reduction techniques are applied. The last

step in the VSR system is to train the classifier in order to develop a model and finally testing the model with unseen data.

2.10 Brief Description of Proposed VSR System

An overview of the proposed visual speech recognition system is presented, with the flow diagram depicted in Figure 2.6. The proposed system consists of pre-processing, motion tracking, feature extraction and finally the classification.

- **Pre-processing**

Illumination and colour variation induced by the video recording system can create a problem in motion estimation by optical flow. A global illumination normalization technique, Gaussian smoothing [88] is adopted to normalize the sequence of images. Following by noise filtering, the temporal segmentation of the isolated uttered visemes in an image sequence is performed. The main purpose of temporal segmentation is to segment the individual utterances from video data containing multiple utterances in order to locate the start and the end frame. Pair wise pixel comparison method [38] is used to find the start and an end frame of an utterance from a series of isolated utterances. Detailed description of pre-processing is given in Chapter 3. A variety of face detection [89] and lip localization algorithms [90, 91] are available in the literature and hence are not the focus of this work.

- **Motion Tracking**

In order to represent the lip movements efficiently, robust visual features are desired. In this work lip movements are represented by features derived from optical flow estimates. The optical flow is defined as the distribution of apparent velocities of movement of brightness pattern in an image [92]. The features computed by optical flow analysis give real motion information. A complete probabilistic model of optical flow based on statistical learning for both brightness constancy error and spatial properties was adopted. Elaborate discussion of optical flow technique is presented in Chapter 4.

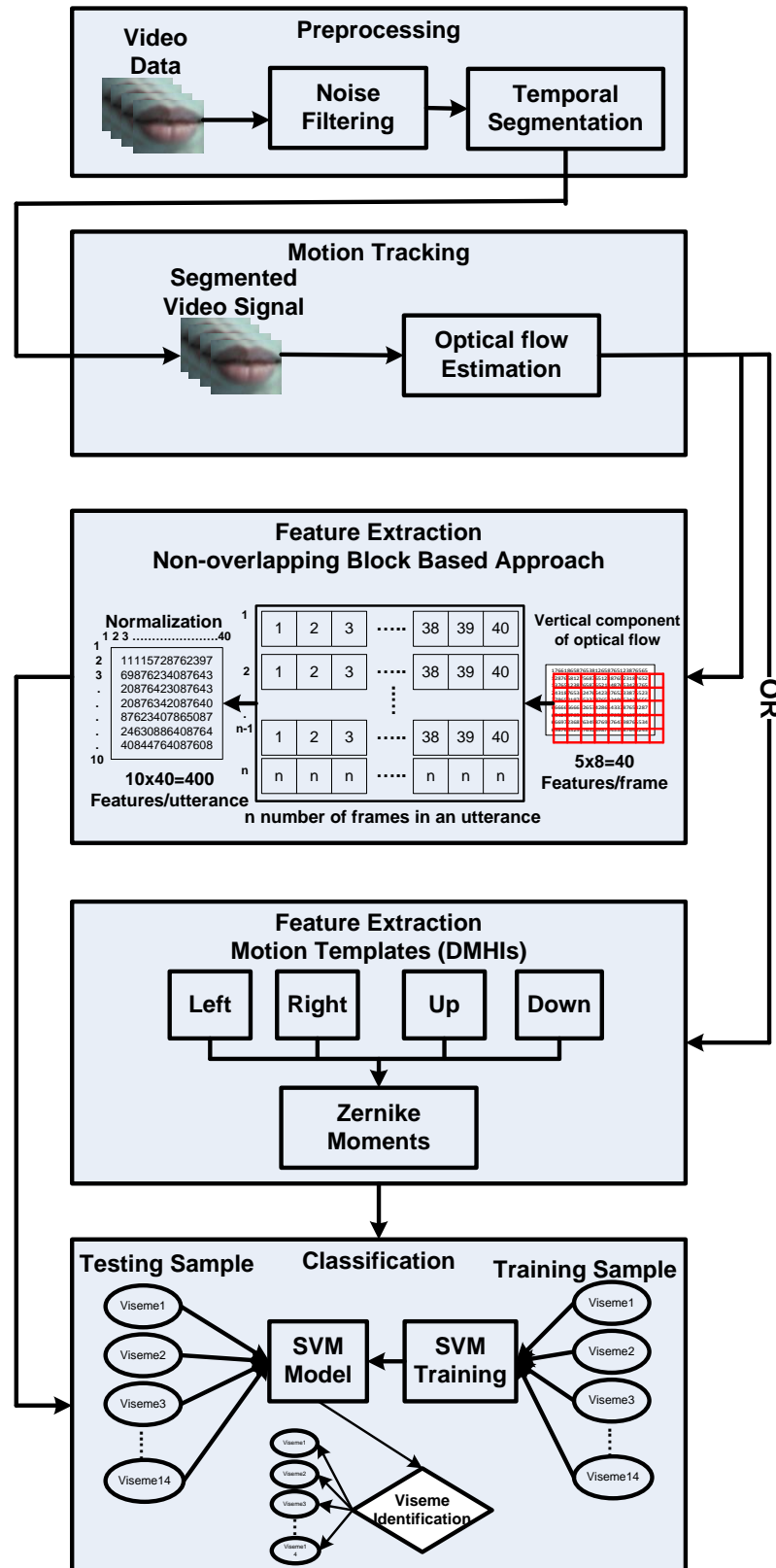


Figure 2.6: A block diagram of the proposed Visual Speech Recognition system.

- **Feature Extraction**

The image resolution of the acquired data set is 240×320 pixels per frame. The optical flow computation provides horizontal and vertical velocity components of the same size of image, so that the feature vector size would be $240 \times 320 \times 2 = 153600$ for every pair of image frame. The feature vector size impacts the classification in terms of computation and accuracy. In this research, in order to produce compact representation of the computed optical flow, two novel approaches are proposed. These compact representations of the lip motions during speech are used to generate a feasible visual speech template for each utterance, so that the obvious inter and intra speech variation in the speed of speaking is compensated. The first approach is block based. In this approach only the vertical component of optical flow is used. The optical flow horizontal component is ignored due to negligible contribution in the viseme classification. The optical flow vertical component is segmented into non overlapping blocks. Each block's mean value is computed and used as feature to the classifier. In second approach four directional motion history images are computed from the optical flow (vertical and horizontal components). Each image represents the consolidated lip movements of an utterance in a particular direction *i.e.*, up, down, left and right for the complete time window of an utterance. Furthermore two rotation and scale invariant features Zernike moments and Hu moments are computed for each image representing an utterance. Detail discussion of each approach is presented in Chapters 4 and 5.

- **Visemes Classification**

In the final stage, training and testing of the features extracted in the previous phase is carried out using binary class and multi class SVM. Another novelty of the work is the use of multiclass classification, used in visual speech recognition for the first time and resulting in an improved performance measure of the proposed VSR system. In addition to this for the performance measure, results are shown not only in traditional accuracy or word error rate (WER) measure, but sensitivity and specificity are used as the performance measures and have increased the credibility of the proposed classification technique. Details of

classification methodology are given in Chapter 6. Experimental Results are presented in Chapter 7.

2.11 Summary

In this chapter a detailed review of AVSR and VSR systems is presented. In addition, a brief review of non speech modalities such as EMA, combined features from ultrasound imagery and standard camera imagery, NAM, EGG, EMG and EEG are also described. These techniques have performed well in many clinical and biomedical applications as well as in voiceless speech recognition. From this review it is concluded that speech recognition systems based on visual cues are more suitable for real time systems when compared to sensor based systems, referred to as silent speech interfaces [44]. Sensor based systems are intrusive and impractical in most scenarios, the visual feature based systems which are non intrusive are shown to be preferred methods. Both persons of normal hearing as well as speech impaired persons use facial movements to augment speech intelligibility [23, 30]. The hardware required for a visual speech recognition system can be a simple hand held mobile phone with built in video camera which are commonly available. One of the interesting phenomena described is the McGurk effect, which proves the bimodality of speech perception *i.e.*, audio-visual. The basic mechanism of human speech production, perception as well as the linguistics associated with the audio and visual modalities is then described. It is also accepted that the main pertinent area for visual speech perception and recognition is the lower part of the face (the mouth and jaw) [60, 83, 86, 87]. At the end of the chapter a brief review of general and proposed visual speech recognition systems is presented.

Chapter 3

Image and Video Preprocessing

This chapter presents a basic review of a visual front end of the visual speech recognition (VSR) system and provides a description of the dataset used in this study. The main result of this chapter is a new temporal segmentation method that detects the start and end frame of an utterance, and is presented in Section 3.3. The building blocks of VSR system were presented in Chapter 2. The first block of a VSR system following the video capturing device is pre-processing. The key component of the pre-processing block is the visual front end. For an automatic visual speech recognition system, the visual front end has to be able to localize the visible speech articulators, this process of the mouth localization is known as the spatial segmentation. It is widely accepted that the pertinent area of visible speech is the region around a speaker's mouth, known as the region of interest (ROI). The visual front end of a VSR system is responsible for localizing the speaker's face, and then locating and keeping track of the ROI. If the spatial segmentation of the ROI is not performed accurately, then erroneous features will be extracted and the overall performance of the system will be degraded. There are some other factors which can affect the performance of a VSR systems such as pose, occlusion and illumination [61]. These are discussed in the following sections.

Much research has been conducted on face detection and ROI localization. Hence this study has not considered these aspects in this work. Furthermore, the dataset used in this research only contains the ROI, so there is no need for ROI localization.

3.1 Visual Front End

Generally, the visual front end is based on three hierarchical processes. Starting with a speaker's face detection, the essential features such as mouth corners, nose and eyes are

then located. In the third step, based on these features, the location of the ROI is estimated. After successful estimation of the ROI, further image pre-processing can be employed on the ROI to minimize the effects of variable lighting conditions on optical flow computation, using techniques such as histogram flattening, Gaussian smoothing and balancing of the left-to-right brightness distribution.

Face orientation is an important factor in VSR systems. Typically, image sequences are recorded in frontal or side profile models. However, frontal profiles have shown better performance than side profile [83]. According to the survey paper of Ming-Hsuan *et al.* [93], the challenges associated with face detection can be attributed to the following factors:

- **Pose.** The images of a face vary due to the relative camera-face pose (frontal, 45 degree, profile, upside down), and some facial features such as an eye or the nose may become partially or wholly occluded.
- **Presence or absence of structural components.** Facial features such as beards, moustaches, or glasses may or may not be present and there is a great deal of variability among these components including shape, colour, and size.
- **Facial expression.** The appearance of faces is directly affected by a person's facial expression.
- **Occlusion.** Faces may be partially occluded by other objects. In an image with a group of people, some faces may partially occlude other faces.
- **Image orientation.** Face images directly vary for different rotations about the camera's optical axis.
- **Imaging conditions.** When the image is formed, factors such as lighting (spectra, source distribution and intensity) and camera characteristics (sensor response, lenses) affect the appearance of a face.

Although the factors above show that the task of face detection is quite complex, some efficient algorithms which provide sufficient accuracy in a controlled environment are available and are discussed below.

3.1.1 Face Detection

In recent years, the demand for face detection algorithms has increased heavily. This is because of their use in several automated systems that take images of the human face as an input. Examples include visual speech recognition, video-based surveillance systems, human face/body tracking systems, fully automatic face recognition systems and perceptual human computer interfaces. Typically, these face detection algorithms are based on the classifiers which estimate the presence of a face given in a particular window of the image.

Most face detection algorithms are either appearance based or feature based. In recent years, appearance based face detection algorithms that exploit statistical estimation and machine learning methods have shown excellent results among all existing face detection methods. Ming-Hsuan *et al.* [93] have classified the single image face detection methods into the following four categories.

- **Knowledge based methods**

These rule-based methods encode human knowledge of what constitutes a typical face. Usually, the rules capture the relationships between facial features. These methods are designed mainly for face localization.

- **Feature invariant approaches**

These algorithms aim to find structural features that exist even when the pose, viewpoint, or lighting conditions vary, and then use these to locate faces. These methods are designed mainly for face localization.

- **Template matching methods**

Several standard patterns of a face are stored to describe the face as a whole or the facial features separately. The correlations between an input image and the stored patterns are computed for detection. These methods have been used for both face localization and detection.

- **Appearance based methods**

In contrast to template matching, the models (or templates) are learned from a set of training images which should capture the representative variability of facial

appearance. These learned models are then used for detection. These methods are designed mainly for face detection.

Appearance based face detection techniques include the AdaBoost algorithm [94, 95], the S-AdaBoost algorithm [96], the FloatBoost algorithm [97], Hidden Markov Models (HMM) [98], neural networks [99, 100], Bayes classifier [101] and Support Vector Machines (SVM) [102, 103]. Viola and Jones [94, 95] developed a robust AdaBoost face detection algorithm, which is computationally efficient and detects faces robustly with high accuracy. The Float-Boost algorithm proposed by Li *et al.* [97] is an improved version of the AdaBoost algorithm, for learning a boosted classifier with minimum error rate. However this method is computationally more demanding than the AdaBoost algorithm. Readers are directed to the paper [89] which contains a survey about recent advancements in face detection.

3.1.1 Spatial Segmentation

The term spatial segmentation is referred to as the finding the ROI in 2D space. Lip segmentation is an important part of the visual front end of an automatic visual speech recognition system. In such a system, the ROI that is area around the lips must be detected in each frame of a video sequence to be processed for the speech recognition. This procedure is normally carried out by fitting a range of colour models to the image and is followed by face detection and extraction of the ROI surrounding the lips.

Early VSR systems performed the lip segmentation in conjunction with the application of artificial markers (lipstick) on the lips [104]. The application of lipstick enables the system to precisely detect the lips in the image data, but this procedure is inappropriate since it is uncomfortable for users and such VSR systems can be operated only in a constrained environment. Thus, the main research efforts have been concentrated in the development of vision-based lip segmentation algorithms. Many studies have shown that colour information can be successfully applied to identify the skin or face in digital images [105]. The main idea behind this approach is to transform the RGB signal into a new representation where the mouth is clearly visible, so that it can be easily segmented.

To this end, a large number of colour representations have been proposed. Coinaize *et al* [106] used the hue component of the HSV representation to highlight the red colours which are assumed to be associated with the lips in the image. Later, the HSV colour space was further used by Zhang and Measereau [107] for lip detection. They used prominent peaks in the hue signal as an indicator to locate the position of the lips, then based on the identified lip area, the interior and exterior lip boundaries were extracted using both colour and spatial edge information using a Markov Random Field (MRF) framework. Other approaches carried out the lip detection task in the YCrCb colour space since the facial skin covers a small area of the CrCb subspace [108, 109].

In 2001, Eveno *et al* [105] proposed a new colour mixture and chromatic transformation for lip segmentation. In their approach, a new transformation of the RGB colour space and a chromatic map was applied to increase the discrimination between the lips and facial skin. They demonstrated that the proposed approach is able to achieve robust lip detection under non-uniform lighting conditions. Later, Eveno *et al* [110] introduced a different method where a pseudo-hue [111] was applied for accurate lip segmentation that has been embedded in an active contour framework. They applied the proposed algorithm for visual speech recognition and the results show significant improvement in terms of accuracy in lip modelling.

Another method for mouth segmentation was proposed by Liew [112] in 2003. In this approach, the colour image is transformed into the CIE-Lab and CIE-Luv colour spaces, and then a lip membership map is computed using a spatial fuzzy clustering algorithm. After morphological filtering, the ROI around the mouth can be identified from the face area.

In 2006, Guan [113] improved the contrast between the lip and other face regions using the Discrete Hartley Transform (DHT). In this paper, lips are extracted by applying wavelet multi-scale edge detection across the C3 component of the DHT which takes both the colour information and the geometric characteristics into account.

Most recently in 2011, Akdemir and Ciloglu [32] proposed a lip localization method in which they use 12 blue markers on the face of a subject. Eight of these markers were marked on the lips and used to extract the shape and position of the lips. Three of them were located on the nose and cheeks to compensate the head movements by aligning them in straight lines. The last mark was on the chin and used to capture the syllabic oscillation of the jaw. A chroma key approach and Auction algorithm were used to locate the position of the blue markers and to track each blue marker through the consecutive images.

Because of the availability of such face and lip detection algorithms and the nature of the dataset used, this research has not considered ROI detection. The following section describes the choice of data set used in this research.

3.2 Choice of Utterances

Visemes are the smallest visually distinguishable facial movements when articulating a phoneme and can be concatenated to form words and sentences, thus providing the flexibility to extend the vocabulary. This is the basic motivation for choosing visemes as the recognition unit. The total number of visemes is less than the English phonemes because different phonemes may have common visible movement [73]. The video of a speaker's face while uttering a phoneme shows the movement of the lips and jaw, whereas the movements of other articulators such as vocal cord and tongue are often not visible. Hence, each viseme can correspond to more than one phoneme, resulting in many-to-one mapping of phonemes to visemes.

3.2.1 Dataset Exploited for this Study

The dataset used in this study was recorded by Yau *et al.* [75] in a typical office environment. Their aim was to collect a dataset that can be used to evaluate the performance of the visual speech recognition system in a real world environment. Publically available audio-visual speech data sets such as M2VTS [114], XM2VTS [115], Tulips 1 database [116], CUAVE database [117] and GRID database [118] were collected in ideal studio environments with controlled lighting. Image sequences are more

affected by illumination noise in an office environment in comparison to an ideal studio environment [75]. The dataset used in this study consists of simultaneous audio and visual recordings of 14 discrete English phonemes/visemes by 7 subjects. Table 3.1 shows the 14 visemes and corresponding phonemes. The visemes of phonemes used in this research are highlighted in bold fonts. Table 3.1 also shows the example words that can be produce by the visemes used in this research. These visemes are originally defined in the MPEG-4 standard and consists of five vowels and nine consonants. A simple webcam was used to capture the videos, in order to analyse the low resolution videos for visual speech recognition. Audio signals were recorded by the inbuilt microphone in the webcam. However the audio signals are not included in this study. The subjects were asked to speak in front of a fixed camera. The distance between the camera and face was kept constant to at 10 cm. The camera was focused on the mouth region of the speaker and was kept stationary throughout the recordings and the frontal profile of the ROI was recorded.

Table 3.1: Fourteen *visemes* defined in MPEG-4 and the average number of frames for each viseme of the used dataset.

	Visemes and Corresponding phonemes	Vowel/Consonant	Example words	Average Number of frames
1	/a/	Vowel	j/a/r	33
2	/ch/ ,/j/,/sh/	Consonant	/ch/ain, /j/oin, /sh/iraz	30
3	/e/	Vowel	/e/gg	35
4	/g/ ,/k/	Consonant	/g/reat, /k/ing	28
5	/th/ ,/D/	Consonant	/th/row ,/th/an	33
6	/i/	Vowel	/i/nk	35
7	/p/,/b/, /m/	Consonant	/p/late, /b/ed, /m/an	33
8	/n/ ,/l/	Consonant	/n/est, /l/ight	38
9	/o/	Vowel	t/o/ne	35
10	/r/	Consonant	/r/ain	37
11	/s/, /z/	Consonant	/s/un, /z/oo	34
12	/t/ ,/d/	Consonant	/t/icket, /d/oor	24
13	/u/	Vowel	p/u/t	35
14	/f/, /v/	Consonant	/f/an, /v/an	40

General Description of the Dataset:

- **Number of subjects:** 7 (4 male, 3 female)
- **Speech data:** 14 isolated visemes given in Table 3.1
- **Repetitions:** 10 epochs of each viseme in a single video

- **Total uttered visemes:** 980 visemes
- **Image resolution:** 240×320 pixels
- **Colour:** RGB
- **Video sampling rate:** 30 frames/second
- **Average number of frames per viseme:** 33.6 frames

Factors such as window size (240×320 pixels), viewing angle of the camera, background and illumination were kept constant throughout the experiments. Figure 3.1 shows example images of all the subjects used in this study with different skin colours.

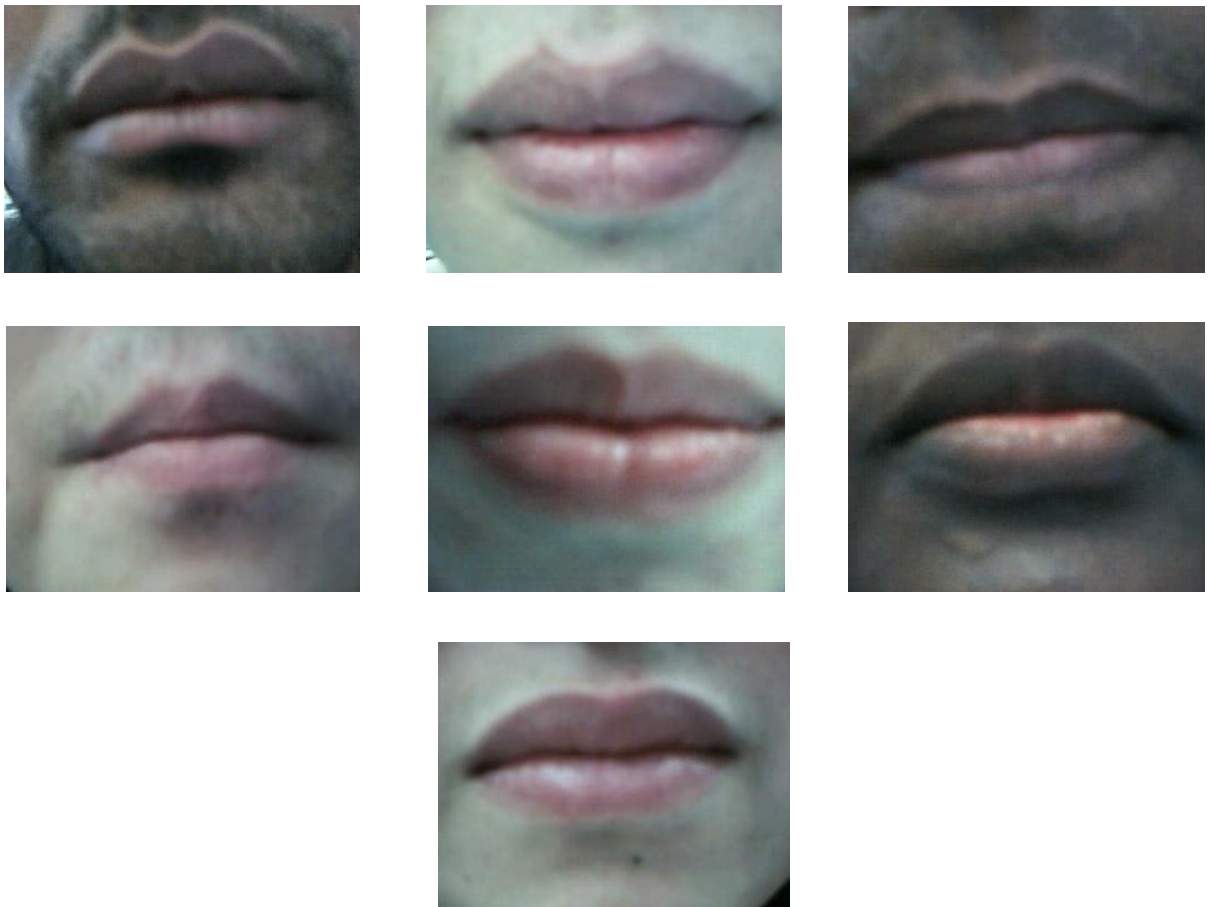


Figure 3.1: Example images of seven subjects with different skin tones.

3.3 Temporal Segmentation

It is important for any end-to-end visual speech recognition system, based either on continuous or discrete speech, to find the start and end frame of the basic unit of visual speech from consecutive image sequence, so that specific features corresponding to that unit are extracted separately. It is difficult to visually segment the continuous speech into basic visemes. However, a recent experimental study by Sell and Kaschak [119] at Florida State University demonstrated that the presence of visual speech information alone is sufficient to allow the learners to segment words from a fluent speech stream. The dataset used in this research consists of discrete utterances, *i.e.*, deliberate little pauses were inserted between the utterances while recording, so that individual utterances can be segmented from the video data. The aim of temporal segmentation is to find the start and end frames of an utterance automatically from an image sequence. The importance of temporal segmentation can be obviated from the data sets such as TULIPS1 and CUAVE. These data sets have been manually segmented and are the basis for most of the current research. However, for real time implementation of these systems, it is important to perform automatic segmentation without human intervention. The earlier works where automatic segmentation was performed have typically considered the combination of the audio and visual data and thus the temporal speech segmentation in AVSR systems is based on audio signals [120]. The amplitude of audio signals provides sufficient cues about the speech and silence.

Temporal segmentation based on a visual signal only is necessary for VSR where the audio signal is not available, or is highly affected by environmental noise. To segment the sequential utterances, this research proposes an *ad hoc* but effective mechanism to detect the start and end frames of non-overlapping utterances. A pair-wise pixel comparison [38] is used for temporal segmentation. It evaluates the differences in intensity of corresponding pixels in two successive frames throughout the video sequence. A little pause is present between every consecutive utterance. This pause period provides important information for visual segmentation from mouth images.

As a first step, the colour images are transformed to gray scale images. Then the average difference square (ADS) of consecutive image frames which represents the magnitude of mouth movement is computed to get the absolute and prominent values using Equation (3.1):

$$ADS = \left[\frac{1}{m \times n} \sum_{x=1}^m \sum_{y=1}^n [I_1(x, y) - I_2(x, y)] \right]^2. \quad (3.1)$$

The pause periods consist of minimal mouth movement and are represented as low magnitude values, whereas the pronunciation of utterances is represented as comparatively large magnitude values. In the consecutive image sequence there are similar subsequent images which provide the zero energy difference even when utterances are pronounced. The basic reason for similar subsequent images in the video data is the frame rate (30 frames/second) at which the videos are recorded. Higher the frame rate, the more number of similar subsequent images. To avoid these zero magnitudes in an utterance, the resulting energy signal is smoothed using a moving average window and further smoothed using Gaussian filtering. Selecting an appropriate threshold leads to the required temporal segmentation.

The steps of the *adhoc* automatic temporal segmentation are shown in Figure 3.2. Figure 3.2 (a) indicates the squared mean difference of intensities of corresponding pixels in accumulative frames (for clarity, only three visemes are shown). The results of average moving window smoothing and further Gaussian smoothing are shown in Figures. 3.2 (b) and 3.2 (c). Finally the result of segmentation (as unit step-pulse-shaped representations) is shown in Figure. 3(d). Each utterance is represented by two unit step-pulses, each representing the opening and closing motion of the mouth while uttering a viseme. It is clear from the figure that the *adhoc* scheme followed is highly effective in viseme segmentation. The results of successful segmentation for 14 visemes of a single user are shown in Figure 3.3. Table 3.2 presents comparative results between manual segmentation and the proposed automatic temporal segmentation. From Table 3.2, it is observed that the overall results are similar for manual and automatic temporal

segmentation and there is no significant loss of information, resulting in no difference in the classification.

Table 3.2: Results of temporal segmentation for 14 visemes (three epochs)

		Epoch 1		Epoch 2		Epoch 3	
		Start frame	End frame	Start frame	End frame	Start frame	End frame
/a/	Manual	24	56	76	107	128	158
	Auto	27	56	75	107	128	158
/ch/	Manual	19	53	70	101	123	157
	Auto	19	53	70	102	122	155
/e/	Manual	49	78	102	133	171	211
	Auto	48	76	102	134	170	209
/g/	Manual	43	86	100	144	165	199
	Auto	43	83	100	142	163	198
/th/	Manual	1	33	58	95	120	155
	Auto	1	34	60	95	119	155
/i/	Manual	22	55	74	107	130	161
	Auto	24	54	73	107	129	161
/m/	Manual	21	55	86	120	154	188
	Auto	22	56	86	119	153	186
/n/	Manual	80	116	136	173	203	241
	Auto	79	116	136	175	204	240
/o/	Manual	26	58	76	114	134	169
	Auto	26	57	75	115	133	167
/r/	Manual	16	58	79	119	142	182
	Auto	19	57	81	118	147	182
/s/	Manual	22	59	80	120	143	181
	Auto	22	58	78	119	146	184
/t/	Manual	16	35	64	85	107	131
	Auto	15	34	63	83	108	132
/u/	Manual	64	101	122	160	180	214
	Auto	63	101	121	159	179	215
/v/	Manual	11	49	61	105	113	153
	Auto	11	48	62	101	113	152

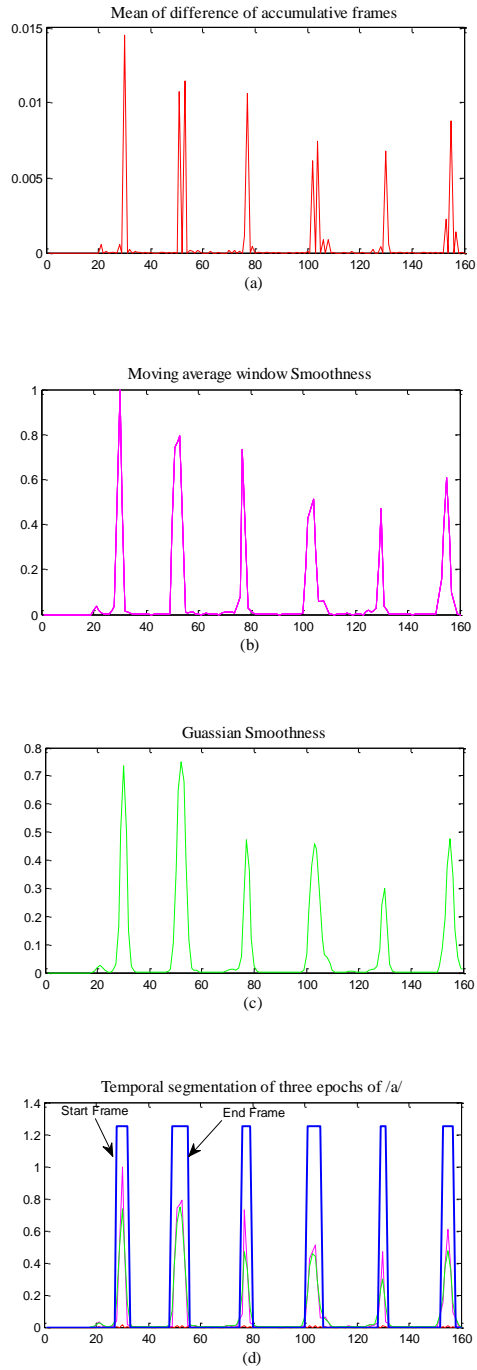


Figure 3.2: Results of Temporal Segmentation (a) Squared mean difference of accumulative frames intensities (b) Result of smoothing data by moving average window (c) Result of further smoothing by Gaussian filtering (d) Segmented data (blue blocks indicate starting and ending points)

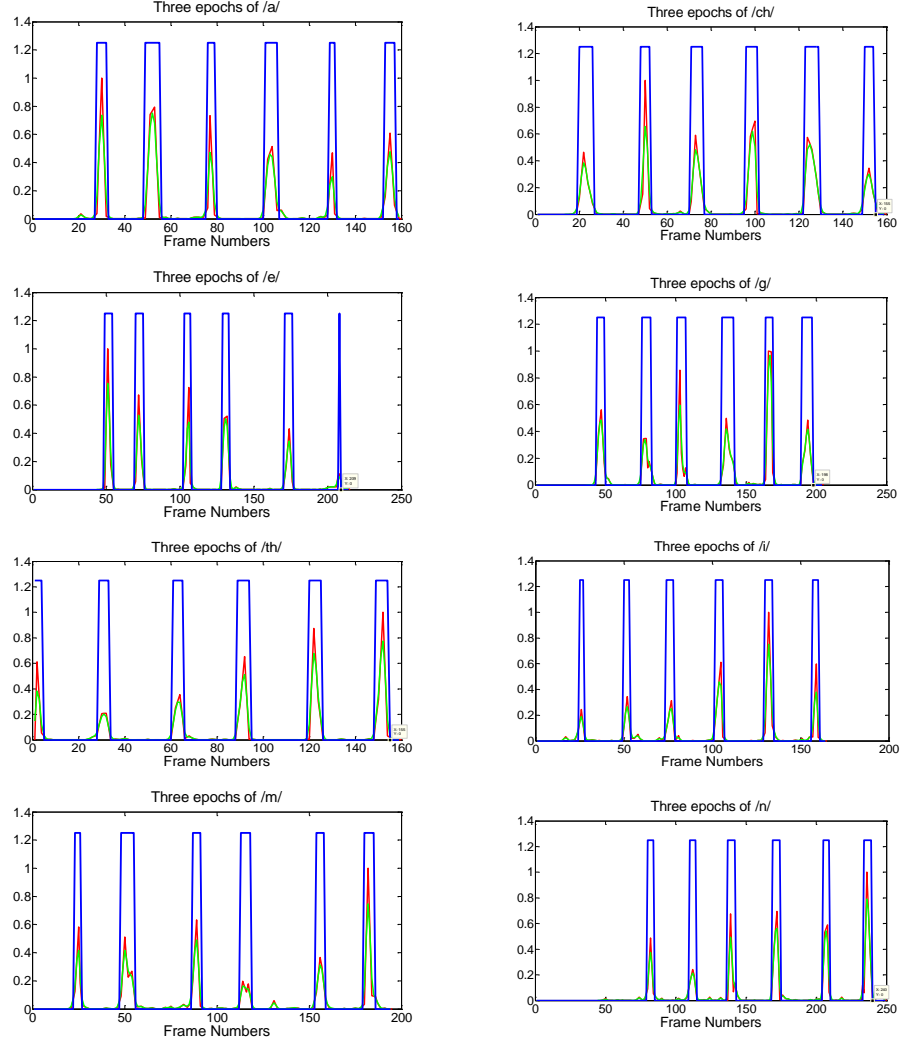


Figure 3.3: Results of Temporal Segmentation of all 14 visemes for a single user using the proposed method.

The results of other subjects were very similar, and the results from all the subjects and all the visemes in terms of starting frame error rate (SFER) and end frame error rate (EFER) have been calculated by Equation (3.2) and (3.3) which are tabulated in Table 3.3 and Table 3.4. From Table 3.3 and Table 3.4, the average error between automatic and manual segmentation for all subjects and all utterances is 2.98 frames/utterance. That is around 1.5 frames on either side of an utterance. It can be seen from Table 3.3 and that subjects 6 and 7 have a comparatively higher frame error rate. By manual observation it was found that both subjects had comparatively larger head movements in the videos during an utterance and had darker skin tones.

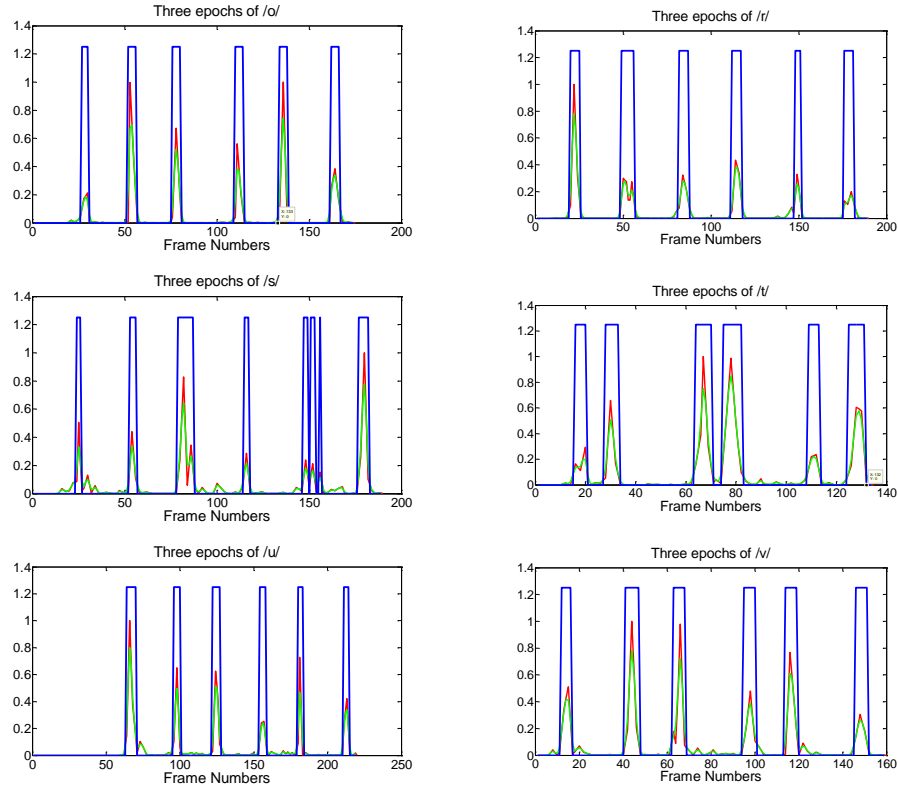


Figure 3.3: Continued

Table 3.3: Average Frame Error Rate of 10 Epochs at Start of an Utterance

Start Frame								
	subject1	subject2	subject3	subject4	subject5	subject6	subject7	Average
/a/	1.3	0.3	0.7	1	1.3	1.3	2	1.13
/ch/	0.3	0.7	0.3	1.7	1	7.7	4.7	2.34
/e/	0.7	2	0.7	0.3	1	1.7	1	1.06
/g/	0.7	1.7	1	0.7	1	2	3.3	1.49
/th/	1	0.7	1	1	1.7	5.7	3.3	2.06
/i/	1	0.7	1.3	1.7	0	3.3	3.3	1.61
/m/	0.7	0.3	1	2.7	0.7	3	1.3	1.39
/n/	0.7	1	1	0.7	1	3	0.7	1.16
/o/	0.7	0.3	0.7	1	0.7	2	5.7	1.59
/r/	3.3	1.7	0.3	1.3	1	5.7	3.7	2.43
/s/	1.7	0.3	0.3	2	1.7	0.3	2	1.19
/t/	1	1.3	1	0.7	7	1	3.3	2.19
/u/	1	1.7	0.7	1.3	4	1	1	1.53
/v/	0.3	0.3	0.3	1	1	1	3.3	1.03
Average	1.03	0.93	0.74	1.22	1.65	2.77	2.76	
Average error of start frame for all subjects, all visemes								1.48

$$SFER = \frac{1}{10} \sum_{i=1}^{10} |Start_Frame_{manual_i} - Start_Frame_{auto_i}| \quad (3.2)$$

$$EFER = \frac{1}{10} \sum_{i=1}^{10} |End_Frame_{manual_i} - End_Frame_{auto_i}| \quad (3.3)$$

Table 3.4: Average Frame Error Rate of 10 Epochs at End of an Utterance.

End Frame								
	subject1	subject2	subject3	subject4	subject5	subject6	subject7	Average
/a/	1	2	0.3	0.7	2.3	1	2.3	1.37
/ch/	1	0.7	0.3	1.3	2	0.7	4.3	1.47
/e/	1.7	1.7	2.7	1	1.3	0.3	1	1.39
/g/	2	0.3	1.3	1.3	0.3	0.7	2.3	1.17
/th/	0.3	0.7	1	0.7	2	1.7	3	1.34
/i/	0.3	2.3	1	0.7	0.7	1.7	2.7	1.34
/m/	1.3	1.7	1.3	1	1.3	1	1	1.23
/n/	1	2	0.3	1	1.3	2.7	2	1.47
/o/	1.3	0.3	0.7	1.3	0.7	3.3	3	1.51
/r/	0.7	2.3	1	2.3	0.3	0.7	3	1.47
/s/	1.7	1	1	1.3	1	2.3	2.3	1.51
/t/	1.3	0.3	0	0.7	0.3	1	2.3	0.84
/u/	0.7	1.3	1	1.7	0.7	3.7	1	1.44
/v/	2	1	0.7	2	1.3	1.7	4.3	1.86
Average	1.16	1.26	0.9	1.21	1.11	1.61	2.46	
Average error of end frame for all subjects, all visemes								1.29

One of the major limitations in using the proposed temporal segmentation technique is that it depends on the mouth motion signal, and each utterance is represented as two peaks. The first peak of an utterance represents the opening movement of the mouth. The second peak corresponds to the closing movement of the mouth, as can be seen in Figure 3.2 (d). It is difficult to differentiate between the start and the end of the utterance. If the start of an utterance has been missed due to the presence of noise in the signals, these segmentation errors will propagate to all the remaining frames in the image sequence. Another limitation of the proposed segmentation technique is that it requires a short pause period between two consecutive utterances to identify the desired frames. Hence we used the term *adhoc* temporal segmentation. Such a visual only temporal

segmentation technique is well suited for isolated utterances, where image sequences consist of multiple utterances separated by short pauses. However, there is a need for robust automatic temporal segmentation that should be able to separate the words and then from words the basic visual speech units, so that robust visual speech recognition can be performed on unknown words.

3.4 Image Noise Reduction

Once the temporal segmentation procedure has captured the start and end frame of an utterance, the corresponding image sequence is further processed for image de-noising. The degradation of an image can be caused by many factors such as movement during image capturing, atmospheric changes and varying illumination conditions. This degradation of an image can affect the optical flow computation. To reduce these effects of noise, the input image sequences are smoothed using Gaussian smoothing [88]. This technique is used in the pre-processing stage in a variety of computer vision applications in order to reduce the noise of images. Gaussian smoothing is performed by convolving each pixel in the input image with a Gaussian kernel, and is then summed up to produce the output smoothed image.

3.5 Issues that Need to be Considered for the Development of Visual Speech Recognition

Visual speech recognition basically depends on the efficient representation of lip movement. There are several issues that need to be considered for the development of a visual speech recognition system.

- Variation in illumination and colour induced by the video acquisition system (camera) and environmental cause image noise these distortions present in the image can affect the optical flow computation results. This unwanted distortion can be avoided by implementing some global illumination normalization techniques before applying feature extraction techniques. In this work a simple

Gaussian smoothing technique [88] is applied to reduce the illumination variations on ROI before optical flow computation.

- It has been commonly observed and has been demonstrated by [4], that there is always inter and intra subject variation in speaking. Repeated utterances of a number, letter or word by the same subject may vary even when all factors are kept constant. These variations in inter speaker affect the performance of lip-reading systems in speaker independent scenarios [121]. To a very great extent, this problem is addressed within the purview of the thesis by proposing robust feature and classification methods. Moreover, the proposed methods are compared using blind testing.
- In the development of a real time visual only speech recognition system, automatic recognition of the start and the end (frame) of an utterance is a challenging task. In continuous speech it is even more difficult to find the visual cues for the start and end frames of a single word or of a unit (viseme). In this work, an *ad hoc* scheme for temporal segmentation is proposed and shown to work on the dataset under consideration.
- Similar utterances can be of different temporal durations. Different utterances may have significantly different temporal durations. Hence, temporal normalization is required.
- Occlusion and self-occlusion of lips during utterance affects the motion templates based method (MHI). Use of DMHI in this work seems to have considerably overcome this issue.
- The projection of movement trajectories of lips depends on the observation view point (Frontal or Profile). In our work, only frontal views are considered.
- The distance between the camera and the subject and the view angle of the camera affect image-based measurements such as orientation, size and position, due to the projection of the utterance on a 2D plane. In this research, by using the Zernike and Hu moments which are the rotation and scale invariant features these issues are fairly addressed.

3.6 Summary

In this chapter the components of the pre-processing block have been reviewed with the main focus on temporal segmentation. A general review of face localization and then lip localization is presented. This chapter has presented information about the data set used in this research. The dataset is based on 14 visemes which can be concatenated to develop the words and sentences and hence are selected for this research.

An *ad hoc* temporal segmentation technique of isolated utterances based on pair-wise pixel comparison has been proposed and demonstrated. This method computes the mouth motion across the entire image sequence. The mouth motion is described using average energy features computed by a pair-wise difference of images. The mouth motion is represented by a one-dimensional motion signal, where high amplitude values represent the speaking signal and the low amplitude values represent a non speaking signal. The experimental results demonstrate the validity of the proposed approach. Finally, important issues relevant to the development of visual speech recognition are discussed.

In the next chapter, a novel feature extraction technique based on non overlapping blocks of the vertical component of optical flow is described.

Chapter 4

Mouth Movement Representation Using Optical Flow Based Motion Template

Motion of the object in a video (sequence of images) provides sufficient information to separate the object from the static background [122]. Motion capturing from the sequence of images and its processing for a variety of applications is a newly emerging technology. Using cutting edge technologies along with motion capturing, cross disciplinary applications can be developed. Some pertinent image processing techniques for motion representation are image differencing, optical flow analysis and background subtraction. As discussed in Chapter 2, the pertinent area for visual speech recognition is the area around the lips. In this region the main component of interest is the motion or movements of the lips. An efficient motion capturing technique is required for robust mouth movement representation. There are three different scenarios causing motion in a given scene:

- the camera is fixed and objects in the scene are moving
- the objects are more or less fixed whereas the camera is moving
- both camera and object are moving

The selection of motion capturing method is based on the static or dynamic background and its' performance varies accordingly. If the background is static, *i.e.*, there is no movement in the background, it is easy to eliminate the background and find the region of interest (ROI). Based on the above mentioned three scenarios, the parameters related to video recording such as distance of the camera and subject, viewing angle of the camera and lighting conditions are important. If these parameters are kept constant, only

the object (mouth) motion is need to be recovered. The dataset used in this research has fixed parameters.

Motion segmentation techniques have been applied on a sequence of images for a variety of applications to separate the ROI from the image sequences. The ROI for the visual speech recognition system is an area around the lips that contains a significant amount of motion during speech. In a real scenario the 3-D mouth movement is represented as 2-D motion on the image sequence recorded by a video system. This apparent motion on a 2D plane has to be recovered from the pixel values of an accumulative image sequence for efficient mouth representation. By identifying this motion of the lips, the stationary parts of the video data which is redundant can be detected. The static elements of the video data can be thus avoided for feature estimation in the subsequent classification stage. This allows for a compact and computationally efficient representation of mouth movement.

The visual speech features can be classified into appearance based and shape-based. The shape-based features are concerned with the representation of the mouth in terms of geometrical shape and dimensions of the lips such as width and height. To ease the extraction of lip contours artificial markers are applied on the face of a speaker [64]. In other approaches 2D or 3D model of the lip contour is used. However, use of artificial markers is not practical for real time system, and model based methods are often computationally demanding and require exact tracking of the lip movements. Another drawback of such approaches is that these require manual annotations on the training samples to develop the lip models. The annotation on the lip contours is sensitive to the facial skin colour and hairs on the face [78].

Appearance-based features are concerned with the low level features. These techniques use the information from the complete pixel values available in ROI. In this approach, the transform coefficients of the image pixels or the direct pixel values are used as appearance based features [14].

Contrary to appearance and shape-based features which describes the static shapes of mouth, motion-based features directly represent the lip motions in an image sequence

[72]. Motion tracking techniques [123-126] describe the facial expressions and human motion more efficiently than the underlying static poses techniques. Goldschen *et al.* [127] demonstrated that motion features for lip-reading are more discriminative when compared to static features.

This chapter describes a novel set of features for lip-reading that identifies visemes from visual data. The technique is based on a robust optical flow analysis that measures the lip movement while speaking. Optical flow is a technique representing the apparent movement in a sequence of images. Only the vertical component of the optical flow is used to extract the features. The vertical component of optical flow is decomposed into multiple non-overlapping fixed scale vertical and rectangular blocks and the statistical features of the each block are computed for the successive video frames of an utterance. A fixed sized temporal motion template of each utterance is developed for classification.

4.1 Introduction

A visual speech recognition technique that identifies visemes from image sequences is presented. The technique is based on lip movements measured by an optical flow analysis. Twenty years ago Mase and Pentland [72] used an optical flow analysis in their automatic lip-reading system to recognize connected English digits. However, due to the computational complexity of the optical flow and less efficient algorithms, it was not a popular method. Recently, the successful development of powerful CPUs and graphics processing units (GPUs) has made its implementation easy in real time systems [128]. The optical flow is defined as the distribution of apparent velocities of brightness pattern movements in an image [92]. The Optical flow technique is insensitive to background noise and lighting conditions and it can be evaluated without prior knowledge of the shape of the object. Therefore, the optical-flow analysis can detect robust visual features without extracting exact lip locations and contours. Other advantages of using optical flow for visual speech recognition are that lip motion features are more logical than the lip shape features, and visual features are independent from the speaker's mouth shape and other attributes of face [129].

This study has explored the contribution of the vertical and horizontal components of optical flow in lip-reading. Preliminary experiments have revealed that the salient motion features are available in the vertical component whereas horizontal features have a much lower contribution in viseme utterance. Hence, optical flow horizontal component is not included in this method. This reduces the number of features and reduces the overall computational burden of the system. The optical flow vertical component is decomposed into multiple non-overlapping fixed scale blocks and statistical features of each block are computed for successive video frames of an utterance. The extracted features from the vertical component are classified using support vector machine (SVM) classifier.

The experiments were conducted on a database (described in Chapter 3) of 14 visemes taken from seven subjects and the accuracy was tested using 5 and 10 fold cross validation for binary and multiclass SVM respectively to determine the impact of subject variations. Unlike other systems in the literature, the proposed method is more robust for inter-subject variations, with high sensitivity and specificity for 12 out of 14 visemes as indicated by the results. The overall viseme classification accuracy of 98.5%, with specificity of 99.6% and sensitivity of 84.2% have been achieved with one-vs-rest SVM classifier. Detailed experimental results and the definitions of the terms sensitivity, specificity and accuracy are given in Chapter 7.

4.2 Development of Optical Flow Based Motion Templates

Motion estimation is the process of finding a displacement (motion) vector of a pixel between two video frames. Motion description between two consecutive images at each pixel is known as an optical flow field. It is evident from the Middlebury optical flow benchmark [130] that the optical flow estimation methods are achieving steady progress by increasing the accuracy of underlying methods. These underlying methods have achieved a considerable level of reliability and accuracy from continuous research for three decades [131-133] and also because of the ever increasing computational ability. Computation of the optical flow motion estimation is presented in the following section. Reviews and evaluation of existing optical flow methods can be found in [134-137] and on the Middlebury optical flow benchmark website [130].

4.2.1 Optical Flow Motion Estimation

Optical flow is a measure of visually apparent motion of objects between two images and measures the spatio-temporal variations of video data. The word apparent implies that the optical flow does not consider the movement of the objects in the real 3D space, but the motion in the image space. Most techniques use two constraint equations to solve the optical flow: brightness constancy and spatial smoothness. The brightness or intensity constancy constraint (data term) implies that a displacement does not affect the intensity values of a point and hence the intensity value of that point remains the same although its position changes [138]. The spatial smoothness constraint (spatial term) comes from the basic idea of Lucas and Kanade [139]. It assumes that neighbouring pixels generally belong to the same surface, so that neighbouring pixels have a spatially constant image motion. Assuming a brightness constancy for optical flow estimation between two image frames which are taken at times t and $t + \delta t$, a pixel at location $P(x, y, t)$ with intensity $I(x, y, t)$ will have moved by $\delta x, \delta y$ in δt between the two image frames, so that:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \quad (4.1)$$

where $(\delta x, \delta y)$ is the spatial displacement during the time period δt and shows that a pixel maintains its intensity value during motion, and corresponding pixels in consecutive frames have the same brightness.

4.2.2 Gradient Based Approach

A differential approximation of the brightness constancy constraint, Equation (4.1), gives the gradient constraint equation. By simplifying the right side of Equation (4.1) using a Taylor series the equation becomes:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x} \frac{\delta x}{\delta t} + \frac{\partial I}{\partial y} \frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} \frac{\delta t}{\delta t} + H.O.T \quad (4.2)$$

where Higher Order Terms (H.O.T) are negligible and can be ignored, so that the above equation will become:

$$\frac{\partial I}{\partial x} \frac{\delta x}{\delta t} + \frac{\partial I}{\partial y} \frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} \frac{\delta t}{\delta t} = 0 \quad (4.3)$$

If we let:

$$\frac{\delta x}{\delta t} = u \text{ and } \frac{\delta y}{\delta t} = v \quad (4.4)$$

where u and v are the x (horizontal) and y (vertical) components of the velocity corresponding to the optical flow of the image intensity $I(x, y, t)$. $\partial I / \partial x$, $\partial I / \partial y$ and $\partial I / \partial t$ are the partial derivatives of the image at $P(x, y, t)$ in the corresponding directions. And these partial derivatives can be represented as

$$I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y} \text{ and } I_t.$$

$$\begin{aligned} I_x u + I_y v &= -I_t \\ \nabla I \cdot \vec{V} &= -I_t \end{aligned} \quad (4.5)$$

The above equation is commonly referred to as the gradient constraint equation, where $\vec{V} = (u, v)$ and $\nabla I = (I_x, I_y)$ represents the spatial gradients of I (intensity). This is a linear equation with two unknown variables so that it cannot be solved independently. It is essential to obtain another equation to compute the unknowns. This problem is known as the *aperture problem*. To solve this problem, researchers exploited the rigidity of surfaces in the scene in various ways. Generally the strategies employed are categorized into two major types: global and local methods.

4.2.3 Global Methods

Horn and Schunk [92] introduced brightness constancy and spatial smoothness constraint for a solution to optical flow estimation. They proposed an iterative gradient-based method combining Equation (4.5) with a global smoothness constraint. Optical flow can be estimated by minimising:

$$\int_D \nabla I \cdot \vec{V} + I_t + \lambda^2 (\nabla u^2 + \nabla v^2) dx, \quad (4.6)$$

where D is the domain of interest, and λ represents the influence of the smoothness constraint, defined as the sum of the square of the Laplacians of u and v .

Global techniques however, are not robust to outliers due to motion boundaries, reflection and occlusion. Black and Anandan [138] attempted to resolve the issue of such outliers but could not obtain a true brightness constancy errors and flow derivatives.

These techniques have another problem in that they implicitly assume a single motion in the scene. Various researchers such as Weickert and Chnorr [140], and Brox *et al.* [141] have extended the Horn and Schunck [92] algorithm in an attempt to overcome these issues. All are computationally intensive for real-time applications on existing computer hardware.

4.2.4 Local Methods

To solve the two unknowns in the gradient constraint equation, Lucas and Kanade [139] proposed a spatially local method, using a least squares estimator as:

$$E(V) = \sum_x W(x) [\nabla I \cdot \vec{V} + I_t]^2 \quad (4.7)$$

where $W(x)$ denotes a weighted spatial window function. This weighting is generally used to increase the influence of the neighbourhood.

The computation of \vec{V} can be achieved by minimizing $E(V)$ with respect to the parameters of $\vec{V} = (u, v)$,

$$\frac{\partial E(V)}{\partial u} = 0 \quad (4.8)$$

$$\frac{\partial E(V)}{\partial v} = 0 \quad (4.9)$$

In matrix form this linear system may be defined as

$$M\vec{V} = b \quad (4.10)$$

where

$$M = \begin{bmatrix} \sum W(x)I_x^2 & \sum W(x)I_xI_y \\ \sum W(x)I_xI_y & \sum W(x)I_y^2 \end{bmatrix} \quad (4.11)$$

and

$$b = - \begin{bmatrix} \sum W(x)I_xI_t \\ \sum W(x)I_yI_t \end{bmatrix} \quad (4.12)$$

From this local optical flow \vec{V} can be computed.

Comparatively, local methods achieve better accuracy than global methods [134]. Whilst, the local technique can be affected at motion boundaries by breakdown of local flow models. However, these effects occur only on local regions and do not influence surrounding optical flow vector estimations. The removal of global regularization of the flow field also improves the efficiency of local methods. In addition, local methods compute the flow across small windows of the surface and do not try to compute the flow across an entire surface. This approach for local methods provides more flexibility. In 2005, Bruhn *et al.*[142] proposed that integrating local velocity constraints with global regularization improved the accuracy of dense optical flow. However, this hybrid approach increases the computational burden significantly.

4.2.5 Optical Flow Computation Used in this Research

A robust optical flow method and a robust way of feature representation provide a better performance which can be seen from the results. Based on above discussed trade-offs of optical flow techniques, in this research a sophisticated regularization method of optical flow proposed by Sun *et al.* [143] was adopted. This is based on statistical learning of both the brightness constancy error and the spatial properties of optical flow and provide a complete probabilistic model of optical flow. Sun *et al.* proposed an estimation of the optical flow between two input images I_1 and I_2 where probabilistic assumption and decomposition of the *a posteriori* probability density of the flow field (u, v) is computed. Equivalently, its negative logarithm is minimized:

$$p(u, v | I_1, I_2; \Omega) \propto p(I_2 | u, v, I_1; \Omega_D) \cdot p(u, v | I_1; \Omega_S) \quad (4.13)$$

where Ω_D and Ω_S are the parameters of the model:

$$E(u, v) = E_D(u, v) + \lambda E_S(u, v). \quad (4.14)$$

In Equation (4.14), E_D is the negative logarithm (*i.e.*, energy) of the data term, E_S is the negative log of the spatial term (the normalization constant is omitted in each case) and λ is a regularization parameter.

Generally, the optimization of such energies is difficult, due to many local optima and also non-convexity. The non-convexity in this approach stems from the fact that the learned potentials are non-convex and also from the warping-based data term used here and in other competitive methods [141]. To limit the influence of false local optima, a series of energy functions is constructed:

$$E_C(u, v, \alpha) = \alpha E_Q(u, v) + (1 - \alpha) E(u, v) \quad (4.15)$$

where E_Q is a quadratic, convex, formulation of E that replaces the potential functions of E by a quadratic form and uses a different λ . Note that E_Q amounts to a Gaussian Markov Random Field (MRF) formulation. The control parameter $\alpha \in [0, 1]$ varies the convexity

of the compound objective and allows a smoother transition from 1 to 0. The combined energy function in Equation (3.10) changes from the quadratic formulation to the proposed non-convex one [144]. During the process, as soon as the solution at a previous convexification stage is computed, the system uses this solution as initialization for the current stage. In practice, it is observed that the use of three stages produces reasonable results. A simple local minimization of the energy was performed at each stage. At a local minimum, it holds that:

$$\nabla_u E_C(u, v, \alpha) = 0 \text{ and } \nabla_v E_C(u, v, \alpha) = 0. \quad (4.16)$$

Since the energy induced by the proposed MRF formulation is spatially discrete, the gradient expressions can be derived. Setting these to zero and linearizing them, the results are rearranged into a system of linear equations, which can be solved by a standard technique. The main difficulty in deriving the linearized gradient expressions is the linearization of the warping step. For this the approach of Brox *et al.* [141] has been followed, using the derivative filters [142].

Large displacements may be caused by sudden movement during fast speech. Standard optical flow techniques are unable to capture such large displacements due to the temporal resolution limitation. To overcome this limitation, image warping technique based on incremental multi-resolution analysis was incorporated [142, 143]. In this approach the optical flow estimated at a coarser level is used to warp the second image toward the first at the next finer level and the flow increment is calculated between the first image and the warped second image. The final result combines all flow increments. At the first stage where $\alpha = 1$, a 4-level pyramid with a down-sampling factor of 0.5 is used. At other stages, only a 2-level pyramid with a down-sampling factor of 0.8 is used, to fully utilize the solution at the previous convexification stage. As identified by Horn and Schunck [92] classical optical flow methods suffer drawbacks such as requiring brightness constancy and spatial smoothness. Other limitations include outliers in the image due to the edges of the frame. Sun *et al.* [143] developed a probabilistic model integrated with the optical flow technique developed by Black and Anandan [138]. This approach overcomes the limitations of requiring spatial smoothness and constant

brightness and hence has been used in our proposed method. Barron *et al.* [134] have demonstrated that the optical flow estimation is more accurate using colour information, so that the optical flow was obtained for each of the three colour components of the image to reduce the optical flow estimation errors caused by illumination variations [145]. An example of optical flow estimation is shown in Figure 4.1. Arrows shows the direction of motion between two images.

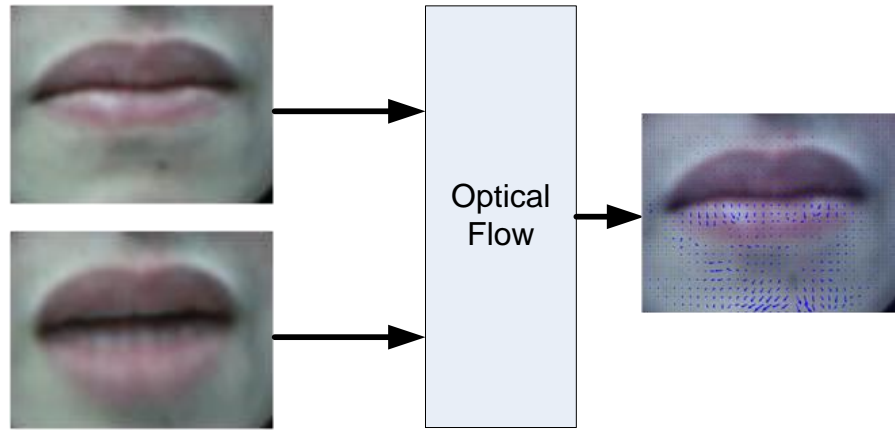


Figure 4.1: Example of optical flow computation using two consecutive images.

4.2.6 Non Overlapping Block Based Approach

In this section, the data-processing performed to develop the feature vector template suitable for the training and testing of SVM classifier is discussed. Once the temporal segmentation of isolated utterances is performed, the cropped video containing a viseme is fed to the system for optical flow computation. As discussed earlier the dataset used in this study contains only the mouth region, resulting in there being no need for the localization of the mouth (ROI). From the computed optical flow field we extracted a set of robust features that represent the amount of vertical movement in the ROI in uttering a viseme. A pilot study was conducted on the collected data to determine the appropriate features which were subsequently used. The results shown in Table 4.1 indicate the average vertical and horizontal movement of an utterance. It can be observed from the table that the horizontal movement of the lips during viseme utterance was insignificant in comparison with the vertical movement and has not contributed significantly in the

performance of the classification of visemes. Therefore, the optical flow corresponding to the horizontal motion was ignored, and only the vertical component of optical flow was considered in this technique. This resulted in a significant reduction in computation complexity during real time implementation.

Table 4.1: Average vertical and horizontal movement for an utterance.

	Horizontal Movement	Vertical movement
Average	0.97	3.12
Standard deviation	0.43	1.33

Unlike the approaches of [72, 129], in which they extracted the features of each frame globally, in this research the optical flow vertical component field is divided into non-overlapping blocks (vertical and rectangular blocks) to retain the salient features. Therefore, this method defines a set of regions around the ROI, and from each region global statistical measures of the optical flow are computed. This approach provides a better representation of motion estimation as compared to global approach. The reason behind the better performance of the block based approach is that visual speech is bi-directional, and each lip always moves in an opposite direction. Thus the vertical component in certain parts of the mouth is cancelled out by averaging, and hence the global features of the complete ROI are not suitable for lip reading, even though the global features of the optical flow can be valuable in some applications. Figure 4.2 shows the system flow diagram of the non-overlapping block based approach. Once, temporal segmentation of an isolated utterance is performed. The corresponding segmented video of 30 frames/sec with a resolution of 240×320 pixels that contains a viseme were given as the input to the visual speech recognition system. Similar subsequent frames which result in zero energy difference between frames are filtered out using the mean square error (MSE) given in Equation (4.17) to reduce the inter and intra subject variation in the speed of speaking.

$$MSE = \frac{1}{m \times n} \sum_{x=1}^m \sum_{y=1}^n [I_1(x, y, t) - I_2(x, y, t)]^2 \quad (4.17)$$

Removing similar subsequent images has an additional advantage of reducing the computational burden while calculating optical flow. The system computes the optical flow between consecutive images. The optical flow computation provides the vertical and horizontal components separately. Each vertical component frame is divided into 8 non overlapping fixed size blocks of size 240×40 pixels as can be seen in Figure 4.3. Each block's statistical property (average intensity) is computed, so that each frame is represented as 8 values in a row. To develop the motion template of corresponding utterance, each row of with 8 values was stacked to develop a matrix of $n \times 8$. There are large inter and intra subject variations in the speed of an utterance and this results in a difference in the number of frames (*i.e.*, rows) for each utterance. The number of frames for each utterance was normalized such that the template size for each of the utterances was the same. This normalization was achieved using a linear interpolation to obtain a constant 10 frames for each utterance. Finally this 10×8 size template *i.e.*, an 80 dimensional feature vector represented a viseme and was used for training and testing of the classifier. The results are presented in Chapter 7 Section 7.1.1.

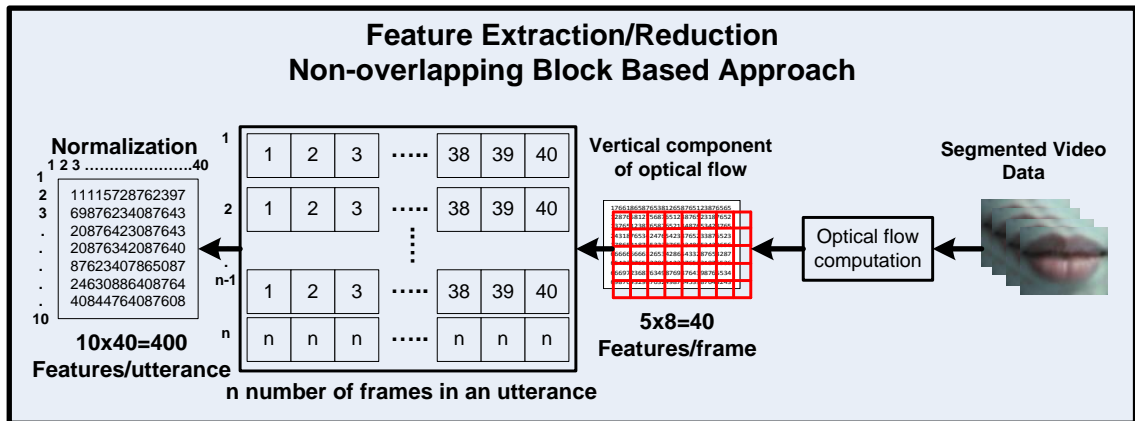


Figure 4.2: System flow diagram of non-overlapping block based approach.

However it was shown that this approach of division into vertical columns does not provide the best results. Further division of ROI into rectangular blocks was performed with the block size being optimized iteratively.

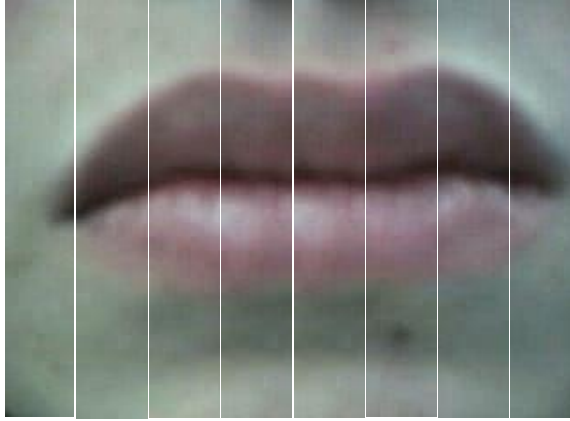


Figure 4.3: Vertical block based approach

4.2.6.1 Block Optimization

The ultimate intention of any simulation based work is to lead towards a real time system, which therefore requires data to be reduced to a minimum. For this purpose, experiments were conducted to optimize the size of the blocks that could be used (Figure 4.4). After experimenting with 7 different block sizes (40×32, 32×32, 48×40, 24×20, 30×40, 30×20, 240×40 pixels) (see Table 4.2), a block size of 48×40 pixels was chosen for all of our experiments as it represents a good compromise between sensitivity, accuracy and the number of features. As a result, each image is divided into blocks of size 48×40 pixels, resulting in 40 blocks (5 rows × 8 columns) per optical flow frame as shown in Figure 4.4.

Table 4.2: Average classification results of seven different block sizes.

	Block Size pixels	Specificity %	Sensitivity %	Accuracy %	No: of Features/ Viseme
1	48x40	99.6	84.2	98.5	400
2	30x40	99.7	83.7	98.6	640
3	40x32	99.7	84.2	98.6	600
4	30x20	99.8	79.5	98.4	1280
5	32x32	99.8	81.1	98.4	1000
6	24x20	99.8	81.1	98.4	600
7	240x40	98.1	66.4	95.9	80

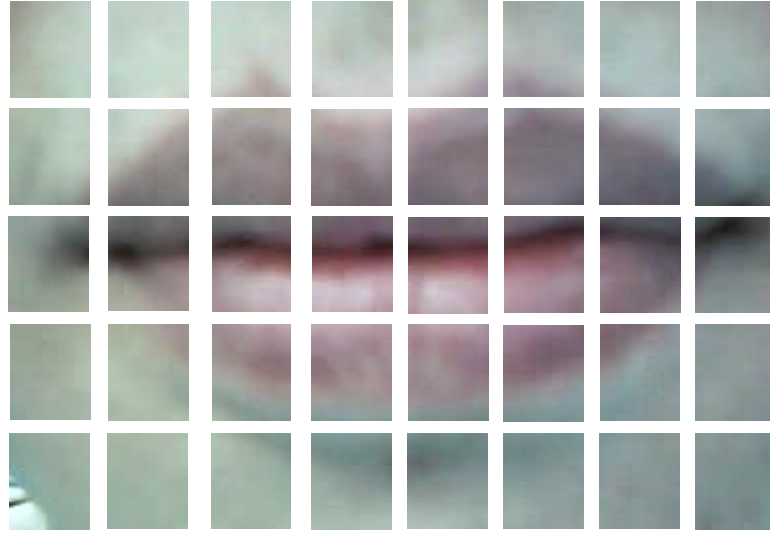


Figure 4.4: Direction of movement of each muscle and the block arrangement used for optical flow feature extraction.

To develop the template for each viseme, each optical flow frame was represented by the average of each block which resulted in an array of 5 rows and 8 columns, in total 40 features per frame (one feature representing each block). Each matrix of 40 elements is converted to a row vector. Rows of subsequent optical flow were stacked to develop the template matrix. To overcome the difference in the speed of speaking, each utterance was normalized (temporally) to 10 frames using a linear interpolation. This resulted in a final feature vector of size 400 (10×40).

4.2.7 Normalization of speed of speech

The speeds of speech between inter and intra subjects vary for each repetition of a phoneme. This variation in the speed of an utterance results in variation in overall duration of an utterance. In videos this variation results in a varying number of frames so that a varying number of visual features will be obtained. It is difficult to model a classifier with large variations in the number of features per utterance. A simple two phase approach to normalize the overall duration of an utterance is adopted. In first phase at the time of the optical flow computation, similar subsequent frames containing

the zero level energy differences are eliminated for optical flow computation. In the second phase each optical flow template with varying number of frames was normalized to 10 frames using linear interpolation so that the no precision is lost.

4.3 Summary

This chapter has provided a novel set of features computed from the optical flow vertical component. A novel block based approach has shown significant improvement when compared to the proposed DMHI approach. A detailed discussion of the results is presented in the Chapter 7. In this chapter detailed description of the optical flow estimation has been given. In addition, the optimal size and number of blocks for feature extraction is discussed. It is concluded that a suitable size of block is required for optimization of the system. Vertical components' vertical blocks have shown less accuracy in comparison to small rectangular blocks. Because the values in the vertical blocks represent both the upper and lower lip motions at the same time, the corresponding positive and negative values cancel each other during the computation of the mean value of the block. The rectangular block size which giving the best results is approximately segmenting the mouth from the centre of the mouth, which separates the upper and lower lip motions.

Chapter 5

Mouth Movement Representation Using Directional Motion History Images

In any video sequence, the change in the consecutive images can be detected by subtracting pixel values of consecutive images that provide the regions with movements. Jain [146] has proposed the Accumulative Difference Picture (ADP) for change detection in dynamic scene analysis. The main goal of motion segmentation in this study is to capture the mouth movement that occurs in a desired temporally segmented image sequence. This chapter discusses another novel and robust mouth motion representation technique. This technique consists of four mouth motion patterns obtained from the vertical and horizontal components of the optical flow rather than from the structural information of movement or the underlying difference of image technique. This method represent the entire space-time dimensions of mouth motion of an utterance by four 2D gray scale images, with each image retaining the essence and temporal structure of the directional movement that is up, down, left and right. This technique is known as the directional motion history images (DMHIs). It is an evolution of the traditional motion history image (MHI) [147]. The MHI is an appearance based technique that preserves the entire space-time dimensions of motion in an image sequence in a single 2D gray scale temporal template that retain pertinent motion information available in video frames.

The gray levels of each directional motion history image describe the measure of motion in respect to time. Thus the intensity value of each pixel of each directional motion history image corresponds to a function which represents the temporal history of motion at particular pixel location in their respective directions. This approach recognizes the

directional motion by computing the apparent motion velocities by the robust optical flow method.

The following section describes the general MHI technique, followed by the development of DMHIs based on optical flow analysis. The experiments compared the performance of the proposed DMHI method with the MHI, computing the Zernike moments from each image.

5.1 Motion Representation Theory

Computer vision based on motion analysis has many applications, including region of interest (ROI) segmentation, object tracking (e.g vehicle or human) and human action recognition, as well as lip-reading from an image sequence. For motion segmentation, frame to frame differencing [148-152] methods have been commonly used. These temporal differencing methods which employ two [149, 151, 152] or three consecutive frames [148, 150] are suitable for dynamic environments, although in general poor relevant features are achieved. To generate the motion history image (MHI) and the motion energy image (MEI), temporal differencing methods are employed [153]. However the optical flow computation around the ROI represented as an alternative to the temporal differencing method to generate the MHIs. Because it directly describes the actual movement in the ROI (mouth), the proposed DMHIs method based on optical flow has outperformed the traditional MHI method.

5.1.1 Basics of Motion History Image (MHI)

Motion History Image (MHI) is an appearance based method used to describe the direction of motion in image sequence. The intensity of each pixel in an image sequence is a function of motion density at that location, and therefore the temporal difference of these pixel values results in MHI, being a temporal template. One of the advantages of the MHI representation is that a range of image frames in several seconds of times may be encoded in a single gray scale image frame and in this way MHI can span the time scale of human visual speech. The resulting single scalar-valued image contains brighter pixels where there is recent movement and darker where the movements are older [154,

155]. Bobick and Davis [156] first proposed a spatiotemporal model for human action representation and recognition by using motion history and energy images (MHI/MEI). The MEI represents a binary motion image that describes where a motion is in an image sequence. The MHI $H_\tau(x, y, t)$ can be obtained from an update function $\Psi(x, y, t)$, which represents the brighter pixels where there is recent movement and darker where the movements are older:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } \Psi(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - \delta) & \text{otherwise} \end{cases} \quad (5.1)$$

where update function $\Psi(x, y, t)$ signals the presence of object (or motion) in the current video image, x , y and t show the position and time, τ decides the temporal duration of MHI, and δ is the decay parameter. To define $\Psi(x, y, t)$ image differencing, optical flow and background subtraction techniques can be used. Figure 5.1 shows the development of two MHI images for the utterances /a/ and /m/. Usually, MHI is developed from a binarized image, computed from frame subtraction [26], using a predefined threshold value, \wp to obtain a motion or no motion classification:

$$\Psi_B(x, y, t) = \begin{cases} 1, & \text{if } Diff(x, y, t) \geq \wp \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

where $Diff(x, y, t)$ with difference distance Δ is as follows:

$$Diff(x, y, t) = |I(x, y, t) - I(x, y, t \pm \Delta)| \quad (5.3)$$

where, $I(x, y, t)$ is the intensity value of pixel located at coordinate (x, y) in the t^{th} frame of the image sequence.

Motion energy image (MEI) is the binary representation of the motion in an image sequence that shows where a motion has occurred in a specific video. In comparison to

MHI, the moving object cleans a particular region of the image and this form can be useful for the determination of motion occurrence [124] in MEI.




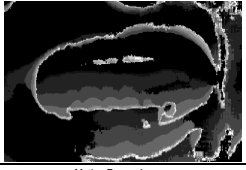






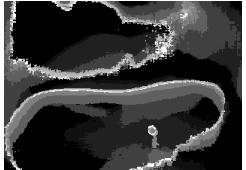

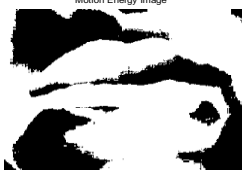

Viseme	First Frame	Middle Frame	Last Frame
/a/			
MHI of /a/			
MEI of /a/			
/m/			
MHI of /m/			
MEI of /m/			

Figure 5.1: Development of MHI images for two utterances /a/ and /m/.

The MEI $E_{\tau}(x, y, t)$ can be represented as:

$$E_{\tau}(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i) \quad (5.4)$$

The MEI can be extracted from the MHI by thresholding the MHI above zero [154].

$$H_{\tau}(x, y, t) = \begin{cases} 1 & \text{if } H_{\tau}(x, y, t) \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.5)$$

As can be seen from Figure 5.1, the gray-scale MHIs are sensitive to direction whereas MEIs do not provide information regarding the direction of motion, and thus the MHI is more suitable for discriminating the motion in opposite directions [157]. However, both the MHI and the MEI images together may provide better discrimination than either of the alone [154], depending on the application. In the following section, procedure for the development of DMHI is described. This method employs the probabilistic model of optical flow, described in chapter 4.

5.1.2 Development of Directional Motion History Images (DMHIs)

Optical flow methods [92, 135, 139, 158-161] can be used for the motion segmentation as well as for MHI development for various applications. Instead of the traditional frame subtraction or background subtraction method used to calculate the update function presented in Equation (5.1), this method employs the probabilistic model of optical flow developed by Sun et al [143] to compute the DMHIs. Computing quality optical flow from an image sequence is a challenging task, considering the conclusion regarding lower accuracies of optical flow presented by Gray et al [162] that non optimized thresholding provides very sparse optical flow fields. To obtain better results in presence of motion and its direction, the probabilistic model of optical flow [143] is employed to reduce outliers, although the optical flow method is computationally expensive and sensitive to noise but performs well in presence of camera motion [163]. The DMHIs constructed by these refined optical flow vectors provide a clearer picture of the presence of motion and its direction (i.e up, down, left and right). Moreover, from a real time perspective, the implementation of optical flow algorithms on FPGAs [128], and the successful development of powerful CPUs makes its implementation easy.

To produce the four DMHIs (up, down, left and right), the optical flow is computed between two consecutive frames and separated into four channels as shown in Figure 5.2. Detailed description of Optical flow computation is given in Section 4.2.5. With reference to the proposed visual speech recognition system presented in section 2.10 and depicted in Figure 2.6 the pre-processing and the temporal segmentation blocks that perform the illumination effect reduction and the temporal segmentation of discrete utterances *i.e.*, to determine the start and end frames of the utterances, are described in the Chapter 3 Sections 3.3 and 3.4.

It has observed that there is inter and intra subject variation in the speed of speech. This variation in speed can give rise to different perceptual impression and can cause inexact viseme recognition. To compensate for the variation in speed of speaking, the mean square error (MSE) given in Equation (5.6) between two subsequent frames was computed and only those frames with nonzero differences were used to compute the optical flow. Removing similar sequential images has the additional advantage of reducing the computational load.

$$MSE = \frac{1}{m \times n} \sum_{x=1}^m \sum_{y=1}^n [I_1(x, y) - I_2(x, y)]^2 \quad (5.6)$$

The optical flow vectors obtained from an image sequence (denoted by $\Psi(x, y, t)$) are first divided into two scalar fields corresponding to the horizontal and vertical components of the flow, Ψ_x and Ψ_y . These components are then half-wave rectified into four non-negative separate channels Ψ_x^+ , Ψ_y^+ , Ψ_x^- , and Ψ_y^- constrained such that:

$$\Psi_x = \Psi_x^+ - \Psi_x^- \quad (5.7)$$

$$\text{and } \Psi_y = \Psi_y^+ - \Psi_y^- \quad (5.8)$$

Figure 5.2 depicts the flow diagram of the four flow vector computations. Based on each of the four directions, each optical flow image is normalized according to the threshold ξ

value, where ξ is computed according to Otsu's [164] global threshold method. Based on these normalized image sequences, four separate optical flow based motion history templates are developed after deriving the four optical flow components.

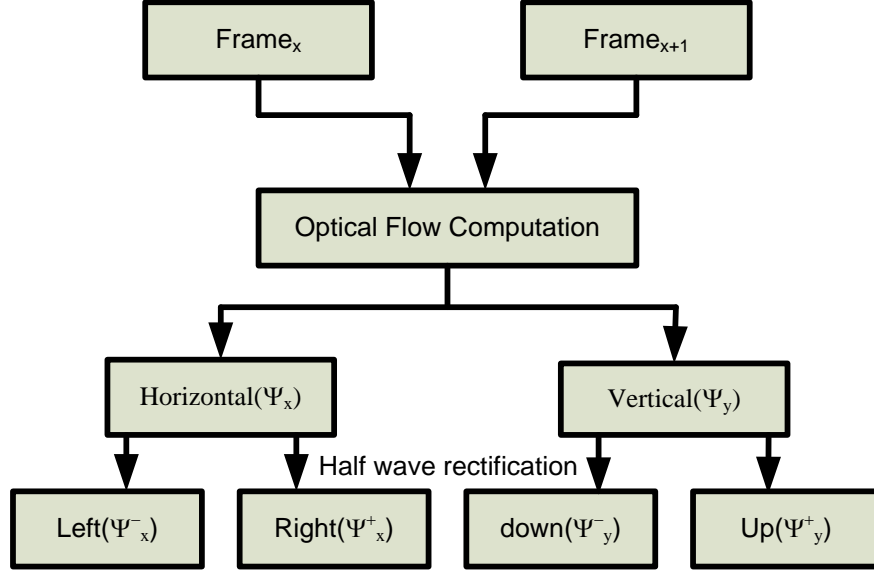


Figure 5.2: Conceptual framework of optical flow vector separation into four directions.

$$\begin{aligned}
 H_{\tau}^{-x}(x, y, t) &= \begin{cases} \tau & \text{if } \Psi_x^{-}(x, y, t) > \xi \\ \max(0, H_{\tau}^{-x}(x, y, t-1) - \delta) & \text{otherwise} \end{cases} \\
 H_{\tau}^{+x}(x, y, t) &= \begin{cases} \tau & \text{if } \Psi_x^{+}(x, y, t) > \xi \\ \max(0, H_{\tau}^{+x}(x, y, t-1) - \delta) & \text{otherwise} \end{cases} \\
 H_{\tau}^{-y}(x, y, t) &= \begin{cases} \tau & \text{if } \Psi_y^{-}(x, y, t) > \xi \\ \max(0, H_{\tau}^{-y}(x, y, t-1) - \delta) & \text{otherwise} \end{cases} \\
 H_{\tau}^{+y}(x, y, t) &= \begin{cases} \tau & \text{if } \Psi_y^{+}(x, y, t) > \xi \\ \max(0, H_{\tau}^{+y}(x, y, t-1) - \delta) & \text{otherwise} \end{cases}
 \end{aligned} \tag{5.9}$$

For positive and negative horizontal directions, $H_{\tau}^{-x}(x, y, t)$ and $H_{\tau}^{+x}(x, y, t)$ are set up as motion history templates, whereas $H_{\tau}^{-y}(x, y, t)$ and $H_{\tau}^{+y}(x, y, t)$ represent the

positive and negative vertical directions (up and down). In this motion separation method, four motion history templates that approximate the directions of the motion vectors are developed.

In DMHI computation of various utterances by multiple subjects results in different number of frames for different utterances; even the same viseme repeated by the same subject has some variation in the number of frames. This can be seen in Table 3.1. Thus, the effect of parameter τ in DMHI computation is crucial, the reason being that if an utterance takes 30 frames (*i.e.*, $\tau = 30$) by one subject, then the maximum value in the produced DMHIs will be 30 (because $\tau = 30$), whereas, if the same utterance is uttered a little slowly either by the same or by a different subject, and taking 38 frames (*i.e.*, $\tau = 38$) then the maximum value for the developed DMHIs will be 38. This intensity variation might produce slightly more isolation in the same utterances. Therefore, this unwanted variation in similar utterances can be mitigated by incorporating intensity normalization in computing the DMHI templates. Consequently, a simple normalization method was employed for different utterances, uttered by multiple subjects with a varying number of frames. The $[\tau_{\min}, \tau_{\max}]$ values were transformed into the range of $[0,1]$. Based on this normalization approach, the produced DMHI images were converted into the range of $[0,1]$ for each of the utterance.

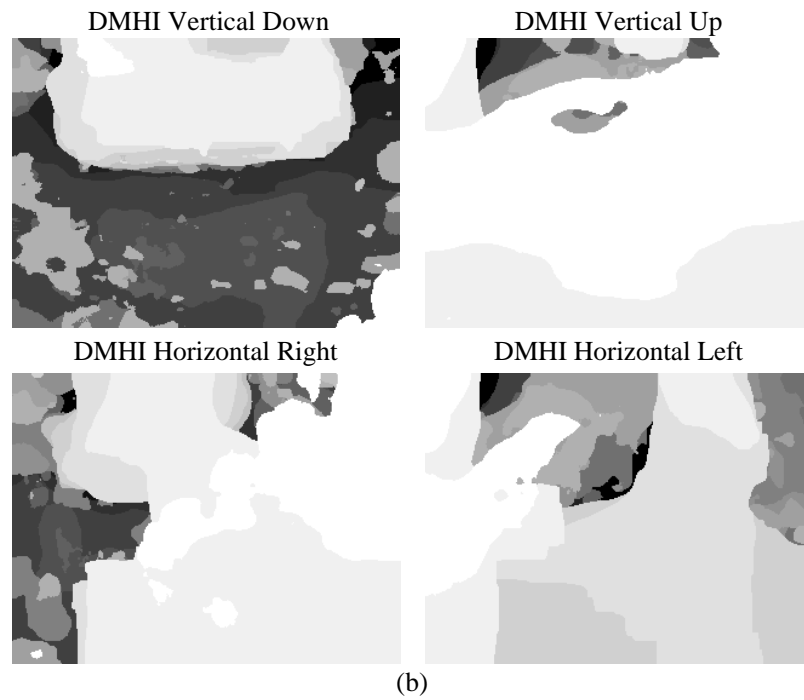
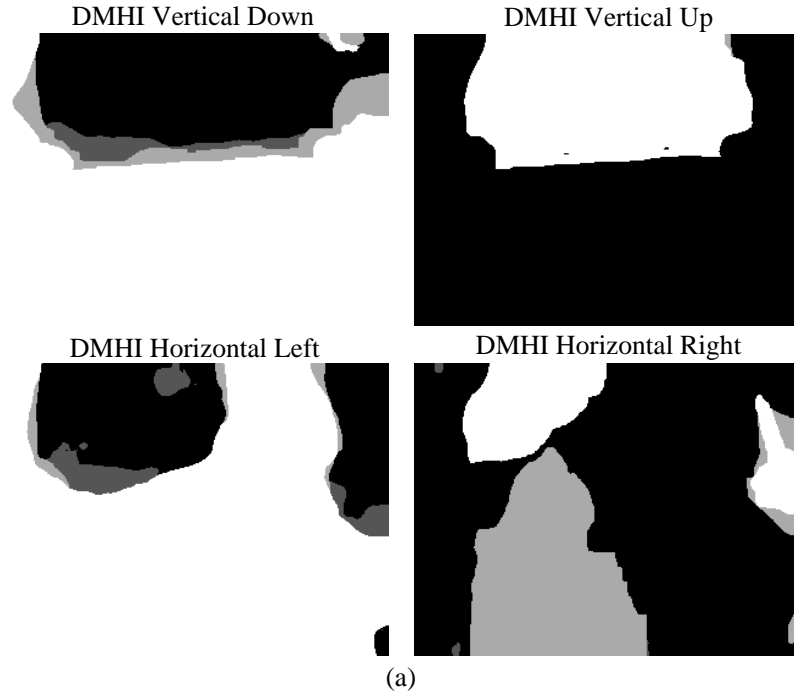


Figure 5.3: (a) Four directional motion history images (DMHIs) of first three frames of an utterance /a/, (b) Complete DMHIs of an image sequence of an utterance /a/.

Figure 5.3 (a) depicts an example of four directional motion history images of the first three frames of an image sequence of an utterance /a/, where more recent movement of pixels is brighter and older pixel movement is darker. It can be seen from the DMHI vertical down image that in the utterance /a/, the lower lip moves down so that most of the lower portion of that image is brighter (white), whereas other grey level pixels show the older movement in the downward direction. Similarly, in the vertical up image the upper part is brighter which shows that the upper lip moves up in uttering /a/ in the initial frames. However, optical flow computation is a global pixel based technique, and the dataset used in this research contains only ROI and there is no static background which separates the static and dynamic portions in an image sequence. In such a type of dataset where all pixels are somehow moving, it is hard to describe such motion history images. Pixel movement would be apparent where there is a static background and objects are moving such as that of human or vehicles in videos with a static background. Figure 5.3 (b) shows an example of four directional motion history images of a complete video of an utterance /a/, where brighter pixels show more recent movement, and vice versa in corresponding directions. Feature vectors are computed from these four history templates by employing Zernike and Hu moments for classification and recognition. Detailed description of feature extraction and classification techniques is given in Chapter 6.

5.2 Advantages and Disadvantages of Directional Motion History Images

One of the major issues with the generic MHI method is motion occlusion caused by other objects, or self occlusion resulting in motion overwriting. Example are human motions such as sitting and standing or in visual speech, long motion sequences such as repeated opening and closing of the mouth. Human actions such as sitting down and standing up at a fixed location have opposite directions. As described earlier, the MHI is based on scalar values, where more recent pixel movement is described brightly and older motion is darker. If a person sits down, the MHI of that action shows the brighter pixels in the lower part of the image. In the same image sequence, if the person stands up the final MHI image contains the brighter pixels in the upper part of the image and

overwrites the previous sit down action so that only the stand up motion is represented in the final MHI. Similarly, in opening and closing of the mouth, the upper and lower lips have opposite actions at their reference points and this causes overwriting in the MHI.

Several methods have been developed to solve the issue of overwriting, in order to represent the multi-directional activities in the form of MHI efficiently. Multilevel MHI(MMHI) [165-167] and Hierarchical Motion History Histogram (HMHH) [71] approaches are proposed to overcome the issue of overwriting. The aim of MMHI is to overcome the problem of motion self occlusion in specified video sequence by obtaining multiple MHIs. In effect, all the MHIs should have a fixed number of history levels n , for that each video is sampled to $(n+1)$ frames.

The MMHI is represented as follows:

$$MMHI_t(x, y, t) = \begin{cases} s * t & \text{if } \psi(x, y, t) = 1 \\ MMHI_t(x, y, t-1) & \text{otherwise} \end{cases} \quad (5.10)$$

where the intensity step between two history levels is $s=(255/n)$.
 $MMHI_t(x, y, t) = 0$ for $t \leq 0$.

From each of these motion templates final multilevel-MHI is computed by iteratively combining all the short term motion templates $t= 1, \dots, n+1$. This method encodes motion which takes place at different time instances on the same location, such that it is uniquely decoded. A simple bit-wise coding scheme is used. If a motion occurs at time t at pixel location (x, y) , it adds 2^{t-1} to the old motion value of the MMHI as follows:

$$MMHI(x, y, t) = MMHI(x, y, t-1) + \psi(x, y, t) \cdot 2^{t-1} \quad (5.11)$$

The proposed MMHI system was demonstrated for the automatic detection of facial actions that show expressions. Due to the bitwise coding scheme, multiple actions occurring at the same position [166] can be separated. This system requires a sensitive registration system, and all the image sequences must have same scale or size and faces

in the video frames should be at the same position. Consequently, MMHI has not clearly demonstrated any superiority when compared to basic MHI. Ahad *et al.* [168] showed that MMHI and HMHH are not efficient when compared to DMHI, by implementing the MMHI with two different datasets. It has also been demonstrated by Ahad *et al.* [168] that the difficulty of self occlusion or motion overwriting in the MHI method can be considerably resolved by using DMHI method. Another possible solution to motion occlusion is proposed by Yau [75] by increasing the single camera views to multi-camera views. However, this is not suitable for real time system and furthermore, the fusing and processing of video inputs from multiple cameras is a complex task.

One of the important features of the motion template based visual speech recognition system is the ability of MT to preserve the short duration dynamic elements (mouth movements) of the image sequence while discarding the static elements [153, 169].

In visual speech recognition the movements of the lips retain important features. The benefit of the proposed directional motion templates based approach is that it represents the entire space-time dimensions of mouth motion by four 2D gray scale images, with each image retaining the essence and temporal structure of the directional movement (up, down, left and right).

The major drawback of all the motion template based systems is the motion overwriting or self occlusion, especially in visual speech when speech is continuous or when long motion sequences are considered. The movements of lips are based on one centre point and repeated with respect to time, but this repetition of lip movement while uttering words can cause overwriting. The proposed DMHI technique has successfully solved this issue of overwriting, but for long motion sequences it has been only partially solved. To resolve this limitation for continuous speech, it is proposed to segment the continuous speech into basic visual units (visemes), where continuous speech can be recognized by concatenating respective visual units.

5.3 Summary

This chapter has given insights into the segmentation of mouth movement from video data by general motion history and by energy images technique. In addition to the MHI and MEI, the main focus is on DMHI, the robust method to overcome the self occlusion problem in MHI. The MHI and MEI methods reduce the dimension of the input video from 3D to a 2D template consisting of greyscale and binary images respectively. These greyscale and binary images show how and where the motion occurs in the video data. The development of optical flow based DMHI is also presented. Rather than a single motion template, DMHIs present the four directional motion images, which contain the information regarding the lip motion in a particular direction *i.e.*, up, down, left and right. This chapter also discusses the advantages and disadvantages of the proposed DMHI technique, elaborating that the DMHI technique is prone to overwriting or self occlusion.

Generally, the size of ROI image is fairly large so that all the pixel values of the generated templates are not suitable to use as feature vector to represent the mouth movement. To represent the MHI or DMHIs with reduced feature set, suitable image descriptors are needed. Chapter 6 describes the feature extraction and classification techniques in visual speech recognition systems used in this work. Further, the issues related to the development of visual speech recognition system are discussed.

Chapter 6

Visual Speech Feature Extraction and Classification

In pattern recognition, *features* are the representation of the given data that provide sufficient variability between two different patterns. Feature extraction is the process of finding the correct representation which will aid in correct classification. The visual features for visual speech recognition are the representation of the given video signals that provide discrimination between the various visemes (visual speech units) whilst providing invariance to similar visemes. The purpose of the feature extraction step in a visual speech recognition system is to yield the robust features which make the task of the classifier trivial (linear if possible). However, due to the various variations in the ROI such as illumination changes, viewing angle of the camera, variation in dimensionality of the ROI and in style of speaking, it is extremely difficult to find a robust set of features which provide accurate discrimination between visemes. Visual speech features should have the following characteristics [170]:

- Robust to environmental variations such as lighting condition, scale and rotation reducing intra-class variation
- Contain the maximum information about the patterns of interest increasing inter-class variation
- Low dimensionality and compact representation allowing real time system implementation

A variety of feature extraction methods have been proposed in literature for video analysis of a mouth while speaking. Generally these features can be broadly categorized

into three types: shape based (contour based), appearance based (global features) and a combination of both. In this Chapter a brief review of these features is presented. Appearance based features have been preferred by many researchers because of their simplicity and analogy to human perception and does not require exact localization and tracking of the mouth. In Section 6.2 of this Chapter, proposed feature extraction methods are explored and finally the classifier employed is discussed in Section 6.4.

6.1 Classification of Visual Feature Extraction

Feature extraction techniques applied to visual speech recognition systems can be classified into the following three categories [14]

- i) Shape based or Contour based.
- ii) Appearance based or Intensity based, and
- iii) A Combination of appearance and shape based features

The following section describes these techniques, their advantages and disadvantages.

6.1.1 Shape Based Feature Extraction

The shape based feature extraction techniques are concerned with the representation of the mouth, in terms of geometrical shape and dimensions of the lips such as height and width, as seen in Figure 6.1(a). This information is encoded with respect to the standard set of mouth shapes already in the system. In 1984, Petajan [41] proposed the first visual lip-reading system based on shape features. In this system Petajan extracted the speaker's mouth height, width, perimeter and area from the binary images as the features for the classifier. In shape based feature extraction, it is desirable to locate the exact location and position of visual speech articulators. This approach has the advantage of low dimensionality of features, however, the requirement of further localization and tracking may have an undesirable effect on the visual speech recognition system [61].



Figure 6.1: Shape based features represent the physical shape of the mouth such as height and width, given in (a). Appearance based features consider the complete ROI, shown in (b).

Generally, researchers [36, 64, 127, 171, 172] have considered the physical measurements such as mouth width, height and area of visual articulators to represent the speaker's mouth. Kaynak *et al.* [173] provided a detailed description and a comparative analysis of lip geometric features. To extract the geometric features such as height, width and contour of the lips, artificial markers have been applied on the speaker's lips [64, 173]. In today's systems, these artificial markers are not suitable in most of the scenarios.

In recent studies model based techniques have been considered for AVSR. In these approaches, a geometric model of the lip contour is used. Typical examples of this technique are active shape models [174], in which the inner and outer contour of the lips are extracted by a labelled set of points on the mouth. Features extracted from model based approaches can be divided into two categories. In the first category, parametric values used to define the lip contours are directly used as a feature vector. In the second category, parameters such as the values of mouth height, width, perimeter, area and ratios between the width and height are considered as a feature vector [24, 175, 176]. Other examples of model based techniques are deformable templates [177], active appearance model (AAM) [17], multi-scale spatial analysis (MSA) [24] and smart snakes [178, 179]. The AAM model is an extension of ASM, which combines a shape model with a statistical model of grey level surface of the ROI. The AAM has been demonstrated to outperform ASM [24]. Wojdel and Rothkrantz [180] introduced a new feature extraction

method that is a model free approach used to describe the shape of the lips. It was based on an image segmentation technique used to detect the pixels belonging to the lips and known as lip geometry estimation (LGE).

The major drawback of AAM and other model based approaches is that it requires manual annotations on the training samples. The performance of AAM decreases if the speaker's sample is not included in the training sample. Model based techniques are also sensitive to the facial skin colour and hairs on the face but are insensitive to illumination variations and image noise [75].

6.1.2 Appearance Based Feature Extraction

The major shortcoming of the shape based feature extraction techniques [17, 174] is that they only consider the geometrical information of the lips to represent mouth shapes. These techniques only analyse the lip contour information and do not consider the other speech articulators, whereas the appearance based representations are concerned with the low level features. Appearance based techniques use the information from the complete pixel values available in consecutive images of the ROI as shown in Figure 6.1(b). In addition to the lips, other speech articulators such as teeth, tongue, jaw and some surrounding muscles of the mouth which are informative about the visual speech [60] are implicitly included in this technique. The raw pixel values in the ROI contain the salient information of speech and can be directly used as a feature vector [181]. However, a high dimensional ROI contains a large number of pixels and can be an overburden to the statistical classifier and would be problematic. In this regard, some image processing/transforming techniques are applied on the ROI image to extract a compact and meaningful feature vector.

Principal component analysis (PCA) is a well-known data dimensionality reduction technique, used for lip-reading by various researchers [61, 63, 181-186] and achieves good results. Another technique that is superficially related to PCA is independent component analysis (ICA) which is also used for lip-reading [162]. However, results

achieved by traditional PCA have outperformed those of the ICA representation for lip-reading [162].

In other approaches, researchers have used linear image transforms such as discrete wavelet transform (DWT) [185], vector quantization [187] and discrete cosine transform (DCT) [33, 63, 184, 188, 189] to compute the features from the ROI. In these data compression techniques, statistical redundancies in the image are removed with respect to transformed coefficients. Computation of these transforms such as Fast Fourier Transform (FFT) become faster when the image size is of a power of 2 or a square image, so can be used in real time implementation of a VSR system [190]. Potamianos *et al.* [35] applied the first and second order derivatives of a DCT on a ROI to capture the visual speech features. The advantage of intensity based feature extraction approaches over the shape based approaches is that they do not need *a priori* statistical lips model. This advantage leads to the development of a computationally efficient VSR system [169].

6.1.3 Hybrid Features

Hybrid feature extraction combines both appearance and shape based representations. The structure of this approach is based on the hypothesis that the high level features that describe the geometric shapes of a mouth and the low level features that assume the pixel level features are complementary to each other. By combining them, performance can be augmented. Luttin *et al.* [191] presented a speech reading system in which they combined the shape based ASM features and intensity based PCA features. A similar combined feature extraction method was used by Chiou and Hwang [182], who used the snake contour model with PCA. Chan [192] described his feature vector using geometric features with PCA features. These approaches concatenate both sets of features into a single feature vector. In addition to the above approaches, the active appearance model (AAM) implicitly extracts both the appearance and shape features [17].

6.1.4 Motion Based Feature Representation

In 1994, Goldschen *et al.* [127] introduced dynamic features to represent the actions of the oral cavity during speech. They demonstrated that motion features are more

discriminative as compared to static features. Besides the seven static features from the oral cavity, such as width, height, area and perimeter, they also extracted the dynamics of each feature by computing the change of each feature between consecutive frames and then calculated the second derivative. Their results indicated that the dynamic features are more discriminative as compared to static features.

Rosenblum and Saldaa (1998) [193] described the static and dynamic visual speech information, where features extracted from static mouth images are static (pictorial) features. The appearance based and shaped based features reside in this category. The dynamic features directly represent the motion or dynamics of the mouth movement with respect to time varying visual information. Recently, Yau *et al.* [26] computed the dynamic features to represent visual speech. In their approach, mouth motion is represented by a motion history image (MHI), computed by image subtraction of consecutive images.

As described in Chapter 4, the motion of the object in an image sequence provides sufficient information. An optical flow motion tracking approach is adopted in this research. It is an alternative method to the static analysis of the input video of a ROI. The motivation for using motion features is that they provide the temporal characteristics of visual speech information or lip positions [194] and describe the actual movement of the mouth. Another reason in favour of using temporal image data is the validation by Rosenblum and Saldaa [193] that time-varying information is important for visual speech perception as compared to time independent information. The optical flow has been used for lip reading since the beginning of this research domain. However, computational complexity and low performance optical flow algorithms have restricted its use. The availability of powerful CPUs/GPUs and very sensitive algorithms have allowed the use of optical flow for real time applications.

In the following section, feature extraction and classification techniques applied on the concise images computed in Chapters 4 and 5 are described. The concise images were computed by means of DMHIs and non-overlapping block based approaches.

6.2 Visual Speech Feature Extraction

The mouth movement is segmented from video data, using the optical flow analysis presented in Chapters 4 and 5, the resulting DMHIs and the templates developed from the optical flow vertical component can be used for matching the movement patterns of visemes. This Section presents the feature extraction technique applied on DMHIs to identify utterances. However, the templates computed from the vertical component of optical flow are directly fed to the classifier for training and testing, because the actual size of the image was reduced by computing the average values of each non-overlapping block.

The pixel values of each directional motion history image and templates developed from the vertical component are the representation of motion history. In DMHIs each image represents the motion of lip movements in a particular direction (up, down, left, right) where more recent movements are represented with brighter pixels. One of the simplest options for recognizing the directional motion history images for a corresponding utterance is to directly use the pixel values of each image as an input feature vector to the classifier that classifies these features into utterances. Nevertheless, analysing visual speech information directly from the intensity values is very difficult due to the large data size and sensitivity to local variations of image intensities [195]. For example, an image of size 100×100 contains 10,000 pixels which is too large in dimension and hinders the robust classification of the image. In this research this calculation would be 57,600 ($240 \times 240 \times 4$) pixels. In image classification tasks, it is highly desirable to have a small feature size that contains most of the relevant cues related to the objects of interest. The raw pixel values can be transformed to a different space to reduce the feature dimensionality, whilst retaining the relevant and meaningful motion information. Regions in an image can be described in terms of the external characteristics such as the boundaries of the regions [196, 197] or internal properties such as the pixel values within the regions [122, 198].

The external representation is used only when the shape or outline of the regions is important for representing the image. Since the pixel intensities of DMHIs contain spatial

and temporal information of mouth movement, global internal features are suitable to efficiently represent the intensity distribution of DMHIs. The gray level of each pixel of DMHI indicates the temporal characteristics of the mouth movement at that particular pixel location and direction. A number of global feature descriptors have been proposed for image representation in the literature such as wavelet transform representation [199], DCT representation [65] and statistical moments such as geometric, Hu moments (HM) [200] and Zernike moments (ZM). In this study two global internal descriptors are examined, to uniquely represent the set of four directional motion history images. These are Zernike moments and Hu moments.

ZM and HM are extracted individually from each DMHI of an utterance to represent the corresponding utterances in a compact form, while redundant information is removed. 64 Zernike moments are computed from each DMHI, so that in total a 256 dimensional feature vector represents an utterance. In a second feature set, a set of seven Hu invariant features are computed from each image, so that each utterance is represented by a 28 dimensional feature vector.

Statistical moments capture the global information of an image and do not require closed boundaries, as compared to boundary based features such as Fourier descriptors. Moments computed from images with different patterns are unique and hence such features are useful for pattern recognition of multiple visemes. The most commonly used moment-based features are the geometric moments that are computed by projecting the image function onto monomial functions. Hu [200] proposed a set of seven nonlinear functions named as moment invariants (MI) or Hu moments (HM) derived from geometric moments of 0 up to 3rd order. MI features are translation, scale and rotation invariant. It is difficult to derive MI that is greater than the 3rd order, thereby limiting the maximum feature size of MI to only seven moments.

ZMs are a type of orthogonal statistical moment proposed by Teague [201] in 1980, and have been recognized as one of the robust region-based shape descriptors in the MPEG-7 standard [202]. The advantages of ZM are that such features are mathematically concise and capable of reflecting not only the intensity distribution of an image, but also the

shape. As opposed to HM that can only be computed up to the 3rd order, ZM can easily be constructed to an arbitrary high order. Another key strength of ZM is the invariance property of these features to rotational changes.

A detailed description of ZM and HM is given in the Sections 6.2.1 and 6.2.2 respectively. Classification is a procedure which tries to assign each input value to one of the pre-defined classes, such as the features extracted by ZM are fed as input to the classifier, classifier attempts to assign it to the corresponding class. Section 6.4 describes the support vector machines (SVM) classifier used for viseme classification.

6.2.1 Zernike Moments

Zernike moments (ZM) are a type of orthogonal image moment commonly used in recognition of image patterns [203]. ZM are independent features due to the orthogonality of the Zernike polynomial V_{nl} [201]. It has many desirable properties. One of the most important property is its simple rotational invariance property [203]. Others are robustness to noise, expression efficiency and multilevel representation for describing the shapes of patterns. In terms of information redundancy, sensitivity to image noise and image representation capability, ZM have been demonstrated to outperform the other image moments such as Legendre moments, geometric moments and complex moments [204]. The orthogonality of the ZM features enables redundancy reduction and enhances the computation efficiency [205]. Zernike moments are computed by projecting the image function $f(x, y)$ onto the orthogonal Zernike polynomial V_{nl} of order n with repetition l defined within a unit circle (*i.e.*, $x^2 + y^2 \leq 1$).

6.2.1.1 Square-to-Circular Image Coordinate Transformation

All the directional motion history images are scaled to a square image of $N \times N$ pixels so that they can be mapped to the unit circle centred at the origin of an individual image. Each image of size $N \times N$ pixels is bounded by a unit circle, the centre of the image is taken as the origin and the pixel coordinates are mapped to the range of a unit circle *i.e.*, $x^2 + y^2 \leq 1$.

Figure 6.1 shows the square-to-circular transformation that maps the square image function $f(i, j)$ to a circular image function $f(\rho, \theta)$ in terms of x-y axes. Each of the images is enclosed within the circular x-y coordinates to ensure that no information is lost in the square-to-circular transformation.

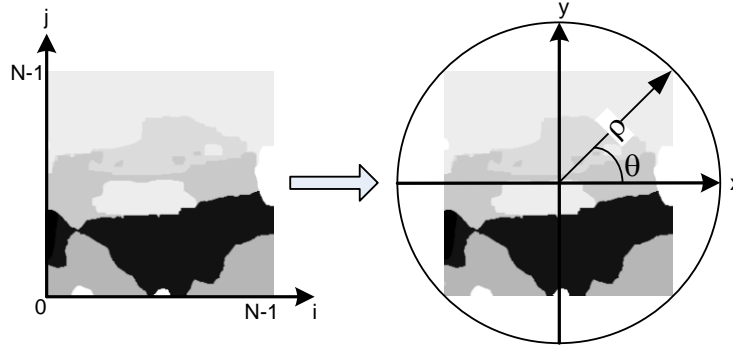


Figure 6.2: The square-to-circular image coordinates transformation of a motion template before Zernike moments computation.

The transformed coordinates are given by:

$$x_i = \frac{\sqrt{2}}{N-1}i + \frac{-1}{\sqrt{2}} \quad (6.1)$$

$$y_i = \frac{\sqrt{2}}{N-1}j + \frac{-1}{\sqrt{2}} \quad (6.2)$$

The radius, ρ and angle, θ after the transformations are given by:

$$\rho_{ij} = \sqrt{x_i^2 + y_j^2} \quad (6.3)$$

$$\theta_{ij} = \tan^{-1}\left(\frac{y_j}{x_i}\right) \quad (6.4)$$

6.2.1.2 Computation of ZM

ZM is computed by projecting an image function, $f(x, y)$ onto the orthogonal Zernike polynomial, V_{nl} . The kernel of ZM is a set of orthogonal Zernike polynomials defined over the polar coordinate space within a unit circle (*i.e.*, $x^2 + y^2 \leq 1$). Zernike moments, Z_{nl} of order n and repetition l are given by

$$Z_{nl} = \lambda \int_0^1 \int_0^{2\pi} V_{nl}(\rho, \theta) f(\rho, \theta) \rho d\rho d\theta \quad (6.5)$$

where $|l| \leq n$ and $(n - |l|)$ is even, $f(\rho, \theta)$ is the intensity distribution of DMHIs mapped to a unit circle of radius ρ and angle θ defined in Equation (6.3) and (6.4). The term λ is a normalizing constant defined as

$$\lambda = \frac{(n+1)}{\pi} \quad (6.6)$$

The Zernike polynomial, V_{nl} is given by

$$V_{nl}(\rho, \theta) = R_{nl}(\rho) e^{-j l \theta}; \hat{j} = \sqrt{-1} \quad (6.7)$$

where R_{nl} is the real-valued radial polynomial, which is given by

$$R_{nl}(\rho) = \sum_{k=0}^{\frac{n-|l|}{2}} (-1)^k \frac{(n-k)!}{k! \left(\frac{n+|l|}{2} - k \right)! \left(\frac{n-|l|}{2} - k \right)!} \rho^{n-2k} \quad (6.8)$$

The integrals in Equation (6.5) are replaced by summations for discrete digital images given by

$$Z_{nl} = \lambda \sum_x \sum_y V_{nl}(\rho, \theta) f(\rho, \theta) \quad (6.9)$$

6.2.1.3 Rotation Invariance of ZM

To illustrate the rotational characteristics of ZM, consider β as the rotation angle of an image (DMHI). The rotated image, f^r is given by

$$f^r(\rho, \theta) = f(\rho, \theta - \beta) \quad (6.10)$$

The mapping of ZM expression from the x-y plane into the polar coordinates can be obtained by changing the double integral of Equation (6.5), given by the general equation

$$\iint \phi(x, y) dx dy = \iint \phi[p(\rho, \theta), q(\rho, \theta)] \frac{\partial(x, y)}{\partial(\rho, \theta)} d\rho d\theta \quad (6.11)$$

$\frac{\partial(x, y)}{\partial(\rho, \theta)}$ defines the Jacobian of the transformation and is also the determinant of the matrix. Since $x = \rho \cos \theta$ and $y = \rho \sin \theta$, the Jacobian becomes ρ . The ZM of the original image (before rotation) is

$$Z_{nl} = \lambda \int_0^{2\pi} \int_0^1 f(\rho, \theta) R_{nl}(\rho) e^{-j l \theta} \rho d\rho d\theta \quad (6.12)$$

The ZM of the rotated image is given by

$$Z'_{nl} = \lambda \int_0^{2\pi} \int_0^1 f(\rho, \theta - \beta) R_{nl}(\rho) e^{-j l \theta} \rho d\rho d\theta \quad (6.13)$$

Let $\alpha = \theta - \beta$. Hence the ZM of the rotated image is

$$\begin{aligned}
Z'_{nl} &= \lambda \int_0^1 \int_0^{2\pi} f(\rho, \alpha) R_{nl}(\rho) e^{-j l (\alpha + \beta)} \rho d\rho d\theta \\
&= \left[\lambda \int_0^1 \int_0^{2\pi} f(\rho, \alpha) R_{nl}(\rho) e^{-j l \alpha} \rho d\rho d\theta \right] e^{-j l \beta} \\
&= Z_{nl} e^{-j l \beta}
\end{aligned} \tag{6.14}$$

Equation (6.14) demonstrates that rotation of images results in a phase shift of ZM. This simple rotational property indicates that the magnitudes of ZM of a rotated image function remain identical to ZM before rotation [203]. The absolute value of ZM is invariant to rotational changes as given by

$$|Z'_{nl}| = |Z_{nl}| \tag{6.15}$$

DMHIs are represented using the absolute value of ZM as visual speech features. An optimum number of ZM needs to be selected to ensure a suitable trade-off between the feature dimensionality and the image representation ability. By including higher order moments, more image information is represented but this increases the feature size. Further, the higher order moments are more prone to noise [204]. The number of moments required is determined empirically. Performance measures of three different numbers of ZM features were computed to obtain a suitable number of ZM features for the viseme classification. Classification results of 49, 64 and 81 ZMs were evaluated by SVM classifier. Table 6.1 shows the average results of three different numbers of ZMs in terms of sensitivity, specificity and accuracy, which are defined in Section 7.1.2.

Table 6.1: Average classification results of three different numbers of ZMs.

Number of Zernike Moments	Sensitivity %	Specificity %	Accuracy %
49	71.7	99.6	97.6
64	75.7	99.7	98
81	72.7	99.7	97.8

Based on the accuracy and sensitivity values, 64 ZMs that comprise of 0th order up to 14th order moments (listed in Table 6.2) are adopted as visual speech features to represent each DMHI.

Table 6.2: Zernike Moments from 0th to 14th order.

Order	Moments	No: of moments
0	$Z_{0,0}$	1
1	$Z_{1,1}$	1
2	$Z_{2,0}, Z_{2,2}$	2
3	$Z_{3,1}, Z_{3,3}$	2
4	$Z_{4,0}, Z_{4,2}, Z_{4,4}$	3
5	$Z_{5,1}, Z_{5,3}, Z_{5,5}$	3
6	$Z_{6,0}, Z_{6,2}, Z_{6,4}, Z_{6,6}$	4
7	$Z_{7,1}, Z_{7,3}, Z_{7,5}, Z_{7,7}$	4
8	$Z_{8,0}, Z_{8,2}, Z_{8,4}, Z_{8,6}, Z_{8,8}$	5
9	$Z_{9,1}, Z_{9,3}, Z_{9,5}, Z_{9,7}, Z_{9,9}$	5
10	$Z_{10,0}, Z_{10,2}, Z_{10,4}, Z_{10,6}, Z_{10,8}, Z_{10,10}$	6
11	$Z_{11,1}, Z_{11,3}, Z_{11,5}, Z_{11,7}, Z_{11,9}, Z_{11,11}$	6
12	$Z_{12,0}, Z_{12,2}, Z_{12,4}, Z_{12,6}, Z_{12,8}, Z_{12,10}, Z_{12,12}$	7
13	$Z_{13,1}, Z_{13,3}, Z_{13,5}, Z_{13,7}, Z_{13,9}, Z_{13,11}, Z_{13,13}$	7
14	$Z_{14,0}, Z_{14,2}, Z_{14,4}, Z_{14,6}, Z_{14,8}, Z_{14,10}, Z_{14,12}, Z_{14,14}$	8

6.2.2 Hu Moments

For the image analysis and pattern recognition the use of moments was inspired by Hu [36] and Alt [5]. Hu derived a set of *invariant* moments (IM) which has the desirable properties of being invariant under image rotation, translation and scaling. The invariants of the object in the pattern recognition are the set of measurable quantities which describes the object. These are insensitive to particular deformations and provide sufficient discrimination power to distinguish objects belongs to different classes [206].

The two dimensional $(p+q)^{\text{th}}$ order moment is defined as follows:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \quad (6.16)$$

$$p, q = 0, 1, 2, 3, \dots$$

If the image is greyscale with pixel intensities $I(x,y)$, moments are calculated by

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y) \quad (6.17)$$

Hu stated that if $f(x,y)$ is a piecewise continuous bounded function, and has nonzero values only in a finite region of an (x,y) plane, then the moment sequence (m_{pq}) is uniquely determined by $f(x,y)$, and conversely, $f(x,y)$ is also uniquely determined by (m_{pq}) .

Considering the fact that an image segment has finite area, or in the worst case is piecewise continuous, moments of all orders exist and a complete moment set can be computed and used uniquely to describe the information contained in the image. However, to obtain all of the information contained in an image requires an infinite number of moment values. Therefore, selecting a meaningful subset of the moment values that contains sufficient information to characterize the image uniquely for a specific application becomes very important.

It can be noted that the moments in Equation (6.16) may not be invariant when $f(x,y)$ changes by translating, rotating or scaling. The invariant features can be achieved using central moments, which are defined as follows:

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q f(x, y) dx dy \quad (6.18)$$

$$p, q = 0, 1, 2, 3, \dots$$

where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$

The pixel point (\bar{x}, \bar{y}) is the centroid of the image $f(x,y)$. The centroid moments μ_{pq} computed using the centroid of the image $f(x,y)$ is equivalent to the m_{pq} , whose centre has been shifted to the centroid of the image. Therefore, the central moments are

invariant to image translations. Scale invariance can be obtained by normalization. The normalized central moments are defined as follows.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}, \quad \gamma = \frac{(p+q)}{2} + 1 \quad (6.19)$$

where γ is the normalization factor. The set of absolute moment invariants consists of a set of nonlinear combinations of central moments that remain invariant under rotation. Based on normalized central moments, Hu defined the following seven functions, computed from central moments through order three, that are invariant with respect to object scale, translation and rotation:

$$\phi_1 = \mu_{20} + \mu_{02}, \quad (6.20)$$

$$\phi_2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2, \quad (6.21)$$

$$\phi_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - 3\mu_{03})^2, \quad (6.22)$$

$$\phi_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2, \quad (6.23)$$

$$\begin{aligned} \phi_5 = & (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] + \\ & (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right], \end{aligned} \quad (6.24)$$

$$\begin{aligned} \phi_6 = & (\mu_{20} - \mu_{02}) \left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] + \\ & 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}), \end{aligned} \quad (6.25)$$

$$\begin{aligned} \phi_7 = & (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] \\ & - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right]. \end{aligned} \quad (6.26)$$

The functions ϕ_1 through ϕ_6 are invariant with respect to rotation and reflection while ϕ_7 changes sign under reflection.

6.3 Classifier for Lip-reading

As discussed, three novel feature sets of 14 visemes are extracted for classification, these are:

- i) Features from the optical flow vertical component, based on a statistical property of non overlapping blocks.
- ii) Zernike moments from four DMHIs
- iii) Hu moments from four DMHIs

Classification can be defined as the process of assigning new inputs to one of the predefined discrete classes (utterances). Achieving accurate classification performance is quite difficult. Generally, the task of a classifier in visual speech recognition is to determine the probability of the input features of an utterance with each of the possible utterances in the system [207]. This can be achieved through supervised classification that creates a function or model from the training examples which consist of pairs of input (feature vectors) and output (class labels of utterances). The task of a trained classifier is to predict the label of new features.

In the literature, a variety of classifiers is used for visual speech features. Classification of visual speech features into multiple visemes can be accomplished using generative or discriminative models. Statistical models for features generated through random processes are known as generative models. These models can be represented by the parameters derived from the statistical properties of the input features. Markov, Gaussian and hidden Markov models are examples of generative models. The most widely used classifier for modelling and recognizing audio and visual speech data has been the hidden Markov model (HMM). HMM provides a mathematical framework that is suitable for modelling time varying signals. HMM is useful in finding patterns that appear over a

space of time and has been successfully implemented as a classifier in applications such as gesture recognition [208] and bioinformatics [209].

Discriminant models such as support vector machines (SVM) [210] and artificial neural networks (ANN) [211] are non-parametric models which classify features without assuming a priori knowledge of the data. These classifiers create decision functions that classify input data into one of the predefined classes based on the training samples. Heckmann *et al.* [212] developed a combination of both the ANN and HMM to form a hybrid ANN-HMM classifier to improve the performance of AVSR system in varying noise conditions. Similarly, Bregler *et al.* [181] and Duchnowski *et al.* [63] devised a hybrid ANN-DTW (dynamic time warping) classifier. Gowdy *et al.* [33] and Saenko *et al.* [22] have proposed the use of Dynamic Bayesian Networks (DBNs) for AVSR.

Though all of the above mentioned classifiers have shown success in their respective AVSR systems, the SVM is the choice of this dissertation because the nature of the computed features is not time varying signal. In all three approaches of feature extraction, the number of features is fixed. Moreover, in earlier work, feed-forward multilayer perceptron (MLP), artificial neural networks (ANN) with back propagation [213] and Hidden Markov Models (HMM) [75] have already been investigated using the same dataset with different feature extraction techniques. The advantages of SVM are:

- able to find a globally optimal solution,
- can produce good generalization and
- performs well with a relatively small number of training data.

SVM can generate a globally best solution, as opposed to neural network training that is susceptible to local maxima. The following section evaluates the SVM classification technique for classification of the three types of feature sets discussed earlier.

6.4 Support Vector Machine

Support vector machine (SVM) is a supervised binary classifier that differentiates the input data into two possible classes. SVM developed by Vapnik [28] is a state-of-the-art

classifier that has been successfully exploited for various pattern recognition applications. The mechanism of SVM works by projecting the data on to a sparse high-dimensional space and then finding the optimal separating hyper-plane between the classes. One of the key strengths of SVM is the generalization obtained by tuning the trade-off between structural complexity of the classifier and empirical error. It also performs well, even with a small number of training data.

6.4.1 Optimal Separating Hyper-plane

Considering the difficulty of separating the set of training vectors belongs to two different classes which are labelled as -1 and +1 for either of the classes. The sample is $\{x^t, \gamma^t\}$, where $\gamma^t = +1$ if $x^t \in C_1$ and $\gamma^t = -1$ if $x^t \in C_2$

The linear decision function of SVM is given by

$$f(x) = w^T x^t + b \quad (6.27)$$

The interest is to find w and b such that

$$w^T x^t + b \geq +1 \text{ for } \gamma^t = +1 \quad (6.28)$$

$$w^T x^t + b \leq -1 \text{ for } \gamma^t = -1 \quad (6.29)$$

Equations (6.28) and (6.29) can be rewritten as

$$\gamma^t (w^T x^t + b) \geq +1 \quad (6.30)$$

The real valued $f(x)$ output is converted to a positive or negative label using the *signum* function. The linear decision function, $f(x)$ partitions the input space into two parts by creating a hyper-plane given by

$$w^T x^t + b = 0 \quad (6.31)$$

This hyper-plane lies between two bounding planes given by

$$w^T x^t + b = +1 \text{ and } w^T x^t + b = -1 \quad (6.32)$$

The distance from the hyper-plane to the bounding planes on either side is called the margin. By maximizing this margin, the generalization error can be minimized. By using the hypothesis class of lines, the optimal separating hyper-plane is the one that maximizes the margin. Figure 6.3 shows the two linearly separable classes, separated by optimal hyper-plane. Data points which are closest to the separating hyper-planes are known as support vectors. The distance from the hyper-plane to the instances closest to it on either side is called the *margin*, which should be maximum for the best generalization.

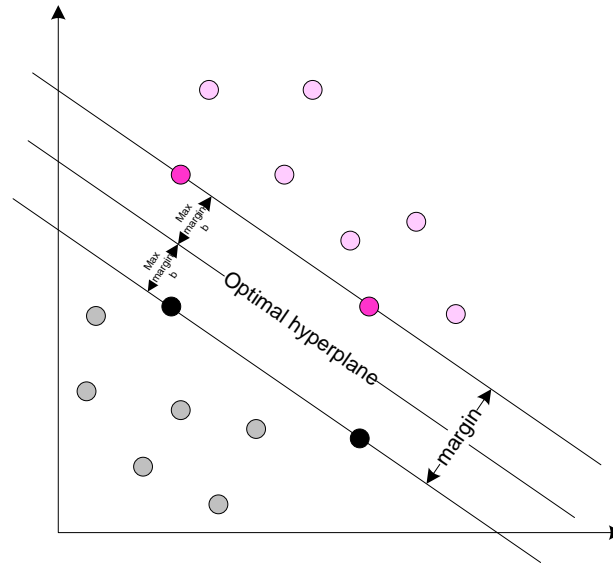


Figure 6.3: Representation of linearly separable data of two classes, class 1 is represented by grey dots and class 2 is represented by pink dots. An optimal hyper plane separates the two classes. The four support vectors are shown as black and dark pink dots.

The distance of a point x^t to the discriminant is given by

$$\frac{(w^T x^t + b)}{\|w\|} \quad (6.33)$$

when $\gamma^t \in \{-1, +1\}$ Equation (6.33) can be rewritten as

$$\frac{\gamma^t(w^T x^t + b)}{\|w\|} \geq \rho, \forall t \quad (6.34)$$

Simple vector geometry shows that the margin is equal to $\frac{1}{\|x\|}$ so that $\rho\|w\|=1$.

Minimizing $\|w\|$ is equivalent to minimizing $\frac{1}{2}\|w\|^2$ and the use of this term makes it possible to perform Quadratic Programming (QP) optimization later on, so that it can be defined as

$$\min \frac{1}{2}\|w\|^2 \text{ subject to } \gamma^t(w^T x^t + b) \geq +1, \forall t \quad (6.35)$$

This is a standard quadratic optimization problem, and can be solved to find w and b . In finding the optimal hyper-plane, the optimization problem can be converted to a form whose complexity depends on N , the number of training instances not on d , the input instances. To get the new formulation, Equation (6.35) is written as an unconstrained problem using Lagrange multipliers α^t :

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{t=1}^N \alpha^t [\gamma^t(w^T x^t + b) - 1] \quad (6.36)$$

$$L_p = \frac{1}{2}\|w\|^2 - \sum_t \alpha^t \gamma^t(w^T x^t + b) + \sum_t \alpha^t \quad (6.37)$$

L_p must be minimized with respect to w and b which requires the gradient of L_p to vanish with respect to w and b . Hence the condition:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_t \alpha^t \gamma^t x^t \quad (6.38)$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_t \alpha^t \gamma^t = 0 \quad (6.39)$$

Substituting Equations (6.38) and (6.39) into Equation (6.37), gives a new formulation which, is dependent on α , and known as dual function,

$$\begin{aligned}
L_d &= \frac{1}{2} (w^T w) - w^T \sum_t \alpha^t \gamma^t x^t - b \sum_t \alpha^t \gamma^t + \sum_t \alpha^t \\
L_d &= -\frac{1}{2} (w^T w) + \sum_t \alpha^t \\
L_d &= -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s \gamma^t \gamma^s (x^t)^T x^s + \sum_t \alpha^t
\end{aligned} \tag{6.40}$$

This is maximized with respect to α^t only, subject to the constraints

$$\sum_t \alpha^t \gamma^t = 0 \text{ and } \alpha^t \geq 0, \forall t \tag{6.41}$$

This can be solved by using quadratic optimization methods. The size of the dual depends on N , sample size, and not on d , the input dimensionality. It will return α and then from Equation (6.38) w can be calculated. b can be calculated from Equation (6.39) which is a support vector x^t and lie on the margin.

$$\gamma^t (w^T x^t + b) = 1 \tag{6.42}$$

Using this fact, b can be calculated from any support vector as

$$b = \gamma^t - w^T x^t \tag{6.43}$$

For numerical stability, it is recommended to consider the average of all the support vectors. The discriminant thus found is called the support vector machine (SVM).

6.4.2 The Non-Separable Data: Soft Margin Hyper-plane

If the given data is not linearly separable, the above mentioned algorithm is not suitable. For such a type of data when there is no hyper-plane to separate the two classes, a soft margin method is introduced to use the optimization criterion in SVM training for classifying non separable data so that the least number of errors occur. The soft margin

method defines slack variables, $\xi^t \geq 0$, which store the deviation from the margin or measures the degree of misclassification. Relaxing Equation (6.30) by using a slack variable,

$$\gamma^t (w^T x^t + b) \geq 1 - \xi^t \quad (6.44)$$

If $\xi^t = 0$, there is no problem with x^t . If $0 < \xi^t < 1$, x^t is correctly classified but in the margin. If $\xi^t \geq 1$, x^t is misclassified, so that a soft error is defined as $\sum_t \xi^t$ by adding this as a penalty term:

$$L_p = \frac{1}{2} \|w\|^2 + C \sum_t \xi^t \quad (6.45)$$

Where C is the penalty factor in any regularization scheme trading off complexity, to penalize the misclassified points and also the ones in the margin for better generalization, adding the constraints, the Lagrangian of Equation (6.37) then becomes

$$L_p = \frac{1}{2} \|w\|^2 + C \sum_t \xi^t - \sum_t \alpha^t [\gamma^t (w^T x^t + b) - 1 + \xi^t] - \sum_t \mu^t \xi^t \quad (6.46)$$

where μ^t is the new Lagrange parameter with $\alpha^t, \mu^t \geq 0$. In order to find the w , b and ξ^t , L_p is differentiated with respect to w , b and ξ^t and derivatives are set to zero.

6.4.3 Kernel Trick

The previously discussed techniques of optimal separating hyper plane and soft margin hyper plane use higher dimensional decision surfaces that are linear. Both techniques are suitable for linear data. However, in most cases the dataset is non linear and inseparable. In this case, a non-linear decision function is required. Instead of using a non linear decision function, the non linear data can be mapped to a new space by doing a non linear transformation, choosing suitable basis functions and then using a linear model in this new space (feature space). The linear model in the new space corresponds to a non linear

model in the original space \mathfrak{R} . The transformation is $z = \phi(x)$ where $z_j = \phi_j(x)$ $j = 1 \cdots k$, mapping the data points from the input space \mathfrak{R} to feature space Z , where the discriminant is written as

$$\begin{aligned} g(z) &= w^T z \\ g(x) &= w^T \phi(x) \\ &= \sum_{j=1}^k w_j \phi_j(x) \end{aligned} \quad (6.47)$$

b is not used separately and it is assumed that $z_1 = \phi_1(x) \cong 1$. The dual form of the decision function is defined in the Equation (6.48):

$$L_d = \sum_t \alpha^t - \frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s \phi(x^t)^T \phi(x^s) \quad (6.48)$$

subject to,

$$\sum_t \alpha^t r^t = 0 \text{ and } 0 \leq \alpha^t \leq C, \forall t \quad (6.49)$$

The kernel machines are replaced by the inner product of basis functions, $\phi(x^t)^T \phi(x^s)$, by a kernel function, $K(x^t, x^s)$, between instances in the original input space.

$$L_d = \sum_t \alpha^t - \frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s K(x^t, x^s) \quad (6.50)$$

The kernel function can be replaced in the discriminant

$$\begin{aligned} g(x) &= w^T \phi(x) = \sum_t \alpha^t r^t \phi(x^t)^T \phi(x) \\ &= \sum_t \alpha^t r^t K(x^t, x) \end{aligned} \quad (6.51)$$

Once the kernel function is known, the projection of the data is done implicitly. For any valid kernel, there is a corresponding mapping function, but it may be much simpler to

use $K(x', x)$ rather than calculating $\phi(x')^T \phi(x)$ and taking the dot product. Thus the linear separation of data can be performed on the high dimensional feature space Z , and is equivalent to non-linear classification in the original space \mathfrak{R} .

Many algorithms have been kernelized. These can be derived by selecting functions that satisfy certain mathematical properties. Selection of a suitable kernel function for the data is an important process for further training of SVM.

The most popular, general purpose kernel functions used in SVM classifiers are

- Polynomials of degree q :

$$K(x', x) = (x'^T x + 1)^q \text{ where } q \text{ is defined by the user.}$$

- Radial basis functions (RBF):

$$K(x', x) = \exp \left[-\frac{\|x' - x\|^2}{2s^2} \right]$$

It defines a spherical kernel, where x^t is the centre and s defines the radius given by the user. The RBF is the most demanding kernel type used in support vector machines, due to its performance

- Sigmoid function:

$$K(x', x) = \tanh(2x'^T x + 1)$$

- Linear Kernel:

$$K(x', x) = x'^T x$$

6.4.4 Multiclass Kernel Machines

Inherently, SVM is a binary class classifier. Generally when there are $K > 2$ classes, the common method one-vs-rest is implied to use a binary classifier. In one-vs-rest, each class is trained against all other classes combined and K support vector machines are learned. In training, examples of Class-1 (C_1) are labelled as +1 and examples of all other classes (C_K), $k \neq 1$ are labelled as -1, whilst in testing, all $g_i(x)$, $i = 1, \dots, K$ are calculated.

In another approach, instead of building K two-class SVM classifiers to separate one from all the rest, a one against one (pair-wise separation) multiclass SVM is proposed. For $K > 2$ classes, $K(K-1)/2$ pair-wise classifiers are built, with each $g_{ij}(x)$ taking examples of C_i with the label +1, examples of C_j with label -1, and not using examples of the other classes. Separation of classes in pairs is trivial and has the additional advantage of faster optimization because it uses less data.

In general, both one-vs-rest and pair-wise separation are special cases of the error-correcting output codes [214], which decompose a multiclass problem to a set of two class problems.

In yet another approach, Weston and Watkins [215] proposed to write a single multiclass optimization problem involving all classes

$$\begin{aligned} \min & \frac{1}{2} \sum_{i=1}^K \|w_i\|^2 + C \sum_i \sum_t \xi_i^t \\ \text{subject to} & \end{aligned} \quad (6.52)$$

$$w_{z^t} x^t + w_{z^t 0} \geq w_i x^t + w_{i0} + 2 - \xi_i^t, \forall i \neq z^t \text{ and } \xi_i^t \geq 0$$

where z^t contains the class index of x^t and C is the usual regularization parameter.

The SVM multiclass implementation used in this dissertation is publicly available, and is based on a multiclass formulation described by Crammer and Singer [216], an enhanced

version of Weston and Watkins [215]. To solve the problem of optimization SVM^{multiclass1} uses an algorithm based on structural SVMs [217]. As far as author is aware, this is the first time a fully implemented multi class classification has been attempted in visual speech recognition.

6.5 Summary

In this chapter, visual features used for visual speech recognition are reviewed. It is hypothesized that appearance based features provide a better representation of visual speech compared to shape based and model based approaches. Appearance based features also do not require further localization of lip features throughout the image sequence as in contour and combination based techniques. Advantages and limitations of both the appearance based and shape based features are discussed.

Computer based lip-reading studies have indicated that the important visual information lies in the temporal change of a mouth [194] and motion features are more discriminative compared to static features for computer based lip-reading [127]. Based on the above studies, this research uses appearance based motion features, computed by optical flow estimation.

Optical flow based DMHIs are developed. To represent these spatio-temporal templates, global internal region based descriptors are selected. In this research, two region based feature descriptors, ZM and HM, are evaluated. ZM are orthogonal moments which are capable of reflecting the shape and intensity distribution of DMHIs. ZM and HM have good rotation property and are invariant to changes of mouth orientation in the images. The number of ZMs is determined empirically, while the 7 HMs are computed from each DMHI.

Finally, the chapter concludes with a thorough discussion of the SVM classifier. SVM is a discriminative classifier that classifies features without knowing the priori information of data. It is able to find a globally optimal solution. The discussion includes the theory

¹ http://svmlight.joachims.org/svm_multiclass.html

behind the SVM binary class and multiclass classifiers, with details of optimal separating hyper-plane and linearly non-separable hyper-plane with soft margin separation using the slack variables. SVM kernels are also described, such as linear, polynomial, radial basis function and sigmoid.

Chapter 7

Experimental Results

This chapter reports on the experiments conducted to evaluate the performance of motion templates computed by the optical flow vertical component and the directional motion history images (DMHIs) technique based also on optical flow.

The experimental work consists of two parts. Section 7.1 reports solely on the optical flow vertical component based technique that investigates the viseme classification in terms of accuracy, sensitivity and specificity. The vertical component of optical flow contains most of the information of a visual speech viseme utterance. However, to capture the mouth motion while a subject smiles or laughs, the horizontal component cannot be ignored. The optical flow vertical component is divided into multiple non-overlapping blocks and the statistical features of each block are used as a feature of an utterance. These features of an utterance were classified using a support vector machine classifier. For recognition and further performance evaluation of the proposed features, SVM multi-class classification is performed. The performance of multiple block sizes was evaluated empirically. The detailed theoretical frame work of the feature extraction and classification techniques have been explained in Chapters 4 and 6 respectively.

Section 7.2 describes the DMHI based viseme classification. Two types of image features examined for DMHI were Zernike moments (ZM) and Hu moments (HM). These features were classified using a SVM classifier. The detailed theoretical description of ZM and HM has been given in Chapter 6. In addition, the proposed DMHI technique is compared with the traditional motion history images. For better representation of the results, performance evaluation is described in terms of accuracy, specificity and sensitivity. All experiments in this dissertation were conducted using leave-one-out mechanism.

7.1 Experimental Setup 1: Performance Evaluation of an Optical Flow Based Motion Template

The proposed lip-reading technique is tested on a viseme based speech model. Recognition units such as digits, phonemes, words and phrases in various languages have been used as vocabulary in lip-reading applications.

7.1.1 Methodology for Classification

The optical flow vertical component of size 240×320 pixels was initially divided into eight vertical columns as shown in Figure 7.1. The average intensity of each vertical column was computed, so that each image of a vertical component of optical flow was represented by eight pixels. As described in Chapter 4, similar subsequent images in the video sequence were ignored for optical flow computation. Ignoring optical flow computation has dual benefits - firstly, it reduces the computation burden on the system and secondly it is useful in compensating the inter subject variation in speed of speech.

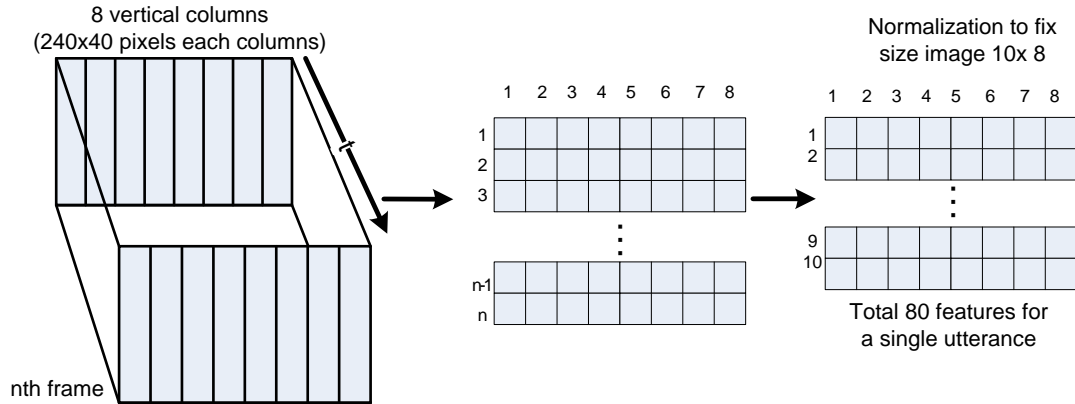


Figure 7.1: Development of optical flow vertical component based motion template (block size is 240×40).

Consecutive values from each vertical component field are stacked such that the matrix of size $n \times 8$ is developed, where n is the number of frames in an utterance. To compensate the variation in speed of speech due to the way people speak, this $n \times 8$ size matrix of features is normalized to a 10×8 matrix by using a linear interpolation method for each

utterance. Finally each utterance is represented by $10 \times 8 = 80$ pixel values. After the feature vectors are obtained, the state-of-the-art SVM classifier is used for classification.

For classification the following schemes can be employed [163]:

- the re-substitution method (training and test sets are the same);
- the holdout method (half the data is used for training and the rest of the data is used for testing);
- the leave-one-out method;
- the rotation method or N -fold cross validation (a compromise between the leave-one out method and the hold out method, which divides the samples into P disjoint subsets, $1 \leq P \leq N$. Use $(P - 1)$ subsets for training and the remaining subset for test); and
- the bootstrap method for partitioning scheme [218]. In most cases, the leave-one-out cross validation scheme is used for the partitioning scheme. This means that out of N samples from each of the classes per database, $N-1$ of them are used to train (design) the classifier and the remaining one to test it [203]. This process is repeated N times, each time leaving a different sample out. Therefore, all of the samples are ultimately used for testing. This process is repeated and the resultant recognition rate is averaged. Usually, this estimate is unbiased.

In the first experiment, 5 fold cross validation is adopted including all subjects together. The one-vs-rest SVM classification technique was adopted to separate the visual speech features into 14 visemes. The LIBSVM tool box [219] was used in the experiments to design the SVM classifier model. Four kernel functions, *i.e.*, linear, sigmoid, polynomial of order three and radial basis function (RBF) were tested on the extracted features. Based on these experiments RBF kernel was found to produce the best results and was selected for the classification. The gamma parameter and the error term penalty parameter C , of the RBF kernel function were optimized using preliminary experiments (grid search).

The classification performance of the SVM was tested using the leave-one-out method. Each repetition of the experiments used 784 training samples and 196 test samples (2 samples from each class of each subject). This was repeated five times for each viseme, every time different training and test data were used. The average rate of the performance measures - sensitivity, specificity and accuracy of the five repetitions of the experiments was computed.

7.1.2 Viseme Classification Results

In the first part of the experiments the performance of the vertical component of optical flow computing 80 features per utterance was investigated. The sensitivity and specificity are the statistical measures of the performance of the two-class problems, where class one is represented as +1 (TP) and second class is represented as -1 (TN). These statistical measures are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7.1)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (7.2)$$

$$Specificity = \frac{TN}{FP + TN} \times 100\% \quad (7.3)$$

where TP=True Positive, TN=True Negative, FP=False Positive, and FN=False Negative and the term *accuracy* is the measure of actual (true) positives and negatives which are correctly recognized as positives and negatives. *Sensitivity* measures the proportion of actual positives which are correctly identified as positives. *Specificity* measures the proportion of negatives which are correctly identified as negatives. The results are summarized in Table 7.1

Table 7.1: Average Classification Results of 14 Visemes in Terms of Specificity, Sensitivity and Accuracy, block size 240×40. (All values in %)

	Visemes	Specificity	Sensitivity	Accuracy
1.	/a/	98	65.7	95.7
2.	/ch/	99.2	85.7	98.3
3.	/e/	95.7	54.3	92.8
4.	/g/	97.7	61.4	95.1
5.	/th/	98.4	65.7	96
6.	/i/	96.9	60	94.3
7.	/m/	99.9	80	98.5
8.	/n/	97.5	52.9	94.3
9.	/o/	97.5	65.7	95.2
10.	/r/	97.9	61.4	95
11.	/s/	99	52.9	95.7
12.	/t/	99.1	74.3	97.3
13.	/u/	97.8	61.4	95.2
14.	/v/	99.7	88.6	98.9
	Average	98.2	66.4	95.9

Though the sensitivity is not up to the mark, the classification results are acceptable. By analysing the adopted methodology, it was observed that the division of the optical flow vertical component into vertical columns as shown in Figure 7.1 is problematic. It nullifies the important features during the average computation of each block. The reason behind this is that during speaking the movement of the two lips is always in opposite directions and this motion is represented by positive and negative values in the optical flow vertical component which shows the direction of motion. Hence while computing the average values of each column, the values in opposite directions nullify the important features of motion.

The solution to this problem is the use of non-overlapping rectangular blocks, as shown in Figure 7.2. As can be seen from the Figure, the upper and lower lip motions are separated so that the nullifying effect is eliminated. To optimize the block size, experiments were conducted to evaluate the size of the blocks that could be used. After experimenting with six different block sizes (40×32, 32×32, 48×40, 24×20, 30×20 pixels) a block size of 48×40 pixels was chosen as the optimized block size. It represents a good

compromise between sensitivity, accuracy and the number of features. As a result, each image of the optical flow vertical component is divided into blocks of size 48×40 pixels, resulting in 40 blocks (5 rows×8 columns) per optical flow, as shown in Figure 7.2.

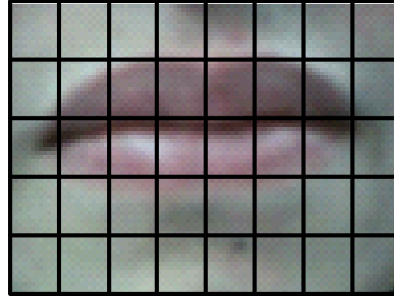


Figure 7.2: Representation of vertical component division in rectangular blocks.

To develop the template for each viseme, each optical flow frame was represented by the average of each block which resulted in a matrix of 5 rows × 8 columns = 40 values, then each matrix were converted into a row matrix. Row matrices of subsequent optical flow frames were stacked to develop the template matrix. To overcome the difference in the speed of speaking, each utterance was normalized (temporally) to 10 frames using linear interpolation. This resulted in a final feature vector of size 400 (10×40). The procedure of this motion template development is shown in Figure 7.3.

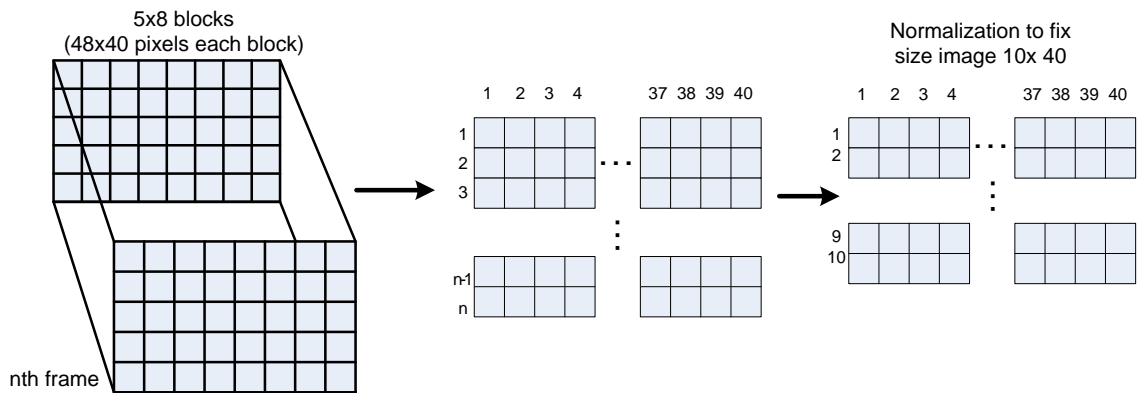


Figure 7.3: Development of optical flow vertical component based motion template (block size is 48×40).

Again the one-vs-rest binary class SVM technique was adopted for the classification. Results are shown in Table 7.2. A wide variety of feature extraction and classification algorithms have been suggested to date. It is quite difficult to compare the results as they are rarely tested on a common audio visual dataset. However it can be observed from Table 7.2 that the accuracy figure of 98.5% compares favourably to the techniques presented in [14, 34, 43, 67, 68, 72], in a visual-only scenario. The average specificity and sensitivity values of 99.6% and 84.2% respectively indicate that the proposed method is very efficient.

Table 7.2: Average classification results of individual one-vs-rest binary class SVM for 14 visemes (All values in %)

	Visemes	Specificity	Sensitivity	Accuracy
1.	/a/	99.9	82.9	98.6
2.	/ch/	99.6	92.9	99.1
3.	/e/	99.5	84.3	98.4
4.	/g/	99.1	77.1	97.6
5.	/th/	99.6	82.9	98.4
6.	/i/	99.2	80	97.9
7.	/m/	100	100	100
8.	/n/	99.8	78.6	98.3
9.	/o/	100	87.1	99.1
10.	/r/	99.2	77.1	97.7
11.	/s/	99.1	72.9	97.2
12.	/t/	99.5	75.7	97.8
13.	/u/	100	94.3	99.6
14.	/v/	100	92.9	99.5
	Average	99.6	84.2	98.5

The work by Mase *et al.* [72] is the closest reported to date to the proposed method. While that work uses a time warping based classification, this work presents the use of Support Vector Machines (SVM), and normalization is achieved by intensity quantization and the use of a fixed number of frames for an utterance to overcome the issue of difference in speed of speaking. A further difference is that Mase *et al.* had experimented on digit recognition while this work reports viseme recognition (visemes being fundamental units of visual speech). This difference is important because the optical flow computed from digits contains more information compared to visemes which are shorter

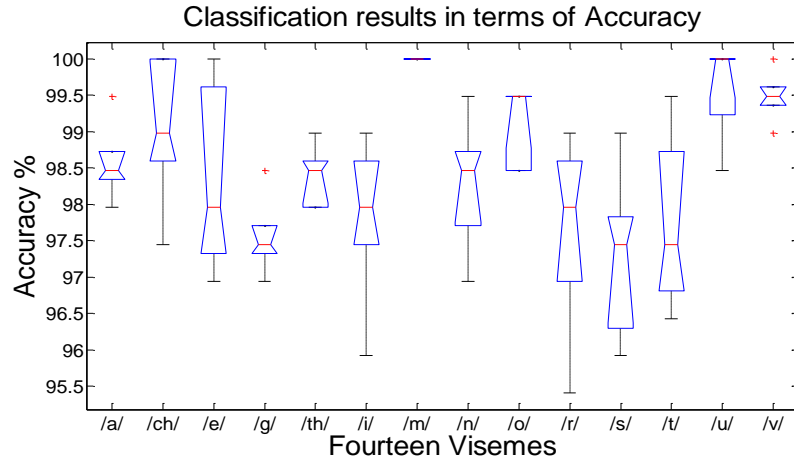
and there are less differences between different visemes. The proposed method will eventually lead to the development of continuous visual speech recognition with limited vocabulary and digit recognition would also be achieved.

Figure 7.4 shows the cross-validation process for accuracy, specificity and sensitivity values. As can be seen, the accuracy and specificity values do not have outliers and the results are highly consistent. The plot for sensitivity shows an increase in the standard deviation for visemes /g/, /r/, /s/ and /t/ (compared with individual data) even though the overall results are impressive. This can be attributed to the fricatives and stop consonants such as /g/, /t/ and /r/ that are the most difficult to identify in visual speech recognition because these sounds are not only based on lip movement, but also on the movement of the tongue that is not observed in the video data. It is proposed that these types of errors can be corrected using contextual information and cannot be achieved using only visual speech data.

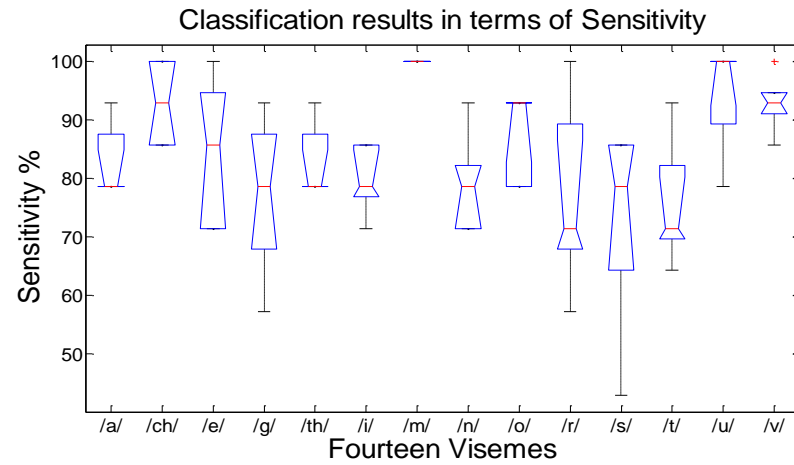
In addition to the above experiments, a hierarchical structure for classification is used to realize the multiclass classification. The SVM^{multiclass} (V2.13)² implementation is used for multiclass classification purpose. It uses the multiclass formulation described by Crammer and Singer [216]. But to avoid the problem of optimization, SVM^{multiclass} uses an algorithm based on Structural SVMs [217].

In multiclass classification the first classifier is binary and classifies /a/ vs ~a/ (not /a/). If the result of the classifier is +1, the result is declared as /a/, otherwise it checks the next classifier /c/ vs ~c/. Again, if the result is +1, the result is declared as /c/, otherwise it checks the next classifier and so on. As the proposed system is basically designed to work in real time, computationally expensive schemes such as Directed Acyclic Graph SVM (DAGSVM) which warrants the use of many more one-vs-one or one-vs-rest classifiers are avoided. Although it is at a cost, the results indicate that the proposed hierarchical process meets the criterion set out for a real time system. The SVM kernel function used was a radial basis function (RBF). The gamma parameter and the trade-off parameter C of the kernel chosen were optimized using iterative experiments (grid search).

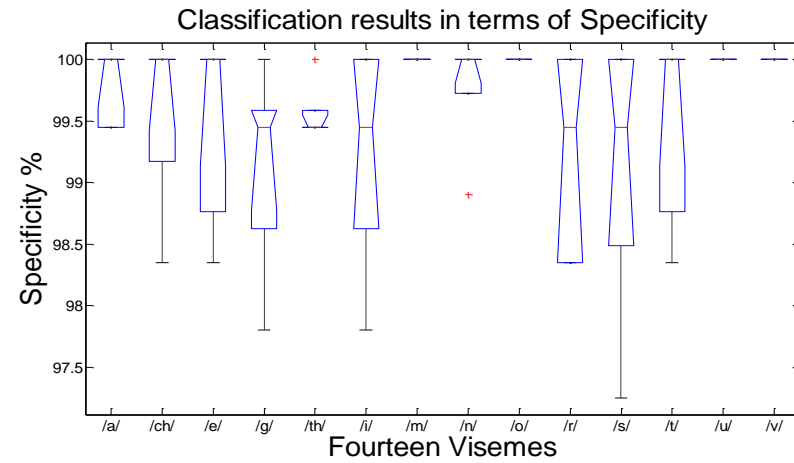
² http://svmlight.joachims.org/svm_multiclass.html



(a)



(b)



(c)

Figure 7.4: Cross validation Results (a) Accuracy, (b) Specificity and (c) Sensitivity.

The SVMs were trained with 882 training samples and were tested using the 98 remaining samples (one sample from each viseme) from all seven speakers. This process was repeated ten times for each viseme using standard cross validation method (10-fold).

To improve the understanding of the error generated and a possible solution, a confusion matrix was generated using a simple hierarchical structure for realizing multi-class SVM. The confusion matrix of multi-class classification is given in Table 7.3. Each row corresponds to a correct class, and columns represent the predicted classes. For instance, four examples of viseme /e/ are misclassified as /i/ and vice versa. The lowest accuracies in Table 7.3 are observed for visemes /th/ and /n/ which are interchangeably misclassified. It has to be noted that both /th/ and /n/ are dental consonants and their visual appearance is identical but the sound is different due to other sources of sound such as tongue, teeth and pallet. These types of errors cannot be solved by the choice of alternate features or classifiers but can be corrected by adding contextual information as carried out in state-of-the-art speech recognizers[220]. The multiclass classification results demonstrate that these features, when classified using SVM, can identify (85%) the visemes, and the misclassifications are not localized but spread across all visemes. As a future work, it is intended to use this data for implementing a contextual based classifier to further improve the results.

Table 7.3: Confusion Matrix for 14 visemes using hierarchical multi class SVM

	/a/	/ch/	/e/	/g/	/th/	/i/	/m/	/n/	/o/	/r/	/s/	/t/	/u/	/v/	Accuracy %
/a/	59	1	1	2	1	3	0	1	0	1	0	1	0	0	84.3
/ch/	1	61	0	2	0	0	0	1	1	0	1	2	1	0	87.1
/e/	3	0	56	1	1	4	0	1	0	1	1	2	0	0	80
/g/	2	2	2	58	0	1	0	1	1	0	0	2	1	0	82.9
/th/	1	0	4	0	53	1	0	6	0	2	2	0	0	1	75.7
/i/	0	0	4	0	0	58	0	2	0	0	5	1	0	0	82.9
/m/	0	0	1	0	1	0	64	0	1	0	0	0	2	1	91.4
/n/	0	0	2	3	4	2	0	53	0	2	1	2	1	0	75.7
/o/	1	2	0	0	0	1	0	1	62	2	0	0	1	0	88.6
/r/	3	0	1	1	1	0	0	0	3	61	0	0	0	0	87.1
/s/	0	0	0	0	4	3	0	2	0	2	57	0	0	2	81.4
/t/	2	0	3	3	1	1	0	0	0	2	0	58	0	0	82.9
/u/	0	1	0	0	0	0	0	0	1	1	0	0	67	0	95.7
/v/	0	0	0	0	1	0	0	2	0	0	0	1	0	66	94.3

7.2 Experimental Setup 2: Performance Evaluation of DMHIs

In Chapter 5 and 6, the development of the proposed DMHIs and the feature extraction techniques are discussed. This section presents the performance evaluated from the proposed approach by using two different feature extraction techniques which are compared with the traditional MHI.

Four directional motion history images (DMHIs) for each viseme are developed from the horizontal and vertical components of optical flow. The procedure to develop DMHIs has been described in Chapter 5. Each image of size 240x240 pixels represents the integrated motion of a mouth during an utterance in four directions (up, down, left and right). Varying facial movements during articulation of the 14 multiple visemes resulted in 4 motion templates of different patterns. In order to classify these images from a dataset, the image features are desired which represent the particular image with an optimized number of features. Features should have sufficient discriminating power and noise immunity for retrieval from the large image dataset. Two types of image features were evaluated to investigate the proposed technique; these were Zernike moments (ZM) and Hu moments (HM). ZM are image moments or features having the desired properties such as expression efficiency, robustness to noise, rotation invariance and multilevel representation for describing the shapes of patterns [203]. The optimum number of ZM features which is required for classification of the fourteen visemes was determined empirically. Hu [200] derived seven moment functions from the regular moments, which are also rotation, scaling and translation invariant. The seven Hu moments were computed from each DMHI. The experiments have compared the performance of both the features by using the SVM classifier and verified the robustness of ZM comparing to HM.

The SVM classification results for 980 utterances using ZM features and HM features are tabulated in Table 7.4 and Table 7.5. The results demonstrate the promising performance of using motion templates for recognition of visemes. The results indicate that both ZM and HM features are efficient descriptors to represent DMHIs. However the smaller number of Hu moments show less accuracy which is obvious from the number of

Table 7.4: Classification results of individual one-vs-rest class SVM using ZM (All values in %)

Zernike Moments of DMHIs				
	Visemes	Specificity	Sensitivity	Accuracy
1.	/a/	99.9	74.3	98.1
2.	/ch/	99.7	71.4	97.7
3.	/e/	99.5	77.1	97.9
4.	/g/	99.8	70	97.7
5.	/th/	100	74.3	98.2
6.	/i/	99.5	75.7	97.8
7.	/m/	99.8	90	99.1
8.	/n/	99.5	71.4	97.4
9.	/o/	99.9	80	98.5
10.	/r/	99.6	72.9	97.7
11.	/s/	99.6	70	97.4
12.	/t/	99.6	72.9	97.7
13.	/u/	99.7	81.4	98.4
14.	/v/	99.9	78.6	98.4
Average		99.7	75.7	98

Table 7.5: Classification results of individual one-vs-rest class SVM using Hu moments (All values in %)

Hu Moments of DMHIs				
	Visemes	Specificity	Sensitivity	Accuracy
1.	/a/	99.7	71.4	97.7
2.	/ch/	100	65.7	97.6
3.	/e/	99.8	67.1	97.4
4.	/g/	99.6	58.6	96.6
5.	/th/	100	74.3	98.2
6.	/i/	100	65.7	97.6
7.	/m/	99.9	90	99.2
8.	/n/	98.9	64.3	96.4
9.	/o/	99.8	80	98.4
10.	/r/	99.6	68.6	97.3
11.	/s/	99.6	58.6	96.6
12.	/t/	99.5	70	97.3
13.	/u/	99.9	80	98.5
14.	/v/	99.8	74.3	97.9
Average		99.7	70.6	97.6

features. The seven Hu moments features provide only the course shape of the image pattern and are insufficient for complicated pattern matching applications. Another shortcoming of geometric moments is the non-orthogonality of the features resulting in redundancy. The high success rates attained is also attributed to the ability of the RBF kernel SVM to correctly classify the non-linearly separable data.

7.2.1 Comparing the Performance of DMHI vs MHI

The experiments also compared the performance of DMHIs with traditional MHI. MHI was compared with the proposed DMHI technique using the same number of features consisting of 64 Zernike moments. However, the total number of Zernike moments of DMHIs is four times that of MHI because of four motion templates. Comparative results are presented in Table 7.6. It shows the average accuracy of identifying the visemes for all the 7 subjects for 14 different visemes using DMHIs and MHI. It can be observed from the results that the performance of DMHIs has outperformed the MHI using the ZM features in identifying the utterance on all accounts as it can address the overwriting problem significantly. While the average accuracy (98% and 93.66%) and specificity (99.7% and 99%) of the two techniques was comparable, the average sensitivity of DMHI was much better than that of MHI, with sensitivity of DMHIs being 75.7% while that of MHI was 24.4%. Thus, from the results, it is evident that the proposed DMHIs have outperformed the MHI in recognizing the lip movements for different phonemes.

The results indicate that the proposed method using DMHI is more sensitive in recognizing the correct viseme and leads to lower false negatives. The proposed method is based on advanced optical flow analysis [143] in which a standard incremental multi-resolution technique is used to estimate flow fields with large displacement. The optical flow estimated at a coarse level is used to warp the second image toward the first at the next finer level, and a flow increment is calculated between the first image and the warped second image. In building the pyramid each level is recursively down-sampled from its nearest lower level. The method employs robustness against lighting changes. The direction of motion of the lips is an important feature which is provided by the optical flow based DMHIs. Compared to DMHI, the standard MHI is the gray scale

representation of the difference of successive binary images of a video. In representing the MHI by ZM features the important information about their direction is lost which is critical in visual speech recognition. Hence, comparing sensitivity values in Table 7.6 suggests that the use of ZMs of DMHIs is successful in representing the lip movement, vindicating the hypothesis of this work. The sensitivity and unique property of rotational and scale invariance of ZMs ensures that the feature representation is independent of subject and the style with which they speak.

The use of motion templates generated by the optical flow vertical component and optical flow based DMHIs has eliminated the need for temporal modelling of visemes, and hence a static classifier such as SVM is able to classify the visemes reliably.

Table 7.6: Comparison of DMHI vs MHI (All values in %).

		DMHI			MHI		
	Visemes	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
1.	/a/	99.9	74.3	98.1	99.7	15.7	93.7
2.	/ch/	99.7	71.4	97.7	96.7	28.6	91.8
3.	/e/	99.5	77.1	97.9	99.8	15.7	93.8
4.	/g/	99.8	70	97.7	99.7	11.4	93.3
5.	/th/	100	74.3	98.2	98.7	21.4	93.2
6.	/i/	99.5	75.7	97.8	98.7	35.7	94.2
7.	/m/	99.8	90	99.1	98.4	52.9	95.1
8.	/n/	99.5	71.4	97.4	99.3	18.6	93.6
9.	/o/	99.9	80	98.5	98.6	25	93.6
10.	/r/	99.6	72.9	97.7	100	17.1	94.1
11.	/s/	99.6	70	97.4	99	24.2	93.6
12.	/t/	99.6	72.9	97.7	98.9	25.1	93.6
13.	/u/	99.7	81.4	98.4	99.1	24.7	93.8
14.	/v/	99.9	78.6	98.4	99.1	25.6	93.8
Average		99.7	75.7	98	99	24.4	93.7

7.3 Summary

In this chapter the performance of the proposed techniques based on optical flow vertical component motion templates and DMHIs was evaluated. A high level of viseme

recognition rate was obtained by each method. The optical flow vertical component of each image of a video was divided into non-overlapping blocks to compute the feature vector. Optimal size of the block was obtained by experimenting seven different block sizes. A block size of 48×40 pixels was chosen as the optimized block size. Average intensity value of each block was computed for a complete image sequence of an utterance and used as feature vector. The mean recognition rate of 98.5% with specificity of 99.6 % and sensitivity of 84.2 % has been achieved with one-vs-rest SVM classifier. To improve the understanding of the error generated multiclass SVM was adopted, however, the average accuracy of recognition reduces to 85%. The proposed system is computationally inexpensive with only 40 features required to represent each frame, and 400 features to represent each utterance. The system is independent of speed of speech of the subjects.

In other experimental setup, optical flow based four DMHIs are developed to represent a viseme. The performance of DMHIs for viseme recognition was evaluated. ZM and HM based features computed from each DMHI have produced the average recognition rates of 98% and 97.6% using the SVM classifier. The classification results of the proposed DMHI technique was compared with the traditional MHI technique using the ZMs as features and SVM as classifier. The results indicated that DMHIs have outperformed MHI in identifying the utterance on all accounts as it can address the overwriting problem in MHI significantly. While the average accuracy (98% and 93.66%) and specificity (99.7% and 99%) of the two techniques were comparable. The average sensitivity of DMHIs was achieved 75.7% while that of MHI was 24.4%.

Chapter 8

Conclusions and Future Directions

8.1 Conclusions

The main focus of this thesis was to develop a robust visual speech recognition system. Robust features are the desired characteristics to improve existing systems. Once robust visual features are extracted, their modelling with any classifier is straight forward.

This thesis has described two novel methods for the extraction of visual speech features and their modelling by support vector machines for visual speech recognition. The proposed feature extraction techniques are based on optical flow analysis, which is defined as the distribution of apparent velocities of brightness pattern movements in an image [92]. This gives the actual motion of the speaker's mouth movement as opposed to approximating the motion by computing the first and second order derivatives or difference of images in consecutive image sequence. Although optical flow analysis is very computational demanding, recent advances in high speed processor have resolved this issue.

Using visual information only, the system obtained performance levels on the recognition of 14 visemes defined in MPEG4. The described visual feature vectors consisting of motion and intensity information present two novel approaches to the representation of visual speech information. In the first approach, pure motion features obtained from the optical flow vertical component are used and have outperformed the other approaches. A SVM classifier was applied to classify the optical flow based feature vector and an overall 98.5% accuracy was obtained, and for viseme identification a hierarchical structure for classification is used to realize the multiclass classification, the overall 85% accuracy was obtained. It is concluded that the optical flow approach performed

according to expectations. However, some of the loss in performance is due to inaccuracies of the optical flow detection algorithm and not to the limited information content of the features. In addition, it is very difficult to compute the accurate optical flow of the speaker's face when occlusions by tongue and teeth suddenly appear. What is more rewarding is that the optical flow vertical component was investigated which contains more important features as compared to the horizontal component. In addition to that, the varying speeds of speech in inter and intra speakers were compensated to make the system suitable for subject independent scenarios.

In the second approach, motion features computed by optical flow analysis were used to develop four directional motion history images, in which motion features are mapped by the integer values in each image and are not the exact representation of motion and hence results in slight reduction in overall accuracy. The advantages of an optical flow based motion computation are that it does not require artificial markers for training and provides pure motion of the mouth, analogous to human perception.

Mouth movements represented by DMHIs were classified using the features extracted from each DMHI image. Two different image descriptors, ZM and HM, were investigated. A support vector machine classifier was used to classify the ZM and HM, where average accuracies of 98% with Zernike moments and 97.6% with Hu moments were achieved. The results have demonstrated that DMHIs are an efficient representation of spatial and temporal information of an utterance and are reliable for phoneme recognition in subject independent scenarios. To evaluate the performance of the proposed technique, ZM features of DMHIs were compared with the ZMs of the traditional MHI technique. The results indicated that the DMHI have outperformed the MHI technique, as it can address the overwriting problem in MHI significantly.

Using the DMHI technique, the average sensitivity was promising with 75.7% when compared to the average sensitivity of basic motion history image that is 24.4%. The advantage of ZM and HM based techniques is that the feature vector is reasonably low in dimension. Also they have important properties like invariance to translation, scale, and rotation which provide the robustness to the view angle and distance variations of

camera. However, the dimension of the feature vector of DMHI is four times that of the MHI but when compared to the HM based features the dimension of the HM feature vector is considerably lower. But HM computation have drawback of increasing complexity with increasing order of moments.

This thesis also describes a video-based *ad hoc* temporal segmentation of isolated utterances. It was used to detect the start and the end frame of an utterance from an image sequence. The technique is based on a pair-wise pixel comparison method. The efficiency of the proposed technique was tested on the available data set with short pauses between each utterance. The average error between automatic and manual segmentation for all subjects and all utterances is 2.98 frames/utterance. That is around 1.5 frames on either side of an utterance which is negligible. The limitation of the proposed technique is that it is suitable only for the data which have small pauses between each utterance.

Because of the nature of the computed features in all three scenarios, the SVM was the choice of classifier. This has performed well because it is suitable for a relatively a small number of features. The success of SVM in experiments is attributed to the use of a fixed number of features that eliminate the need of a temporal classifier. This is one of the first studies in its domain, which used the SVM multiclass classification and achieved good results.

Finally, it is concluded that progress of lip reading is increasing steadily but it requires more research to reach a human perceptual level. However, the advancement in processing speed and advanced algorithms in computer vision provides a larger momentum. This coupling assures that in the future, lip reading will be a major addition to human machine interaction.

8.2 Future Work

Optical flow based visual speech recognition techniques have demonstrated the ability to produce promising performance. A more robust visual speech recognition system can be obtained if the face and mouth detection algorithms are integrated with the proposed methods. Viola-Jones [95] face detection algorithm is one of the fast and efficient face

detectors available in literature that can be implemented for face detection. Even better results can be obtained if the rectangular division of the vertical component of optical flow is automated, and the division is performed exactly from the centre line of the lips by using mouth corner detection algorithms.

So far, current VSR systems are research level setups and datasets are recorded under carefully controlled conditions. Parameters such as head pose, distance and view angle between the camera and subject and lighting are fixed. However, variations in these parameters can have adverse affects on visual speech recognition. Processing of videos in these setups is performed offline. For future work it is desirable to build a limited vocabulary prototype, which is practically deployable on existing computers or mobiles, so that real time limitations can be observed.

To examine the feasibility of proposed optical flow based motion template for emotion recognition from the facial expressions of a subject. Facial expressions can indicate whether a subject is angry or happy, study of psychology has reported that the facial expressions contribute 55% of the effect of a communicated message while language and voice contribute 7% and 38% respectively [221]. The human perception about the emotions is based on movements of multiple facial features, so the proposed method can be suitable for human emotion recognition as it is solely based on motion tracking.

In Chapter 4, a simple speed of speech normalization technique based on linear interpolation has been adopted. In addition to that, DTW is a well-known technique to find an optimal alignment between variable length (time dependent) sequences. The overall distortion between signals is based on a sum of local distances between sample points. However, the particular optimal alignment minimizes the overall distortion to match the sequences. Initially, DTW has been used to compare different speech patterns in audio speech recognition. Later on, DTW has been used to automatically handle different speed and time deformation associated in the time dependent data in the fields of information and data mining. More work is required to compare the performance of the proposed linear interpolation method to the DTW in order to find out which approach

is more suitable for visual speech recognition, specifically when dealing with word recognition instead of viseme recognition.

Furthermore, almost all work to date in the visual domain has focused on a limited vocabulary dataset, and these datasets are pre annotated to find the word boundaries. Automatic annotation of visual speech recordings at the viseme level is a fundamental requirement in visual speech recognition. Identification of visemes' boundaries in continuous speech is a starting point of continuous visual speech recognition. Typically manual or audio signal based approaches has been adopted. These approaches are not suitable for continuous speech recognition in a visual domain. HMM is a finite state model that represents signals as transitions between a numbers of states. Each state is associated with a probability distribution. HMM assumes that the speech signals contains short time segments that are stationary. HMM models these short periods where the signals are steady and describes how these segments changes to subsequent segments. The changes between states are represented as transitions of states in HMM. The temporal variations within each of these segments are assumed to be statistical. Characteristics of HMM to represent the transition of states can be utilized to identify the word boundaries in continuous speech and then can be further used to segment the visemes.

Automatic temporal segmentation of continuous visual speech into basic visual units and the definition of new standard visemes is the biggest knowledge gap which should be addressed. This is a challenging part in the visual speech recognition domain and will solve the major problems of continuous visual speech recognition.

References

- [1] J. R. Pomerantz, "Perception: Overview," in *Encyclopedia of Cognitive Science*, ed: John Wiley & Sons, Ltd, 2006.
- [2] E. Hazard, *Lipreading for the Oral Deaf and Hard-of-hearing Person*: Thomas, 1971.
- [3] J. Jeffers and M. Barley, "Lipreading (speechreading)," *Charles C. Thomas, Springfield, IL*, 1971.
- [4] J. Jeffers and M. Barley, *Speechreading (lipreading)*: Charles C. Thomas Publisher, 1980.
- [5] A. Potamianos, *et al.*, "Adaptive categorical understanding for spoken dialogue systems," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 321-329, 2005.
- [6] M. A. Walker, "An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email," *Journal of Artificial Intelligence Research*, vol. 12, pp. 387-416, 2000.
- [7] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," 2004, pp. 351-es.
- [8] C. Xiaodong and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 1161-1172, 2005.
- [9] T. M. Sullivan and R. M. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, pp. 91-94.
- [10] X. Haitian, *et al.*, "Robust Speech Recognition by Nonlocal Means Denoising Processing," *Signal Processing Letters, IEEE*, vol. 15, pp. 701-704, 2008.
- [11] T. Zheng-Hua and B. Lindberg, "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 798-807, 2010.
- [12] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech & Language*, vol. 1, pp. 109-130, 1986.
- [13] R. Stern, *et al.*, *Signal processing for robust speech recognition*: Norwell, MA: Kluwer Academic Pub, 1997.

- [14] G. Potamianos, *et al.*, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, pp. 1306-1326, 2003.
- [15] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, pp. 341-353, 2010.
- [16] S. Oviatt, "Mutual disambiguation of recognition errors in a multimodel architecture," 1999, pp. 576-583.
- [17] G. Papandreou, *et al.*, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 423-435, 2009.
- [18] P. S. Aleksic, *et al.*, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1213-1227, 2002.
- [19] I. Arsic and J. Thiran, "Mutual information engenlips for audio-visual speech," in *14th European Signal Processing Conference*, Italy, 2006.
- [20] M. Gurban and J. P. Thiran, "Audio-visual speech recognition with a hybrid SVM-HMM system," in *13th European Signal Processing Conference*, 2005.
- [21] A. V. Nefian, *et al.*, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1274-1288, 2002.
- [22] K. Saenko and K. Livescu, "AN ASYNCHRONOUS DBN FOR AUDIO-VISUAL SPEECH RECOGNITION," in *Spoken Language Technology Workshop, 2006. IEEE*, 2006, pp. 154-157.
- [23] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746-748, 1976.
- [24] I. Matthews, *et al.*, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 198-213, 2002.
- [25] S. Kate, "Visual Speech Recognition with Loosely Synchronized Feature Streams," 2005, pp. 1424-1431.
- [26] W. C. Yau, *et al.*, "Visual speech recognition using dynamic features and support vector machines," *International Journal of Image & Graphics*, vol. 8, pp. 419-437, 2008.

- [27] D. Yu, *et al.*, "A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition," presented at the Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology, Tokyo, Japan, 2008.
- [28] V. N. Vapnik, *The nature of statistical learning theory*: Springer Verlag, 2000.
- [29] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions: Biological Sciences*, pp. 71-78, 1992.
- [30] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, pp. 212-215, 1954.
- [31] D. Reisberg, *et al.*, "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by eye: The psychology of lip-reading.*, B. Dodd and R. Campbell, Eds., ed: Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1987, pp. 97-113.
- [32] E. Akdemir and T. Ciloglu, "Bimodal automatic speech segmentation based on audio and visual information fusion," *Speech Communication*, vol. 53, pp. 889-902, 2011.
- [33] J. N. Gowdy, *et al.*, "DBN based multi-stream models for audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, pp. I-993-6 vol.1.
- [34] K. Iwano, *et al.*, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, p. 4, 2007.
- [35] G. Potamianos, *et al.*, "Towards practical deployment of audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, pp. iii-777-80 vol.3.
- [36] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, pp. 9-21, 2001.
- [37] X. Zhang, *et al.*, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1228-1247, 2002.
- [38] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal processing: Image communication*, vol. 16, pp. 477-500, 2001.
- [39] B. E. Dodd and R. E. Campbell, *Hearing by eye: The psychology of lip-reading*: Lawrence Erlbaum Associates, Inc, 1987.

- [40] D. Huggins-Daines, *et al.*, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. I-I.
- [41] E. Petajan, "Automatic lipreading to enhance speech recognition," in *IEEE Global Telecommunications Conference*, Atlanta, GA, USA, 1984, pp. 265-272.
- [42] S. Kumar, *et al.*, "EMG based voice recognition," in *Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004. Proceedings of the 2004*, 2004, pp. 593-597.
- [43] S. P. Arjunan, *et al.*, "Unspoken Vowel Recognition Using Facial Electromyogram," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, 2006, pp. 2191-2194.
- [44] B. Denby, *et al.*, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270-287, 2010.
- [45] M. Fagan, *et al.*, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, pp. 419-425, 2008.
- [46] T. Hueber, *et al.*, "Ouisper: corpus based synthesis driven by articulatory data," in *16th International Congress of Phonetic Sciences 2007*, pp. 2193-2196.
- [47] Y. Nakajima, *et al.*, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. V-708-11 vol.5.
- [48] M. Otani, *et al.*, "Vocal tract shapes of non-audible murmur production," *Acoustical science and technology*, vol. 29, p. 195, 2008.
- [49] V. A. Tran, *et al.*, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, pp. 314-326, 2010.
- [50] V. A. Tran, *et al.*, "Predicting F0 and voicing from NAM-captured whispered speech," in *Proc. Speech Prosody*, Campinas, Brazil, 2008.
- [51] Y. Nakajima, *et al.*, "Non-audible murmur (NAM) recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, pp. 1-8, 2006.
- [52] P. Heracleous, *et al.*, "Analysis and Recognition of NAM Speech Using HMM Distances and Visual Information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 1528-1538, 2010.

- [53] L. C. Ng, *et al.*, "Denoising of human speech using combined acoustic and EM sensor signal processing," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, pp. 229-232 vol.1.
- [54] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, pp. 36-43, 1992.
- [55] I. R. Titze, *et al.*, "Comparison between electroglottography and electromagnetic glottography," *The Journal of the Acoustical Society of America*, vol. 107, p. 581, 2000.
- [56] T. F. Quatieri, *et al.*, "Exploiting nonacoustic sensors for speech encoding," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 533-544, 2006.
- [57] M. Walliczek, *et al.*, "Sub-word unit based non-audible speech recognition using surface electromyography," *Proc. Interspeech, Pittsburgh, PA*, pp. 1487-1490, 2006.
- [58] P. Suppes, *et al.*, "Brain wave recognition of words," *Proceedings of the National Academy of Sciences*, vol. 94, p. 14965, 1997.
- [59] M. Wester and T. Schultz, "Unspoken speech-speech recognition based on electroencephalography," *Master's thesis, Universität Karlsruhe (TH), Karlsruhe, Germany*, 2006.
- [60] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by eye: The psychology of lip-reading.*, B. Dodd and R. Campbell, Eds., ed: Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1987, pp. 3-51.
- [61] P. J. Lucey, "Lipreading across multiple views," PhD Thesis, Speech, Audio, Image and Video Technology Laboratory, School of Engineering Systems, Queensland University of Technology, Brisbane, 2007.
- [62] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94.*, 1994, pp. 669-672
- [63] P. Duchnowski, *et al.*, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proceedings of the International Conference on Spoken Language and Processing*, Yokohama, Japan, 1994, pp. 547-550.

- [64] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 150, pp. 461-472, 1996.
- [65] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, 2003, pp. 1293-1296.
- [66] J. Huang, *et al.*, "Audio-visual speech recognition using an infrared headset," *Speech Communication*, vol. 44, pp. 83-96, 2004.
- [67] G. Potamianos, *et al.*, "Audio-visual automatic speech recognition: An overview," *Issues in Visual and Audio-Visual Speech Processing*, 2004.
- [68] A. Sagheer, *et al.*, "Appearance feature extraction versus image transform-based approach for visual speech recognition," *International Journal of Computational Intelligence and Applications*, vol. 6, pp. 101-122, 2006.
- [69] G. Zhao, *et al.*, "Lipreading with local spatiotemporal descriptors," *Multimedia, IEEE Transactions on*, vol. 11, pp. 1254-1265, 2009.
- [70] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," *International Journal of Computer Vision*, vol. 67, pp. 21-51, 2006.
- [71] H. Meng, *et al.*, "Motion information combination for fast human action recognition," in *Proc: Computer Vision Theory and Applications*, Spain, 2007, pp. 21-28.
- [72] K. Mase and A. Pentland, "Automatic lipreading by optical-flow analysis," *Systems and Computers in Japan*, vol. 22, pp. 67-76, 1991.
- [73] T. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1082-1089, 2006.
- [74] P. Lucey, *et al.*, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. 10th Australian Int. Conf. Speech Science and Technology.*, 2004.
- [75] W. C. Yau, "Video Analysis of Mouth Movement Using Motion Templates for Computer-based Lip-Reading," PhD Thesis, School of Electrical and Computer Engineering Science, Engineering and Technology, RMIT University, Melbourne, 2008.

- [76] A. J. Goldschen, *et al.*, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 150, pp. 505-518, 1996.
- [77] A. Rogozan, "Discriminative learning of visual data for audiovisual speech recognition," *International journal on artificial intelligence tools*, vol. 8, pp. 43-52, 1999.
- [78] J. Luettin, *Visual speech and speaker recognition*: Citeseer, 1997.
- [79] E. Ju and J. Lee, "Expressive Facial Gestures From Motion Capture Data," *Computer Graphics Forum*, vol. 27, pp. 381-388, 2008.
- [80] L. Bernstein, *et al.*, "Speech perception without hearing," *Attention, Perception, & Psychophysics*, vol. 62, pp. 233-252, 2000.
- [81] L. E. Bernstein, *et al.*, "Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults?," *Journal of speech, language, and hearing research*, vol. 44, pp. 5-18, 2001.
- [82] E. T. Auer Jr and L. E. Bernstein, "Enhanced visual speech perception in individuals with early-onset hearing impairment," *Journal of speech, language, and hearing research*, vol. 50, p. 1157, 2007.
- [83] C. Benoit, *et al.*, "Which components of the face do humans and machines best speechread?," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 150, pp. 315-330, 1996.
- [84] M. McGrath and Q. Summerfield, "Intermodal timing relations and audio visual speech recognition by normal hearing adults," *The Journal of the Acoustical Society of America*, vol. 77, pp. 678-685, 1985.
- [85] N. M. Brooke and Q. Summerfield, "Analysis, synthesis, and perception of visible articulatory movements," *Journal of Phonetics*, vol. 11, pp. 63-76, 1983.
- [86] K. E. Finn, "An investigation of visible lip information to be used in automated speech recognition," PhD Thesis, Georgetown University, 1986.
- [87] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *The Journal of the Acoustical Society of America*, vol. 73, p. 2134, 1983.
- [88] C. Yu, *et al.*, "Illumination normalization based on 2D Gaussian illumination model," in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, 2010, pp. V3-451-V3-455.

- [89] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Citeseer2010.
- [90] X. J. Zhang, *et al.*, "Finding lips in unconstrained imagery for improved automatic speech recognition," presented at the Proceedings of the 9th international conference on Advances in visual information systems, Shanghai, China, 2007.
- [91] G. Fanelli, *et al.*, "Hough transform-based mouth localization for audio-visual speech recognition," 2009.
- [92] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, pp. 185-203, 1981.
- [93] Y. Ming-Hsuan, *et al.*, "Detecting faces in images: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 34-58, 2002.
- [94] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2002.
- [95] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518 vol.1.
- [96] J. L. Jiang and L. Kia-Fock, "S-AdaBoost and pattern detection in complex environment," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2003, pp. I-413-I-418 vol.1.
- [97] S. Z. Li and Z. Zhenqiu, "FloatBoost learning and statistical face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 1112-1123, 2004.
- [98] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*: Prentice Hall, 1993.
- [99] H. A. Rowley, *et al.*, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 23-38, 1998.
- [100] K. Curran, *et al.*, "Neural network face detection," *Imaging Science Journal, The*, vol. 53, pp. 105-115, 2005.
- [101] H. Schneiderman and T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, 1998, pp. 45-51.

- [102] E. Osuna, *et al.*, "Training support vector machines: an application to face detection," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 130-136.
- [103] S. Peichung and L. Chengjun, "Face detection using discriminating feature analysis and support vector machine in video," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 407-410 Vol.2.
- [104] M. Heckmann, *et al.*, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002, pp. 1260-1273, 2002.
- [105] N. Eveno, *et al.*, "New color transformation for lips segmentation," in *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, 2001, pp. 3-8.
- [106] T. Coianiz, *et al.*, "2D deformable models for visual speech analysis," 2002.
- [107] X. Zhang and R. M. Mersereau, "Lip feature extraction towards an automatic speechreading system," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, 2000, pp. 226-229 vol.3.
- [108] J. Chaloupka, "Automatic lips reading for audio-visual speech processing and recognition," *skin*, vol. 1, p. 1.
- [109] N. Tsapatsoulis, *et al.*, "Efficient face detection for multimedia applications," 2000, pp. 247-250 vol. 2.
- [110] N. Eveno, *et al.*, "Accurate and quasi-automatic lip tracking," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, pp. 706-715, 2004.
- [111] A. C. Hurlbert and T. A. Poggio, "Synthesizing a color algorithm from examples," *Science*, vol. 239, p. 482, 1988.
- [112] A. W. C. Liew, *et al.*, "Segmentation of color lip images by spatial fuzzy clustering," *Fuzzy Systems, IEEE Transactions on*, vol. 11, pp. 542-549, 2003.
- [113] G. Ye-Peng, "Automatic Extraction of Lip Based on Wavelet Edge Detection," in *Symbolic and Numeric Algorithms for Scientific Computing, 2006. SYNASC '06. Eighth International Symposium on*, 2006, pp. 125-132.
- [114] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," 1997, pp. 403-409.
- [115] K. Messer, *et al.*, "XM2VTSDB: The extended M2VTS database," 1999, pp. 965-966.

- [116] J. Movellan, "Visual speech recognition with stochastic networks," *Advances in Neural Information Processing Systems*, pp. 851-858, 1995.
- [117] E. K. Patterson, *et al.*, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, pp. II-2017-II-2020.
- [118] M. Cooke, *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, p. 2421, 2006.
- [119] A. J. Sell and M. P. Kaschak, "Does visual speech information affect word segmentation?," *Memory & cognition*, vol. 37, pp. 889-894, 2009.
- [120] J. Ma, *et al.*, "Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of divisive motion capture data," *Computer Animation and Virtual Worlds*, vol. 15, pp. 485-500, 2004.
- [121] X. Zhang, "Automatic Speechreading for Improved Speech Recognition and Speaker Verification," PhD Thesis, Georgia Institute of Technology, 2002.
- [122] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*: Addison Wesley Publishing Company, 1992.
- [123] R. Polana and R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, 1994, pp. 77-82.
- [124] I. A. Essa and A. P. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, 1995, pp. 360-367.
- [125] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, 1995, pp. 374-381.
- [126] J. M. Siskind, "Grounding language in perception," *Artificial Intelligence Review*, vol. 8, pp. 371-391, 1994.
- [127] A. J. Goldschen, *et al.*, "Continuous optical automatic speech recognition by lipreading," in *Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, 1994, pp. 572-577.
- [128] J. Diaz, *et al.*, "FPGA-based real-time optical-flow system," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, pp. 274-279, 2006.

- [129] S. Tamura, *et al.*, "A robust multimodal speech recognition method using optical flow analysis," *Spoken multimodal human-computer dialogue in mobile environments*, pp. 37-53, 2005.
- [130] S. Baker, *et al.*, "A database and evaluation methodology for optical flow," 2007, pp. 1-8.
- [131] H. Zimmer, *et al.*, "Complementary optic flow," 2009, pp. 207-220.
- [132] A. Wedel, *et al.*, "Structure- and motion-adaptive regularization for high accuracy optic flow," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1663-1668.
- [133] A. Wedel, *et al.*, "An improved algorithm for TV-L 1 optical flow," *Statistical and Geometrical Approaches to Visual Motion Analysis*, pp. 23-45, 2009.
- [134] J. L. Barron, *et al.*, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43-77, 1994.
- [135] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys (CSUR)*, vol. 27, pp. 433-466, 1995.
- [136] H. Liu, *et al.*, "Accuracy vs Efficiency Trade-offs in Optical Flow Algorithms," *Computer Vision and Image Understanding*, vol. 72, pp. 271-286, 1998.
- [137] B. McCane, *et al.*, "On Benchmarking Optical Flow," *Computer Vision and Image Understanding*, vol. 84, pp. 126-143, 2001.
- [138] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, pp. 75-104, 1996.
- [139] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *In: Proc. Int. Joint Conf. on Artificial Intelligence*, 1981, pp. 674-679.
- [140] J. Weickert and C. Schnörr, "Variational Optic Flow Computation with a Spatio-Temporal Smoothness Constraint," *Journal of Mathematical Imaging and Vision*, vol. 14, pp. 245-255, 2001.
- [141] T. Brox, *et al.*, "High accuracy optical flow estimation based on a theory for warping," *Computer Vision-ECCV 2004*, pp. 25-36, 2004.
- [142] A. Bruhn, *et al.*, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, pp. 211-231, 2005.

- [143] D. Sun, *et al.*, "Learning Optical Flow," in *Computer Vision – ECCV 2008*. vol. 5304, D. Forsyth, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2008, pp. 83-97.
- [144] A. Blake and A. Zisserman, "Visual reconstruction," 1987.
- [145] B. Liao, *et al.*, "Color optical flow estimation based on gradient fields with extended constraints," in *Networking and Information Technology (ICNIT), 2010 International Conference on*, 2010, pp. 279-283.
- [146] R. Jain, "Difference and accumulative difference pictures in dynamic scene analysis," *Image and vision computing*, vol. 2, pp. 99-108, 1984.
- [147] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 928-934.
- [148] J. R. Bergen, *et al.*, "A three-frame algorithm for estimating two-component image motion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, pp. 886-896, 1992.
- [149] R. T. Collins, *et al.*, *A system for video surveillance and monitoring*: Citeseer, 2000.
- [150] Y. Kameda and M. Minoh, "A human motion estimation method using 3-successive video frames," 1996, pp. 135–140.
- [151] A. J. Lipton, *et al.*, "Moving target classification and tracking from real-time video," in *Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on*, 1998, pp. 8-14.
- [152] C. Wang and M. S. Brandstein, "A hybrid real-time face tracking system," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, pp. 3737-3740 vol.6.
- [153] W. Yau, *et al.*, "Lip-Reading Technique Using Spatio-Temporal Templates and Support Vector Machines," *Progress in Pattern Recognition, Image Analysis and Applications*, pp. 610-617, 2008.
- [154] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 257-267, 2001.
- [155] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image and vision computing*, vol. 21, pp. 729-743, 2003.

- [156] A. Bobick and J. Davis, "An appearance-based representation of action," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on, 1996*, pp. 307-312 vol.1.
- [157] J. W. Davis, "Appearance-based motion recognition of human actions," *Master's Thesis, Massachusetts Institute of Technology, 1996*.
- [158] A. Del Bimbo and P. Nesi, "Real-time optical flow estimation," in *Systems, Man and Cybernetics, 1993. 'Systems Engineering in the Service of Humans', Conference Proceedings., International Conference on, 1993*, pp. 13-19 vol.3.
- [159] N. Papenberg, *et al.*, "Highly accurate optic flow computation with theoretically justified warping," *International Journal of Computer Vision*, vol. 67, pp. 141-158, 2006.
- [160] A. Talukder, *et al.*, "Real-time detection of moving objects in a dynamic scene from moving robotic vehicles," in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on, 2003*, pp. 1308-1313 vol.2.
- [161] J. Wei and N. Harle, "Use of temporal redundancy of motion vectors for the increase of optical flow calculation speed as a contribution to real-time robot vision," pp. 677-680 vol. 2.
- [162] M. S. Gray, *et al.*, "Dynamic features for visual speechreading: A systematic comparison," *Advances in Neural Information Processing Systems*, pp. 751-757, 1997.
- [163] M. A. R. Ahad, *et al.*, "Motion history image: its variants and applications," *Machine Vision and Applications*, pp. 1-27, 2010.
- [164] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, pp. 62-66, 1979.
- [165] M. Pantic, *et al.*, "Learning spatio-temporal models of facial expressions," 2005.
- [166] M. Valstar, *et al.*, "Motion history for facial action detection in video," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004*, pp. 635-640 vol.1.
- [167] M. Valstar, *et al.*, "Facial action unit recognition using temporal templates," in *13th IEEE International Workshop on Robot and Human Interactive Communication, ROMAN. , 2004*, pp. 253-258.
- [168] M. Ahad, "Analysis of motion self-occlusion problem due to motion overwriting for human activity recognition," *Journal of Multimedia*, vol. 5, pp. 36-46, 2010.

- [169] W. C. Yau, *et al.*, "Visual recognition of speech consonants using facial movement features," *Integrated computer-aided engineering*, vol. 14, pp. 49-61, 2007.
- [170] P. Scanlon, "Audio and Visual Feature Analysis for Speech Recognition," PhD Thesis, University College Dublin, 2005.
- [171] M. Heckmann, *et al.*, "A hybrid ANN/HMM audio-visual speech recognition system," 2001.
- [172] P. Teissier, *et al.*, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 629-642, 1999.
- [173] M. N. Kaynak, *et al.*, "Lip geometric features for human-computer interaction using bimodal speech recognition: comparison and analysis," *Speech Communication*, vol. 43, pp. 1-16, 2004.
- [174] J. Luettin, *et al.*, "Active shape models for visual speech feature extraction," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 150, pp. 383-390, 1996.
- [175] A. Chitu, *et al.*, "Automatic Lip Reading in the Dutch Language Using Active Appearance Models on High Speed Recordings," in *Text, Speech and Dialogue*. vol. 6231, P. Sojka, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 259-266.
- [176] K. Kumar, *et al.*, "Profile View Lip Reading," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-429-IV-432.
- [177] M. E. Hennecke, *et al.*, "Using deformable templates to infer visual speech dynamics," in *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, 1994, pp. 578-582 vol.1.
- [178] M. Lievin, *et al.*, "Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme," in *Multimedia Computing and Systems, 1999. IEEE International Conference on*, 1999, pp. 691-696 vol.1.
- [179] A. Salazar, *et al.*, "Automatic Quantitative Mouth Shape Analysis," in *Computer Analysis of Images and Patterns*. vol. 4673, W. Kropatsch, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 416-423.
- [180] J. C. Wojdel and L. J. M. Rothkrantz, "Visually based speech onset/offset detection " in *Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia)*, Belgium 2000, pp. 156-160

- [181] C. Bregler, *et al.*, "Improving connected letter recognition by lipreading," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, pp. 557-560 vol.1.
- [182] G. I. Chiou and H. Jenq-Neng, "Lipreading from color video," *Image Processing, IEEE Transactions on*, vol. 6, pp. 1192-1195, 1997.
- [183] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *Multimedia, IEEE Transactions on*, vol. 2, pp. 141-151, 2000.
- [184] L. Luhong, *et al.*, "Speaker independent audio-visual continuous speech recognition," in *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, 2002, pp. 25-28 vol.2.
- [185] G. Potamianos, *et al.*, "An image transform approach for HMM based automatic lipreading," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 1998, pp. 173-177 vol.3.
- [186] M. J. Tomlinson, *et al.*, "Integrating audio and visual information to provide highly robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 821-824 vol. 2.
- [187] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, pp. 337-351, 1996.
- [188] M. Heckmann, *et al.*, "DCT-based video features for audio-visual speech recognition," in *Proceedings of International Conference on Spoken Language and Processing*, Denver, CO, USA, 2002, pp. 1925-1928.
- [189] P. Lucey and S. Sridharan, "Patch-based representation of visual speech," in *Proceedings of the HCSNet workshop on Use of vision in human-computer interaction - Volume 56*, Canberra, Australia, 2006, pp. 79-85.
- [190] J. H. Connell, *et al.*, "A real-time prototype for small-vocabulary audio-visual ASR," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, 2003, pp. II-469-72 vol.2.
- [191] J. Luettin, *et al.*, "Speechreading using shape and intensity information," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 58-61 vol.1.
- [192] M. T. Chan, "HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features," in *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, 2001, pp. 9-14.

- [193] L. D. Rosenblum and H. M. Saldaña, "Time-varying information for visual speech perception," *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, pp. 61-81, 1998.
- [194] C. Bregler, *et al.*, "A hybrid approach to bimodal speech recognition," 1994, pp. 556-560 vol. 1.
- [195] R. Goecke, "Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input," 2006, pp. 25-32.
- [196] S. Gurbuz, *et al.*, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," 2001, pp. 177-180 vol. 1.
- [197] S. Giannarou and T. Stathaki, "Shape signature matching for object identification invariant to image transformations and occlusion," 2007, pp. 710-717.
- [198] L. da Fontoura Costa and R. M. Cesar, *Shape analysis and classification: theory and practice*: CRC, 2001.
- [199] M. Kubo, *et al.*, "Content-based image retrieval technique using wavelet-based shift and brightness invariant edge feature," *International Journal of Wavelets Multiresolution and Information Processing*, vol. 1, pp. 163-178, 2003.
- [200] M. K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, pp. 179-187, 1962.
- [201] M. R. Teague, "Image analysis via the general theory of moments*," *JOSA*, vol. 70, pp. 920-930, 1980.
- [202] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern recognition*, vol. 37, pp. 1-19, 2004.
- [203] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, pp. 489-497, 1990.
- [204] C. H. Teh and R. T. Chin, "On image analysis by the methods of moments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 10, pp. 496-513, 1988.
- [205] Y. H. Pang, *et al.*, "Palmprint verification with moments," *Journal of WSCG*, vol. 12, pp. 325-332, 2004.
- [206] J. Flusser, *et al.*, *Moments and moment invariants in pattern recognition*: Wiley Online Library, 2009.

- [207] R. O. Duda, *et al.*, "Pattern Classification, New York," NY: Wiley InterScience, 2000.
- [208] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, pp. 884-900, 1999.
- [209] Z. Lingyun and W. Zhengzhi, "Predicting Transmembrane Topology of β -barrel Membrane Proteins with A Hidden Markov Model," in *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, 2007, pp. 145-148.
- [210] S. Abe, *Support vector machines for pattern classification*: Springer-Verlag New York Inc, 2010.
- [211] A. D. Kulkarni and P. Byars, "Artificial neural network models for image understanding," 1991, p. 512.
- [212] M. Heckmann, *et al.*, "A hybrid ANN/HMM audio-visual speech recognition system," in *International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark,, 2001, pp. 189–194.
- [213] W. Yau, *et al.*, "Voiceless speech recognition using dynamic visual speech features," in *Proc. of HCSNet Workshop on the use of Vision in HCI*, Canberra, Australia,, 2006, pp. 93-101.
- [214] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Arxiv preprint cs/9501101*, 1995.
- [215] J. Weston and C. Watkins, "Multi-class support vector machines," 1998.
- [216] K. Crammer, *et al.*, "On the algorithmic implementation of multi-class SVMs," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.
- [217] I. Tsochantaridis, *et al.*, "Support vector machine learning for interdependent and structured output spaces," 2004, p. 104.
- [218] A. K. Jain, *et al.*, "Statistical pattern recognition: A review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 4-37, 2000.
- [219] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001.
- [220] M. W. Kim, *et al.*, "Speech Recognition by Integrating Audio, Visual and Contextual Features Based on Neural Networks," in *Advances in Natural Computation*. vol. 3611, L. Wang, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 155-164.

[221] A. Mehrabian, *Nonverbal communication*: Aldine, 2007.