Robust Recognition of Human Behavior in Challenging Environments

by

Rajitha Navarathna, B.Sc Eng (Hons)

PhD Thesis

Submitted in Fulfilment

of the Requirements

for the Degree of

Doctor of Philosophy

at the

Queensland University of Technology

Speech, Audio, Image and Video Technologies Laboratory

Faculty of Science and Engineering

December 2013

Keywords

Human Behavior, Visual Information, Face Registration, Lucas & Kanade, Fourier Domain, Fourier Lucas & Kanade, Active Appearance Models, Fourier Active Appearance Models, Audio-Visual Automatic Speech Recognition, Lip Reading, Audience Engagement, Audience Ratings

Abstract

Being able to automatically and objectively measure human behavior ubiquitously would have profound implications within the educational, marketing, advertising, security/surveillance, human-computer-interaction and behavioral science domains. A key component in pursuing this lofty goal is to utilise visual modality. However, the task of measuring human behavior via videos is challenging due to variations caused by illumination, pose, scale, resolution and occlusion changes. Additionally, gaining reliable annotations of human behavior to train automatic systems can be problematic as it is often subjective, time-consuming and costly when continuously monitoring humans over long periods of time (i.e. many hours). This thesis focuses on robust methods which can deal with: a) visual variability, and b) long-term monitoring of people.

As the face is the most identifiable part of a person, as well as conveying the majority of behavioral information, accurate face registration is central to the success of automatically measuring human behavior. A plethora of work has been conducted in face registration with notable progress being made. Depending on the application and resolution of the input image, face registration is typically performed using either a *coarse* (e.g. bounding-box) or dense (e.g. Active Appearance Models (AAMs)) alignment approach. For conditions which do not vary over time, adequate solutions for face registration have already been formulated. However, when conditions vary over time (such as illumination), such approaches fail. The first contribution of this thesis

is a novel alignment algorithm based on applying the Lucas-Kanade (LK) algorithm within the Fourier domain, which is robust to illumination variation. With respect to AAMs, the *Fourier AAM* (FAAM) interprets the joint alignment across filter responses as a form of the weighted AAM algorithm. This particular method empirically shows the substantial improvement in person specific AAM fitting performance over canonical LK inspired fitting algorithms, when using the proposed Fourier variants. With faces of a small resolution, a similar approach is used to track the bounding box of a person's face using a template update approach, which we call the *Fourier Template Update algorithm*.

In this thesis, we used two real-world data sources which are representative of scenarios where automatic systems could be effectively deployed in the future. These were: 1) people driving in an automotive environment and 2) a group of people watching a movie. As these environments had a specific target application with respect to measuring human behavior (i.e. speech recognition and audience engagement), as well as being in challenging environments (e.g. poor and fluctuating lighting conditions, varying head-pose, resolution and long in duration) - we were able to evaluate the effectiveness of automatic approaches in these domains. Consequently, the second contribution of this work was to show that we could improve speech recognition by tracking a person's mouth region from multiple camera-views.

The third contribution stemming from this thesis, is a method of representing audience behavior through facial and body motions from a single video stream, and using these features to objectively measure and summarize audience reactions to feature-length movies. As the movie viewing environment is dark and contains views of many people at different scales and viewpoints, we introduce an IR based test-bed. We use the FLK algorithm to register and stabilize audience faces, and generate flow-profile of each person, contained within their local 3D temporal volume via optical flow. We then use an "entropy of pair-wise correlations" method, which compares short-term

motion features across audience members to obtain an objective measure of audience “coherency”. Additionally, we utilize an off-line change-point detection algorithm to temporally cluster and summarize audience behaviors into a series of interest segments, as well as to learn behaviors associated with good and bad movies using crowd-sourced audience ratings from *rottentomatoes.com*. As audience interest segments are highly synchronized and occur over very small periods of time (i.e. a number of frames), we show that our approach can outperform human-annotated labels, which do not pick up on these fine details.

Contents

List of Tables

List of Figures

Acronyms & Abbreviations

2D	2-dimensional
3D	3-dimensional
AAM	Active Appearance Model
ADC	Analog-to-Digital Converter
API	Application Programming Interface
ASM	Active Shape Model
AUC	Area Under the Curve
AVASR	Audio Visual Automatic Speech Recognition
AVICAR	Audio-Visual Speech Corpus in a Car Environment (database)
CL	Center Left
CP	Change Point
CR	Center Right
CUAVE	Clemson University Audio Visual Experiments (database)
DCT	Discrete Cosine Transform
DSB	Delay-Sum Beamforming
FAAM	Fourier Active Apperence Model
FLK	Fourier Lucas Kanade
FMN	Feature Mean Normalisation
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
HOOF	Histogram of Oriented Optical Flow
HTK	HMM Toolkit (software)

IC	Inverse Compositional
IOS	iPhoneOS
IR	Infrared
KITT	Knight Industries Two Thousand
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LK	Lucas Kanade
M2VTS	Multi Model Verification for Teleservices and Security Applications (database)
MBSS	Multi-Band Spectral Subtraction
MFCC	Mel-Frequency Cepstral Coefficients
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
PDM	Point Distribution Model
PLP	Perceptual Linear Predictive
PTZ	Pan Tilt Zoom
RANSAC	Random Sample Consensus
RMS	Root Mean Square
ROI	Region Of Interest
SHMM	Synchronous HMM
SHORE	Sophisticated High-speed Object Recognition Engine
SIFT	Scale Invariant Feature Transform
SL	Side Left
SNR	Signal-to Noise Ratio
SR	Side Right
SSD	Sum-of the Square Difference
SVM	Support Vector Machine
VIRAT	Video and Image Retrieval Analysis Tool
VJ	Viola-Jones
XM2VTS	Extended M2VTS (database)

Certification of Thesis

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

Signed: ______________________

Date: ______________________

Acknowledgments

Work on this journey, while interesting, was not easy and would not have been possible without support and guidance from a lot of people along the way. First and foremost, I would like to express my utmost gratitude to my Amma and Appachi who have been, and still are, my biggest supporters. Their never ending support, as well as their guidance, comfort, compassion and perspective have allowed me to achieve more than I could ever imagine. I am eternally grateful to them for everything they have done for me and they will never know how much of a positive influence they have been on my life.

I would like to express my warm and sincere gratitude to Professor Sridha Sridharan, who is the director of the Speech Audio Image Video Technology (SAVIT) laboratory at Queensland University of Technology (QUT), for all the support and guidance he has given me in order to overcome numerous obstacles in this journey. I am really grateful to him for providing an excellent work environment at the SAIVT lab as well as for the opportunities he has made for me, in working with world-class engineers and scientists and attending international conferences. I would like to thank him for his effort, patience and excellence guidance during my PhD journey.

I wish to further extend my sincere thanks to my associate supervisors, Associate Professor Clinton Fookes and Dr David Dean, for their valuable support and advice throughout this journey. I would also like to thank Associate Professor Simon Lucey for his excellence guidance and teaching of important technical concepts. Also, I wish to convey my gratitude to QUT for providing me this opportunity and also to all the administrative staff in HDR research coordination. My appreciation is further extended to the Co-operative Research Centre for Advanced Automobile Technology (AutoCRC) for providing me financial assistance during my candidature.

I was very fortunate to have Dr Patrick Lucey at Disney Research Pittsburgh during my PhD. His excellent approach to tackling a research question with real-world applications was always interesting to me. His unrelenting pursuit of the question “visualization?” has been a great source of inspiration for me, to think ’out of the box’. I convey my gratitude for his effort, patience and guidance and of course, for our discussions about cricket.

I was lucky enough to visit Disney research Pittsburgh for 6 months internship during my course. I would like to thank Professor Jessica Hodgins, Dr Iain Matthews, Dr Patrick Lucey and Professor Sridha Sridharan for providing me this opportunity to work at Disney research. I would also like to thank Dr Peter Carr from the computer vision lab at Disney research for his excellence support during my stay there. My time at Disney proved to be one of the most rewarding experiences in my life.

I wish to acknowledge the members of the SAIVT laboratory; Dr Simon Denman, Dr Ruan Lakemond, Alina Bialkowski, Sabeshan Sivaplan, Chris Chew, Afsaneh Ghasemi, Shahram Kalantarri, Kien Nguyen, Xin Shane, Ahilan Kanagasundaram, Kaneswaran Anatharajah, Daniel Chen, JingXin and the remainder of our SAIVT colleagues, also YingYing Zhu, Jack Valmadre and Hilton Bristow from CI2CV lab at CSIRO. I would like to express my warm and sincere gratitude to Palmo Thinley, Chandima Gunawardena, Madara Karunaratne, Dhanushka Krishnajith, Rupika Bandara, Kanchana Rathnayake, Rinku Tuli, Rakkitha Thilakrathne, Dinesha Chathurani, Anuruddha Ratnayake, Gawri Edussuriya for their constant support and valuable help. This support and help during the ups and downs of PhD candidature, was extremely

important to me.

I would also like to thank my sister for her support and laughter, given to me over the years. Her encouragement and support is always there to make me be a better person. Last but not least, I would like to express my gratitude to all my relatives, school and university teachers, friends and others who have helped me directly or indirectly along this journey.

RAJITHA NAVARATHNA

Queensland University of Technology
December 2013

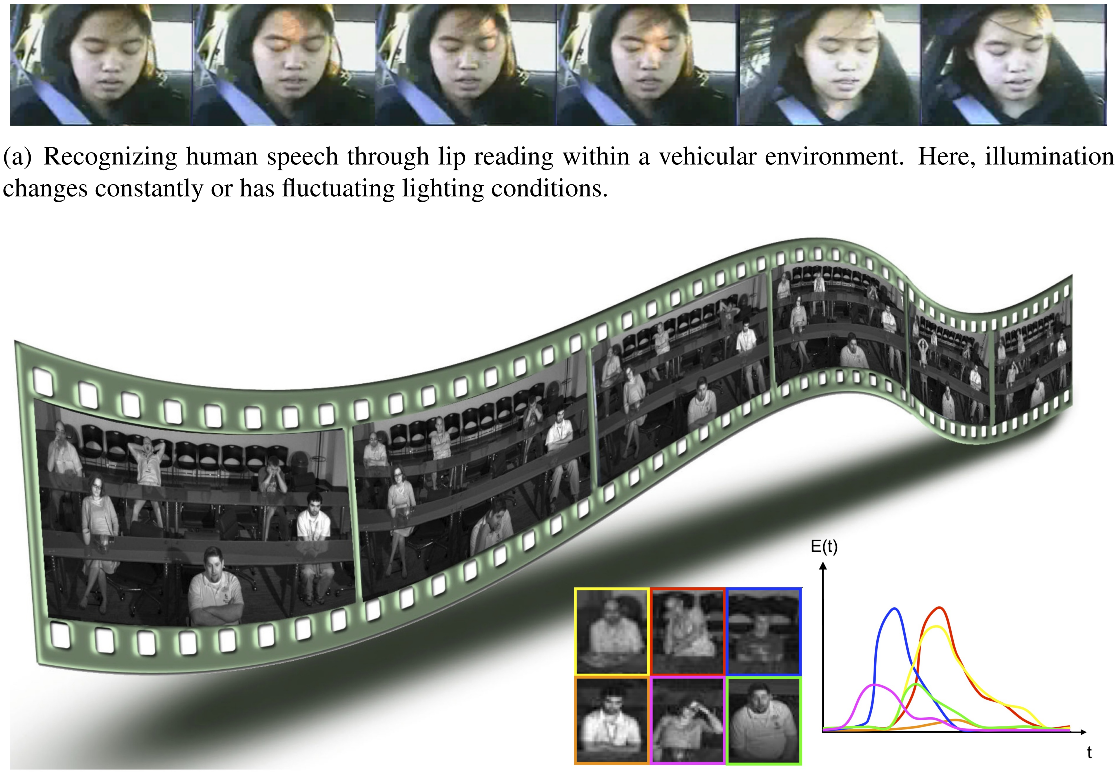
To my loving *Amma, Appachi* and *Nangi*

Chapter 1

Introduction

1.1 Motivation and Overview

The recent dramatic improvement in video technology has opened up a plethora of potential assistive tools which have the possibility of making people's day-to-day lives safer and more efficient. As video is passive and non-invasive, information stemming from the visual modality is often seen as the solution in making these tools robust to real-world scenarios. An example of this is in speech recognition within a vehicle, which is acoustically very noisy (see Figure 1(a)), where using the visual information from around a speaker's mouth can help improve the speech intelligibility. Another example is measuring the emotion state or response of a person or group of people watching a movie by automatically measuring facial expressions or movements (Figure 1(b)). Even though the deployment of these applications is quite exciting, major bottlenecks exist - mostly due to the variability or noise that injects itself in the video signal over long-periods of time (such as illumination variation, head-pose, occlusion and resolution). Additionally, gaining reliable annotations of human behavior to train automatic systems can be problematic as it is often subjective, time-consuming and



(a) Recognizing human speech through lip reading within a vehicular environment. Here, illumination changes constantly or has fluctuating lighting conditions.

(b) Recognize audience collective/uninterested synchronized behavior in very dark environment through detecting facial expressions and body gestures.

Figure 1.1: Real world challenging examples of recognizing human behaviour using rich amount of visual information in very challenging lighting conditions.

costly when continuously monitoring humans over long periods of time (i.e. many hours). This thesis focuses on robust methods which can deal with: a) visual variabilities, and b) long-term monitoring of people.

Central to the success in automatically measuring human behavior is the accurate registration of a person's face as it is conveys the majority of behavioral information. Face registration has got to the stage where it is widely used in a host of applications such as smart-phones, online applications (picasa, iPhoto etc.), video games (e.g. Xbox Kinect), face retargeting [167] and marketing and advertising [127]. These approaches are typically performed using either a *coarse* (e.g. bounding-box) or dense (e.g. Active Appearance Models (AAMs)) alignment approach. These approaches work well but when conditions vary over time (such as illumination), such approaches fail. The first contribution of this thesis is the development of a novel alignment algorithm which can deal with this variation. It is based on the Lucas-Kanade (LK) algorithm within the Fourier domain, which can be applied to both AAMs and bounding-box approaches. We also show the utility of this approach on videos of people driving with fluctuating illumination environments (see Figure 1(a)). Stemming from this work, we show another contribution which is improving speech recognition by tracking a person's mouth region from multiple camera-views.

The final contribution of this thesis is to learn, summarise and predict audience behaviors over long-periods of time (Figure 1(b)). As automatic systems rely on human annotation to train classifiers, the quality of the classifier heavily depends on the reliability and quality of annotations. However, when dealing with large volumes of video data across long-periods of time - the reliability, quality and quantity of annotations are low due to the monotonous nature of the task in addition to the cost and time involved with the task. In this thesis, we use the Fourier Lucas-Kanade (FLK) algorithm to register and stabilize audience faces, and generate flow-profile of each person contained within their local 3D temporal volume via optical flow in a very dark environment. We then use a novel "entropy of pair-wise correlations" method, which compares short-term motion features across audience members to obtain an objective measure of audience "coherency". Additionally, we utilize an off-line change-point detection algorithm to temporally cluster and summarize audience behaviors into a series of interest segments, as well as to learn behaviors associated with good and bad movies using crowd-sourced audience ratings from *rottentomatoes.com*[1]. As audience interest segments are highly synchronized and occur over very small periods of time (i.e. a number of frames), we show that our approach can outperform human-annotated labels which do not pick up on these fine details.

[1] `rottentomatoes.com`

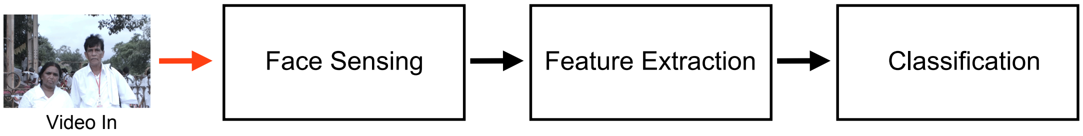


Figure 1.2: This thesis attempted to remove the expenses associated with the lighting variations and recognise human behavior using visual modality. The normal pipeline would be to first sense the face and then extract out features which contains information and finally do the classification task. In this dissertation all these stages will be investigated.

1.2 Scope of Thesis

Recognising human behavior through visual information is a challenging task. A major reason stymieing the full deployment of such a system in "real-world" applications, is the lack of research being conducted that focuses on unwanted variabilities that lie within the visual domain, such as *various illumination conditions* (i.e fluctuating, very dark lightings etc...). The key problem is obviously sensing the face in these illumination conditions as the face contains a rich amount of behavioral information. In an attempt to remedy this situation, the work in this thesis has concentrated on researching and developing methods to recognize human behavior through visual information by focusing on face sensing, feature extraction and classification stages (See Figure 1.2). Within this multi-tasked problem, the scope of this thesis was constrained to the following objectives:

1. Remove the expenses associated with the lighting variations for the task of face sensing which limits the use of computer vision theory for many real-world applications.

2. Recognise human speech when multiple frontal or near-frontal views of speakers' faces are available in an automotive environment which has fluctuating lighting conditions, to improve the audio-visual automatic speech recognition

(AVASR) performance.

3. Develop an automatic real-time objective measure of audience in very dark lighting conditions through analysing the collective facial and body movements.

All the work contained in this thesis is designed to address each of these novel and previously unsolved problems.

1.3 Outline of Dissertation

The remainder of this thesis is structured as follows:

Chapter 2 presents a background review of human behavior recognizing focusing on the effect of visual information. This Chapter has given insights on the transition process of human behavior to a computer-vision perspective. A brief history of three specific tasks (i.e facial expression recognition, visual speech recognition and activity recognition) of human recognising using visual information especially focusing on face and body are described.

Chapter 3 gives comprehensive evaluation of existing face alignment technologies relevant to the work in this thesis. This Chapter has given insights into publicly available coarse-type of face alignment methods (i.e *Viola-Jones, Apple face detector and "Fraunhofer" face detector*) and the well known fine alignment method called AAMs. The model construction of person-specific AAMs and generic AAMs are briefly detailed and two common fitting algorithms (i.e Simultaneous and Project-out algorithm) are described.

Chapter 4 presents a novel solution to overcome the problem of poor fitting performance with varying illumination. When unaccounted appearance variations are encountered due to a change in the environment (e.g.,illumination or camera change), person specific AAMs perform poorly. The Chapter introduces the use of filters with AAM fittings in the Fourier domain which improves the fitting performance. In the beginning of the Chapter, the LK algorithm and FLK algorithm are reviewed. A detailed comparison of these two algorithms is outlined. In addition to that this Chapter shows the substantial improvement in person specific AAM fitting performance, over canonical LK inspired fitting algorithms (i.e Simultaneous and Project-out algorithm).

Chapter 5 presents review of an audio-visual speech recognition (AVASR) system in a real world automotive environment. Each component of AVASR system (i.e acoustic feature extraction, visual front-end, visual feature extraction, speech modeling) is deeply outlined, and single-view, multi-view and in-car audio-visual databases are highlighted. Audio-visual speech recognition has previously been shown to provide a considerable improvement over acoustic-only approaches in noisy environments, but most audio-visual speech recognition approaches have only been examined in relatively clean conditions and have rarely dealt with the visual variabilities such as head movement, poor/varying illumination and poor video resolution/quality. The Chapter reviews advanced speech enhancement techniques for improved audio-only speech recognition in an automotive environment. Another avenue for improving AVASR in real world conditions is to take advantage, if possible, of multiple views of the visual-speech information, or lip-movements, of the active speaker. This Chapter extended upon the established audio-visual speech recognition literature to show that, in a real-world automotive environment, further improvements in speech recognition accuracy over traditional single-camera AVASR approaches can be obtained when multiple frontal or near-frontal views of speakers' faces are available. It also investigates the usefulness of the visual information from different camera angles.

Chapter 6 proposes the novel problem of an automatic real-time objective measure of audience engagement, through automatically analysing the collective/uninterested synchronized behaviour in a very dark environment in detecting facial expressions and body gestures. In addition to introducing a new problem to the field of face and gesture analysis as well as a solution on how to capture such data, there are numerous technical challenges that are highlighted in this Chapter for which solutions are presented. The first part of the Chapter highlights various kinds of problems with illumination and reflection from the test-screen and a novel solution has been proposed to capture smooth data in a dark environment. A new set-up of data has been collected while audiences of various sizes are watching movies, live-sports matches and public presentations. Next the Chapter reviews the individual audience behavior and group behavior via smiles and optical flow energy measurements. Experiments are conducted while an audience engage with movies, live sports events and live presentations. Finally it proposes an *entropy of pair-wise correlations* measure to give an indication of audience *coherency*. Additionally, it proposes an off-line *change-point* detection algorithm to temporally cluster and summarize audience behaviors into a series of interest segments.

Chapter 7 summaries the work contained in this thesis and outcomes are highlighted. In addition to the contributions, this Chapter also suggests future research avenues which can be taken to improve the research conducted in this dissertation.

1.4 Original Contribution of Thesis

The summarised key contributions from the work presented in this thesis are as follows:

(i) Chapter 3 illustrates the behavior of coarse-type of face alignments methods and fine registration methods. The performance of face alignments methods is also compared with different lighting conditions and demonstrates that it degrades with illumination, especially in low-light conditions when coarse-type of face alignments methods are used.

(ii) In Chapter 4.5 a novel method is demonstrated of how LK inspired AAM fitting gives identical performance in the spatial and Fourier domains. Further, we demonstrate how the effect of multiple filter responses can be re-interpreted as a diagonal weighting matrix in the Fourier domain leading to substantial computational savings when performing inverse compositional simultaneous fitting across multiple filter responses.

(iii) We demonstrate the process of applying the inverse compositional project-out algorithm in the Fourier domain by showing how: (i) Fourier transform to the current image, and (ii) the application of multiple filter responses can be completely pre-computed offline (Chapter 4.5.1). This contribution is of key importance to person specific AAM face fitting as it provides an extremely computationally efficient method that affords both invariance to both expression and environmental variations.

(iv) Chapter 4.6 presents empirically the substantial improvement in person specific AAM fitting performance, over canonical LK inspired fitting algorithms (i.e. simultaneous and project out), when using our proposed Fourier variants. For all our experiments we employed biologically motivated Gabor filter banks.

(v) We synchronised the audio and visual streams of the phone-number portion of the AVICAR database, allowing audio-visual experiments to be conducted and introduce a novel audio-visual protocol in Chapter 5.3.1 for the AVICAR database for the task of speaker-independent speech recognition.

(vi) We provide a comparison of the recognition performance of single channel and multi-channel enhanced speech (in Chapter 5.4.2) with the performance of audio-visual speech using data from a challenging automotive environment (AVICAR [98]), which introduces a number of visual challenges, including changes in illumination and speaker pose as well as severe audio impairment arising from car engine, wind and road noise. Chapter 5.7 shows that visual speech recognition results within a vehicle-environment are obviously diminished from what is obtained in ideal laboratory conditions.

(vii) We extend this study to also demonstrate that the complementary nature of visual information and enhanced audio observed in [45] still holds true when using multi-channel speech enhancement algorithms and state-of-the-art middle integration techniques (i.e synchronous hidden Markov model (SHMM)) for audio-visual fusion. Experimental results in Chapter 5.8 show that the combination of acoustic speech enhancement and SHMM-based AVASR can provide further gains in accuracies.

(viii) Chapter 5.8.2 presents that further improvements in speech recognition accuracy over traditional single-camera AVASR approaches can be obtained when multiple frontal or near-frontal views of speakers' faces are available in a real-world automotive environment.

(ix) Chapter 5.6 re-examines the current state of-the art visual HMM by comparing off-the shelf face detector with FLK approach.

(x) Audience environments and test-screenings are very dark and suffer from reflections from the screening. To counter these issues, we employ a hardware solution which gives us a uniform smooth signal in Chapter 6.3.

(xi) Chapter 6.3.2 introduces a labelled dataset of audiences of varying sizes watching movies. A key insight from this data collection effort is the lack of movement/actions, which highlights the sensitivity of this task.

(xii) Even though the movie viewing environment is very dark and contains views of people at different scales and viewpoints, we can measure audience behavior by improving smile detection by using the FLK algorithm to register audience members faces. This overcomes instances when there is abrupt change in illumination caused by sudden movement. Chapter 6.4 shows that improved smile detection is possible using this method. In addition to smiles Chapter 6.5 introduces a method to obtain an indicator for audience *engagement* or *disengagement* using standard optical flow. We generate a *flow-profile* of each person contained within their local 3D temporal volume via optical flow which is aggregated into a collective *stillness* measure.

(xiii) Chapter 6.6 shows that the proposed unsupervised approach outperforms human-annotated labels, which do not pick-up these fine details. Using the audience ratings from *rottentomatoes.com*, we are able to learn to differentiate between good and bad movies based on these interest segments. The introduced method showed that we can give a reasonable approximation of audience behavior compared to *rottentomatoes.com* ratings.

1.5 Notations

1.5.1 General Notation

Vectors are always represented in lower-case bold (e.g., $\mathbf{a}$). Matrices are always expressed in upper-case bold (e.g., $\mathbf{A}$). Scalars in lower-case (e.g. a). Images in this thesis shall always be expressed in capitalized form A. Warp functions $\mathcal{W}(\mathbf{x};\mathbf{p})$ will be used throughout this paper to denote a warping of a $2D$ coordinate vector $\mathbf{x} = [x, y]^T$ by a warp parameter vector $\mathbf{p} \in \mathcal{R}^P$, where p is the number of warp parameters, back to a fixed base coordinate system. This base coordinate system is defined when $\mathbf{p} = \mathbf{0}$ such that $\mathcal{W}(\mathbf{x};\mathbf{p}) = \mathbf{x}$. An abuse of notation is entertained in this paper for when an image A is warped by the warp parameter vector $\mathbf{p}$, such that $A(\mathbf{p}) = [A(\mathcal{W}(\mathbf{x}_1;\mathbf{p})), \ldots, A(\mathcal{W}(\mathbf{x}_D;\mathbf{p}))]^T$. In this instance $A(\mathbf{p})$ is a d dimensional vector of image intensities, where d denotes the number of discrete coordinates in the base coordinate system. The steepest descent matrix of an image $A(\mathbf{p})$ is defined as $\frac{\partial A(\mathbf{p})}{\partial \mathbf{p}}$. This $p \times d$ matrix is formed by combining image gradients of $A(\mathbf{p})$ with the Jacobian of the warp function $\mathcal{W}(\mathbf{x};\mathbf{p})$, more details on the formation of this matrix can be found in [120]. Finally, the notation $\| \mathbf{a} \|^2_{\mathbf{Q}}$ to represent the quadratic form $\mathbf{a}^T\mathbf{Q}\mathbf{a}$, and $\mathbf{Q}$ is a symmetric, positive semi-definite weighting matrix.

1.5.2 Fourier Notation

This thesis also borrows concepts from signal processing. A 2D convolution operation is represented as the $*$ operator. A $\hat{}$ applied to any vector denotes the 2D Discrete Fourier Transform (DFT) of a vectorized 2D image $A(\mathbf{p})$ or signal $\mathbf{a}$ such that $\hat{A}(\mathbf{p}) \leftarrow \mathbf{F}\, A(\mathbf{p})$ and $\hat{\mathbf{a}} \leftarrow \mathbf{F}\mathbf{a}$. $\mathbf{F}$ is the $D \times D$ matrix of complex basis vectors for mapping to the Fourier domain for any D dimensional vectorized image/signal. We have chosen to employ a Fourier representation in this thesis due to its particularly useful ability to represent convolutions as a Hadamard product in the Fourier domain. Additionally, we take advantage of the fact that $\text{diag}(\hat{\mathbf{g}})\hat{\mathbf{a}} = \hat{\mathbf{g}} \circ \hat{\mathbf{a}}$, where $\circ$ represents the Hadamard product, and diag() is an operator that transforms a D dimensional vector into a $D \times D$ dimensional diagonal matrix. The role of filter $\hat{\mathbf{g}}$ or signal $\hat{\mathbf{a}}$ can be interchanged with this property. Any transpose operator ^{T} on a complex vector or matrix in this paper additionally takes the complex conjugate in a similar fashion to the Hermitian adjoint [137].

1.6 Publications Resulting from Research

The following fully-reviewed publications have been produced as a result of the work outline in this dissertation.

1.6.1 International Journal Publications

(i) **R. Navarathna**, P. Lucey, A. Ghasemi, P. Carre, S. Sridharan, and I.Matthews "Did you Engage with the Movie: Automatically Analysing and Summarisng Audience Behaviour" *In IEEE Transaction on Affective Computing*, 2013 (to be submitted).

(ii) **R. Navarathna**, D. Dean, S. Sridharan, and P.Lucey "Multiple Cameras for Audio-Visual Speech Recognition in an Automotive Environment" *In Computer Speech and Language (CSL)*, 2012.

(iii) S.Lucey, **R. Navarathna**, A. Ashraf, and S. Sridharan, "Fourier Lucas-Kanade Algorithm" *In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

1.6.2 International Conference Publications

(i) **R. Navarathna**, P. Lucey, A. Ghasemi, P. Carre, S. Sridharan, and I.Matthews "Automatically Analysing Audience Behavious in Movies" *In Proceedings of International Conference in Computer Vision (ICCV)*, 2013 (submitted).

(ii) S.Kalantari, **R. Navarathna**, D. Dean, and S. Sridharan "Visual Front-End Wars: VJ Face Detector vs FLK" *In Proceedings of 11th International conference on Auditory-Visual Speech Processing (AVSP)*, 2013.

(iii) **R.Navarathna**, D.Dean, S.Sridharan, C.Fookes, and P.Lucey, "Visual Voice Activity Detection using Frontal vs Profile" *In Proceedings of International Conference on Digital*

Image Computing: Techniques and Applications (DICTA), pp 134-139, Australia 2011.

(iv) **R.Navarathna**, S.Lucey and S.Sridharan "Fourier Active Appearance Models" *In Proceedings of International Conference in Computer Vision (ICCV)*, pp 1919-1926, 2011.

(v) **R.Navarathna**, T.Kleinschmidt, D.Dean, S.Sridharan and P.Lucey "Can Audio-Visual Speech Recognition outperform Acoustically Enhanced Speech Recognition in Automotive Environment" *In Proceedings of 12th Interspeech Conference*, pp 2241-2244, Italy 2011.

(vi) **R.Navarathna**, D.Dean, P.Lucey, S.Sridharan and C.Fookes "Recognising Audio-Visual Speech in Vehicles using the AVICAR Database" *In Proceedings of 13th Australasian International Conference on Speech Science and Technology (SST)*, pp 110-113, Melboune, Australia 2010.

(vii) **R.Navarathna**, D.Dean, P.Lucey and S.Sridharan "Cascading Appearance-Based Features for Visual Voice Activity Detection" *In Proceedings of 9th International conference on Auditory-Visual Speech Processing (AVSP)*,pp 3-7, Japan 2010.

(viii) **R.Navarathna**, P.Lucey, D.Dean, C.Fookes and S.Sridharan "Lip Detection for Audio-Visual Speech Recognition in-Car Environment" *In Proceedings of 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp 598-601, Malaysia 2010

Chapter 2

Understanding Human Behavior via Faces and Bodies

2.1 History of Recognising Human Behavior

The understanding and recognition of human behaviour is a very broad research field. It spans multiple research specialties (i.e. speech, vision, behavioral science) and is concentrated in sub-fields (e.g. vision - activity recognition, facial expression recognition etc.).

The work by Charles Robert Darwin (1809 1882) concerning genetically determined aspects of behavior can be identified as one of earliest works in understanding human behavior [52]. In the book *The Expression of the Emotions in Man and Animals* he seeks to trace the animal origins of human characteristics. According to Austrian neurologist Sigmund Freud (1856-1939), human beings are just mechanical creatures, whom he views as prisoners of primitive instincts and powers. He states that our purpose is to control/understand these behaviors [1]. Understanding human behavior involves analysis and recognition of motion patterns and well as the production of high-level description of actions and interactions between or among objects [184]. According to the recent Video and Image Retrieval Analysis Tool (VIRAT) project human behavior can be classified into two categories (i) *events* and (ii) *activities* [29]. An event refers to a single low-level spatio temporal entity that cannot be further decomposed such as speaking and walking. On the other hand, an activity refers to a composition of multiple events such as a person loitering [29].

According to the survey conducted by Pantic et al. [140] the scientific and engineering challenges related to the realization of machine sensing and understanding of human behaviors can be described as follows [140]:

- **Which types of messages are communicated by behavioral signal?** The term behavioral signal is usually used to describe a set of temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (a blink) to minutes (talking) or hours (sitting). This question is related to psychological issues pertaining to the nature of behavioral signals and the best way to interpret them.

- **Which human communicative cues convey information about a certain type of behavioral signal?** This issue shapes the choice of different modalities to be included into an automatic analyser of human behavioral signals.

- **How are various kinds of evidence to be combined to make inferences about shown behavioral signals?** This question is related to issues such as how to distinguish between different types of messages, how best to integrate information across modalities, and what to take into account in order to realize context-aware interpretations.

[1] `http://library.thinkquest.org/26618/`

All human interactive modalities such as audio, tactile and visual modality should include automated analys of human behavior and should analyse all verbal and nonverbal interactive signals (speech, body gestures, facial and vocal expressions, and physiological reactions) [140]. The recent advent of non-intrusive sensors and wearable computers can be seen as possibilities for including tactile modality into automatic analyzers of human behavior [144]. However, the visual model which carryies facial expressions and body gestures can be identified as the most important in the human judgment of behavioral cues [5]. Ambady et al. [5] found that humans seem to be most accurate in their judgment when they are able to observe the face and the body. They found that ratings based on the face and the body were 35% more accurate than the ratings based on the face alone.

The ability to recognise human behaviors by vision, is key for a machine to interact intelligently with a human. [75]. Over the past decades extensive research has been conducted by psycho-physicists, neuro-scientists and engineers on various aspects of human behaviors [34]. Psychological studies on visual analysis of body movement show that human movement differs from other movements [16]. Several studies from psychology have focused on the relationships between emotion and movement qualities [26, 28]. This has been the focus of many areas such as visual surveillance [51], event mining [135], event detection and recognition [134], and motion of the human body [154].

Simple approaches to recognise human behaviour are the use of templates [23, 147]. In these methods, human behaviors are characterised by patterns. Then, a set of features are extracted and matched with pre-defined patterns to recognise the person's behavior. These methods are computationally inexpensive and sensitive to the variance in different patterns of the same activity and to noise in the observations. Other techniques use deterministic models such as finite state machines [11] to recognise human behaviour. The advantage of these models is they do not account for uncertainty, which is essential for modeling complex behaviors. Due to this fact, probabilistic models in human behavior recognition are much more widely investigated [168, 200]. Bayesian network [143], the Markov network [32], the Dynamic Bayesian Network [58] and the Hidden Markov Model [155] are some of the probabilistic models used in behavior recognition.

Some works have been proposed to study the affect of human behavior by combining several modalities such as face and body. Ambady et al. [5] suggests that the combined face and body are the most informative for the analysis of human expressive behavior [5]. Most of these works use the rich information from face to recognise behavior, however, detection of face using vision based approaches is quite challenging and the most challenging/interesting problem in the vision community.

2.2 Importance of Face Sensing

The most important stage in the use of the visual modality for recognising human behavior is to reliably track and detect the persons ROI (i.e area around the face). The success of the entire system depends on the reliability of this stage, as the face contains a rich amount of behavioral information.

Given an arbitrary image, the goal of face detection is to determine whether there are any faces that are exist. Humans are very capable of recognising the faces and also recognising the behavior such as facial expression and lip-reading. With the rapid increase of computational powers and availability of modern sensing computers are becoming more intelligent. However this task is still more challenging to computers due to pose, change of illumination occlusions etc as shown in Figure 2.1. When computers can understand a face well in those conditions, then they begin to understand human behavior such as recognising facial expressions, lip reading in noisy environ-

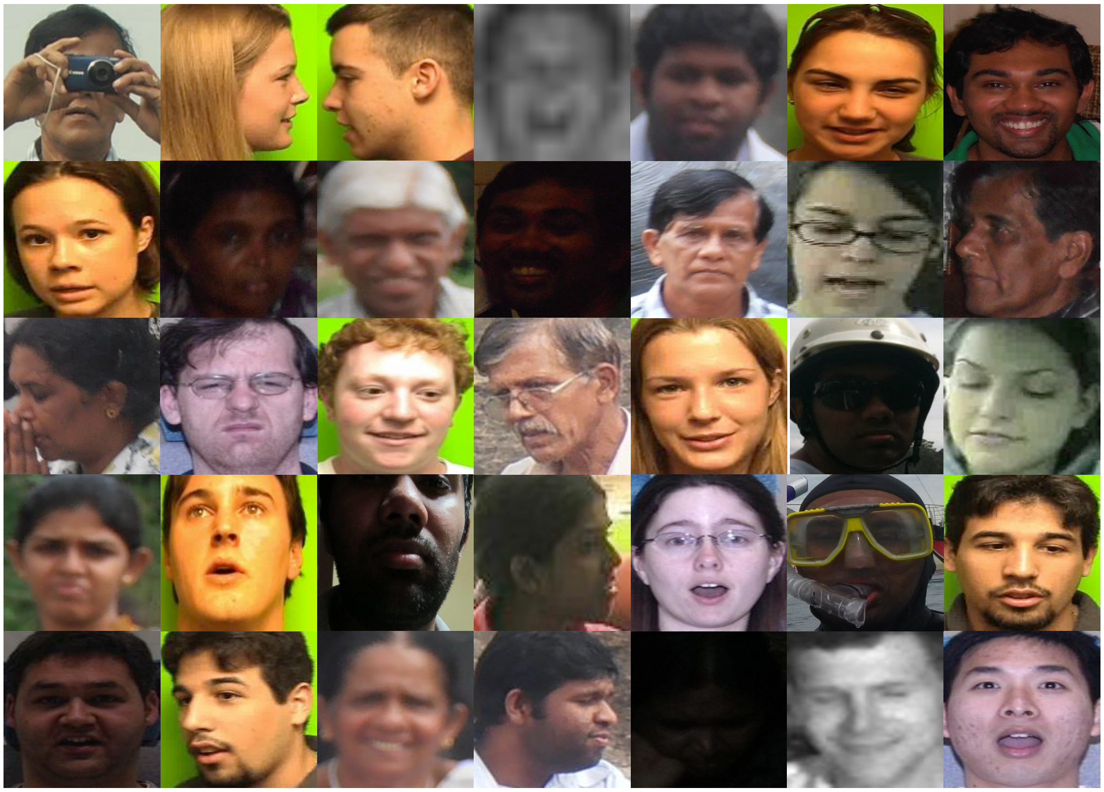


Figure 2.1: Examples of face images. Variations of pose, illumination, facial expressions, occlusion, image orientations and presence of structural components are some of the challengers for face detections.

ments well. If the face detection or face sensing stage is not accurate, it will have a detrimental effect on the remaining stages of the recognising pipeline (i.e feature extraction and classification).

2.2.1 Why Face Sensing is Difficult?

As mention in Chapter 2.2 to recognise and reliably track the face using computer vision has many challenges. Yang et al. [203] has identified the following challenges which associated with face sensing. According to Yang et al. [203] these factors can be identified as follows:

- **Pose:** The images of a face vary due to the relative camera-face pose such as frontal, different angle views, 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, and 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 a person's face can vary due to their facial expressions, such as smile, surprise, anger, disgust, fear, sadness and contempt.
- **Occlusion:** Faces may be partially occluded by other objects (ex: own hands, newspapers etc). 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 varying lighting conditions or poor lightings (i.e very dark illumination) and camera characteristics affect the appearance of a face.

The factors listed above show that the task of face sensing through a computer is quite a complex task. Most of the work has focused on data that has been collected in ideal laboratory conditions. However, as mentioned in Chapter 1, this thesis is focused on overcoming the problems associated with illumination variations.

2.3 Facial Expression Recognition

A facial expression or emotional expression of the brain can be identified as motions of the muscles of the face. These motions express the emotional state of a person.

These expressions can be identified as conveying social information between humans in an action of nonverbal communication. The book *Mcanisme de la Physionomie Humaine* from Guillaume Duchenne [60] in 1862 illustrates which muscles are used for which expression. Thirteen primary emotions were identified by Guillaume, all of which are controlled by one or two muscles. Humans tighten the muscles around eyes and expose teeth for expressions of anger and there is also a major effect around the mouth and chin area for laughter/smiles. Charles Darwin (1809 - 1882) illustrated that emotions illuminate not only the expressions of humans, but also of animals. This was highlighted in the book *The Expression of the Emotions in Man and Animals* [52].

One of the major works in literature is the work conducted by psychologists Ekman and Friesen [66]. This work has a large influence on the development of automatic facial expression recognisers. From this work, Ekman and Friesen [66] identified a total of six universal facial expressions known as: (i) happiness, (ii) disgust, (iii) anger, (iv) fear, (v) sadness, and (vi) contempt (recently added). Following this initial motivation, the *Facial Action Coding System* (FACS) [67] was developed by Ekman and Friesen in 1977.

Facial expressions have been studied by social psychologists, medical practitioners and artists. With the advance of computer vision technology, computer scientists started showing interest in the study of facial expressions. Earlier work on automatic facial expression recognition was the work conducted by 1978 by Suwa et al [182] in which they attempted to recognise facial expression using a sequence of images. Since the 1990s, research on automatic facial expression recognition has become a very broad research field [68, 141]. A facial expression detection system consists of face detection, feature extraction and classification. A plethora of work has been conducted and a survey of recent work in automatically measuring a person's behavior using vision-based approaches can be found in [207]. Much of this work has centered on recognizing an individual's facial expression, with notable progress made in the areas of smile detection in consumer electronics [197], pain detection [114, 115].

Emotional responses to multimedia content have been studied in the research community [83, 172, 188]. In 2011, Teixerira et al. [185] demonstrated that joy is one of the states and to analyse engagement with commercials through facial expression such as smiles would be a significant indicator in evaluating joy. Recognising emotions induced by video has also been studied in the affective computing community [86, 93, 105, 175, 183].

2.4 Visual Speech Recognition

Speech is multi-model. Humans can understand speech better if they can see the speaker [22]. It provides complimentary information about the place of articulation, such as tongue, teeth and lips. This is of particular benefit to those who have poor hearing because they would normally use this lip-reading information as the primary source of speech information. Humans use visual speech information to improve speech intelligibility from a very young age [7]. Dodd [59] noticed that toddlers at the age of 19 months actually perform lipreading. The majority of visual speech stems from the areas around the lips, even though visual speech is located throughout the human face [96]. The research conducted by McGrath et al. [124] showed that the human lips alone carry more than half the visual information compared with that of the face. Benoít et al. [20] illustrates that a combined lip/jaw model gives higher performance over a lip only model.

Once we have tracked the face and lip area we can use this information to derive features from the lip area. Even though a plethora of research has been conducted within the field of visual feature extraction for lipreading, it is still not clear which approach is best. Potamianos et al. in their review paper [151], highlight two very important points in the argument towards appearance based features. Firstly, their use is well motivated by human perception studies of visual speech as these contain information about the visible articulators (such as tongue, teeth, muscles around the jaw etc.), which are not contained just by the contours of the lips [20]. In the perception studies cited, perception of the mouth using the entire mouth ROI was far superior to just the lip movement [179]. Secondly, appearance based features can be computed very quickly, which lends itself to real-time implementation. This point is probably the most important in terms of deploying a real-world lipreading system. Another point also very important, is that the appearance based features are generic and can be applied to mouth ROIs of any viewpoint compared to the contour based approaches as specific contours have to be developed for the many views which may be a very cumbersome and exhaustive task.

The current state-of-the-art in visual feature extraction is that of multi-stage cascade of appearance features devised by Potamianos et al. [153]. For both frontal and non-frontal poses, the same process applies. Following ROI extraction, feature mean normalisation (FMN) is applied, which consists of removing the mean ROI across the utterance. This step removes a lot of speaker redundancies and non-speech related variabilities. A two-dimensional, separable, discrete cosine transform (DCT) is then applied on the resulting mean-removed ROI (This process is known as “static-feature” extraction stage).

Visual speech is represented by the movements of the visual articulators. The best features for representing visual speech are generally considered focus on the movement of the features, rather than the features within each frame. The simplest method to extract dynamic features is through the use of time-derivative-based delta and acceleration coefficients. These coefficients are used in addition to the original static features [152] which result in a higher feature vector. Recently one technique which has shown good performance is the use of linear discriminant analysis (LDA) to extract the relevant dynamic speech features from the ROI. In order to incorporate dynamic speech information the static features from the static feature extraction stage are concentrated before speech-class based LDA is performed based on a known transcription. The linear transforms such as DCT and LDA assume that the information contained in the image is high dimensional and try to preserve as much information as possible whilst gaining a low dimensional representation. Once the features are extracted hidden Markov models (HMMs) [205] can be applied to understand visual speech.

2.5 Activity Recognition

The goal of activity recognition is to recognise ongoing activities - one or a group of agents - from a series of observations. Using computer vision to recognise activities has been active in research areas since 1980s. The ability to recognise human activities from videos enables many real-world applications such as surveillance systems in airports and subway stations.

Most of the activity recognition work is focused on analysing patterns of motion. Aggarwal et al. [4] categorise human activities into four different categories: (i) gestures, (ii) actions, (iii) interactions and (iv) group activities. They have defined gestures as elementary movements of a person's body part such as *stretching an arm* and *raising a leg*. On the other hand actions were categorised as single agent activities that may consist of multiple gestures such as *walking* and *punching*. Thirdly, interactions were defined as human activities that involve two or more persons and/or objects such as *hand-shaking* and *two persons fighting*. Finally, they defined group activities as the activities performed by a group of multiple agents, such as a *group having a meeting*.

In the late 1990's, Cutler and Davis [48] presented a system which can detect and classify periodic motions. Bobick and Davis [24] presented a system to capture not only motion but also the shape. Global template [64] and bag-of-words models [170] have been shown to be effective in activity recognition tasks. One of the well-know techniques, in order to analyse the relation of the motion field, is optical flow. In 2009, Wang et al. [196] presented a system to recognize single person activities such as run, jump, wave in a surveillance system with optical flow features by representing a compact motion representation. Chaudhry et al. [33] proposed a method with a histogram of oriented optical flow (HOOF) which is independent from the the direction of motion to identify human actions. Recently, in Mahbub et al. [119] optical flow was employed to present the direction of motion and random sample consensus (RANSAC) which is used to further localization. They extend this work to recognize human actions using frequency domain features [118].

2.6 Chapter Summary

Monitoring the behaviors of single and multiple people using vision-based approaches is an interesting task. Understanding and recognition of human behavior spans multiple research specialties (i.e. speech, vision, behavioral science) and is concentrated in sub-fields (e.g. vision - activity recognition, facial expression recognition etc.). Initially the Chapter introduces a holistic view of understanding human behavior. This chapter presented the holistic view associated with recognising human behavior with visual information. The major works associated with this field are listed. The review then gives insights on transition process of human behavior to a computer-vision perspective.

The major component of this pipeline is face, as it contains rich amounts of behavioral data and this chapter highlights the importance of face and technical challenges of face detection. The later part of this chapter presents a brief history of three specific tasks of human recognising using visual information specially focusing on face and body. The three tasks are facial expression recognition, visual speech recognition and activity recognition.

Ekman and Friesen [66] identified a total of six universal facial expressions known as: (i) happiness , (ii) disgust, (iii) anger, (iv) fear, (v) sadness, and (vi) contempt (recently added). Following this initial motivation FACS was developed by Ekman and Friesen in 1977. Since then Some of the subsequent major works in facial expression recognition are highlighted in this Chapter. In addition to understanding facial expression humans can understand speech better if they can see the speaker. It provides complimentary information about the place of articulation, such as tongue, teeth and lips. Humans use visual speech information to improve speech intelligibility from very young age [7]. Understanding speech using vision-based approaches is also highlighted in this Chapter. Finally, the Chapter concludes with a brief review of activity recognition. According to Aggarwal et al. [4] human activities are divided into four different groups: (i) gestures, (ii) actions, (iii) interactions and (iv) group activities. They have defined gestures as elementary movements of a person's body parts such as *stretching an arm* or *raising a leg*. On the other hand, actions were catergorised as single agent activities that may consist of multiple gestures, such as *walking* or *punching*. Thirdly, interactions were defined as human activities that involve two or more persons and/or objects, such as *hand-shaking* or *two persons fighting*. Finally, they defined group activities as the activities performed by group of multiple agents, such as *a group having a meeting*. These three major activities of understanding human behavior using face and body are presented in this Chapter.

Chapter 3

Face Alignment: Background

3.1 Introduction

The human face contains a rich amount of information and it is a unique feature of human beings. This feature is one of the major reasons for widespread applications such as surveillance. Humans are highly capable of detecting a face pattern by casual inspection of the scene. As an example, people accurately detect the face even when the face is occluded. Also humans are capable of detecting the faces in a variety of conditions such as bad lightings, different variation of poses and far distance. Human face detection mechanisms are very strong [164]. However this task is still more challenging to computers due to pose, change of illumination occlusions etc as shown in Figure 2.1. The last few decades has seen a plethora of work conducted in face detection using machine learning and computer vision techniques. Yang et al. [203] categorise locating a person's face and facial features from a single image or colour image into four broad categories. According to Yang et al. [203] these can be grouped as follows:

1. **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.

2. **Feature invariant approaches:** The aim of these algorithms is to find structural features that exist even when the viewpoint, illumination or pose of the person varies and then use these to locate faces. Such features include facial features, skin colour, texture, size, shape and edge information.

3. **Template matching methods:** In this approach, several standard patterns of a face are stored to describe the face as a whole or the facial feature separately. The correlations between input and the stored patterns are computed for detections. These methods have been used for both face localization and detection.

4. **Appearance-based methods:** In contrast to template matching, the models (or templates) are learnt from a set of training images which should capture the representative variability of facial appearance. These learned models are then used for detection.

Knowledge-based methods are based on predefined set of rules. For example, a face within an image has two eyes, which are symmetrical to each other, a mouth and a nose. The relationship between those features can be represented by their distance. Earlier work on this approach can be seen in research conducted by Yang et al. [202]. They defined three levels of rule. At the top level, all possible face regions are found by scanning a window over the given image and applying set of rules at each location. This information is more likely to be *What is a face looks like*. In the middle level, local histogram equalization is performed on the face regions received from the top level followed by edge detection. The bottom levels rely on details of facial features. Finally, the successful face regions from level 2, are examined through set of rules that respond to facial features.

An abundance of research has been conducted using facial features, texture information, edge information and skin colour. Using a Canny detector [30], Sirohey proposed a method to extract a face from a cluttered background [174]. In 1995, Graf et al. [77] proposed a method to locate facial features and faces in gray-scale images. Leung et al. [99] introduced a method based on local feature detectors and random graph matching. The goal behind this work was to find certain facial features that were (i.e two eyes, nose/lip) most likely to be a pattern of a face.

In the template matching method, for a given input image the correlation values with standard patterns are computed for facial features. These correlation values are used to determine whether there is a face existing, or not. These methods can be divided in two sub groups: (a) Predefined templates and (b) Deformable Templates. Sakai et al. [163] attempted to detect faces using predefined templates. They used several sub templates for eyes, nose, mouth and face contours to model a face. Using a Sobel filter, Craw et al. [46] presented a face detection method based on a shape template of a frontal-view. The generated edges from the Sobel filter is grouped together to search for the template of a face. Samal et al. [165] presented a method using face silhouettes. A set of these features were obtained using principal component analysis (PCA) on faces. Silhouette is represented by an array of bits, then these eigen-silhouettes are used with a Hough transform for localisation. Miao et al. [126] introduced a method for face detection using a hierarchical template matching. These methods work well for single face images compared with multiple face images.

Using deformable templates, Yuille et al. [206] first applied a template matching method for mouth and eye localisation using appearance and shape models. In this approach, deformable template of the eyes and labial contour is fitted to an intensity model, by calculating a cost function based on the grayscale intensity edges around the template boundary. Unfortunately, this approach has poor performance due to the heuristic nature of the shape models and intensity models when applied across a large number of subjects. Cootes et al.[43] introduced a similar method for building a deformable template incorporating texture and shape, known as an active shape model (ASM). The ASM was able to statistically learn allowable variations in shape of an object pre-labeled object shapes in a point distribution model (PDM) [41, 43]. Edwards et al. [63] extended this approach using Kalman filters to estimate the shape-free intensity parameters. A major improvement in deformable face modelling are the Active Appearance Models (AAMs) proposed by Cootes et al. [41]. These models have shown themselves to be an accurate method of aligning a predefined shape model that also has linear appearance, to a previously unseen source image that is the object of interest. Although deformable template approaches have been shown to be useful for face/eye detection they are highly sensitive to initialisation and do not guarantee convergence.

In appearance-based methods, the models are learnt from a set of training images which should capture the representative variability of facial appearance. Generally these methods use statistical analysis and machine learning to find the relevant features of face and non-face. Eigenface [190], distribution method [181], neural network [160], SVM [103], Naive Bayes classifier [169], hidden Markov model [131] and information theoretical approach [40, 100] are some of appearance based methods which can be found in literature. Recently Viola-Jones [194] introduced a rapid object detection algorithm. The main principle of the algorithm, which was based on a boosted cascade of simple classifiers, is to scan sub windows within an image to detect objects of interest (ex:faces). It provides a quick and accurate framework, which can be used in real-time object detection applications.

The importance of good alignment is heightened, as any misalignment will greatly affect detection of human behavior such as facial expressions (ex: smiles, anger, happy, disgust) and visual speech recognition (Two real-world applications are highly detailed in Chapter 5 & 6). This Chapter provides a review of existing face alignment technologies relevant to the work in this thesis. A complete review of face alignment

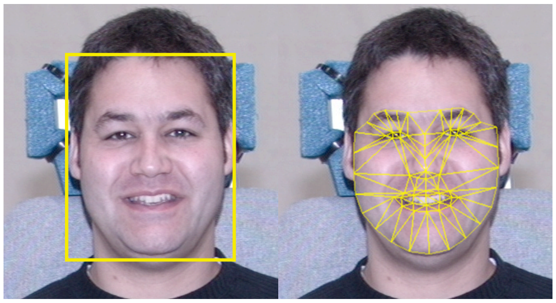


Figure 3.1: An example of face alignment methods to obtain rich amount of behavioral information in faces. Clockwise from left (a) Coarse-type of registration and (b) Fine registration

technologies is beyond the scope of this work.

The Chapter mainly centered around two main methods of face alignment (Refer to Figure 3.1):

(a) Coarse type of alignment (bounding box)

(b) Deformable face modelling

Chapter 3.2 presents an overview of coarse type of face detection methods. Specifically, Chapter 3.2 presents three publicly available off-the-shelf face detectors namely:

(a) Viola-Jones face detector [194]: which comes with the OpenCV libraries [1]

(b) The Apple Face detector which is in built with apple frameworks [2]

(c) The Fraunhofer Face Engine. [3]

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

[2] `http://developer.apple.com/library`

[3] `http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore/`

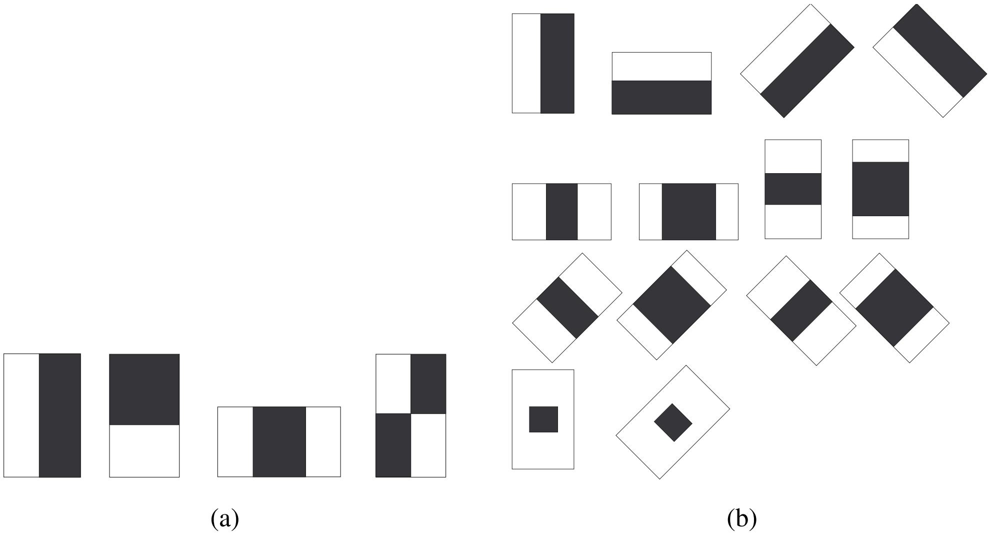


Figure 3.2: Comparison of the haar-like feature sets used by: (a) Viola and Jones with the original four features; and (b) Lienhart and Maydt with their extended set of features including their rotated features with an angle of 45°.

Chapter 3.3 briefly highlights a well-known deformable face modelling technique AAMs. The central motivation of this Chapter is to review, how well these methods perform in noisy lighting conditions. The Chapter concludes with extensive comparison of these methods.

3.2 Coarse-type of Alignment

Coarse type of alignment, (bounding box) detects the face and facial features such as eyes and mouth region coarsely. The Chapter reviews the above three different publicly available face detection methods in controlled, fluctuating and low-light illumination conditions.

3.2.1 The Viola-Jones Algorithm

The Viola-Jones algorithm is a rapid object detection algorithm proposed by Viola and Jones in 2001 [194]. The main principle of the algorithm, which is based on a boosted cascade of simple classifiers is to scan sub windows within an image to detect objects of interest (ex:faces). It provides a quick and accurate framework, which can be used in real-time object detection applications. The main steps of the Viola-Jones algorithm are described briefly in the following sub sections.

Feature representation

The Viola-Jones algorithm uses a “Haar-like” feature representation of the images instead of pixels. The original features were extended to fourteen features by Lienhart and Maydt [104] by introducing a new set-up of features which are rotated by 45°. The original features and the extended features are shown in Figure 3.2.

Each haar-like’ feature consist of black and white rectangles. The value of the haar-like feature is the subtraction of the sum of the pixel values in white rectangles from the sum of the pixel values in black rectangles. An object of size 16×16 pixels with an image, can have over 100,000 haar-like features. These features can be computed rapidly using the concept of integral image [194]. However, calculating over 100,000 features is a time consuming process. The Viola-Jones algorithm overcomes this problem by selecting the features using the “AdaBoost” algorithm [70].

AdaBoost algorithm

The Viola-Jones algorithm uses an efficient and effective learning algorithm called the AdaBoost algorithm. The main concept of the algorithm is to produce a strong classifier which has a high detection performance by linearly combining weak classifiers. The steps of the algorithm are:

1. Given n example images $(x_1, y_1), \ldots, (x_n, y_n)$ where x is the sub-window of the entire image and $\mathbf{y}_i = 0, 1$ for negative and positive examples respectively.
2. Initialize weights $\mathbf{w}(1, i) = \frac{1}{2m}, \frac{1}{2l}$, for $\mathbf{y}_i = 0, 1$ respectively, where m is the number of negative examples and l is the number of positive examples.
3. For $\mathbf{t} = 1, \ldots, T$
   - Normalize the weights at each stage, so that w_t is a probability function

     $$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \tag{3.1}$$

   - For each feature j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to w_t,

     $$\epsilon_j = \sum_j w_i |h_j(x_i) - y_i| \tag{3.2}$$

   - Choose the classifier with h_t, with the lowest error ϵ_t
   - Update the weights :

     $$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i} \tag{3.3}$$

   where $e_t = 0$ if example x_i is classified correctly, $e_t = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$
4. The final classifier is :

   $$h(x) = \begin{cases} -1 & : \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & : otherwise \end{cases} \tag{3.4}$$

   where $\alpha_t = \log \frac{1}{\beta_t}$

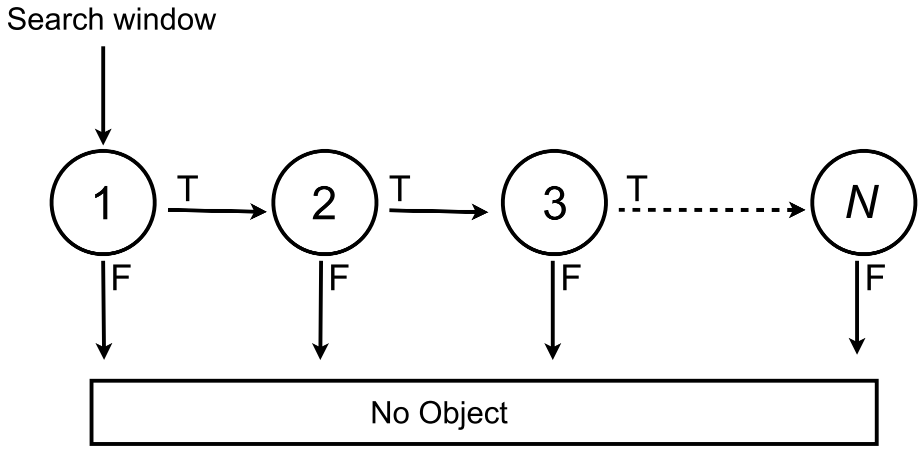


Figure 3.3: Block diagram of a N cascade of classifiers to detect the interest of object in a given search window.

The AdaBoost algorithm updates the weights of testing data in every stage. Mainly, it gives higher weight for the misclassified data and lower weight for correctly classified data in the next stage. The process is iterate for T times to generate the T weak classifiers. Finally, the generated classifiers are linearly combined to produce a strong classifier.

Cascading the classifications

Viola et al. [194] proposed the use of a cascade of weak classifiers instead of a single strong classifier to detect objects. The complexity of each stage increases in the cascade and dramatically increases the detection speed of the object. A key innovation in having a cascade of classifiers is the ability to reject the majority of sub windows that are not likely to contain the object at early stages rather than allowing them to go to complex stages as shown in Figure 5.15.

3.2.2 The Apple Face Detection Framework

With the face detection Application Programming Interface (API) included within the Core Image library in the iPhoneOS (IOS) 5.0 and Lion 10.7, face detection along with the locations of the eyes and mouth is straightforward. This API was mainly built to detect faces with the IOS. Even though the implementation of this face detector is not publicly available, it is easy to use with two adjustable levels of accuracy (i.e high/low accuracy of detection). The face detection capability is available in IOS 5.0 and later (with Mac operating system).

3.2.3 The Fraunhofer Face Detector

The "*Fraunhofer*" face engine [161] is another publicly available API which can be used to detect faces and the position of the eyes, nose and corners of the mouth. The developed algorithms are embedded in a library called *Sophisticated High-speed Object Recognition Engine* (SHORE). The "*Fraunhofer*" face engine can be used for either simple face detection or for more complex tasks such as gender classification, facial expression recognition, classification of eyes that are opened and closed and the analysis of facial features using different set-ups. This commercial face engine provides a black-box framework which can be configured easily. However the details of the algorithms are not publicly available and are hidden by the framework interfaces.

3.3 Fine Registration: Active Appearance Models

In order to obtain a fine registration, deformable model approach where a dense of 60-70 points on the face is used. This method is ideal in situations where there is a lot-of head movement, especially out-of-plane rotations. AAMs [41, 120] have been widely

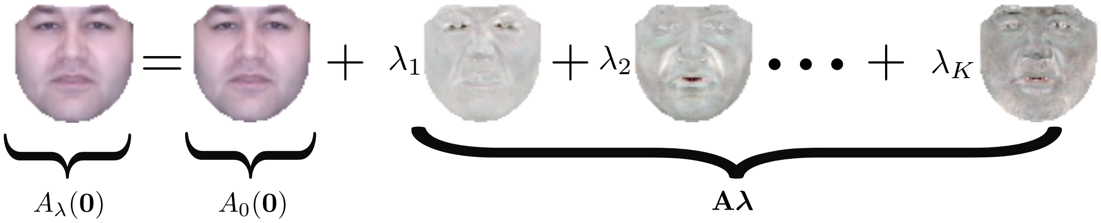


Figure 3.4: Construction of the appearance $A_{\boldsymbol{\lambda}}(\mathbf{0})$ of an AAM. This can be represent using mean appearance $A_0(\mathbf{0})$ plus a linear combination of K orthonormal appearance vectors $\mathbf{A}\boldsymbol{\lambda}$.

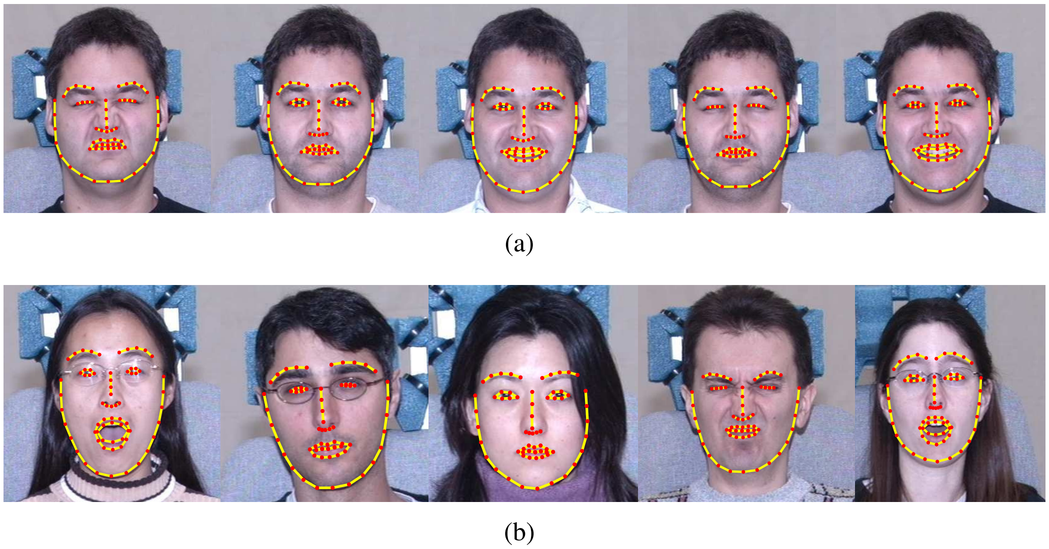
used and succeed the fine registration method in facial expression recognition [9, 10, 114, 116] therefore this Chapter focused AAMs. This approach needs manual labelling of the training sequence (up to 5%-10% of key frames).

The AAMs [41, 120] employ a paradigm of inverting a synthesis model (or in machine learning terms, a generative model) of how an object can vary in terms of shape and appearance. As a result, the ability of AAMs to register an unseen object image is intrinsically linked to how well the synthesis model can reconstruct the object image.

The AAMs are usually constructed from a set of training images with the AAM mesh vertices hand-labelled on them [42]. The training mesh vertices are first aligned with procrustes analysis. Then principal component analysis (PCA) is used to build a 2D linear model of shape variation [42]. The shape $\mathbf{s}$ of an AAM is described by a 2D triangulated mesh. The 2D shape $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$ can be represented as a base shape $\mathbf{s}_0$ plus a linear combination of P shape vectors $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{P} p_i \mathbf{s}_i \tag{3.5}$$

where $\mathbf{p} = [p_1, \ldots, p_P]^T$ is the shape parameter vector. The AAM model of appearance variation is obtained by first warping all the training images onto the mean shape and then applying PCA on the shape normalized appearance images. The appearance of an AAM $A(\mathbf{0})$ is an image vector defined over the pixels $\mathbf{x} \in \mathbf{s}_0$ inside the base mesh $\mathbf{s}_0$ when $\mathbf{p} = \mathbf{0}$. The appearance $A_{\boldsymbol{\lambda}}(\mathbf{0})$ can be represented as a mean appear-



(a)

(b)

Figure 3.5: Comparision of construction of an AAM using hand annotated (68 points) ground-truth images. (a) Person specific AAMs are construct using 5%-10% key frames using a single subject with different expression in a given video sequence and(b) Generic AAMs are construct across many subjects with many expressions.

ance $A_0(\mathbf{0})$ plus a linear combination of K orthonormal appearance vectors $A_j(\mathbf{0})$:

$$\begin{aligned} A_{\boldsymbol{\lambda}}(\mathbf{0}) &= A_0(\mathbf{0}) + \sum_{j=1}^{K} \lambda_j A_j(\mathbf{0}) \\ &= A_0(\mathbf{0}) + \mathbf{A}\boldsymbol{\lambda} \end{aligned} \tag{3.6}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^T$ is the appearance parameter vector and $\mathbf{A} = [A_1(\mathbf{0}), \ldots, A_K(\mathbf{0})]$ is the matrix of concatenated appearance vectors. A visual example is given in Figure 3.4

3.3.1 Person Specific vs Generic AAMs

As mentioned above, AAMs are usually constructed from a set of training images with the AAM mesh vertices hand-labelled on them. The AAMs can be constructed in two different ways. The most successful method is to construct an AAM using a single subject to model the variation in the appearance pose, illumination and expression as shown in Figure 3.5(a). Such a person specific AAM might be useful for facial expression recognition in consumer electronics [197] and pain detection [114, 115]. However, this approach is not practical, where manually labelling of frames is prohibitive. Alternatively, one can overcome this problem by constructing an AAM model across many subjects, expressions as shown in Figure 3.5(b). However, the generic AAMs are very poor in registration performance compared with person specific AAMs, [79] which limit the use of generic AAMs for real-world applications.

3.3.2 AAM fitting

A number of approaches have been proposed in literature for fitting AAMs [42, 120]. The most notable and popular approach is to minimize the sum of squared distances (SSD) between the input image $\mathbf{I}(\mathbf{p})$ and AAM model $\mathbf{A}(\boldsymbol{\lambda})$ [120]. In this approach one can pose AAM fitting as minimizing the following objective function:

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \parallel I(\mathbf{p}) - A_{\boldsymbol{\lambda}}(\mathbf{0}) \parallel^2_{\mathbf{Q}} \tag{3.7}$$

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \parallel I(\mathbf{p}) - A_0(\mathbf{0}) - \mathbf{A}\boldsymbol{\lambda} \parallel^2_{\mathbf{Q}} \tag{3.8}$$

where $I(\mathbf{p})$ represents the warped input image using the warp specified by the parameters $\mathbf{p}$.

The central task of the objective function described in Equation 3.8 is to find the shape $\mathbf{p}$ and appearance $\boldsymbol{\lambda}$ that minimizes the weighted SSD between the warped input image and the AAM. For most AAM fitting problems the weight matrix $\mathbf{Q}$ is assumed to be an identity matrix $\mathbf{I}$ (i.e. unweighted SSD).

Generally, the objective function in Equation 3.8 is difficult to solve as there is a non-linear relationship between the shape $\mathbf{p}$, and appearance $\boldsymbol{\lambda}$ parameters. A key insight, stemming from Lucas & Kanade (LK) [110], was that a linear approximation can be made between $\mathbf{p}$ and $\boldsymbol{\lambda}$ through the judicious use of image gradients and the chain rule to form steepest descent matrices (i.e. $\frac{\partial A(\mathbf{p})}{\partial \mathbf{p}}$). The LK algorithm is briefly described in Chapter 4.2. This section briefly reviews two common approaches in AAM fitting.

(a) Simultaneous algorithm

(b) Project-out algorithm

3.3.3 Simultaneous Algorithm

The simultaneous algorithm [120] linearizes the objective function in Equation 3.8 such that:

$$\arg \min_{\Delta \mathbf{p}, \Delta \boldsymbol{\lambda}} \parallel I(\mathbf{p}) - A_{\boldsymbol{\lambda}}(\mathbf{0}) - \frac{\partial A_{\boldsymbol{\lambda}}(\mathbf{0})}{\partial \mathbf{p}} \Delta \mathbf{p} - \mathbf{A} \Delta \boldsymbol{\lambda} \parallel_{\mathbf{Q}}^2 \quad . \tag{3.9}$$

Instead of solving for the shape $\mathbf{p}$ and appearance $\boldsymbol{\lambda}$ parameters directly, through the linearization step in 3.9 we iteratively solve for the updates $\Delta \mathbf{p}$ and $\Delta \boldsymbol{\lambda}$. The objective function in Equation 3.9 takes advantage of a computationally efficient extension to the LK algorithm referred to as the inverse compositional (IC) algorithm [120]. The IC algorithm linearizes the template image $A_{\boldsymbol{\lambda}}(\Delta \mathbf{p})$, with respect to $\Delta \mathbf{p}$, instead of the source image $I(\mathbf{p} + \Delta \mathbf{p})$. The rationale for this switch shall be examined more closely in the next section concerning the project-out algorithm.

A consequence for this switch is that the update to the the current warp parameters are updated by the inverse (as we want to update the source image not the template) of the warp update $\mathbf{p} \leftarrow \mathbf{p} \circledcirc \Delta \mathbf{p}^{-1}$. The operation $\circledcirc$ represents the composition of two warps (e.g. for an affine warp this is represented as a matrix multiplication). The update to the appearance parameters, however, remain additive such that $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}$. The explicit solution to $\Delta\mathbf{p}$ and $\Delta\boldsymbol{\lambda}$ can be found "simultaneously" such that:

$$\begin{bmatrix} \Delta\mathbf{p} \\ \Delta\boldsymbol{\lambda} \end{bmatrix} = \mathbf{H}_{sim}^{-1}\mathbf{J}_{sim}^{T}\mathbf{Q}[I(\mathbf{p}) - A_{\boldsymbol{\lambda}}(\mathbf{0})] \tag{3.10}$$

where the pseudo simultaneous Hessian matrix is defined as:

$$\mathbf{H}_{sim} = \mathbf{J}_{sim}^{T}\mathbf{Q}\mathbf{J}_{sim} \tag{3.11}$$

The simultaneous Jacobian matrix is defined as:

$$\mathbf{J}_{sim} = \begin{bmatrix} \frac{\partial A_{\boldsymbol{\lambda}}(\mathbf{0})}{\partial \mathbf{p}} \\ \mathbf{A}^{T} \end{bmatrix} \quad . \tag{3.12}$$

Empirically, the simultaneous algorithm has been noted to have excellent fitting performance compared other LK inspired methods to AAM fitting. A major problem, however, with the simultaneous algorithm occurs with respect to computational efficiency. Specifically, as a consequence of the update step $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}$ the appearance image $A_{\boldsymbol{\lambda}}(\mathbf{0})$, Jacobian matrix $\mathbf{J}_{sim}$, and Hessian matrix $\mathbf{H}_{sim}$ must be re-estimated at each iteration.

3.3.4 Project-out Algorithm

The project-out algorithm [120] circumvents the computational limitations of the simultaneous algorithm by attempting to "project-out" appearance variation:

$$\arg\min_{\Delta\mathbf{p}} \parallel I(\mathbf{p}) - A_{\mathbf{0}}(\mathbf{0}) - \frac{\partial A_{\mathbf{0}}(\mathbf{0})}{\partial \mathbf{p}}\Delta\mathbf{p} \parallel^{2}_{\mathbf{Q}_{\perp}} \quad . \tag{3.13}$$

where $\mathbf{Q}_{\perp}$ is the modified weight matrix where the appearance basis $\mathbf{A}$ has been projected out:

$$\mathbf{Q}_{\perp} = \mathbf{Q} - (\mathbf{A}\mathbf{A}^T)\mathbf{Q}(\mathbf{A}\mathbf{A}^T) \ . \tag{3.14}$$

Equation 3.13 is an approximation to the simultaneous algorithm in Equation 3.9 as it is no longer updating the appearance template $A_{\boldsymbol{\lambda}}$ (note $A_{\mathbf{0}}$ is instead used, which is the mean appearance template).

The computational advantages of inverse compositional inspired fitting are readily apparent for the project-out algorithm. Specifically, one can solve the objective function in Equation 3.13 such that:

$$\Delta\mathbf{p} = \mathbf{B}[I(\mathbf{p}) - A_0(\mathbf{0})] \tag{3.15}$$

where $\mathbf{B}$ can be completely pre-computed and the current warp parameters are iteratively updated by the inverse (as we want to update the source image not the template) of the warp update. The update matrix is defined as:

$$\mathbf{B} = \mathbf{H}_{po}^{-1}\mathbf{J}_{po}^T\mathbf{Q}_{\perp} \tag{3.16}$$

where the pseudo project-out Hessian matrix is $\mathbf{H}_{po} = \mathbf{J}_{po}^T\mathbf{Q}_{\perp}\mathbf{J}_{po}$. The project-out Jacobian matrix is defined as $\mathbf{J}_{po} = \frac{\partial A_{\mathbf{0}}(\mathbf{0})}{\partial \mathbf{p}}$. For the project-out algorithm the Jacobian and Hessian matrices remain static across all iterations, thus allowing $\mathbf{B}$ to remain static. To gain an insight into the speedup that is afforded by pre-computing $\mathbf{B}$, implementations of project-out AAM face fitting have been reported in literature [120] running at real-time (i.e. 30 fps) on a modern PC (Intel core dual 3.0 GHz).

3.4 Experiments: Off-the-shelf Face Detectors and AAMs

Face alignment experiments are conducted using (a) Off-the-shelf face detectors (Chapter 3.2) and (b) Person-specific AAMs (Chapter 3.3) in varying illumination conditions.

3.4.1 Off-the-shelf Face Detection Results

As the central theme of this thesis is to recognize human behavior in noisy environments, initially different coarse-type face detectors are evaluated in controlled, fluctuating and low-light illumination conditions.

In order to calculate the face detection performance, this work followed a similar approach described by Vidit et al. [192]. To represent the score of a match between a detection region d_j and groundtruth region g_i, we employ the commonly used ratio of intersected areas to joined areas as follows,

$$S(g_i, d_j) = \frac{area(g_i) \cap area(d_j)}{area(g_i) \cup area(d_j)} \tag{3.17}$$

and the score $S(g_i, d_j)$ is $0 < S(g_i, d_j) < 1$.

Initially, ground-truth coordinates were manually labelled for 1200 images in controlled, fluctuating and low-light illumination conditions. Figure 3.6 shows the true-positive rate for the three face detectors in controlled, fluctuating and low-light illumination conditions for a given false positive rate. Even though the off-the-shelf face detectors work reasonably well in controlled and fluctuating conditions, it is interesting

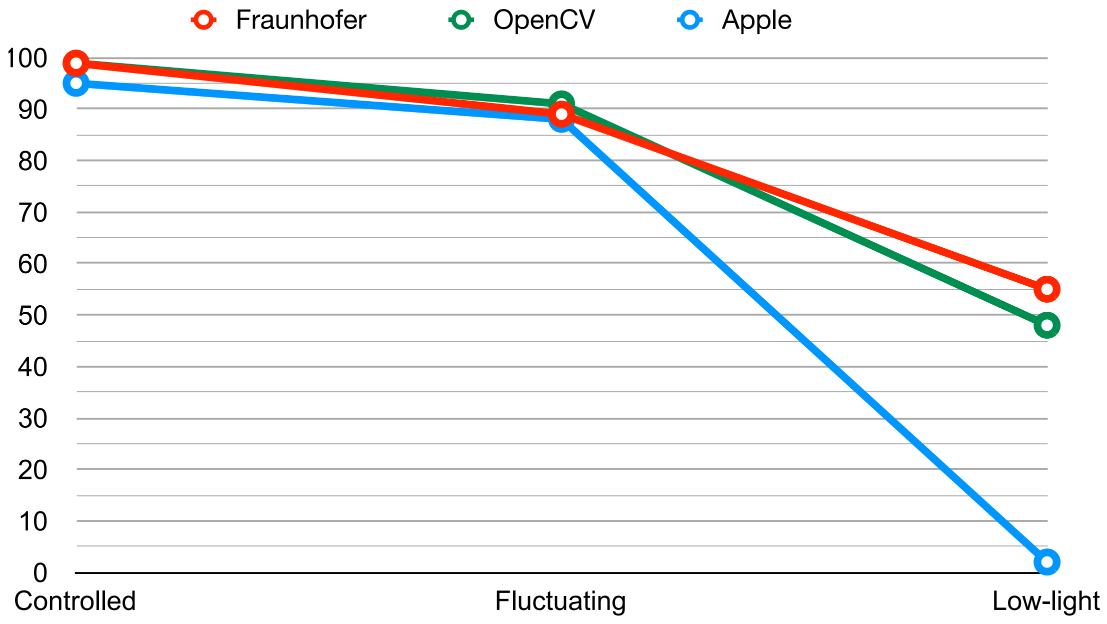


Figure 3.6: The Face detection performance using the off-the-shelf face detectors. All the face detectors work reasonably well in fluctuating lighting conditions, even though the hit-rate is less than controlled lighting conditions. In all the face detectors, the performance in low-light conditions (i.e an audience environment) dramatically drops. The performance of the Apple face detector is very poor in low light conditions .

to observe that the performance dramatically decreases in all the three face detectors in low-light conditions (i.e an audience environment) as shown in Figure 3.6. The performance of the Apple face detector in low-light condition is poorest (less than 2%).

3.4.2 AAM Face Registration Results

The experiments were conducted with the MultiPIE face image dataset [80]. The MultiPIE database consisted of 19 illumination conditions (i.e., 18 variations of flash firing and without flash) with a range of facial expression including *neutral, smiles, surprise, squints, disgust and screams*. Examples of this variation can be seen in Figure 3.7. All images were hand annotated with 68 points. More details about the MulitiPIE database can be found in [80].

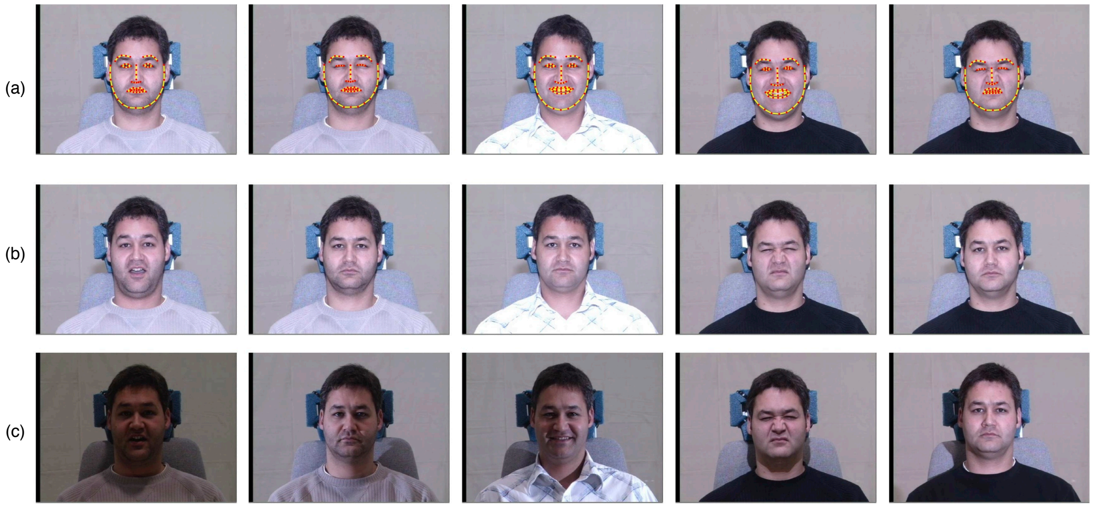
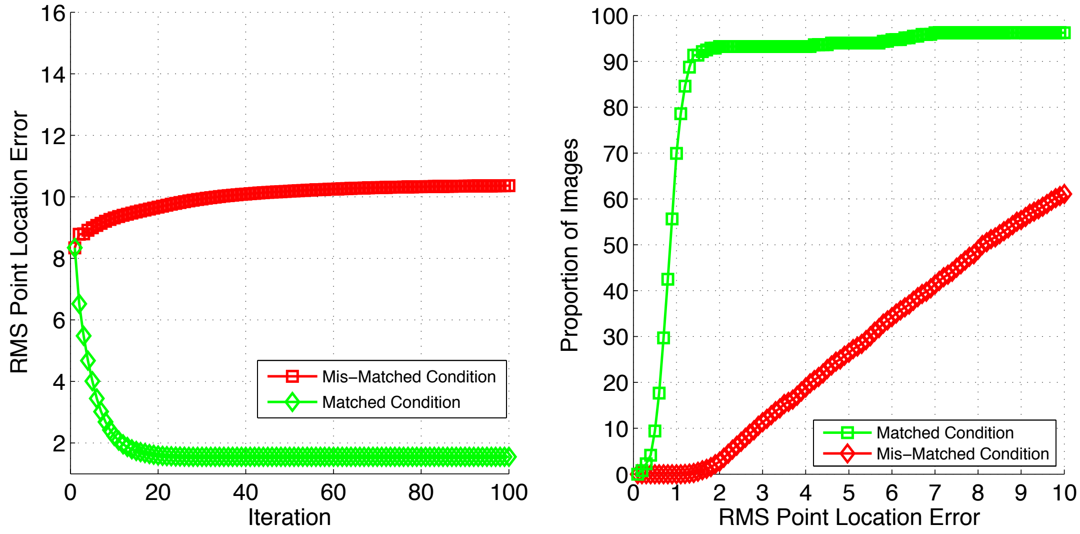


Figure 3.7: (a) Five of the hand annotated (68 points) ground-truth images for a specific subject used to construct the AAM model. (b) Sample of testing images with the matched illumination condition. (c) Sample of testing images with the mismatched illumination condition.

Person specific AAM fitting experiments were conducted as they have shown better fitting performance than Generic AAMS [79]. For all our experiments, two types of fitting performance were measured: (a) matched (i.e illumination in training and testing images are same) and (b) mismatched illumination (i.e illumination in training and testing images are different). We measured fitting performance in terms of root mean square error (RMS) between the 2D mesh location of the current fit results and the ground-truth 2D mesh coordinates with respect to the base mesh. Results were calculated for (a) and (b) when the initialized shape was randomly perturbed from ground-truth.

Simultaneous results

Figure 3.8(a) depicts the average RMS mesh location error against iterations for simultaneous variants of AAMs (a) matched and (b) mismatch illumination. Similarly, Figure 3.8(b) depicts the number of converged trials as a function of the RMS error



(a) Average convergence rates

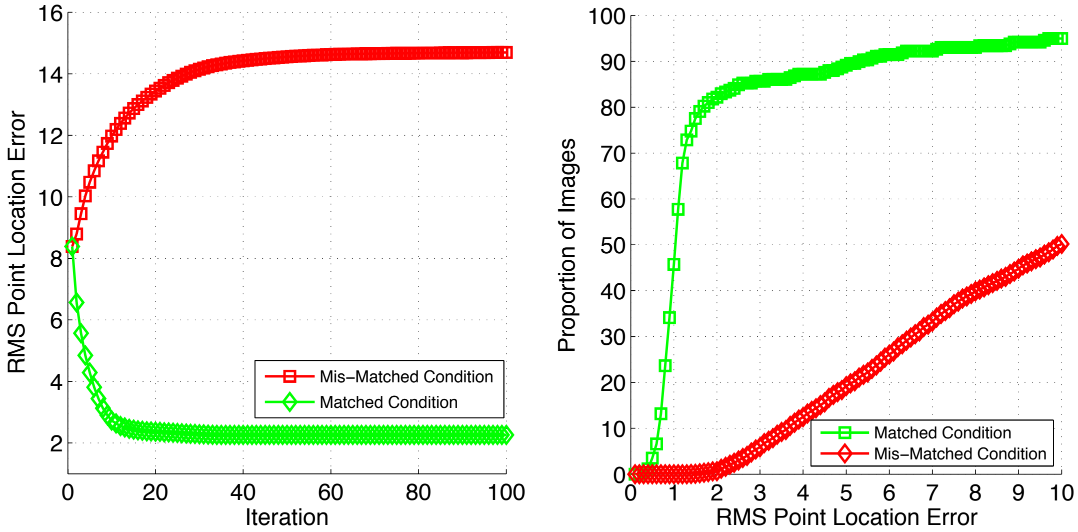
(b) Fitting performance

Figure 3.8: Performance comparison using simultaneous algorithm when the input and training images have the same illumination, and varying illumination condition. When the illumination change AAM diverge.

threshold for matched and mismatched illumination conditions. While the illumination matched person-specific AAM shows better performance, however in the presence of mismatched illumination AAM fitting is divergent, which results in poor fitting performance.

Project-out results

The performance in terms of convergence rate and fittings for project-out algorithm is shown in Figure 3.9. The observation was similar compared with the simultaneous algorithm. Even though AAMs have good registration performance in matched illumination, once the illumination changes the performance is degraded. Note that the fitting performance and convergence rate using the simultaneous algorithm is much better than the project-out algorithm as shown in Figure 3.8.



(a) Average convergence rates

(b) Fitting performance

Figure 3.9: Performance comparison using project-out algorithm when the input and training images have the same illumination, and varying illumination condition. When the illumination change AAM diverge.

3.5 Comparison: Coarse vs Fine Alignment

This section briefly describes the advantages and disadvantages of the described face alignment methods. Generally, coarse-type of alignment (i.e off-the-shelf face detection methods) is more robust to unseen objects compared with AAMs. These state-of-the art methods perform reasonably well in controlled or fluctuating lighting conditions, however performance dramatically decreases in low light conditions such as audience environment.

A major drawback to person specific AAMs stems from their tendency to generalize unseen objects. When unaccounted appearance variations are encountered due to a change in the environment (e.g., illumination or camera change), person specific AAMs perform poorly. In situations, where manually labelling of frames is prohibitive such as marketing, health-care off-the-shelf face alignment is a practical solution as they are more robust to unseen objects.

Systems which can detect unseen subjects with high accuracy are required for real-world applications such as facial expression recognition, pain recognition systems in health-care and visual speech recognition in automotive environments. The validation results in this Chapter show that, even though fine alignment methods such as person-independent AAMs are highly accurate, in practice, this type of dense alignment has so far been impossible to achieve in a generic sense. On the other hand, off-the-shelf face detectors are more robust to unseen subjects, which have the potential to be applied for real-world applications with less accuracy of face alignment.

3.6 Chapter Summary

This chapter has given insights into publicly available, coarse-type of face alignment methods (i.e *Viola-Jones, Apple face detector and "Fraunhofer" face detector*) and the well-known fine alignment method called AAMs. Three major off-the-shelf face detectors are detailed, which have influenced this field of research. In the second part of the chapter, widely used AAMs are described in great detail. The model construction of person-specific AAM and generic AAMs are briefly detailed and two common fitting algorithms (i.e Simultaneous and Project-out algorithm) are reviewed.

The Chapter experimentally validates the behaviour of coarse-type of alignments and AAMs fittings in different lighting conditions. The Chapter concludes with off-the-shelf face detection methods that are more robust to unseen objects compared with AAMs with less alignment accuracy.

A major drawback to person specific AAMs stems from their tendency to generalize unseen objects. When unaccounted appearance variations are encountered due to a change in the environment (e.g., illumination or camera change), person specific AAMs perform poorly. Motivated from this work, the next chapter describes a novel method to overcome the problem with illumination in AAMs.

Chapter 4

Face Alignment in Fourier Domain

4.1 Introduction

Generic non-rigid face fitting is still an ongoing topic in computer vision with notable theoretical inroads being made [47, 106, 166]. However, none of these approaches can provide the level of registration accuracy or computational efficiency achieveable through a person specific AAM [39, 79]. As a result, person specific AAMs are still the method of choice in a number of applications where users are willing to provide subject specific images and labels. Notable applications of person specific AAMs in literature can be found in areas such as expression classification, avatar synthesis, and visual speech synthesis [39].

Even though state-of-the art person specific AAM face fitting outperforms generic non-rigid face fitting methods, significant problems still remain. A major drawback to person specific AAMs stems from their capacity to only generalize to small amounts of appearance variation (essentially appearance variation that can be expressed as a linear combination of the training instances, e.g. expression variation). However, as

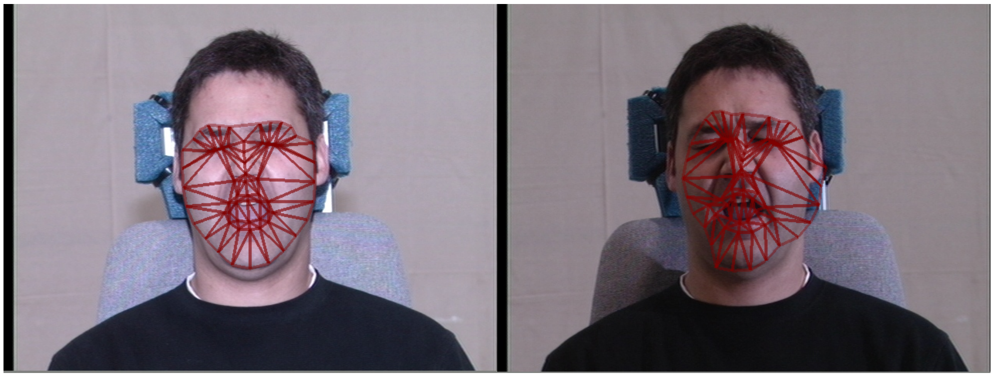


Figure 4.1: Example of person specific AAM fitting in the presence of: Clockwise from left AAM output in the same illumination condition as the training images, (b) output in a different illumination condition compared to training images. In this chapter a computationally efficient and accurate solution that provides environmental invariance to overcome the problem in (b) is introduced.

shown in Chapter 3, AAMs are very sensitive to illumination. When illumination conditions during the testing are significantly varied with illumination with the training images, AAMs tend to suffer as shown in Figure 4.1. This effect severely limits the usefulness of person specific AAMs, as one either needs to: (i) ensure the environment is strictly controlled, or (ii) collect and label training examples of the subject in the new environment.

To overcome this problem one can: a) manually label landmarks of the subject in the different illumination environments, or b) preprocess the image to encode illumination invariance through a bank of filters. However, the fundamental drawback of the first approach is the requirement of multiple examples of the face under changing illumination conditions as labelling ground-truth for every illumination condition can be impractical. Filter-banks have been shown to be useful in gaining invariance to spectral distortions such as those encountered in the presence of illumination variations on approximately Lambertian surfaces [95]. Unfortunately, in the spatial domain the complexity is a direct function of the number of filter banks being applied, greatly increasing the computational and memory requirements of image matching on raw pixels.

Recently, Ashraf et al. [8] proposed image alignment in Fourier domain using a bank of filters which are referred to as Fourier LK (FLK). The LK and FLK algorithms are briefly described in Chapter 4.2 and Chapter 4.4. Using this as our motivation this Chapter introduces a novel framework by employing the AAM in the Fourier domain with a bank of filters. We refer to this method as the Fourier AAM (FAAM) algorithm. The FAAM can handle substantial illumination variations, poor in standard AAM and it is much more efficient computationally with a bank of filters.

4.1.1 Related work

Gaining invariance to environmental variations such as camera and illumination variations has been a muchwell investigated topic in AAM fitting literature [44, 79, 187]. Notably, Gross et al. [79] modelled illumination variation by using an abundance of examples from different illumination conditions. As discussed earlier, this approach is unattractive in practice as one has to collect multiple images/labels of the subject from a wider variety of environmental conditions. In 2006 Theobald et al. [187] demonstrated the usefulness of robust-error functions for AAM fitting for dealing with previously unseen appearance variations. Although successful, this approach is problematic as it requires a re-computation of the Hessian for each iteration of fitting, irrespective of the approach employed (i.e., simultaneous and project-out). This problem is particularly limiting for the project-out algorithm as it dramatically slows down its real-time performance. The work described in the Chapter differs as it reformulates the problem with a constant matrix, which results in for a constant Hessian for each iteration of fitting.

Filter-based solutions have also been utilized in the past to gain environmental invariance in AAM fitting. Of particular note is the work of Cootes and Taylor [44] where the authors explored the use of multiple filter (specifically orientated gradients) responses for fitting. Although exhibiting impressive results, the approach is problematic as it requires the explicit computation of multiple image filter responses at each iteration of AAM fitting. The work present in the Chapter differs to the work presented in [44] by proposing a novel method for completely pre-computing the effect of multiple filter responses, such that the online portion of the AAM fitting algorithm operates solely and efficiently on raw pixels.

4.2 Lucas & Kanade Algorithm

The goal of the LK algorithm [111] is to find the parametric warp $\mathbf{p}$ that minimizes the sum of squared difference (SSD) between a template image T and a warped source image I. The error term can be written as,

$$\arg\min_{\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) \parallel^2 \tag{4.1}$$

where $I(\mathbf{p})$ represents the warped input image using the warp specified by the parameters $\mathbf{p}$, while $T(\mathbf{0})$ represents the un-warped template image.

The LK algorithm finds an effective solution to Equation 4.1, by iteratively solving for $\Delta\mathbf{p}$ and refining the parameters at each iteration till converge such that $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$.

The non-linear Equation 4.1 can be linearised by performing Taylor series expansion,

$$\arg\min_{\Delta\mathbf{p}} \parallel I(\mathbf{p}) + \mathbf{J}\Delta\mathbf{p} - T(\mathbf{0}) \parallel^2 \tag{4.2}$$

where the Jacobain matrix $\mathbf{J} = \left(\frac{\partial I(\mathbf{p})}{\partial \mathbf{p}}^T\right)$. The explicit solution for $\Delta\mathbf{p}$, that minimizes the linearized objective function in Equation 4.2:

$$\Delta\mathbf{p} = \mathbf{H}^{-1}\mathbf{J}^T \left[T(\mathbf{0}) - I(\mathbf{p})\right] \tag{4.3}$$

where the pseudo Hessian matrix is defined by,

$$\mathbf{H} = \mathbf{J}^T\mathbf{J} \tag{4.4}$$

The fundamental problem with canonical LK formulation referred as the forwards additive (FA) algorithm [13] is the requirement of the re-estimation of the Hessian matrix at each iteration which greatly impacts computational efficiency.

4.2.1 Inverse Compositional Algorithm

In 2003, Baker and Matthews [12] introduced a computationally efficient algorithm to minimize the SSD objective function described in Equation 4.1 which they referred to as inverse compositional (IC) algorithm. The main change is this algorithm linearizes $T(\Delta\mathbf{p})$ rather than $I(\mathbf{p} + \Delta\mathbf{p})$ resulting in the following objective function,

$$\arg\min_{\Delta\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) - \frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}^T \Delta\mathbf{p} \parallel^2 \tag{4.5}$$

Since $\frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}$ needs only to be computed once, irrespective of the current value of $\mathbf{p}$, one can then solve the linearized objective function,

$$\Delta\mathbf{p} = \mathbf{B}\left[I(\mathbf{p}) - T(\mathbf{0})\right] \tag{4.6}$$

where matrix B,

$$\mathbf{B} = \mathbf{H}_{ic}^{-1}\frac{\partial T(\mathbf{0})}{\partial \mathbf{p}} \tag{4.7}$$

and the pseudo Hessian matrix can be defined as $\mathbf{H}_{ic} = \frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}\frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}^T$. The current warp parameters are iteratively updated by the inverse of the incremental update warp $\mathbf{p} \leftarrow \mathbf{p} \circ \Delta\mathbf{p}^{-1}$. The operation $\circ$ represents the composition of two warps.

4.3 Fitting with Filter Responses

The employment of filter banks as a pre-processing step in many tasks in vision involving illumination variations is motivated by two widely accepted assumptions about human vision: (i) human vision is mostly sensitive to scene reflectance and mostly insensitive to the illumination conditions, and (ii) human vision responds to local changes in contrast rather than to global brightness levels [78]. These two assumptions are closely related since local contrast is a function of reflectance. A natural way to encode local contrast is through the employment of a bank of filters that encode local intensity differences at different orientations and scales.

Linear filters are widely used to extract feature representations. Gabor wavelets [72] are the most popular filter which can be used to extract the useful features due to their biological relevance and computational properties [53, 54, 55, 69]. However, any type of filter that encodes relative intensity differences across many different orientations and scales is suitable.

4.3.1 Filter Responses in Spatial Domain

The reformulation of the LK algorithm to entertain fitting across multiple linear filter responses can be written as,

$$\arg\min_{\mathbf{p}} \parallel \{\mathbf{g}_i * I(\mathbf{p})\}_{i=1}^{M} - \{\mathbf{g}_i * T(\mathbf{0})\}_{i=1}^{M} \parallel^2 \quad . \tag{4.8}$$

Where $\mathbf{g}_i$ is i-th filter with M filters in total, while $\{.\}_{i=1}^{M}$ represents the concatenation operation i.e. $\{\mathbf{x}_i\}_{i=1}^{M} = [\mathbf{x}_1^T \ldots \mathbf{x}_M^T]^T$.

Computational cost

As previously pointed out by [15, 17, 107] a particular problem with Equation 4.8 is the inherently large memory and computational overheads required for representing images in this over-complete filter response domain. The main fundamental problems when applying to the LK framework are:

- If there are M filters in the bank, and D pixels in the input image, we need to do M 2D convolutions involving images containing D pixels each.
- The number of columns in the Jacobian $\mathbf{J}$ matrix increases from PD to PMD, where P is the number of warp parameters. For the special case of the simultaneous algorithm P refers to the number of warp & appearance parameters.
- The computational cost for building Hessian $\mathbf{H}$ matrix increases from P^2D to P^2MD.

As a result of these computational overheads, the idea of doing object alignment with even a modest number of Gabor filter banks (e.g., 9 scales times 8 orientations, i.e. $M = 72$, as employed in [107]) becomes prohibitively expensive and impractical when employing the forward additive algorithm. Even for the inverse compositional algorithm, where the Jacobian and Hessian matrices can be pre-computed to form $\mathbf{B}$, the additional cost of estimating the overcomplete image representation $\{\mathbf{g}_i * I(\mathbf{p})\}_{i=1}^{M}$ and M-fold increase in the size of the pre-computed matrix $\mathbf{B}$ remains. For smaller filter bank sizes authors in literature have resorted to methods for approximating the full response vectors such as: (i) downsampling of filter responses [107], (ii) employing filter responses at certain fiducial positions within the image [198], (iii) the employment of feature selection methods to select the most discriminative filter responses [15], and most recently (iv) where individual classifiers are learnt for each filter response and a fusion strategy employed to combine the outputs in a synergistic manner [101].

4.4 Fourier Lucas & Kanade

Recently, Ashraf and Lucey [8] proposed an extension to the LK algorithm for fitting a template across multiple filter responses in the Fourier domain, which they referred to as the Fourier Lucas-Kanade (FLK) algorithm, that can be written as follows,

$$\arg\min_{\mathbf{p}} \sum_{i=1}^{M} \parallel \mathbf{g}_i * [I(\mathbf{p}) - T(\mathbf{0})] \parallel^2 \quad . \tag{4.9}$$

which can further written as,

$$\arg\min_{\mathbf{p}} \parallel \hat{I}(\mathbf{p}) - \hat{T}(\mathbf{0}) \parallel^2_{\mathbf{S}} \tag{4.10}$$

where,

$$\mathbf{S} = \sum_{i=1}^{M} \text{diag}(\hat{\mathbf{g}}_i)^T \text{diag}(\hat{\mathbf{g}}_i) \tag{4.11}$$

and $\hat{I}, \hat{T}, \hat{\mathbf{g}}_i$ are the $2D$ Fourier transforms of vectorized $I, T, \mathbf{g}_i$ respectively. The matrix $\mathbf{S}$ is a *diagonal* matrix that can be *precomputed* and is *independent* of the number of filters being applied. We also know that the operation of a $2D$ Fourier transform can be replaced by pre-multiplying a signal (of length D) by a $D \times D$ matrix $\mathbf{F}$ containing the Fourier basis vectors. This can be seen by subsuming Equation 4.10 into the weighted LK objective function,

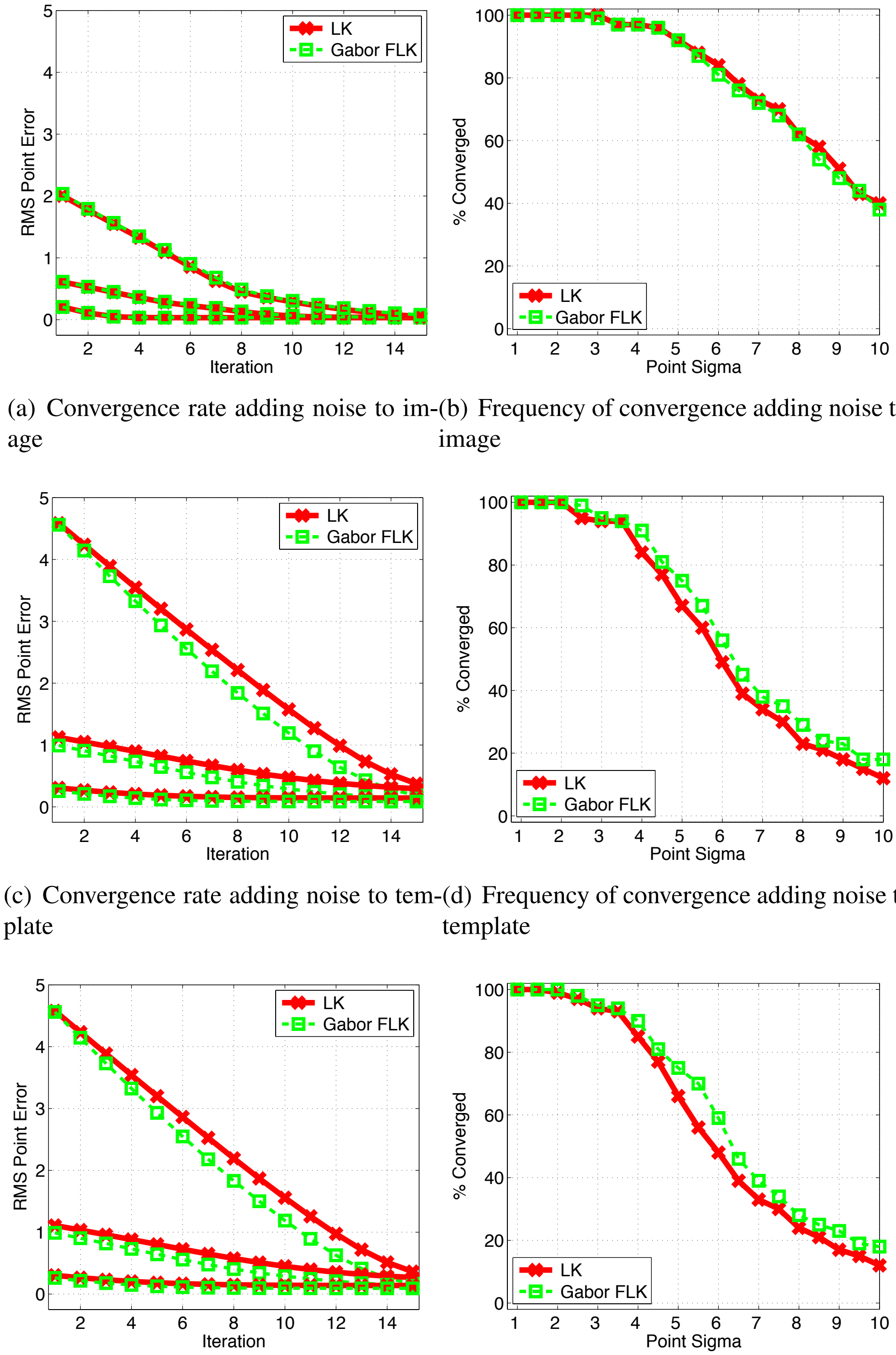
$$\arg\min_{\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) \parallel^2_{\mathbf{F}^T\mathbf{S}\mathbf{F}} \quad . \tag{4.12}$$

4.4.1 Fourier Inverse Compositional Algorithm

The derivation of the FLK IC algorithm follows a similar approach as described in Section 4.2.1. The difference to the LK algorithm defined in Equation 4.7 is that the matrix $\mathbf{B}$ is defined as,

$$\mathbf{B} = \mathbf{H}^{-1}_{flk(ic)}(\mathbf{F}\mathbf{J}_{ic})^T\mathbf{S}\mathbf{F} \tag{4.13}$$

As with all instances of the inverse compositional approach, the Jacobian $\mathbf{J}_{ic}$ depends only on the template image i.e. $\mathbf{J}_{ic} = \left(\frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}\right)^T$ remains constant across all iterations. Consequently, the pseudo-Hessian $\mathbf{H}_{flk(ic)} = \mathbf{J}^T_{ic}\mathbf{F}^T\mathbf{S}\mathbf{F}\mathbf{J}_{ic}$ also remains constant for all iterations.



(a) Convergence rate adding noise to image

(b) Frequency of convergence adding noise to image

(c) Convergence rate adding noise to template

(d) Frequency of convergence adding noise to template

(e) Convergence rate adding noise to both images

(f) Frequency of convergence adding noise to both images

Figure 4.2: The effect of intensity noise on the rate of convergence rates and frequency of convergence. For all the cases white gaussian noise with standard deviation 16.0 grey levels is added to face images. Gabor FLK is sightly more robust for the case where there is noise added.

4.4.2 Registration: LK vs FLK

A series of synthetic experiments were conducted on images: (i) without noise, (ii) noise added to the image, (iii) noise added to the template image and (iv) noise added to both the image and template image (white gaussian noise with standard deviation 16.0 grey levels is added to face images). The performances were computed and the convergence rate and frequency of convergence with the LK-IC and FLK-IC algorithms for different added noise levels compared to ground-truth point locations. The weighting matrix $\mathbf{S}$ for FLK is defined using a bank of Gabor filters (9 scales times 8 orientations). We refer to this as Gabor FLK algorithm. As shown in Figure 4.2 Gabor FLK-IC is slightly more robust for the case where there is noise added.

Then we compared the tracked/detected mouth regions with the ground-truth annotations for FLK and Viola-Jones (VJ) approaches and the root mean square (RMS) point location error is calculated. Figure 4.3 shows the RMS of point location errors for the two approaches performed on consecutive images. As this figure shows, the FLK approach starts tracking the mouth region, with a better fitting accuracy than VJ. However, as it continues tracking, it reaches the RMS of VJ at around 400^{th} frame and then again it continues to degrade. This suggests that the template image needs to be updated at a particular frame to avoid failing in the subsequent frames (Refer to Figure 4.4).

4.5 AAMs with Filter Responses

A major drawback to person specific AAMs stems from their ability to generalize unseen objects. When unaccounted appearance variations are encountered due to a change in the environment (e.g., illumination or camera change), person specific AAMs perform poorly. An overview of AAMs and AAM fitting algorithms are given

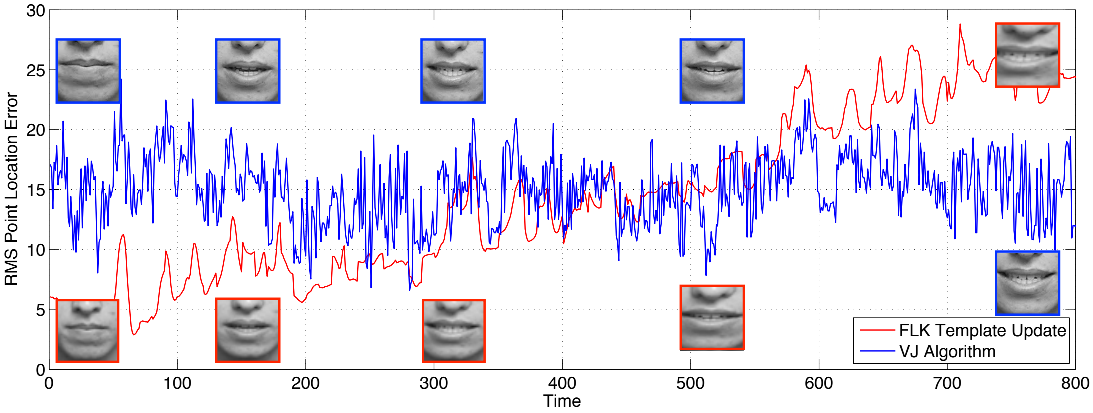


Figure 4.3: FLK mouth detection degradation based on RMS of point location errors over time

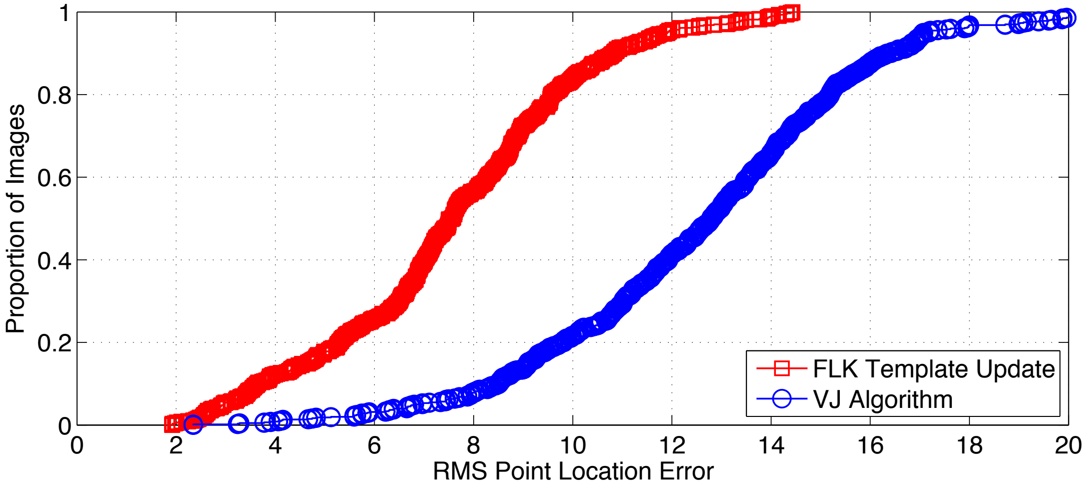


Figure 4.4: Proportion of images that have RMS of point location errors of mouth region less than specified values

in Chapter 3.3.

Fitting an AAM in Equation 3.8 across multiple linear filter responses can be represented as minimizing the following objective function,

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \sum_{i=1}^{M} \| \mathbf{g}_i * [I(\mathbf{p}) - A_{\boldsymbol{\lambda}}(\mathbf{0})] \|^2 \quad . \tag{4.14}$$

Exploiting the fact that convolution becomes a Hadamard (i.e., element-by-element) product in the Fourier domain, and employing Parseval's relation [137] (energy content is preserved as we move from the spatial to the Fourier domain), we may write the error in Equation 4.14 as follows:

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \parallel \hat{I}(\mathbf{p}) - \hat{A}_{\boldsymbol{\lambda}}(\mathbf{0}) \parallel^2_{\mathbf{S}} \tag{4.15}$$

where,

$$\mathbf{S} = \sum_{i=1}^{M} (\text{diag}(\hat{\mathbf{g}}_i))^T \text{diag}(\hat{\mathbf{g}}_i) \tag{4.16}$$

and $\hat{I}(\mathbf{p}), \hat{A}_{\boldsymbol{\lambda}}(\mathbf{0}), \hat{\mathbf{g}}_i$ are the $2D$ Fourier transforms of vectorized images $I(\mathbf{p}), A_{\boldsymbol{\lambda}}(\mathbf{0})$ and filters g_i respectively. The matrix $\mathbf{S}$ is a *diagonal* matrix that can be *precomputed* and is *independent* of the number of filters being applied.

As described in Chapter 4.4, Equation 4.15 can be described with a matrix $\mathbf{F}$ which contains the Fourier basis vectors, resulting in the following FAAM objective function,

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda}} \parallel I(\mathbf{p}) - \mathbf{A}_{\boldsymbol{\lambda}}(\mathbf{0}) \parallel^2_{\mathbf{F}^T\mathbf{S}\mathbf{F}} \tag{4.17}$$

4.5.1 Fourier Simultaneous and Project-Out

An immediate consequence of Equation 4.17 is that it now becomes possible to apply the canonical simultaneous and project-out AAM fitting algorithms, described in Chapter 3.3.3 and 3.3.4, by setting the weight matrix to:

$$\mathbf{Q} = \mathbf{F}^T\mathbf{S}\mathbf{F} \tag{4.18}$$

where $\mathbf{S}$ (Equation 4.16) is determined by the choice of filters being used. Moreover we can also see that FLK and LK inspired fitting strategies become equivalent when $\mathbf{S} = \mathbf{I}$

since $\mathbf{F}^T\mathbf{F} = \mathbf{I}$. It should be noted that in many practical formulations of a 2D-DFT $\mathbf{F}^T\mathbf{F} = c\mathbf{I}$, where c is a constant. Typically, $c = D$ where D is the dimensionality of the feature space.

In practice, however, one never explicitly computes $\mathbf{Q}$, instead applying efficient DFTs to the source and appearance images directly. For the simultaneous algorithm, this has the small drawback of having to perform a DFT at each iteration of the algorithm adding to its already sizable computational cost.

For the project-out algorithm, however, the entire role of $\mathbf{Q}$ and its Fourier transform $\mathbf{F}$ can be completely pre-computed incurring *no* additional computational cost. This result is one of the major contributions, as it allows one to obtain the favorable properties of multiple filter responses in the project-out algorithm without the need to explicitly compute those multiple responses.

4.5.2 Weighted PCA

The appearance basis $\mathbf{A}$ is traditionally found using unweighted principal component analysis (PCA) to find the first K eigenvectors from raw pixel shape normalized training images. However, for the case when $\mathbf{Q} \neq \mathbf{I}$ the weighting matrix must be included in the canonical PCA objective function:

$$\arg\max_{\mathbf{A}} \text{tr}(\mathbf{A}^T\mathbf{V}\mathbf{C}\mathbf{V}^T\mathbf{A}) \text{ subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{4.19}$$

where $\mathbf{C}$ is the scatter matrix of the training images and $\mathbf{V}$ is the decomposition of the positive semi-definite weighting matrix $\mathbf{Q} = \mathbf{V}\mathbf{V}^T$

Step	Complexity
Warp I with $\mathbf{p}$ to compute $I(\mathbf{p})$	$O(nN)$
Compute the error image: $I(\mathbf{p}) - \mathbf{A}_{\boldsymbol{\lambda}}(\mathbf{0})$	$O(mN)$
Compute FFT of the error image	$O(NlogN)$
Compute the steepest descent images	$O((n+m)N)$
Compute the Jacobian	$O((n+m)N)$
Compute FFT for the Jacobian	$O((n+m)NlogN)$
Compute the Hessian $\mathbf{H}_{sim}$	$O((n+m)^2N$
Compute the inverse of the Hessian	$O((n+m)^3N)$
Compute $\Delta\mathbf{q}$	$O((n+m)^2)$
Update $\mathbf{p} \leftarrow \mathbf{p} \circledcirc \Delta\mathbf{p}^{-1}$	$O(n^2)$
Update $\lambda \leftarrow \lambda + \Delta\lambda$	$O(m)$

Table 4.1: The computation cost of the Gabor FAAM algorithm.

4.5.3 Computational Concerns

By casting the AAM algorithm in the Fourier domain, we have shown that it is equivalent to the AAM with a weighting matrix $\mathbf{Q} = \mathbf{F}^T\mathbf{S}\mathbf{F}$. In practice, however, one never explicitly computes $\mathbf{Q}$, instead applying efficient DFTs to the source and appearance images directly. For the simultaneous algorithm, this has the small drawback of having to perform a DFT at each iteration of the algorithm adding to its already sizable computational cost. However, what makes this approach computationally feasible is that we can replace the matrix form of the Fourier transform $\mathbf{F}$ which has a cost of $O(N^2)$ with a computationally feasible Fourier transform which is $O(NlogN)$ [137], where N is the number of pixels.

Computational cost of most of the steps depends on (i) n number of warp parameters and (ii) m number of appearance parameters. The computational cost is independent from the number of Gabor filters. Table 4.1 shows the summary of the computational cost.

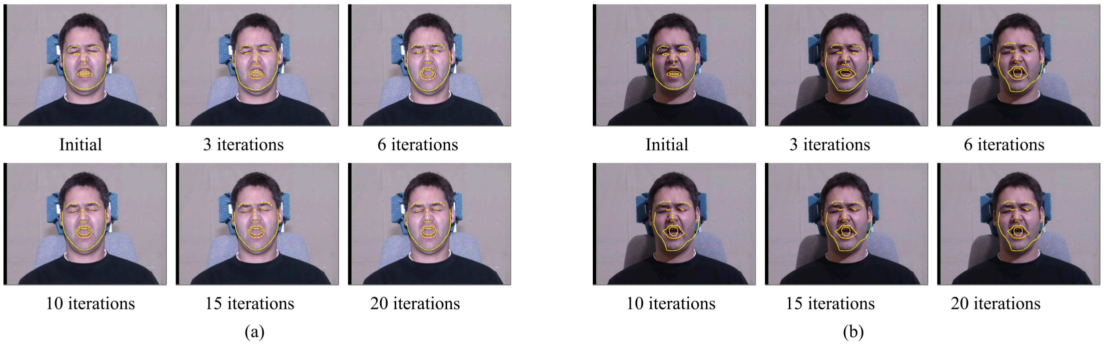


Figure 4.5: Examples of tracking with the Euclidean AAM: (a) Iteration frames with the matched illumination condition, (b) Iteration frames with the mismatched illumination condition.

4.6 MultiPIE Experiments

Throughout this section we will be comparing AAM fitting algorithms for two different weighting matrices: (i) $\mathbf{Q} = \mathbf{I}$, and (ii) $\mathbf{Q} = \mathbf{F}^T\mathbf{S}\mathbf{F}$ where $\mathbf{S}$ is defined through a bank of Gabor filters (9 scales times 8 orientations, see [8] for more details). We shall refer to all variants of (i) and (ii) as *Euclidean Active Appearance Models (AAM)* and *Gabor Fourier Active Appearance Models (FAAM)*.

4.6.1 Measuring Fitting Performance

For all our experiments, a person specific AAM was estimated using MultiPIE database [80] for frontal illumination. Two types of fitting performance were measured: (a) matched and (b) mismatched illumination. We measured fitting performance in terms of root mean square error (RMS) between the 2D mesh location of the current fit results and the ground-truth 2D mesh coordinates with respect to the base mesh. Results were calculated for (a) and (b) when the initialized shape was randomly perturbed from ground-truth.

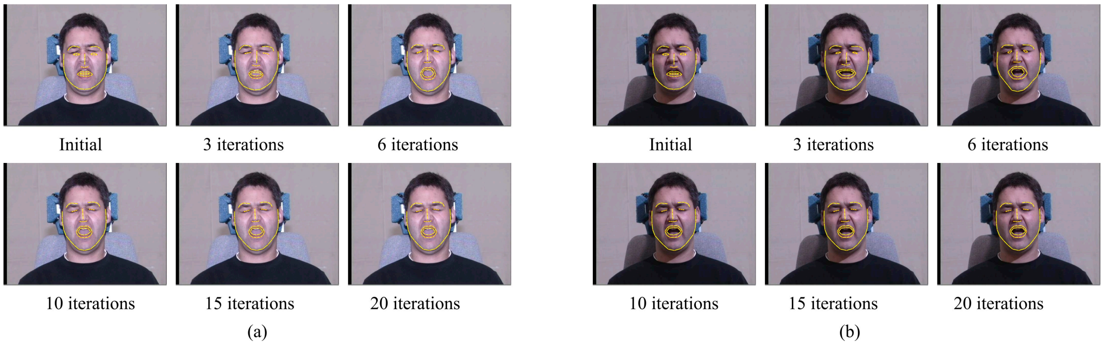


Figure 4.6: Examples of tracking with the Gabor FAAM: (a) Iteration frames with the matched illumination condition, (b) Iteration frames with the mismatched illumination condition.

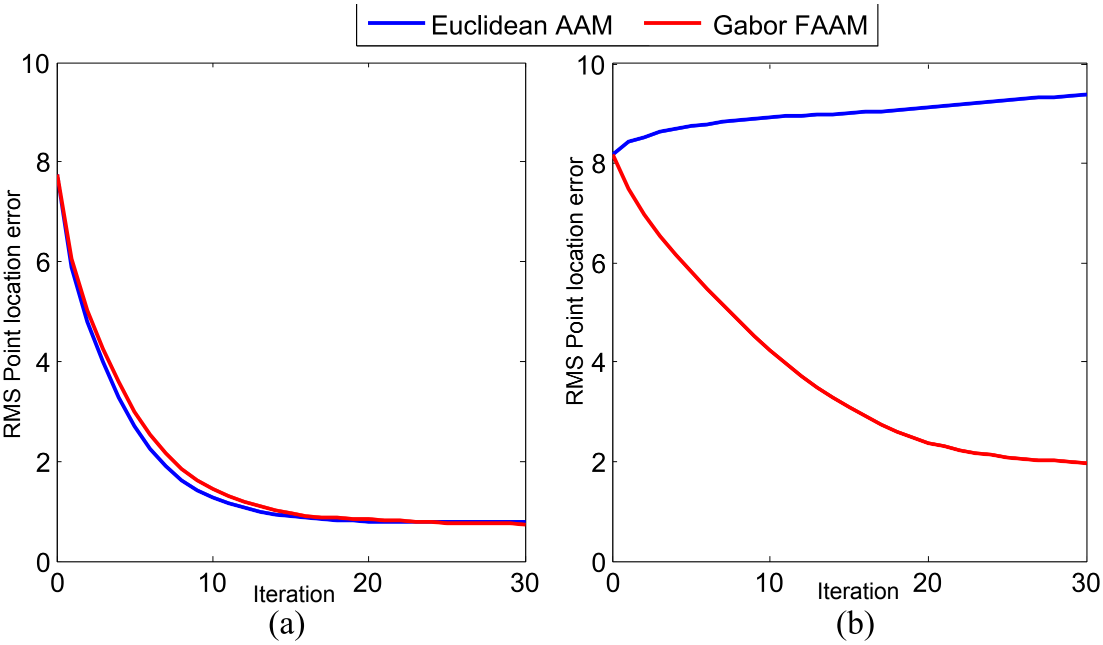
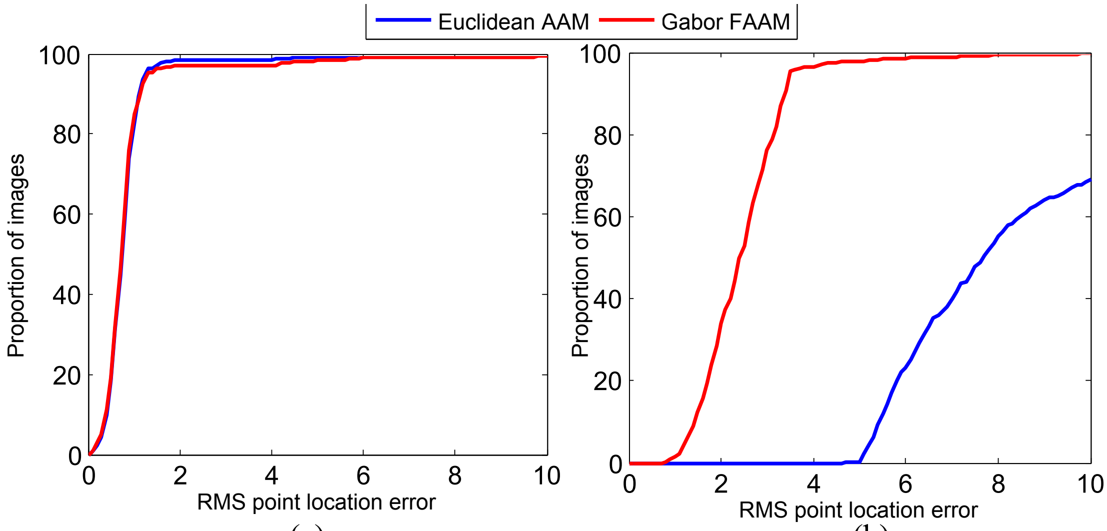


Figure 4.7: Average convergence rates for simultaneous algorithm: (a) when the input and training images have the same illumination conditions, both algorithms perform equally well. (b) when the illumination of the input image changes, the Gabor FAAM algorithm is still able to do the fitting, while the Euclidean AAM diverge.



(a) (b)

Figure 4.8: Fitting performance curves for simultaneous algorithm using Euclidean AAM and Gabor FAAM: (a) when the input and training images have the same illumination, (b) when the input and training images have the mismatched illumination.

4.6.2 Simultaneous Results

Figure 4.7 depicts the average RMS mesh location error against iterations for simultaneous variants of Euclidean AAMs and Gabor FAAMs for (a) matched and (b) mismatch illumination. Visual examples of fitting performance can be seen in Figures 4.5 and 4.6 for Euclidean AAM and Gabor FAAM respectively. Similarly, Figure 4.8 depicts the number of converged trials as a function of the RMS error threshold for (a) and (b). For (a) Euclidean AAM and Gabor FAAMs obtain almost identical performance. However, for (b) in the presence of mismatched illumination there is a clear advantage in using a Gabor FAAM.

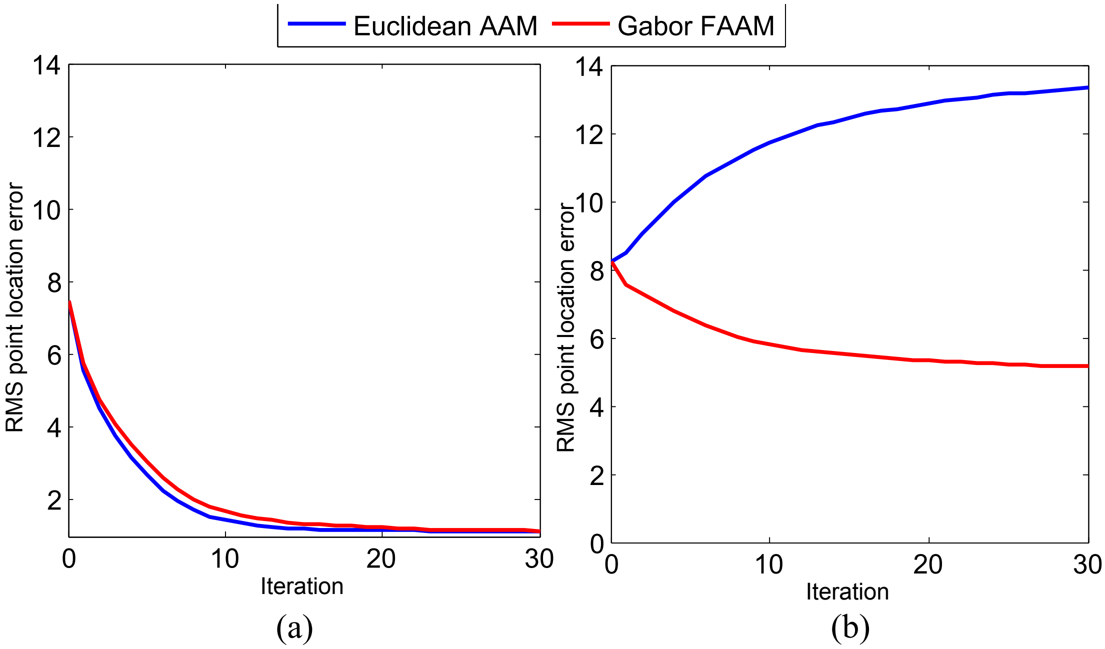


Figure 4.9: Average convergence rates for project-out algorithm: (a) when the input and training images have the same illumination conditions, both algorithms perform equally well, (b) when the illumination of the input image changes, the Gabor FAAM algorithm is still able to do converge, while the Euclidean AAM diverge.

4.6.3 Project-out Results

Figure 4.9 depicts the average RMS mesh location error against iterations for project-out variants of Euclidean AAM and Gabor FAAM for (a) matched and (b) mismatch illumination. Figure 4.10 depicts the number of converged trials as a function of the RMS error threshold for (a) and (b). As expected the results in Figures 4.9 and 4.10 for the project-out algorithm, are poorer than the simultaneous results depicted in Figures 4.7 and 4.8. In a similar fashion to the simultaneous results, however, (a) obtains almost identical performance to Euclidean AAM and Gabor FAAM. In the presence of substantial illumination mismatch (b) the Gabor FAAM outperforms Euclidean AAM by a substantial margin with no additional computational burden during online fitting.

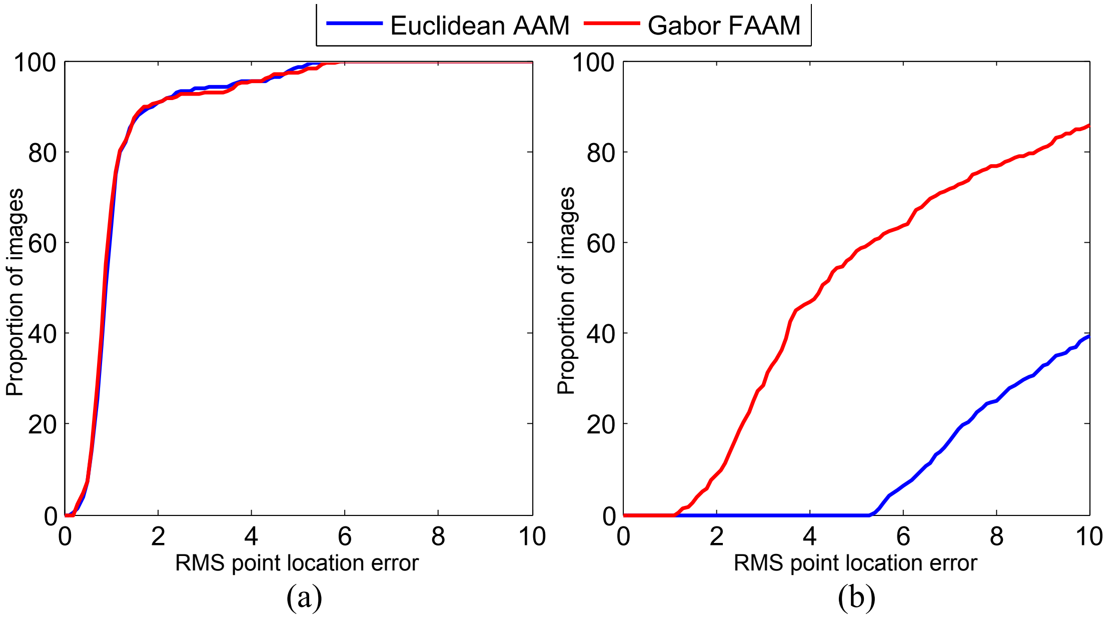


Figure 4.10: Fitting performance curves for project-out algorithm using Euclidean AAM and Gabor FAAM: (a) when the input and training images have the same illumination, (b) when the input and training images have the mismatched illumination.

4.7 Tracking Experiments

We conducted a tracking experiment on video sequences containing substantial variations in illumination over time. The sequence was obtained in a laboratory setting. Ground-truth for the first-frame was given for both the Euclidean AAM and Gabor FAAM. Results in terms of RMS error from ground-truth can be seen in Figure 4.11 showing a substantial benefit to Gabor FAAM in person specific face tracking tasks. Examples of tracking sequence in a real world car environment and an indoor environment are shown in Figure 4.12(a) and Figure 4.12(b) respectively. The tracking performance with Gabor FAAM is much better compared with the AAM tracking results.

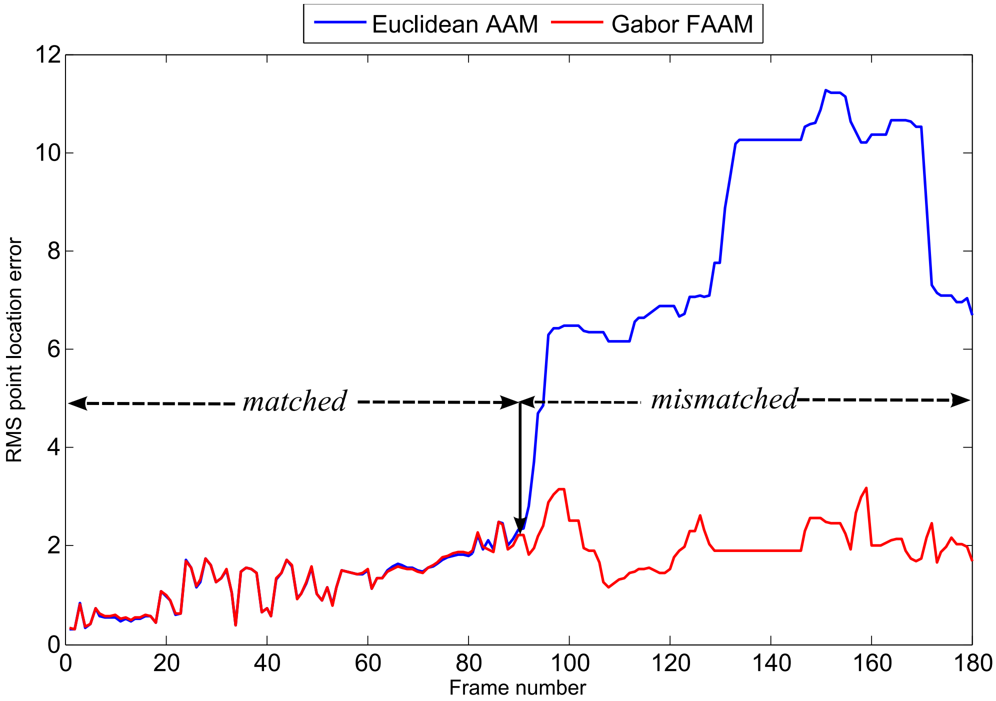
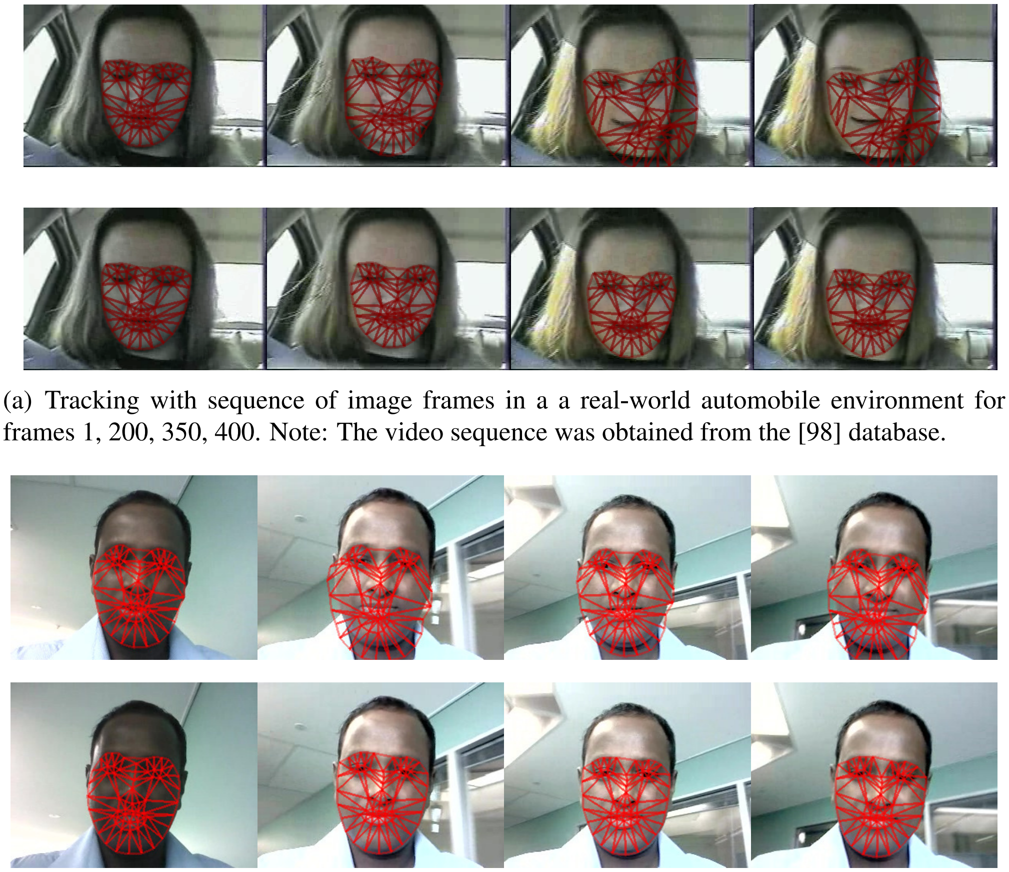


Figure 4.11: Example of a tracking with the Euclidean AAM and the Gabor FAAM in a video sequence. Illumination is changing over the time using the 3 different flashes. Euclidean AAM and the Gabor FAAM showed smiler results in the initial frames, but only Gabor FAAM showed good tracking results when the illumination changing over the time.

4.8 Chapter Summary

This Chapter presented a novel extension to the AAM fitting algorithm which allows for them to be equivalently cast in the Fourier domain. The algorithm is referred to as Fourier AAM. This formulation allows us to interpret the joint alignment across filter responses as a form of the weighted AAM algorithm. We have shown that doing image & object alignment in the high dimensional multiple filter response space is mathematically equivalent to doing alignment in a lower dimensional image intensity space, if appropriate weightings are applied in the Fourier domain.

The key contributions of this Chapter include: (i) show how LK inspired AAM fitting gives identical performance in the spatial and Fourier domains. Further, we demon-



(a) Tracking with sequence of image frames in a a real-world automobile environment for frames 1, 200, 350, 400. Note: The video sequence was obtained from the [98] database.

(b) Tracking with sequence for frames 1, 70, 100, 130 of person who is walking along a passage in a building.

Figure 4.12: Examples of tracking in a real world applications. Gabor FAAM showed better tracking results when the illumination changing over the time. Top Row : tracking sequence with the AAM, Bottom row: tracking sequence with the Gabor FAAM

strate how the effect of multiple filter responses can be re-interpreted as a diagonal weighting matrix in the Fourier domain leading to substantial computational savings when performing inverse compositional simultaneous fitting across multiple filter responses, (ii) demonstrate the process of applying the inverse compositional project-out algorithm in the Fourier domain by showing how: (a) the Fourier transforms to the current image, and (b) the application of multiple filter responses can be completely pre-computed offline incurring no additional computational cost. This contribution is of key importance to person specific AAM face fitting as it provides an extremely computationally efficient method that affords both invariance to both expression and environmental variations, (iii) empirically show the substantial improvement in person specific AAM fitting performance over canonical LK inspired fitting algorithms (i.e. simultaneous and project out), when using the proposed Fourier variants. For all our experiments we employed biologically motivated Gabor filter banks with person-specific AAMs.

Chapter 5

Analysis I: Visual Speech Recognition in Varying Lighting Conditions

One of the primary aims of this dissertation is to demonstrate that human behavior in noisy environments can be identified using face information, as the face contains a rich source of information. However, implementing a human recognizing framework requires a system able to accurately locate and track the person's face in real-time. In Chapter 3, a review of face alignment technologies are presented and problems are identified. A novel solution for the problem with varying illumination is presented in Chapter 4. This Chapter presents a comprehensive study of recognizing human behavior (i.e speech recognizing) in vehicles which have fluctuating illumination conditions.

Recognizing human speech within a vehicular environment has the potential of reducing driver distraction. With the addition of in-vehicle navigational, phone and other operational systems, this problem is heightened. Giving a driver the ability to interact with this technology via voice-only commands (i.e. hands-free) has the potential to reduce these safety concerns. Vehicles, by nature, create very acoustically noisy environments with the amount of noise varying with respect to the speed of the vehicles;

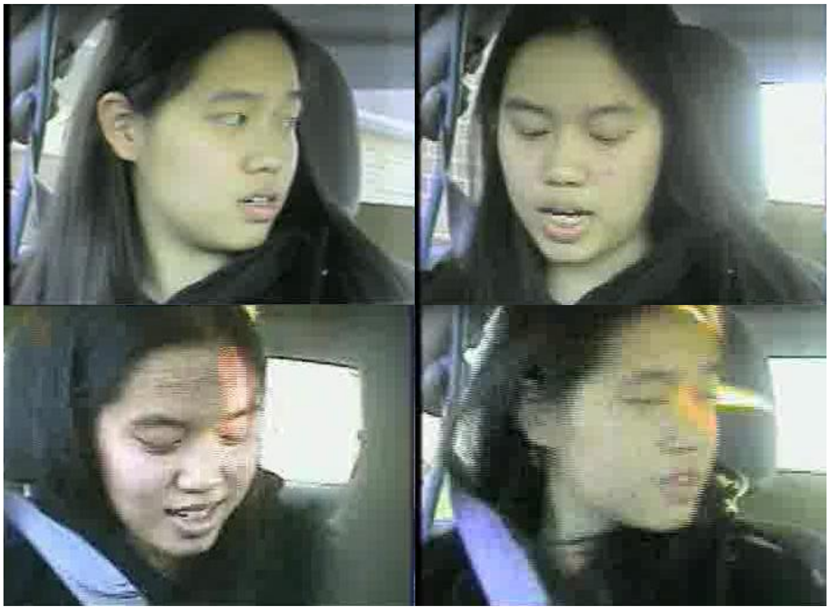


Figure 5.1: Examples of the challenging environments encountered within an automotive environment. Clockwise from top-left (a) head position, (b) head scale, (c) illumination, and (d) resolution/image quality.

whether the windows are down or up; people are speaking in the vehicle; the radio is on/off and the quality of the insulation of the vehicle. Acoustically, car cabins are extremely noisy and as audio-only automatic speech recognition (ASR) systems are susceptible to these conditions; poor performance is normally obtained, which inhibits its use within a vehicle. The chapter addresses the above mentioned real-world application and provides potential solutions by identifying the speech using visual modality, which is somewhat immune to these variabilities and a potential method to improve the robustness of these conditions in conjunction with the audio stream.

5.1 Introduction

One of the major issues affecting driver safety is the continuing increase in the complexity of in-vehicle navigational and other operating systems [156]. The use of voice recognition technology has the potential to provide solutions to this problem by providing voice based control, as a less distracting alternative to manual control, for the operation of such in-vehicle systems. A vision of an in-vehicle voice recognition system in action was depicted the 1980's television series "Knight Rider" [1]. In this show, *KITT* (Knight Industries Two Thousand) which is a fictional computer that controls the high-tech black Pontiac Firebird automobile, is essentially capable of conducting a natural conversation with the driver as well as enacting any command given by the driver. Unfortunately, current voice recognition systems are a long way from achieving this vision.

One of the main reasons for this is that they rely solely on the audio channel for input which can be corrupted by many environmental factors, some examples of which are road, wind and/or engine noise. One possible solution is to make use of the bimodal nature of speech, by incorporating visual-speech information from the driver's face, to improve speech intelligibility in noisy conditions [38]. The field of recognizing speech using both audio and visual inputs is known as audio visual automatic speech recognition (AVASR) [152].

Notable progress has been made with AVASR technology in the last few decades and continuous research in this field has been ongoing [27, 38, 116, 133, 150, 153, 189]. Over this period of time, the need for the visual modality in voice recognition systems has been established theoretically and prototype systems have been built that have demonstrated improved performance over audio-only systems under clean recording conditions. One major reason behind the lack of progress in getting a real-world AVASR system deployed is that most research has neglected addressing variabilities in the visual domain such as viewpoint, lighting conditions, out-of-plane face movements and occlusions [116, 153] that could be expected in a vehicular scenario as shown in Figure 5.1. This is mainly due to the restricted position of speaker's faces in early AVASR databases, but as more research is being concentrated in more 'real-world' conditions, such as meeting rooms or automotive environments, audio-visual speech data has recently become available to allow AVASR to be conducted where the speaker has more freedom to move their head naturally.

One avenue for improving AVASR in real world conditions is to take advantage, if possible, of multiple views of the visual-speech information, or lip-movements, of the active speaker. Limited work in this area has begun by focusing on recognising visual speech from profile views [204], and where both frontal and profile views of the visual speech are available, combining them to show that visual speech recognition can be improved when multiple views of the speaker are available. [116]. However, these limited studies have only been conducted in relatively clean conditions by combining quite distinct frontal and profile views. By taking advantage of the multiple cameras available in the AVICAR database [98] which captures challenging visual variabilities in a vehicular scenario, this Chapter proposes that multiple near-frontal views of a speaker can provide for improved AVASR performance over traditional single-camera approaches.

The Chapter presents AVASR experiments using four multiple cameras using the AVICAR dataset and investigates the usefulness of the visual information from different camera angles. A series of visual speech recognition experiments are conducted on the four-camera AVICAR database to demonstrate that the best visual speech performance can be obtained using the side and central orientated cameras in a four-stream visual synchronous hidden Markov model (SHMM). The combination of the four visual streams with a single audio stream in a five-stream audio-visual SHMM demonstrates even better performance when compared to any single camera audio-visual SHMM and improves upon the acoustic-only HMM approach across every noise condition of the AVICAR database.

5.2 AVASR System

Recognizing speech in noisy environments has been a topic of interest for engineers since the 1890s [178] and has continued more recently into speech understanding by computers. The main challenge of speech recognition in noisy environments is the presence of a number of environmental factors such as acoustic noise. During the 1940s and 1950s, with the rapid growth in military and civil aviation, an important application of interest to engineers working in the field was improving ways in which air traffic controllers could communicate with pilots.

All of this interest led to the first work on audio-visual speech processing, published by Sumby and Pollack in 1954 [177]. In this work, Sumby and Pollack examined the contribution of visual factors to oral speech intelligibility as a function of the *speech-to-noise* (SNR) ratio and the size of the vocabulary. The first actual implementation of an AVASR system was developed by Petajan in 1984 [145], where simple black and white images of a speaker's mouth were extracted and the mouth height, width, perimeter and area were used as visual features. The next major progress in AVASR was when Bregler and Konig [27] in 1994, published their work using *eigenlips*. This work was further extended by Duchnowski et al. [61] in 1994 by employing linear discriminant analysis (LDA) to improve the visual feature extraction.

In the summer of 2000, IBMs Human Language Technologies Department at the T.J. Watson Research Center coordinated a workshop at the John Hopkins University in Baltimore, USA, where leading researchers from around the world converged to collect audio-visual database and to further improve techniques associated with AVASR [133].

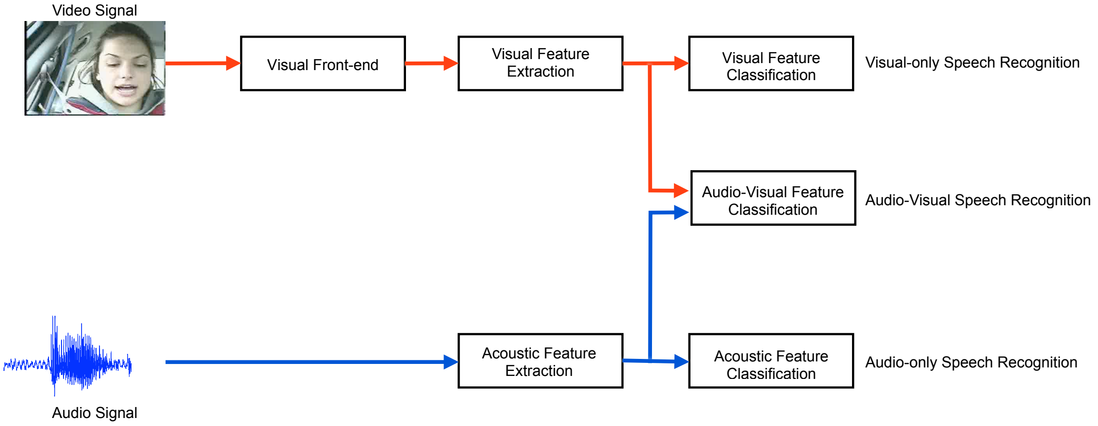

Video Signal
Visual Front-end
Visual Feature Extraction
Visual Feature Classification
Visual-only Speech Recognition
Audio-Visual Feature Classification
Audio-Visual Speech Recognition
Acoustic Feature Extraction
Acoustic Feature Classification
Audio-only Speech Recognition
Audio Signal

Figure 5.2: Block diagram comparing an AVASR system with uni-modal acoustic and visual speech recognition.

In 2003, Potamianos et al. [150] conducted AVASR experiments in typical office and automobile environments. In this work, they found that the performance degraded in both modalities by more than twice their respective word error rates, however, the visual modality still remained beneficial in recognizing speech. As the main benefit of using the visual modality in speech recognition systems is to counteract the problems associated with real-world acoustic environments, it is interesting to note that the majority of research conducted in AVASR has not yet focused on real-world environments, mostly due to the time and the cost, which it takes to capture, rather than simulate, typical real-world audio and visual degradation [133].

A block diagram comparing a typical AVASR system with traditional uni-modal acoustic and visual speech recognition systems is shown in Figure 5.2. An overview of the each subsystem of the AVASR system will be presented in the remainder of this section, followed by a brief overview of AVASR databases.

5.2.1 Acoustic Feature Extraction

The main aim of the feature extraction stage is to reduce the dimensionality of the incoming acoustic signal and to obtain useful features for speech recognition. The incoming audio signal is sampled by an analog-to-digital (ADC) converter to obtain a digital signal, typically at a sampling rate of between 8 and 16 kHz [199]. The two most common audio feature extraction methods are Mel-frequency cepstral coefficients (MFCCs) or perceptual linear predictive (PLP). MFCCs are motivated speech representation based on Fourier transform and filter bank analysis, first proposed by David and Mermelstein in 1980 [56]. Lockwood et al. [108] showed that MFCC features are superior to the alternative feature extraction methods in clean speech and more robust to background noise, especially in a vehicle environment. Therefore, MFCC feature representation is used extensively in this thesis work.

MFCC Features

MFCCs are motivated speech representation based on Fourier transform and filter bank analysis, first proposed by David and Mermelstein in 1980 [56]. Psychophysical studies have shown that the human ear resolves non-linearly across the speech spectrum [205]. This behavior can be represented by a triangular filter bank spaced across the speech spectrum using the Mel-frequency scale such that,

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{5.1}$$

where f is the linear frequency and f_{mel} is the perceived Mel frequency. The energies for each filter bank are then calculated and summed up together. Finally, the Mel-frequency cepstral coefficients can be obtained by applying the Discrete Cosine

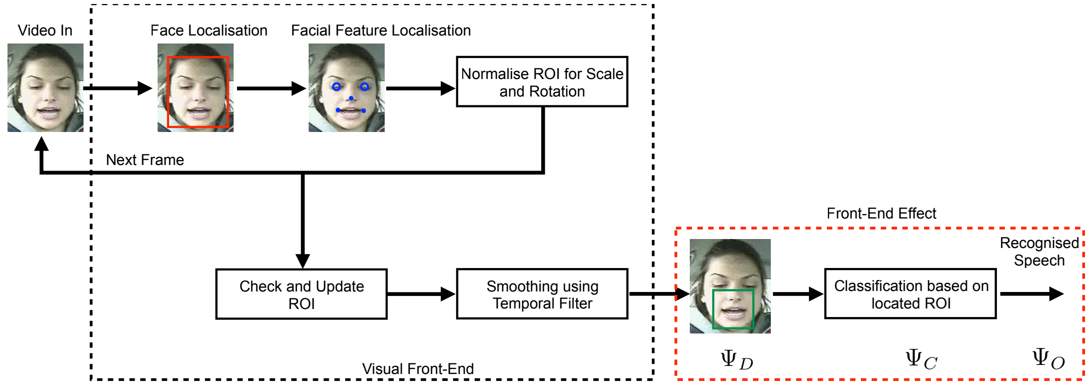

Video In
Face Localisation
Facial Feature Localisation
Normalise ROI for Scale and Rotation
Next Frame
Check and Update ROI
Smoothing using Temporal Filter
Visual Front-End
Front-End Effect
Recognised Speech
Classification based on located ROI
Ψ_D
Ψ_C
Ψ_O

Figure 5.3: Block diagram of visual front-end system for visual speech recognition and the cascading of the front-end effect. The output ROI from the visual front-end system has a detrimental effect on the remaining stages of the AVASR.

Transform (DCT) and logarithmic compression as follows,

$$C_i = \sum_{j=1}^{N} \log_{10}(m_i) cos\left(\frac{(2j+1)\pi i}{2N}\right) \tag{5.2}$$

where m_i is the filter bank energy of the $i^t h$ filter bank, N is number of cepstral coefficients and C_i are the cepstral coefficients.

5.2.2 Visual Front-End

In AVASR, the most important stage in the use of the visual modality is to reliably track and detect the speaker's region of interest (ROI). The majority of these visible articulators emanate from the region around a speaker's mouth. The success of the entire system depends on designing a robust visual front-end which will be able to locate and track the speaker's face and facial features across many variables (i.e. illumination and head pose). If the visual front-end system is not highly accurate, it will have a detrimental effect on the remaining stages of the AVASR system [112]. This error will cascade throughout the system and will most likely recognize the visual speech incorrectly. This effect is known as the *front-end effect*. The *front-end effect* can be written as,

$$\Psi_O = \Psi_D \times \Psi_C \tag{5.3}$$

where Ψ_D is the probability that the ROI has been located, Ψ_C is the decision probability given the located ROI and Ψ_O is the overall probability that the system will recognize the correct speech. An overview of the visual-front-end process with the *front-end effect* is depicted in Figure 5.3.

The selection of visual front-end systems is dependent on the type of application being used. In lip reading literature, appearance based approaches have been widely used in visual front-end systems, [102, 201] they being well suited to many different objects (face, eyes, nose etc.) under varied conditions due to their probabilistic nature and having shown good performances compared with other approaches.

Matthews et al. [122] used AAMs to fit a lip shape model to an image containing a mouth. Although AAMs and FAAM which were introduced in Chapter 4, have been shown to be useful for face and facial feature tracking the main problem is the need for a massive amount of annotated training data. The accuracy of registration is highly dependent on this training data. This problem was highlighted in the experiments conducted by Matthews et al. [122], where the results show that the AAMs often failed to follow small facial motions, which results in poor visual speech recognition results [122]. This is very problematic for real world applications such as an automotive environment due to long term driver's face monitoring and occurrence of a lot of unseen subjects. As the main motivation of this dissertation is to recognise human behaviour in very noisy environments, and due to the problems associated with deformable face modellings as shown in previous research [122], we therefore im-

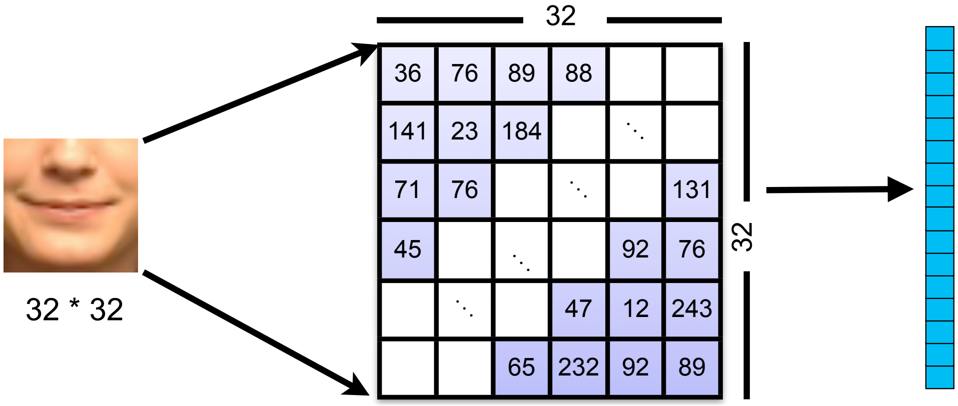


Figure 5.4: The visual feature extraction step seeks to find representations of the observation (ex: 32 × 32 lip region).

plemented the visual front end system for this AVASR system using coarse type of registration which can generalised well for unseen subjects. A good comparison of coarse-type of registration and AAMs is given in Chapter 3.

The most common coarse-type of approaches is the haar-like feature matching approach of the off-the-shelf Viola Jones (VJ) object detector (available in the OpenCV image processing libraries [194]). The VJ algorithm [194], which is based on a boosted cascade of simple classifiers will be used to develop a baseline visual face detection system as it showed a reasonable detection rate in fluctuating conditions as shown in Chapter 3.

5.2.3 Visual Feature Extraction

While acoustic feature extraction for speech recognition is relatively mature, in that MFCCs and PLP have been extensively verified in practical experiments, visual speech feature extraction is still a developing research area. Visual speech is best discriminated by the movements of the visual articulator (i.e mouth, lip and jaw movement) [153]. The visual feature extraction step seeks to find representations of the given observations as shown in Figure 5.4 that provide discrimination between the various speech units whilst providing invariance to irrelevant transforms on the observations that are in the same class.

Various sets of visual features for visual speech recognition have been proposed in the literature [151] and can generally be separated into three groups:(a) appearance-based, (b) contour-based, or (c) combinations of both.

Appearance based features utilise the entire ROI to extract the visual features [133, 150, 153], while the contour based representation are concentrated on capturing the geometric parameters of the lip region, such as mouth height and width [3, 35, 76] as depicted in Figure 5.5. Conversely, the AAM creates a single model of both shape and appearance. The main disadvantage of this approach is that it requires an extremely large number of manually annotated points for the training examples and does not perform well for unseen subjects. Matthews et al. [122] showed that the appearance features are outperformed AAMs using the task of large vocabulary speaker independent AVASR.

In general, appearance-based methods are preferred by most researchers as they are perceptually motivated by human perception studies and do not require finer localisation and tracking, reducing the impact of the ‘front-end effect’ (i.e the impact for visual-speech recognition system on having the visual speech articulators successfully located). For all these reasons, the appearance-based approach is preferred and will be the focus in this dissertation.

Cascading appearance-based features, devised by Potamianos et al. [153] have been established as the state of the art for AVASR visual feature extraction as they contain information about the visible articulators such as tongue, teeth, and the muscles around the jaw and can be computed very quickly, lending themselves to real-time implementation. Essentially, this process is broken into two sections:

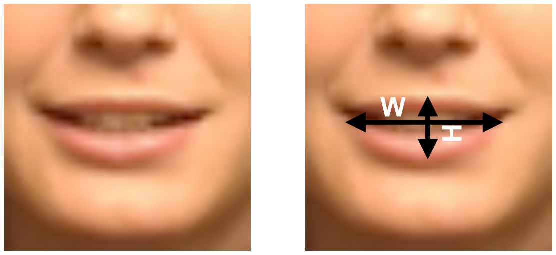


Figure 5.5: Comparison of Appearance based features and Contour based features. (left) Appearance based features utilize the entire ROI to extract visual features. (right) Contour based features based on the geometric parameters of the lip region such as mouth height, width.

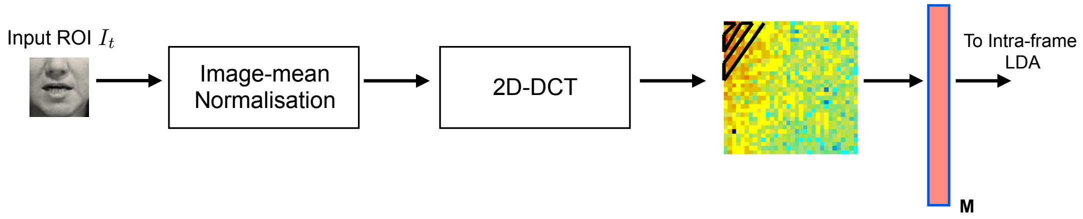


Figure 5.6: Block diagram of showing static feature extraction

(a) Static feature extraction

(b) Dynamic feature extraction

Static visual features

The goal of the static feature extraction stage is to maximise the amount of relevant visual speech information contained within each frame of the ROI within the least number of features. Typically, following a ROI tracking stage, the ROI images are converted to grayscale and image-mean normalization is performed to help attenuate any irrelevant information, such as illumination or long-term variations in speaker appearance. This process is performed by subtracting a mean image calculated over the entire utterance from every incoming frame in the utterance. Then a two-dimensional, separable, DCT is applied to the mean-removed image. The top M higher energy

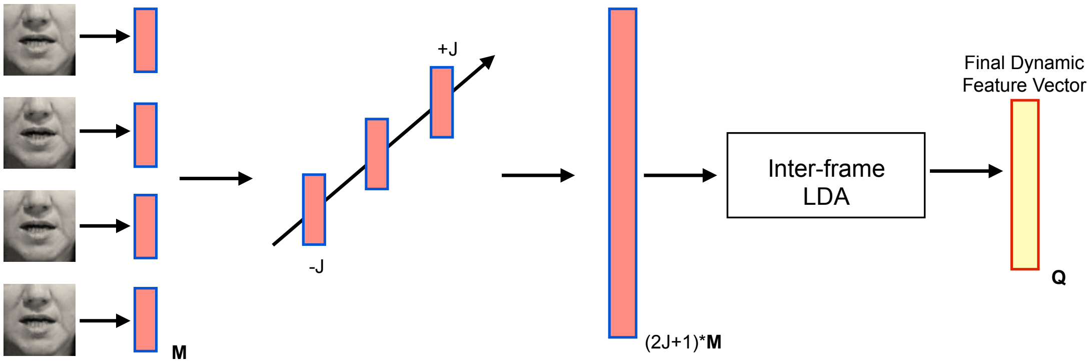


Figure 5.7: Block diagram of showing dynamic feature extraction stage

components according to a zig-zag pattern from the top-left, are then used as static features to represent the visual speech information within the image, which contains the most variability in the tracked ROI. An overview of the static feature extraction stage is given in Figure 5.6.

Dynamic visual features

Visual speech is represented by the movements of the visual articulators [153]. The best features for representing visual speech are generally considered to be focus on the movement of the features, rather than the features within each frame. The simplest method to extract dynamic features is through the use of time-derivative-based delta and acceleration coefficients. These coefficients are used in addition to the original static features [152] which result in a higher feature vector. Recently one technique which has shown good performance is the use of LDA to extract the relevant dynamic speech features from the ROI. In order to incorporate dynamic speech information the static features from the static feature extraction stage are concentrated before speech-class based LDA is performed based on a known transcription.

Such an approach is shown in Figure 5.7. It can be seen in this figure that the transformation matrix is found from the concatenation of $\pm J$ frames surrounding the current frame. Each input frame to the LDA step can be represented as follows and the resulted feature vector is size of $(2J+1)\mathbf{M}$.

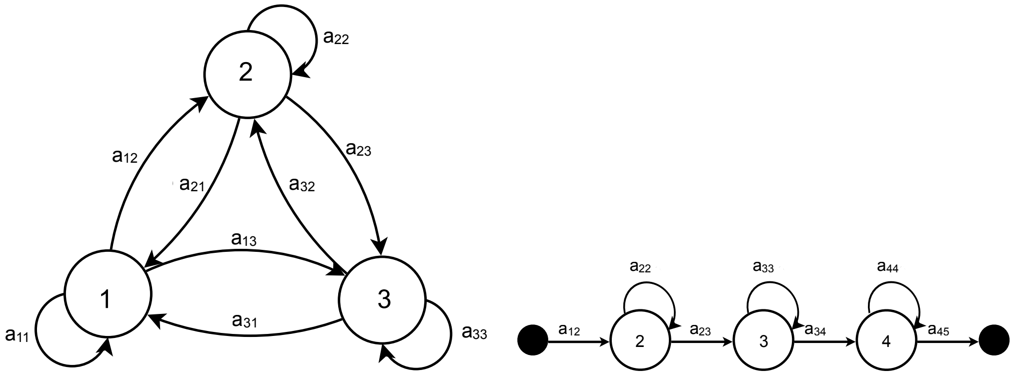
$$\mathbf{O}^{C}t = [(\mathbf{O}^{s}_{t-J})', \ldots, (\mathbf{O}^{s}_{t})', \ldots, (\mathbf{O}^{s}_{t+J})']' \tag{5.4}$$

The obtained static features (input feature vectors span across multiple frames and not just within the frame) can then be projected via an inter-frame LDA stage, where the LDA transformation is trained on acoustically-aligned subword units, to yield a $\mathbf{Q}$ dimensional 'dynamic' visual feature vector. Earlier work by Neti et al. [132] and Potamianos et al. [151], found that using $J = 2$ (i.e 5 adjacent frames) gave optimal results.

5.2.4 Audio-Visual Speech Modelling

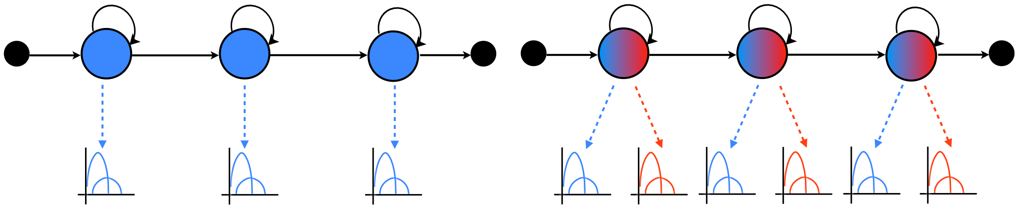
The most widely used classifier for modelling and recognising in audio-only and visual-only speech recognition is the hidden Markov models (HMMs), due to the natural ability to model temporal signals [205]. Here, the internal states are hidden compared with a standard Markov process where the states are known. Figure 5.8(b) represents a left-to-right topology of a HMM with five states (i.e entry, 3 emitting states and exit state). In this left-to-right topology, the transactions between the states can be either, move the next state to the right or bounce back to the same state compared with an ergodic model as shown in Figure 5.8.

While a number of alternative modelling techniques have been proposed and demonstrated for audio-visual modelling, the state-of-the-art is generally considered to be the middle-integration synchronous hidden Markov model (SHMM) [57, 150] approach which couples the acoustic and visual observations at every frame. It has the ability to reliably weight each modality (i.e. audio features and visual features) on an individ-



(a) (b)

Figure 5.8: Comparison of the markov models. The probabilistic transactions are denoted as a_{ij} (a) ergodic model; and (b) left-to-right HMM topology with entry, 3 emitting states and exit state.



(a) (b)

Figure 5.9: HMM modelling approaches for AVASR. Emission densities for acoustic and visual modality is shown in blue and red in respectively. (a) Unimodal (acoustic) HMM and (b) Synchronous HMM.

ual basis. This approach can be seen to be similar to a unimodal acoustic (or visual) HMM, but with multiple observation-emission Gaussian mixture models (GMMs) for each feature-stream in each HMM state as is depicted in Figure 5.9.

While any number of feature streams can be supported by SHMM modelling, AVASR approaches typically have two streams, one for audio and one for video. One of the major advantages of the SHMM approach over alternative approaches is that each stream can be weighted on an individual basis, and by allowing the streams to be treated independently, the SHMM model is more flexible and can generally provide improved AVASR performance [57, 150].

Typically the observation-emission score of an individual state u in a SHMM is given in terms of the acoustic stream weighting parameter α as

$$P(\mathbf{o}_t|u) = P(\mathbf{o}_{a,t}|u)^{\alpha} P(\mathbf{o}_{v,t}|u)^{1-\alpha}, \tag{5.5}$$

where $P(\mathbf{o}_{s,t}|u)$ is the probability of stream s having an feature observation vector $\mathbf{o}_{s,t}$ at time t in HMM state u. For example audio and visual observation vectors are represented as, $\mathbf{o}_{a,t}$ and $\mathbf{o}_{v,t}$ respectively. The parameters α and 1- α are the audio stream and visual stream weighting parameter respectively and $0 \leq \alpha \leq 1$. The SHMM observation-emission score can be written in the more general form of S streams as follows:

$$P(\mathbf{o}_t|u) = \prod_{s=1}^{S} P(\mathbf{o}_{s,t}|u)^{w_s} \tag{5.6}$$

where w_s is the stream weight for stream s and $\sum_s w_s = 1$. It can be seen that Equation 5.6 is equivalent to Equation 5.5 for the case where $S = 2$, $w_a = \alpha$ and $w_v = \alpha - 1$.

The choice of the stream weighting parameter w_s is typically taken by maximising the speech recognition performance on an evaluation session.

5.2.5 Audio Visual Databases

Single-view audio-visual databases

To date, a number of interesting databases have been developed in the AVASR research community, but most are focused on single-camera views of speakers in relatively clean conditions. Coinciding with the first ever AVASR system, Petajan [145] collected a database consisting of a single subject uttering 2-10 repetitions of 100 isolated English words. Since then, similar single-subject databases have been collected [3, 37, 76, 81, 157, 176, 186, 189]. Apart from the single subject audio-visual databases many multiple speaker databases have been collected over the last years. However, due to the cost of capturing, storing and distributing, these have only been concerned with small vocabulary tasks [35, 121, 128].

In recent years, the multi-speaker databases have been extended to include many more speakers. However, most of these databases are still concerned with small vocabulary tasks in very clean conditions [125, 142, 146]. One of the most popular databases in the late 1990's was the M2VTS database, [146] which consisted of 37 speakers. This database was later extended to form the XM2VTS database [125], which contains 295 subjects in fully frontal with four sessions. The XM2VTS database is currently the largest publicly available audio-visual database in terms of number of speakers in clean controlled conditions. The CUAVE [142] database is another publicly available audio-visual database which contains speakers talking in frontal and non-frontal poses. This database consists of 36 speakers with two sections: (i) individual utterances and (ii) group section (i.e look at pairs of simultaneous speakers).

Multi-view audio-visual databases

There are a few audio-visual databases that provide multiple close-up views of a speaker's face suitable for performing visual speech recognition, although limited research has taken advantage of this data. The IBM smart-room database [148] was captured using two microphones and three PTZ (Pan Tilt Zoom) cameras and it is not a publicly available database. It consists of 290 fully frontal subjects uttering continuous speech with mostly verbalised punctuation, dictation style. The AMI meeting corpus [31] consists of 100 hours of meeting recordings captured with either two or three cameras. The VACE multi-model meeting corpus [36] was recorded using wireless microphones and stereo-vision cameras. The MM4 audio-visual corpus [123] was cap-

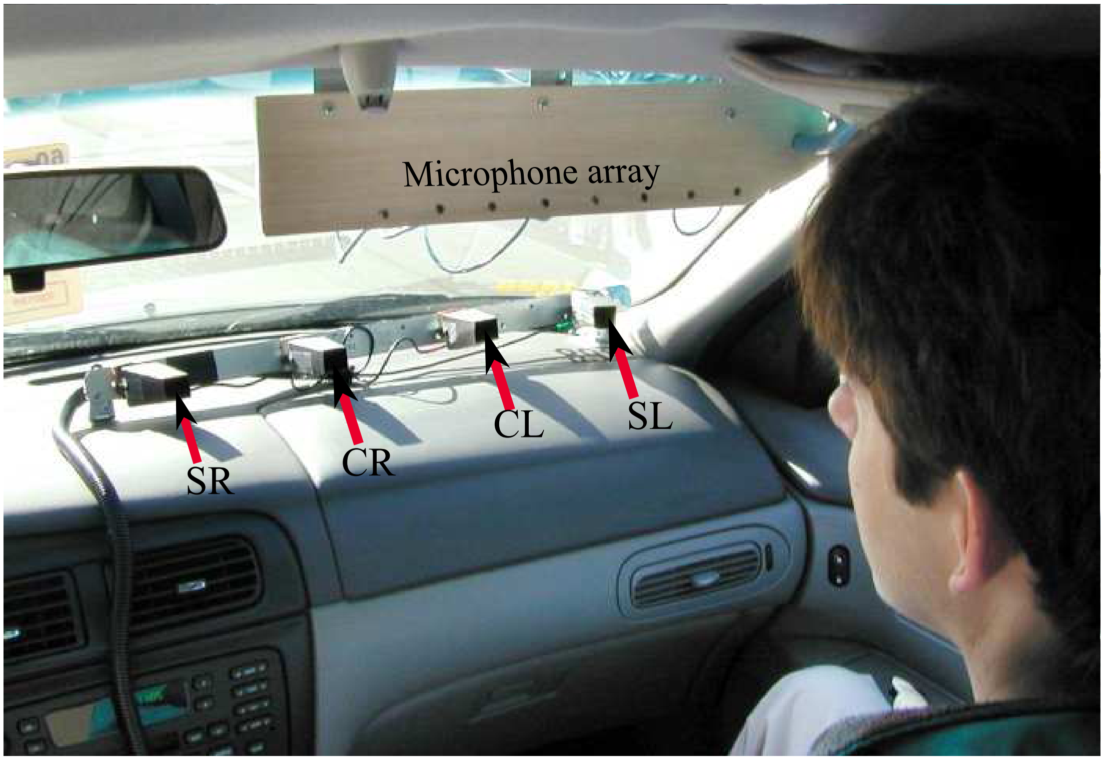


Figure 5.10: Set up of the microphones and the cameras in the AVICAR database. An 8-microphones array is positioned on the passenger's sun-visor and a 4-camera array is positioned on the dash board. For this paper we have labelled the cameras based upon whether they are side (S) or central (C) and left (L) or right (R) according to the cameras' viewpoint. (image from [98])

tured in a single meeting room with high quality miniature lapel microphones and three close-circuit television cameras fitted with an adjustable wide angle lens. The NIST meeting room database [74] was captured using five Sony EUI-D30 motorized NTSC analogue video cameras and four microphones. Finally the AVICAR [98] database captured audio-visual speech in a real-world vehicle environment using four passenger facing cameras and an eight microphones arranged in a linear array.

In-Car audio-visual databases

When studying the behaviour in-car conditions, there are few databases that have been collected [49, 138]. Even though these databases have captured a vehicular environment [49, 138], they: (i) are not publicly available, (ii) are very expensive to acquire (iii) neglect the viewpoint or pose of the speaker or (iv) are not recorded in English words. The AURORA-3J-AV database [49] consists of 58 Japanese speakers with three different driving conditions (idling, city road and express-way). All the recorded speech is in Japanese. The AV@CAR [138] Spanish database was captured using 20 speakers (10 male and 10 female speakers) whose ages range from 25 to 50. The database is composed of seven audio channels, one video channel and information about the speed of the car, the conditions of the road, the weather, the traffic as well as information about the speaker and the lighting conditions.

For the evaluations performed in this Chapter, we focused on the AVICAR database [98], which is publicly available with English recording and captures the types of variabilities that could be expected in a vehicular scenario as shown in Figure 5.1

5.3 The AVICAR Database

The AVICAR database is a publicly available in-car speech corpus containing multi-channel audio and video recordings [98] which was recorded by researchers at the University of Illinois. The collection was designed to examine the performance of speech recognition through combining multi-channel audio and visual speech recognition in adverse conditions.

As shown in Figure 5.10, the audio-visual speech was captured using an array of 8

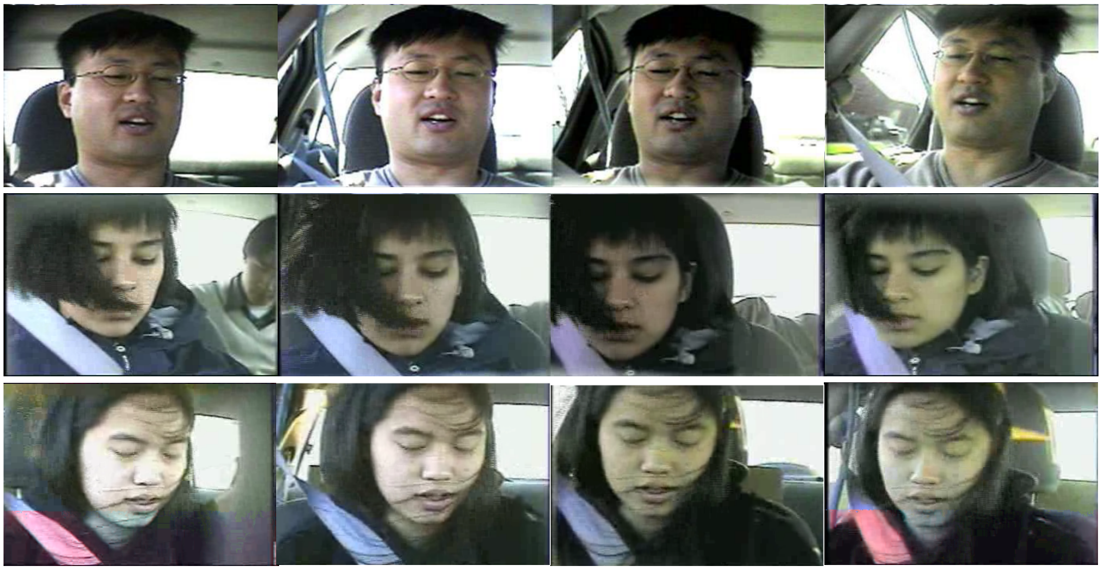


Figure 5.11: Examples of captured visual data from the AVICAR database. Each row shows a simulateous capture from cameras SL, CL, CR and SR (from left to right). The top row shows a example of a person without occlusion, while remainding show examples of occlusions which can occur when the vehicle is moving with windows down.

microphones on the passenger's sun-visor and a 4-camera array positioned on the dash, with each camera aimed towards the passenger to capture the different views of the face. While the cameras were not labelled in the AVICAR database, we have labelled them as SL, CL, CR and SR according to whether they are side (S) or central (C) cameras and left (L) to right (R) according to the cameras' view point. All audio channels were recorded using 16 bit resolution at a sampling rate of 48kHz and down sampled to 16kHz after segmentation of individual utterances. The video streams are combined using a multiplexer in order to be stored in a single file for each utterance with 29.97 frames per second with each camera having an individual resolution of 360×240 pixels. Some examples of a speaker captured simultaneously captured from the four cameras are shown in Figure 5.11.

The AVICAR database consists of audio and video recording of 100 speakers (50 male and 50 female). However, the released portion of the AVICAR database contains less

Noise	Description
35U	Car traveling at 35mph and windows closed
35D	Car traveling at 35mph and windows open
55U	Car traveling at 55mph and windows closed
55D	Car traveling at 55mph and windows open
IDL	Car stopped and engine idling

Table 5.1: Noise Conditions in the AVICAR database

data than documented in [98], with full audio only included for 87 speakers and video for 86. All of the recorded speech is in English. Most of the speakers are American English speakers, with the reminder of speakers from Latin America, Europe, East or South Asia. The AVICAR database is also recorded across five distinct recording conditions, as shown in Table 5.1, based upon the speed the car was traveling and whether the windows were up or down. An idle condition is also provided based upon the car idling in park to serve as a baseline.

For the experiments performed in this Chapter, the audio-visual speech data is taken from the phone numbers portion of the AVICAR database across all noise conditions. This portion consisted of two sessions of 10 digit utterances for each speaker and noise condition. The phone number digit sequences were identical across all subjects with all digits used for each 10 digit phone number. Subjects were instructed to pronounce the digit 0 as ‘zero’ in session 1 and ‘oh’ in session 2.

5.3.1 AVICAR Protocol

The AVICAR database captures the types of variabilities that could be expected in a vehicular scenario, there has been very little work performed on the visual portion of the dataset apart from [71] and no work done on the audio-visual portion. This is due to the poor synchronisation of the audio-visual data which limits its use. A significant contribution for the research community is to resolve these synchronisation issue as the

AVICAR dataset presents itself as an ideal test-bed to evaluate and investigate different strategies. As a result of this work, a novel speaker-independent audio-visual protocol is presented in this section(which has been lacking). This protocol[1] can be used to facilitate further research in the area of AVASR in a vehicular environment.

Audio-video alignment

The main problem with the AVICAR database is that the audio and video streams were not synchronised. Therefore a significant amount of work has been conducted to find the correct timing details in the video files for corresponding phone utterances and this task was only considered in this protocol.

The following steps were followed to align the audio and video information.

1. Locate the positions (phone numbers 0 to 9) of the selected audio files in the full-session audio (i.e. raw audio file).
2. Align the full-session audio beep track with the full-session beep track contained in the video files.
3. Use the previous alignment to determine the position of the video frames that correspond to the audio files.

Development of the audio-visual protocol

Initially the speakers were selected according to the protocol developed by Kleinschmidt et. al [92] by selecting speakers according to the following conditions:

[1] The synchronise timing information is available for the researches by contacting {rajitha.navarathna@student.qut.edu.au}

Condition	Subjects
IDL	AF3, AM4, AM5, BF1, BF2, BF5, BM1, BM2, BM3, BM4 CF5, CM1, CM3, DF1, DF3, DF4, EM2, FM4, FM5, GF1 GF3, GF4, HF1, HF2, HF3, HM4, IF3, JF1, JF4, JM2
35U	BF1, BF5, BM1, BM3, BM4, CF1, DF1, DF3, DF4, DM2 DM3, EF3, EF4, EM1, EM2, EM3, FF2, FM3, GF3, GF4, HF3 HF5, HM1, HM4, IM4, IM5, JF2, JF4, JF5, JM4
35D	AF3, AM2, AM4, AM5, BF2, BF5, BM3, CF1, CM1, DF4 DM2, EF3, EM1, EM3, FF2, FM3, FM5, GF1, GF3, GF4 HF2, HF3, HM1, HM3, HM4, IM5, JF2, JF4, JF5, JM4
55U	AF2, AM2, AM4, AM5, BF2, BF5, BM2, BM3, CF1, CM1 DF1, DM3, EM3, FF2, FM5, GF1, GF3, HF1, HF3, HM1 HM3, HM4, IF1, IF3, IM4, IM5, JF1, JF4, JM2, JM4
55D	AF3, AM3, AM4, AM5, BM2, BM3, BM4, DF1, DF3, DF4 DM2, DM3, EF3, EM3, FF2, FF5, FM2, FM5, HF1, HF2 HF3, HF4, HF5, HM1, HM4, IF3, IM4, JF1, JF4, JM2

Table 5.2: Develpoed AVICAR protocol speaker list

- A single phone number utterance in any noise condition must have all microphones in working condition (i.e. audio exists, and not considered poor due to hardware failure).
- The corresponding raw-audio file should exist and it should be in working condition.

In order to develop an audio-visual protocol, the following condition was added.

- The video should be in working condition. (i.e. video exists, and be of reasonable quality).

The resulting speaker groups are listed in Table 5.2. Since the effect of noise is the most important parameter, speakers with only one condition of data available were still included in the final list. The final list consisted of 3000 phone-number utterances with 150 video files (30 videos per condition).

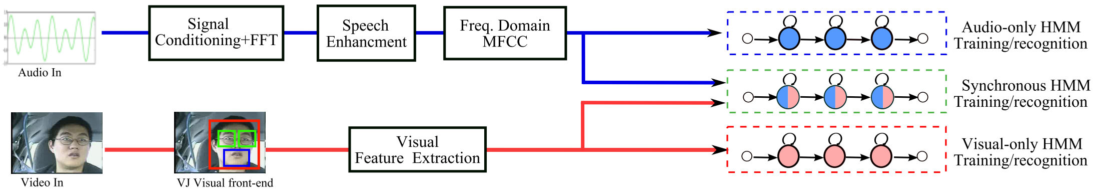

Audio In
Signal Conditioning+FFT
Speech Enhancment
Freq. Domain MFCC
Audio-only HMM Training/recognition
Synchronous HMM Training/recognition
Video In
VJ Visual front-end
Visual Feature Extraction
Visual-only HMM Training/recognition

Figure 5.12: An overview of the experimental design, showing both the acoustic and visual speech recognition systems in combination with the audio-visual SHMM approach.

5.4 Experimental Configuration

In order to evaluate the performance of visual and audio-visual speech recognition when mulitple camera view ports were available, a series of experiments will be performed in Sections 5.7 and 5.8 of this Chapter comparing multiple video-stream based SHMMs with single-stream acoustic and video approaches. A basic overview of this experimental design is shown in Figure 5.12, showing both the acoustic and visual speech recognition systems in combination with the audio-visual SHMM approach. This section will outline the experimental design of the AVASR system, beginning with an overview of the evaluation protocol which was used to conduct the speech recognition experiments. The specifics of the acoustic, visual and audio-visual speech recognition approaches will then be outlined.

5.4.1 Evaluation Protocol

Speech recognition experiments were conducted with native English speakers in the AVICAR database according to the protocol developed in [130]. This protocol divided the available audio-visual speech portions of the database into 6 groups with a non-native English group covering 3000 phone-number utterances across 150 separate video sessions.

Group	Speakers
I	BF5, BM4, CF5, DF1, EF4, EM1, FF2, HF2, HM3, IF1
II	AM3, AM4, BM1, CM1, DM2, EM4, FF5, HF3, IM5, JM2
III	AM5, BM3, CF1, DF4, EM2, FM2, GF1, HF5, JF1
IV	AF2, BF2, DF3, EF3, EM3, FM4, GF5, HF1, HM4, JF5
V	AM2, BF1, EF5, FM5, GF4, HM1, IM4, JF4

Table 5.3: Native-English speaker groups available in the AVICAR audio-visual evaluation protocol [130].

Fold	Training	Evaluation	Testing
1	I, II, III	IV	V
2	I, III, IV	V	II
3	I, IV, V	II	III
4	I, II, III	V	IV
5	II, III, V	IV	I

Table 5.4: Five folds were used across the evaluation protocol to ensure that all speakers were available for testing.

The experiments were conducted using only native-English speakers due to the limited number of non-native-English speakers in the protocol list. The selected native-English subjects were grouped into five groups as presented in Table 5.3. These groups were further divided into five separate folds as shown in Table 5.4, to ensure that all speakers are included at least once in a testing partition. In the Bowen Lee's doctoral dissertation [97], the author has used a five validation fold system by selecting six groups for training two groups for testing and two groups for validation in the audio domain. However, due to the limited number of utterances in the protocol, for each of the five validation folds, three groups of speakers were selected for training, the fourth for validation and system tuning, and the fifth group was used for testing of the AVASR system.

Speech recognition performance measure

All speech recognition results quoted in this paper are HTK-style [205] word accuracies for the small-vocabulary task of connected digits (in %) collated by noise condition, with the average results presented over all folds. The word recognition accuracies are calculated using:

$$\text{Accuracy} = \left(1 - \frac{D + S + I}{N}\right) * 100\% \qquad (5.7)$$

where N is the true number of words, D the number of deleted words, I the number of inserted words and S the number of substitutions.

5.4.2 Acoustic Speech Extraction

AVASR studies to date have generally concentrated on improving the quality of the visual information [152], and inherently assume that adding visual information to the ASR system will improve its robustness under noise. Moreover, it is assumed that the inclusion of visual information will be superior to (or at least comparable to) speech enhancement performed on the audio channel. One of the only examples where this comparison was made directly was in [45] where an AVASR system was presented incorporating spectral subtraction [25]. This system showed significant benefits in combining speech enhanced audio and visual speech information.

Since this study, there have been a number of advances in both speech enhancement and AVASR. Apart from this work, there has been no detailed study reporting AVASR with more modern speech enhancement techniques and audio-visual feature extraction and fusion techniques. Even though the focus of this Chapter is in the visual domain, we wanted to ensure that full use was made of all seven microphones for the acoustic speech extraction, so that the acoustic speech recognition performance provided both a sensible baseline for visual speech recognition, and provided the most improvement in fusion for AVASR. Initially this section reviews speech enhancement techniques which are used in this Chapter followed with acoustic feature extraction

Enhancement techniques can be broadly classified by the number of microphones used. Single-channel techniques are well suited to a number of applications, for example where hardware costs are a key factor. Multi-channel techniques, whilst increasing hardware requirements, can reduce the distortion introduced by single-channel techniques through the use of spatial filtering [19], which consequently improves ASR performance in comparison.

Spectral subtraction

Spectral subtraction (first proposed by Boll [25]) aims to estimate the spectrum of the clean speech signal by subtracting an estimate of the noise spectrum from that of the noise-corrupted speech. Subtraction typically takes place in the magnitude or power spectrum assuming that the noise and speech signals are statistically independent and can therefore be regarded as being added acoustically.

The generalised frequency-domain spectral subtraction rule derived from [21, 25] is defined as:

$$|\hat{S}_t(f)|^\gamma = |Y(f)|^\gamma - \alpha(f)|\hat{D}(f)|^\gamma$$

$$|\hat{S}(f)|^\gamma = \begin{cases} |\hat{S}_t(f)|^\gamma & |\hat{S}_t(f)|^\gamma > \beta|Z(f)|^\gamma \\ \beta|Z(f)|^\gamma & \text{otherwise} \end{cases} \tag{5.8}$$

where $|\hat{D}(f)|$ is the estimate of the noise spectrum, $|Z(f)|$ is either the instantaneous noisy speech signal magnitude $|Y(f)|$ or the noise magnitude estimate, and γ determines the spectrum the subtraction takes place in; either magnitude ($\gamma = 1$) or power ($\gamma = 2$). The frequency-dependent subtraction factors, $\alpha(f)$, compensate for over- or under-estimating the noise spectrum, and β is the noise floor factor which ensures the clean speech spectrum cannot become negative. A number of variations to this generalised subtraction rule have been proposed in literature, including subtraction in the power spectral domain [21], and multi-band spectral subtraction (MBSS) [89]. The MBBS method determines the subtraction factors $\alpha(f)$ using the local signal-to-noise ratio, and an additional subtraction factor, $\delta_b(f)$ is introduced to each pre-defined frequency band, b. This technique was designed to improve speech intelligibility as opposed to ASR performance.

Delay-Sum beamforming

Multi-channel beamforming combines the acoustic signals from all microphones to perform spatial filtering which differentiates the signal of interest from the background noise based on propagation delays between the source and each microphone. Having compensated for the delays, microphone channels are individually weighted and combined in order to reinforce the speech signal. This is referred to as filter-sum beamforming which is represented as:

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} G_n(k)Y_n(k)\exp^{-j2\pi k\Delta_n} \tag{5.9}$$

where N is the number of microphones, $Y_n(k)$ is the signal received at the n^{th} microphone, $G_n(k)$ are the filter coefficients ($G_n(k) = 1$ for Delay-Sum Beamforming (DSB) [19]), and the exponential term is compensation for the delay Δ_n.

The acoustic features were extracted using four speech enhancement techniques based on those described in above and applied as per Figure 5.12, plus a baseline system without speech enhancement. In particular, the 39-dimensional acoustic features used in this experiment were:

(a) Baseline MFCC

(b) MFCC with Spectral Subtraction (SpecSub) according to Equation (5.8)

(c) MFCC with Kamath & Loizou's MBSS [89]

(d) MFCC with Dual-channel delay-sum beamforming (2-ch DSB)

(e) MFCC with 7-channel delay-sum beamforming (7-ch DSB)

For each set of acoustic features listed in the above, all acoustic speech was represented using 39-dimensional audio features (13 MFCC including C_0, plus deltas and accelerations coefficients), captured every 10 ms from 25 ms windows.

5.4.3 The VJ Based Visual-Front End

An efficient visual front end system which is able to track and locate the speaker's face and mouth ROI was developed using the VJ algorithm [194]. Initially, the classifiers for the face, eyes and mouth were developed using the OpenCV libraries [2] as described in Patrick Lucey's doctoral dissertation [112]. The overall visual frond-end system was developed using Microsoft Visual C++ to detect the face and extract the mouth ROI. An overview of the front-end system is presented in Figure 5.13.

Given the video of a speaker, initially the system detects the face using the face classifier. Once the face was located, we then locate the eyes and based on these locations, the face was similarity normalised (i.e. normalised with respect to scale, rotation and translation) based on an inter-ocular distance of 32 pixels. We then applied a mouth classifier and from that we extracted a ROI to be used in visual speech recognition. The extracted mouth region mostly contains jaw and cheeks and it was downsampled

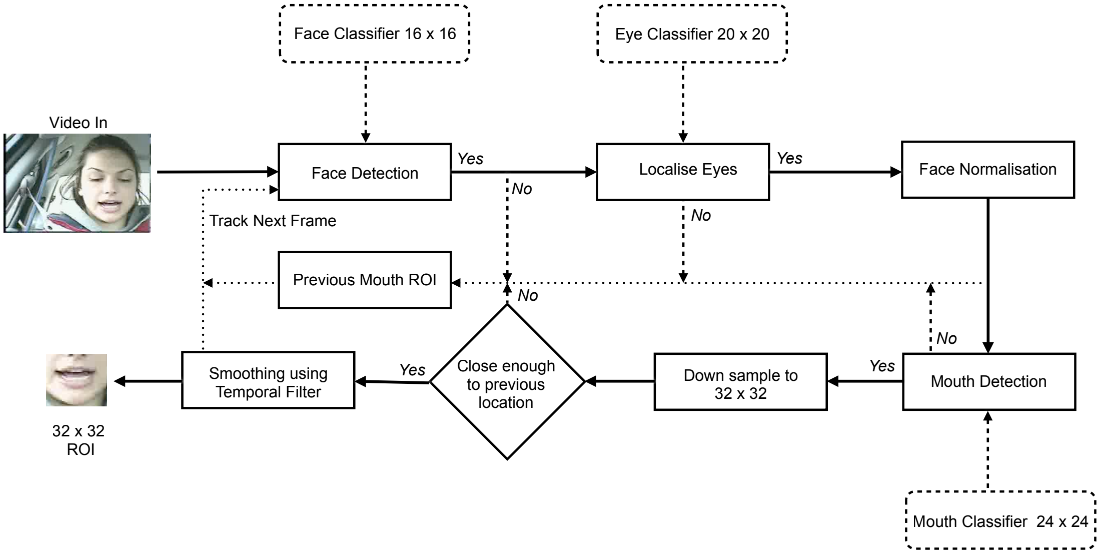
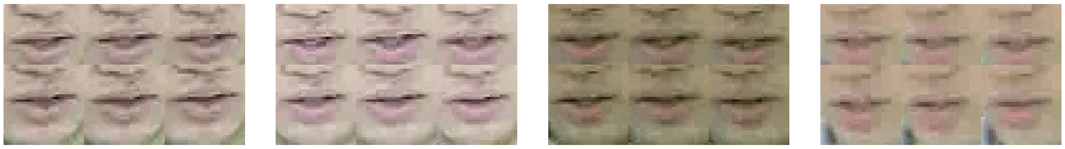


Figure 5.13: Block diagram for the visual front end system to detect the face and mouth region of a speaker.



(a) Camera Side Left (SL)

(b) Camera Centre Left (CL)

(c) Camera Centre Right (CR)

(d) Camera Side Right (SR)

Figure 5.14: Examples of the extracted mouth ROI images from each camera in the AVICAR database.

to 32×32 to keep the dimensionality low[2]. The 32×32 mouth region is smoothed using a mean filter. Following the ROI localisation, this process was performed over consecutive frames. The previous ROI location is used if the detected ROI is too far away from the previous frame. We used the same tracker to capture the mouth region across all noise conditions and cameras. Some examples of tracked lip regions are shown in Figure 5.14.

[2] Jordan and Sergeant [88] observed that no significant effects with image sizes for visual speech recognition performance

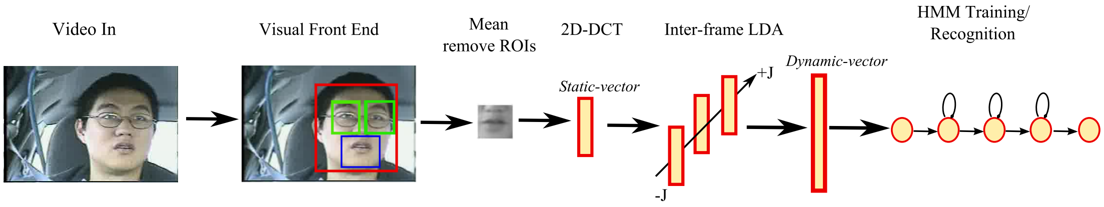


Figure 5.15: An overview of the visual feature extraction system.

5.4.4 Visual Speech Extraction

Following the tracking of the mouth ROI, the visual features were extracted using the cascading appearance based feature extraction process, as shown in Figure 5.15. As described in Chapter 5.2.3 initially, image mean normalization was performed to remove any irrelevant information, such as illumination or speaker variances. Then a 2D-DCT was applied to the mean-removed image and the top 100 higher-energy components were selected in zig-zag pattern to capture the static visual speech information.

Subsequently, in order to incorporate dynamic speech information, seven of these neighbouring static feature vectors over ±3 adjacent frames were concatenated, and were projected via an inter-frame LDA step to yield a 40-dimensional 'dynamic' visual feature vector. The classes used for LDA matrix calculation were HMM states, based on forced alignment of a separately-trained audio-only HMM.

5.4.5 Speech Modelling

Audio and single-camera visual word models were trained using 9-state left-to-right HMMs across all speakers and noise conditions in the training portions of each fold to enable speaker independent speech recognition. Similarly, for the multiple-camera visual and audio-visual speech recognition experiments, multiple-stream 9-state left-to-right SHMMs were trained across all speakers and noise conditions. For both the single and multiple-stream models, each stream was represented by an 8-mixture GMM within each state of the word models.

For the audio-visual experiments, in order to allow the acoustic and visual features to be aligned, the visual features were upsampled to match the acoustic feature rate of 100Hz using nearest neighbour interpolation to synchronise with the audio signal.

Because SHMMs allow for weights to be applied to each individual stream during evaluation, we wished to investigate within this Chapter the effect that weighting the acoustic stream and the various cameras will have on the final visual and audio-visual speech recognition performance. To that end we wanted to investigate the effect of three weighting parameters:

- α_L, the proportional weighting of left cameras in comparison to their right partners,
- α_C, the proportional weighting of central cameras in comparison to their side partners, and
- α_A, the proportional weighting of the acoustic stream in comparison to the video streams.

Each of these proportional weights would be balanced by its partner, such as the proportional weighting of the right-hand camera which would be represented by $(1-\alpha_L)$. By experimentally determining these proportional weighting parameters, the process of which will be outlined later in this Chapter, the final weights in a 5-stream (audio, SL camera, CL camera, CR camera, SR camera) can be determined as follows

$$w_A = \alpha_A, \tag{5.10}$$

$$w_{CL} = (1 - \alpha_A) \times \alpha_C \times \alpha_L, \tag{5.11}$$

$$w_{CR} = (1 - \alpha_A) \times \alpha_C \times (1 - \alpha_L), \tag{5.12}$$

$$w_{SL} = (1 - \alpha_A) \times (1 - \alpha_C) \times \alpha_L, \tag{5.13}$$

$$w_{SR} = (1 - \alpha_A) \times (1 - \alpha_C) \times (1 - \alpha_L). \tag{5.14}$$

For situations where a particular proportional weighting parameter is not relevant, it will not be applied to the weights of that particular SHMM. An example of this would be for visual only experiments, where the $(1 - \alpha_A)$ proportional weight would not be applied to every video stream as it would have no effect on the final visual-only speech recognition performance.

5.5 Acoustic Speech Recognition

The audio-only ASR results with each of the speech enhancement algorithms are shown in Table 5.5. These results demonstrate a clear improvement over baseline MFCC performance by applying speech enhancement for all noise conditions. The 7-channel DSB technique outperforms all other speech enhancement techniques, with the dual-channel system providing the next best overall ASR performance. This result is consistent with the belief that microphone-array based speech enhancement is superior to single-channel techniques which typically distort the desired signal, and have access to less information about the audio signal [19]. It is also important to note that the spectral subtraction algorithm described by Equation (5.8) outperformed Kamath & Loizou's multi-band spectral subtraction [89] when both algorithms were empirically optimised.

Evaluation Algorithm	Word Accuracy (%)				
	IDL	**35U**	**35D**	**55U**	**55D**
Baseline MFCC	61.0	41.9	36.9	32.8	23.7
MFCC with SpecSub	61.9	45.3	43.0	36.7	25.8
MFCC with Kamath	62.4	43.2	42.5	35.7	25.8
MFCC with 2-ch DSB	63.4	48.5	43.4	39.1	27.4
MFCC with 7-ch DSB	64.1	47.8	45.9	41.6	29.8

Table 5.5: Audio speech recognition baseline evaluation results for phone number task

In addition to the above observation, it can be seen that increase in speed causes degradation in the recognition accuracy. The word accuracy of the windows down condition is less compared to the windows up in the same speed condition for all the algorithms. With windows open, greater decreases in accuracy occur as the speed increases. The **IDL** condition shows good performance accuracy, due to there being less acoustic noise. In the **55D** condition the word accuracy is poor. This is mostly due to increases in road and wind friction as vehicle speed increases.

5.6 Visual Speech Recognition using FLK

Initially, this cChapter compared the effect of visual-front using VJ face detector and semi-automatic FLK (by manually inspection we re-update the template to avoid subsequent failures;. this isWe defined this as semi-automatic FLK.) in terms of visual speech recognition. We selected a small subset using CL camera from the AVICAR protocol described in Chapter 5.3.1. We extracted the lip images using VJ and FLK approach. We implemented the FLK inverse compositional template tracking method to find the best match to the template mouth region in every subsequent frame in the given video. The template is updated in every frame and by manually inspecting the extracted ROI, if it is too far from the actual mouth region we manually re-initialise the current template. We defined the weighting matrix S in Equation 4.16 using a bank

	Visual HMM accuracy(%)				
	IDL	**35U**	**35D**	**55U**	**55D**
VJ	37.93	37.41	38.83	36.14	33.57
semi-automatic FLK	40.02	38.22	39.82	37.64	36.48

Table 5.6: Word recognition accuracy of FLK-based vs VJ based front-end on AVICAR database organized into different noise conditions

of Gabor filters with 9 scales times 8 orientations [72].

From the selected subset from AVICAR protocol, 70% were used for training and the remaining 30% were used for testing. We conducted visual speech recognition experiments using 16 states and 8 mixtures. Table 5.6 shows the visual HMM performance using the two approaches. As it can be seen in this table, the FLK-based mouth detection outperformed the VJ method on AVICAR database with an average of 4.51% relative improvement in visual SHMM accuracy. Even though this improvement has an effect for AVASR, the limitation of the FLK based approach is the re-initilisatione of the template frame after the detection fails. This effect will be applicable to the data from the other camera angle too.

5.6.1 Visual Speech Recognition using FLK: Discussion

We can conclude that the FLK approach for tracking mouth region is slightly better than the VJ approach and moreover, it has better fitting performance. Even though the FLK approach is more robust in handling the illumination variation, it should be noted that if the detection fails completely at one frame, then the wrong template will be used in the subsequent frames and the whole tracking process will be failed from that time onwards. This is not the case for VJ, as it performs detection on each frame independently. This is a major limitation of the FLK approach for visual speech recognition. Due to this limitation, we selected the VJ approach for further visual speech recognition experiments and AVASR experiments from single and multiple cameras.

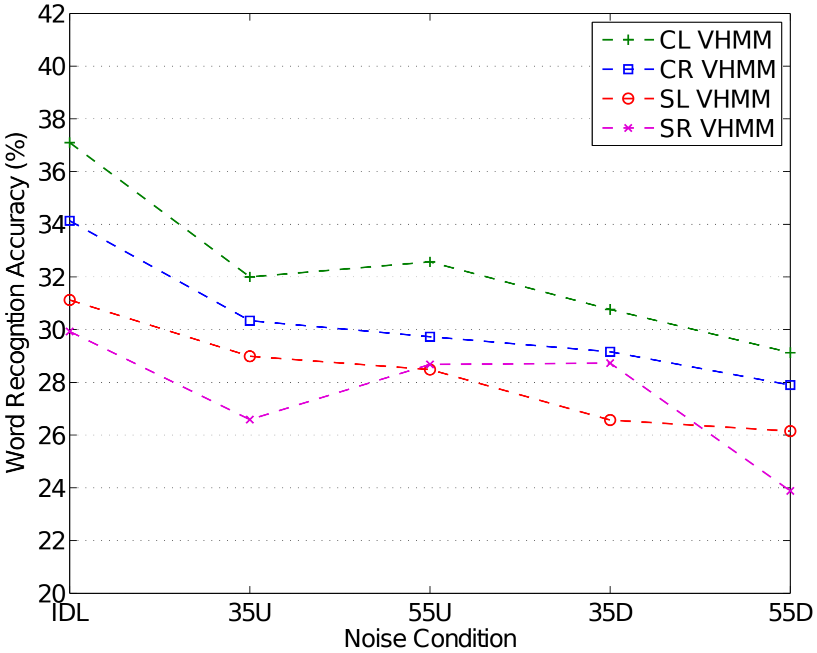


Figure 5.16: Accuracy of single-camera visual HMM speech recognition across all noise conditions, averaged across all validation folds.

5.7 Visual Speech Recognition using Multiple Cameras

Before investigating full audio-visual speech recognition, initially the experiments were conducted to investigate the best visual speech performance that could be obtained from a weighted combination of cameras in a single visual SHMM. In this section, we will begin by investigating the visual speech recognition performance that can be obtained from each single camera, and how the four cameras can be combined to provide the best visual speech recognition performance on the AVICAR database.

5.7.1 Single Cameras

Most approaches to AVASR in the existing AVASR literature [27, 38, 116, 133, 150, 153, 189], have only considered sources of simulated acoustic degradation and used the same video data in every reported noise condition. However, because the noise conditions in the AVICAR database correspond to real-world noise conditions that affect both the acoustic and visual modalities, we have the advantage that we can investigate what effects a typical driving environment would have on the visual as well as the acoustic modality.

The first set of experiments that we ran allowed us to investigate the effect that each driving condition had on visual speech recognition. These experiments trained and evaluated a series of single-stream visual HMMs (using all 40 visual features) for each word in the AVICAR database in order to evaluate speaker independent visual speech recognition across all noise conditions and cameras available in the AVICAR database.

The results of these experiments, shown in Figure 5.16, indicate that the visual-only speech recognition accuracy using the AVICAR dataset differ considerably according to the noise condition, and all results are obviously diminished from ideal laboratory conditions where word level accuracies are typically around 60-70% [149].

By comparing the general visual speech recognition performance across all noise conditions, it can be seen that the visual speech recognition performance is affected by the driving condition, but not to the extent that would be expected in the acoustic modality. All moving conditions provide poorer visual speech recognition performance than the IDL condition, most likely attributable to difficulties in tracking the mouth ROI and changes in illumination as the car is moving and the speaker within it. There is little difference in visual speech recognition performance as the speed is increased from 35 to 55 mph, but having the windows down degrades performances, largely due to self occlusions from subject hair as the wind penetrates the automotive cabin.

In addition to the general degradation in noisier conditions, the visual-only speech recognition results can be seen to also differ according to the location of the cameras from the AVICAR database. While the side cameras still provided a near-frontal view of the speaker's face, the performance of the two side cameras can be seen to be degraded with reference to the central cameras, which had the most frontal view of the speaker's face. While there is no clear difference between the left and right side cameras across all the noise conditions, the left central camera did outperform the right central camera across all noise conditions, suggesting that it had the best view of most speakers' faces throughout the database.

5.7.2 Multiple Cameras

Motivated by the similar performance levels obtained by the single camera visual speech recognition experiments, we wished to determine if the cameras contained complementary information that could be combined in fusion to provide improved visual speech recognition performance when multiple views of the speaker's face are available. However, in order to investigate this question, we first had to determine the weighting parameters w_{SL}, w_{CL}, w_{CR} and w_{SR} that would be optimal for a four-stream SHMM.

In Equations 5.11-5.14, we showed that the individual stream weights for a visual SHMM can be expressed in terms of the proportional weighting parameters α_L showing the proportion of left cameras in comparison to right and α_C indicating the proportion of central cameras in comparison to side (the α_A terms can be discarded as we are only dealing with video).

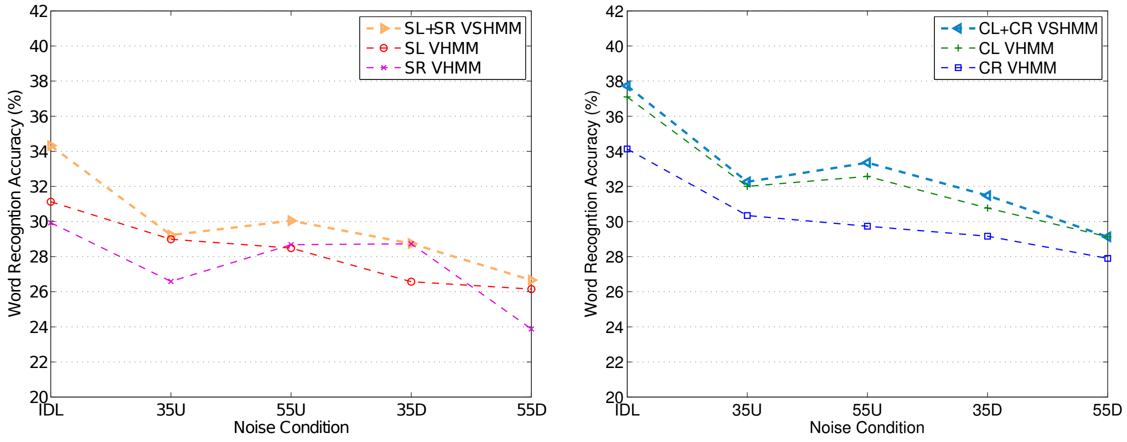
In order to arrive at the final four-stream visual SHMM, we will first investigate two-stream fusion where the left and right pairs are fused into two separate dual-stream central and side visual SHMMs according to the best performing α_L parameter. We will then investigate the proportional weighting parameter α_C between the central and side visual SHMMs to arrive at the final best performing four-stream visual SHMM.

It should be noted that due to limitations in the HMM toolkit software [205], and to limit the processing requirements, only the first 10 video features from each of the streams (out of 40 extracted) were used in the visual fusion experiments. A limited set of development experiments were also performed of visual fusion with 15 and 20 from each stream, but little degradation in speech recognition performance occurred, so 10 features were chosen as the best trade-off between accuracy and processing time for the full-scale experiments.

Two camera fusion

In order to combine the two left and right camera pairs into a side visual SHMM and a central visual SHMM, the proportional weighting parameter between left and right cameras, α_L was determined by performing a set of two-stream visual speech recognition experiments between the central left and right cameras on the first fold of the AVICAR protocol. The best value of α_L, chosen from the set $\{0.0, 0.1, \ldots, 1.0\}$ to maximise the average visual speech recognition accuracy across all noise conditions, was found to be $\alpha_L = 0.5$ which also had the simple advantage of keeping the left and right streams equally weighted.

Experimental results of the combination of visual streams into the side and central two-camera SHMMs are shown in Figure 5.17. While the two-stream fusion approach taken here does appear to provide a small improvement across all conditions on the individual cameras, the improvement is not large, suggesting that this is likely attributable to little complementary information in the left and right views of a speaker's face, when taken from similar angles.



(a) Side cameras

(b) Central cameras

Figure 5.17: Accuracy of two-camera visual SHMM speech recognition in comparison to the single-camera approach using (a) side and (b) central cameras across all noise conditions, averaged across all validation folds.

Four camera fusion

Knowing the best value of the left/right proportional weighting parameter α_L, the only other parameter required for four-camera visual SHMM speech recognition is the proportional weighting between the central and side cameras, expressed as α_C and $(1 - \alpha_C)$ respectively. Similar to the set of development experiments conducted to calculate α_L, a set of four-camera visual SHMM experiments were conducted on the first fold of the AVICAR protocol. The α_L was set to 0.5 and the best value of α_C was chosen from $\{0.0, 0.1, \ldots, 1.0\}$ to maximise the average visual speech recognition accuracy across all noise conditions. The results of this development experiment, shown in Figure 5.18, found that the best speech recognition performance was obtained when $\alpha_L = 0.7$.

By combining the chosen proportional weighting parameters of $\alpha_L = 0.5$ and $\alpha_C = 0.7$, and setting $\alpha_A = 0$ as this a visual-only SHMM, it can be seen that the stream

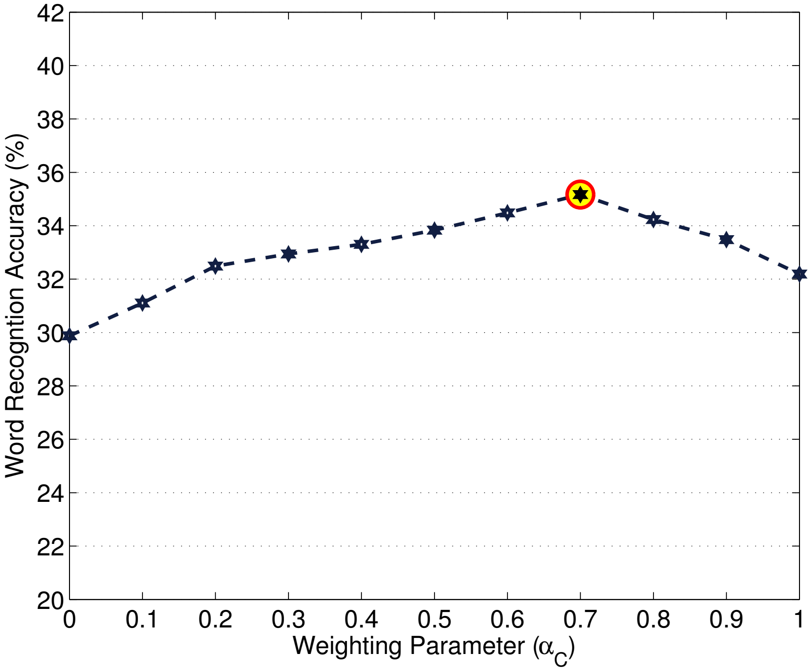


Figure 5.18: Average accuracy across all noise condition of four stream VSHMM speech recognition as the central/side proportional weighting parameter α_C is varied from 0.0 to 1.0 on the validation fold.

weights in the final four stream SHMM will be given by

$$w_{CL} = \alpha_C \times \alpha_L \qquad = 0.35, \tag{5.15}$$

$$w_{CR} = \alpha_C \times (1 - \alpha_L) \qquad = 0.35, \tag{5.16}$$

$$w_{SL} = (1 - \alpha_C) \times \alpha_L \qquad = 0.15, \tag{5.17}$$

$$w_{SR} = (1 - \alpha_C) \times (1 - \alpha_L) \qquad = 0.15. \tag{5.18}$$

Experimental results of the fully weighted four-camera visual SHMM (Figure 5.19), show that in comparison to the small improvement of left/right camera fusion, the fusion of the central and side camera pairs provides a considerable improvement over the both the side and central camera-pair fusion accuracies. These results demonstrate that there is considerable complementary information available in differing views of a speaker's face, even when the difference in viewing angle is relatively small, as it is in the AVICAR database. In particular, these results are showing that even though the side

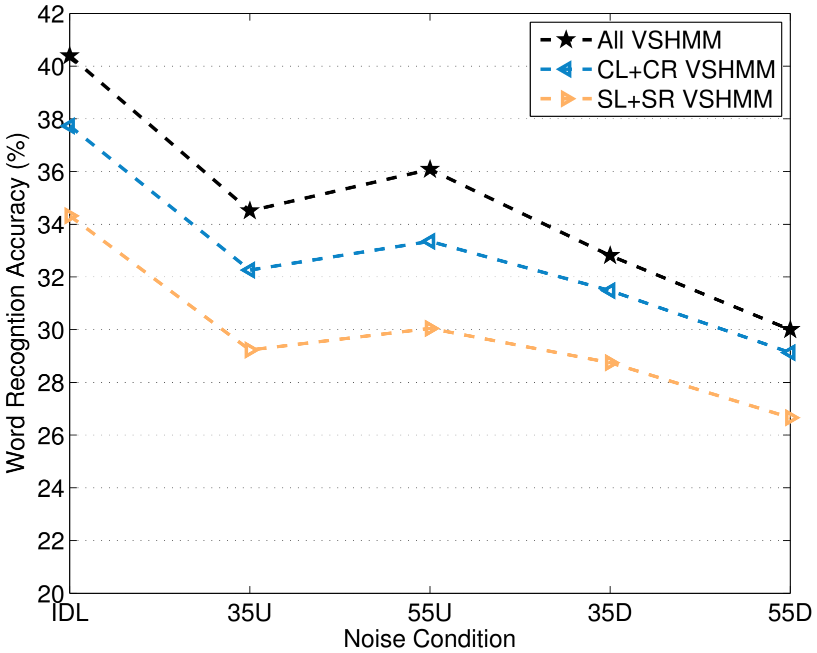


Figure 5.19: Accuracy of four-camera visual SHMM speech recognition in comparison to the two-camera side and central SHMMs across all noise conditions, averaged across all validation folds.

cameras are performing poorly in comparison to the central cameras, there is important visual speech information in the side cameras' views that is not being captured by the central cameras and that can, in fusion, provide improved visual speech recognition performance.

5.8 Audio Visual Speech Recognition using Multiple Cameras

In the previous section, we showed that the best performing visual speech recognition system on the AVICAR database can be obtained from a fusion of all four cameras. However, while the four-camera visual HMM provided much better performance than any of the single camera visual speech recognition experiments, the visual speech per-

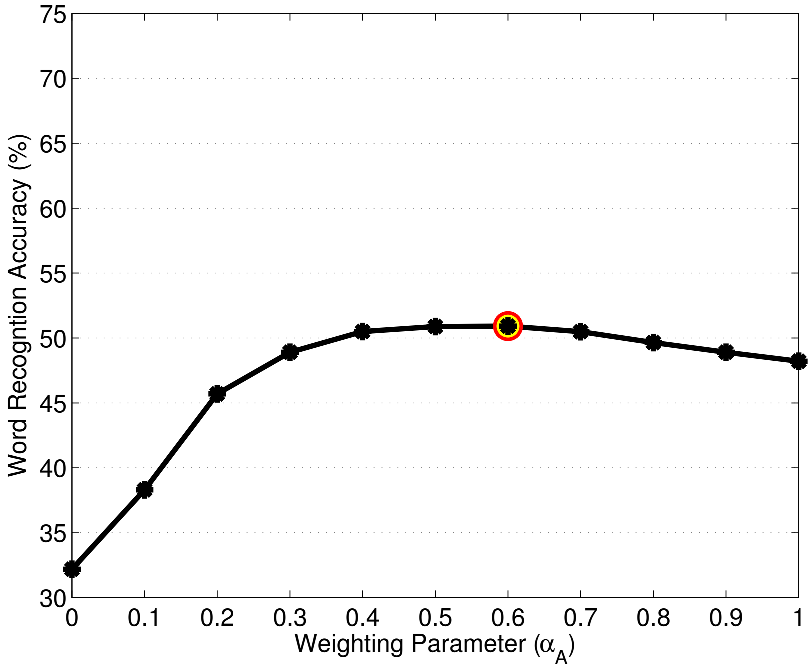


Figure 5.20: Average accuracy across all noise conditions for audio-visual SHMM speech recognition as the audio/visual proportional weighting parameter α_A is varied from 0.0 (video only) to 1.0 (audio only) on a validation folds.

formance is still fairly poor in comparison to what would be expected from a normal acoustic speech recognition system. This section will extend the visual-only speech recognition experiments in the previous section by incorporating audio as fifth stream to the SHMM structure, comprised of the multiple-microphone acoustic signal converted into a single MFCC feature stream through the DSB process.

5.8.1 Single Cameras

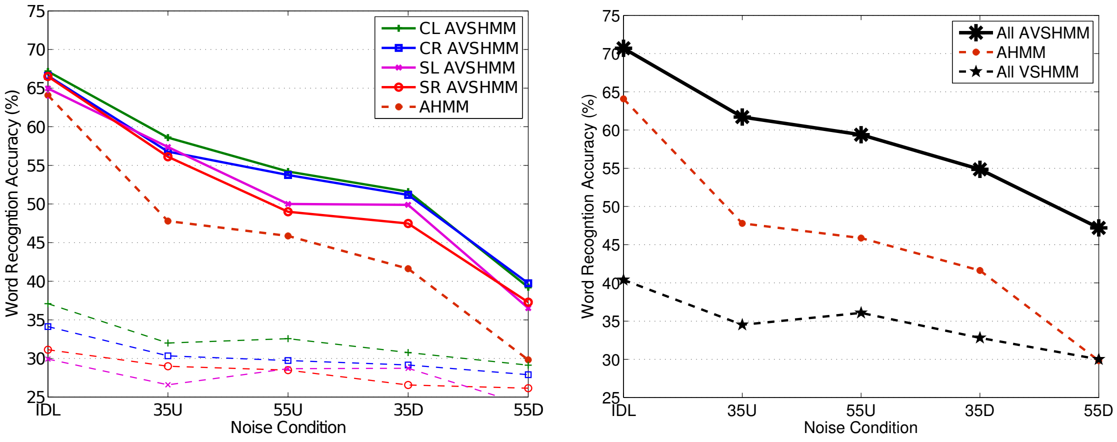
Before investigating the full five-stream audio-visual SHMM, we first wished to investigate the performance that could be obtained in a one-camera two-stream audio-visual SHMM, and investigate the value of the proportional weighting parameter α_A that will provide the best audio-visual speech recognition performance. This investigation was performed by choosing α_A from $\{0.0, 0.1, \ldots, 1.0\}$ to maximise the audio-visual speech recognition performance, averaged across all noise conditions, when the acoustic features are combined with camera CL in a two-stream audio-visual SHMM on the first fold of the AVICAR protocol. In order to demonstrate the robustness to multiple noise conditions, the weighting parameter α was empirically chosen. The results of this experiment, shown in Figure 5.20, found that the best audio-visual speech recognition performance was obtained when $\alpha_A = 0.6$.

The single-camera audio-visual SHMM experiments, with the acoustic streams weighted at $w_A = \alpha_A = 0.6$ and the individual single-camera video streams at $w_V = (1 - \alpha_A) = 0.4$, are shown in Figure 5.21(a). By comparing these results to the audio and single-camera visual only approaches, also shown, it can be seen that even the worst-performing single-camera visual speech recognition systems provide for an improvement on the acoustic-only approach at all noise levels present in the AVICAR database, with the greatest improvement in the noisiest 55-mph, windows-down (**55D**) condition. This finding is in line with previous audio-visual speech recognition experiments [150] that have shown that even poorly performing visual-speech features can provide complementary information to acoustic features when combined using audio-visual SHMMs.

5.8.2 Multiple Cameras

Having chosen an appropriate proportional weighting parameter α_A for the acoustic stream, and showing that each of the single cameras can provide an improvement in audio-visual fusion, we finally had gathered enough information to construct the full five-stream audio-visual SHMM and demonstrate the improvements that can be obtained from multiple cameras for audio-visual speech recognition in noisy environments.

By combining the three chosen proportional weighting parameters of $\alpha_A = 0.6$, $\alpha_C =$



(a) Single camera audio-visual fusion. Visual-only results are shown similar to fusion, but dashed.

(b) Four camera audio-visual fusion.

Figure 5.21: Accuracy of (a) single-camera and (b) four-camera audio-visual SHMM speech recognition in comparison to audio and video-only baselines across all noise conditions, averaged across all validation folds.

0.7 and $\alpha_L = 0.5$, it can be seen that the stream weights in the final five-stream audio-visual SHMM will be given by

$$w_A = \alpha_A = 0.60, \tag{5.19}$$

$$w_{CL} = (1 - \alpha_A) \times \alpha_C \times \alpha_L = 0.14, \tag{5.20}$$

$$w_{CR} = (1 - \alpha_A) \times \alpha_C \times (1 - \alpha_L) = 0.14, \tag{5.21}$$

$$w_{SL} = (1 - \alpha_A) \times (1 - \alpha_C) \times \alpha_L = 0.06, \tag{5.22}$$

$$w_{SR} = (1 - \alpha_A) \times (1 - \alpha_C) \times (1 - \alpha_L) = 0.06. \tag{5.23}$$

Experimental results of the fully-weighted five-stream audio-visual SHMM (Figure 5.21(b)), show that the combination of all four cameras into the single five-stream audio-visual SHMM provides a considerable improvement when compared to the single-camera audio-visual SHMM results presented in Figure 5.21(a). Over all of the noise conditions in the AVICAR database, it can be seen that the multiple-camera audio-visual SHMM approach provides a considerable improvement in word-level speech recognition accuracy over all of the single-camera visual SHMM systems.

These results show that there are useful visual-speech cues available in differing views of speaker's mouth region that prove complementary both to each other and to the acoustic speech information for the audio-visual speech recognition application.

5.9 Chapter Summary

Audio-visual speech recognition has previously been shown to provide a considerable improvement over acoustic-only approaches in noisy environments, but most audio-visual speech recognition approaches have only been examined in relatively clean conditions and have rarely dealt with the visual variabilities such as head movement, poor/varying illumination and poor video resolution/quality. The research presented in this chapter extended upon the established audio-visual speech recognition literature to show that, in a real-world automotive environment, further improvements in speech recognition accuracy over traditional single-camera AVASR approaches can be obtained when multiple frontal or near-frontal views of speakers' faces are available. This Chapter review presents a comparison of the recognition performance of single channel and multi-channel enhanced speech, which was lacking in the audio-visual speech community. In addition to that, this Chapter compares the VJ face detector with the FLK approach in terms of fitting and visual HMM performance.

A series of visual speech recognition experiments conducted on the four-camera AVICAR database demonstrated that the best visual speech performance was obtained through a combination of all four cameras in a four-stream visual SHMM. Finally, combination of the four visual streams with a single acoustic stream in a five-stream audio-visual SHMM demonstrated a relative improvement of between 6% and 17% word-level accuracy over traditional single-camera AVASR, and between 9% and 56% relative improvement in word-level accuracy when compared to the acoustic-only approach.

We hope that this research effort will serve as a motivation for including multiple cameras in effective human-vehicle computer interfaces, which reduce driver distraction over manual alternatives. This research work has further demonstrated the usefulness of the AVICAR database for real-world speech recognition research, and it is hoped that researchers, including ourselves, will be able to continue to use the database as a common benchmark in improving the performance of audio-visual speech recognition in real-world environments.

Chapter 6

Analysis II: Long-Term Audience Analysis

Making a movie is an iterative process where multiple information sources are sought, obtained, analysed and then fed to the director, which they then use to gauge whether any changes to the movie are required before it is released to the public. One of the most important information sources that the director gets is via test screenings. Generally in a test screening, an audience of volunteers gather to watch the movie and after the viewing, each volunteer answers a questionnaire about certain aspects of the movie. Even though the gathered information is useful, these questionnaires are subjective, biased and do not contain specific time information.

This chapter seeks to gain an automatic real-time objective measure of audience through analysing the collective facial and body movements. Due to the complexity and difficulty of this task, no one has previously looked at this problem. In addition to introducing a new problem to the field of face and gesture analysis, as well as a solution on how to capture such data, there are numerous technical challenges highlighted in this chapter, for which solutions are then presented.

6.1 Introduction

The iterative process of releasing a movie to the public requires the accumulation and analysis of information. In the later part of the movie making process, the director uses this information to make improvements to the movie that are more likely to hold the public interest. One of the most important information sources that the director uses is acquired from test screenings. [1]

A test screening refers to a special showing of a movie before its release to the public, in order to gauge audience reaction and generally with an audience of volunteers or selected audience. After watching the movie, volunteers answer a questionnaire about certain aspects of the movie. Typically these questionnaires consist of: (i) How did you engage with the movie? (ii) What are the parts you enjoyed most? (ii) What are the parts you enjoyed least? This test screening is an interactive and iterative process (i.e it can go on for several test screenings).

The feedback from the volunteers can be used to better understand the movie and also can lead to making changes in some of the characteristics of the movie. These feedbacks may cause a very simple change, such as change to the title of the movie. One of the examples which can be found in the film industry is the movie *Licence to Kill*. The director changed the original name *Licence Revoked* to *Licence to Kill* after the test screening. The *Mary Poppins, Final Destination*, and *Titanic* are further examples of where the ending was changed after a test screening. The negative reactions from the test-screening audience, has caused film makers to remove some scenes from movies (eg: *The Pelican Brief, The Mighty Quinn*). The movie time was reduced in the movie

[1] http://everything2.com/title/Hollywood+test+screening+process

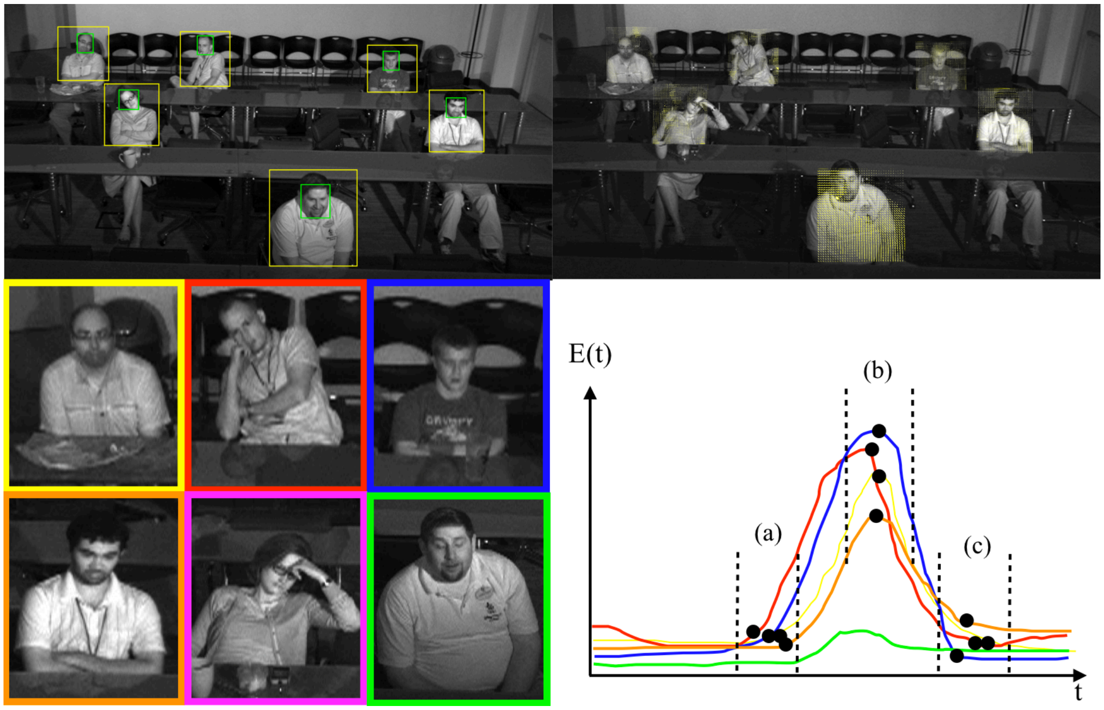


Figure 6.1: In our infra-red illuminated screening room, we use both face (top left) and body motion features (top right) to profile each audience member (bottom left) and find the synchrony or coherence of motion to analyze, summarize and predict audience ratings to movies (bottom right - each curve color corresponds to an audience member).

Clear and Present Danger after the feedback from the test-screening audience[2,3].

While the test-screening is useful, these questionnaires are subjective, biased (e.g. loyalty to the brand). Moreover, these questionnaires cannot be used in larger populations, such as children and people who have limited ability to communicate. In addition, questionnaires or self-report methods do not contain feedback at precise timestamps [171].

Having the ability to automatically and objectively measure group or audience behavior would have profound implications within the educational, marketing, advertising and behavioral science domains. However, due to the volume of data and the complex-

[2] `http://en.wikipedia.org/wiki/Test_screening`

[3] `http://uk.movies.yahoo.com/10-films-drastically-changed-after-test-screenings.html`

ities of the observed environments, the current de-facto standard of measuring audience/group behavior is still via self-report [14]. As self-report measures are subjective, labor intensive and do not provide feedback at precise time-stamps, an automated and objective measure is desirable. In an attempt to provide an objective measure, Madan et al. [117] utilized a wearable device which measured audio and head movement in addition to galvanic skin responses of a group interacting. Eagle and Pentland [62] developed a system to analyze group behavior using a PDA which required continuous user input. While both are interesting approaches, a less invasive and automatic solution is the goal of this Chapter.

In terms of measuring reactions to consumer products, nearly all ratings are via self-report (i.e“likes” or a Likert-type scale). Given enough crowd-sourced ratings (100k’s), useful measures can be obtained, which can be used to predict other scenarios where people may interested in their previous behaviors. Such *recommendation* systems are often based on matrix factorization approaches. *Pandora*[4] (songs), *Netflix*[5] (movies/tv-shows) and *Amazon*[6] (products) are popular examples for content-based and collaborative filtering approaches [94].

For movies, *rotten tomatoes*[7] has both critic and audience ratings which are crowd-sourced. Such information is only useful at a coarse level as it captures the overall global reaction to the stimuli and does not contain any specific local “interest” information. For long continuous time-series signals such as movies, knowing at which parts the audience (or sub-groups of the audience) *engage/disengage*, would be very beneficial to writers/directors/marketers/advertisers. Achieving this through self-report is subjective and difficult as it would require a person to consciously think and document what they are watching and subjects may miss important parts of the movie, due to distractions. Similarly, subjects could be instrumented with a myriad of wearable

[4] `pandora.com`

[5] `netflix.com`

[6] `amazon.com`

[7] `rottentomatoes.com`

sensors, but such approaches are invasive and unnatural and therefore may not result in good indicators of the actual rating.

In this chapter, we use a single camera as our input sensor and use face and body motion features to measure the synchrony and coherency of audience behaviors to a full-length movie (see Figure 6.1). This work is motivated by the famous film editor Walter Murch's book "In the Blink of an Eye" [129], where he speculates that the engagement of an audience can be gauged through the synchrony of audience motion. Apart from the very dark environment, monitoring an audience from a single vantage point for a full-length feature film is a challenging problem because: i) it is across a very long time (typically movies normally range from 90-150 minutes) which is an enormous amount of video data to process, ii) people are at different vantage points and resolutions, iii) we required frame-based measurements to measure synchrony, iv) getting ground-truth labels of activity is subjective and time-consuming. To counter these issues, we use a robust face tracker based on the Fourier Lucas-Kanade (FLK) template update method to locate face and body regions to obtain face and body motion features.

Finally, it proposes an *entropy of pair-wise correlations* measure to give an indication of audience *coherency*. Additionally, this chapter proposes an off-line *change-point* detection algorithm to temporally cluster and summarize audience behaviors into a series of interest segments. We show that the proposed, unsupervised approach outperforms human-annotated labels, which do not pick-up these fine details. Using the audience ratings from *rottentomatoes.com*, we are able to learn to differentiate between good and bad movies based on these interest segments.

6.2 Related Work

In the late 1990's, theories of group behaviours and interaction were developed [14]. In order to identity the group behaviours and interaction, self-report is the current standard measure of affect. In the self-report method people answer a questionnaire about certain aspects and try to scale their feelings [14, 171]. As self-report measures are subjective, labor intensive and do not provide feedback at precise time-stamps, an automated and objective measure is desirable. In an attempt to provide an objective measure, Madan et al. [117] utilized a wearable device which measured audio and head movement in addition to galvanic skin responses of a group interacting. Eagle and Pentland [62] developed a system to analyze group behavior using a PDA which required continuous user input. While both are interesting approaches, a less invasive and automatic solution is desired.

A survey of recent work in automatically measuring a person's behavior using vision-based approaches can be found in [207]. Much of this work has centered on recognizing an individual's facial expression, with notable progress made in the areas of smile detection in consumer electronics [197], pain detection [114, 115], driver fatigue [195], human-computer-interaction [193] and security/surveillance [162]. Even though these aforementioned works all acknowledge, these works are with fully frontal images with controlled lighting conditions. An emerging area of research over the last couple of years is the use of affective computing for marketing and advertising purposes.

It is well-known that when a user watches video clips or listens to music, they may experience certain feelings and emotions [91, 159, 175] which manifest through bodily and physiological cues such as laughter. These emotional responses to multimedia content have been studied in the research community [83, 172, 188]. In 2011, Teixerira et al. [185] demonstrated that joy is one of the states in which to analyze engagement with commercials and that smiles would be a significant indicator evaluating this. Rec-

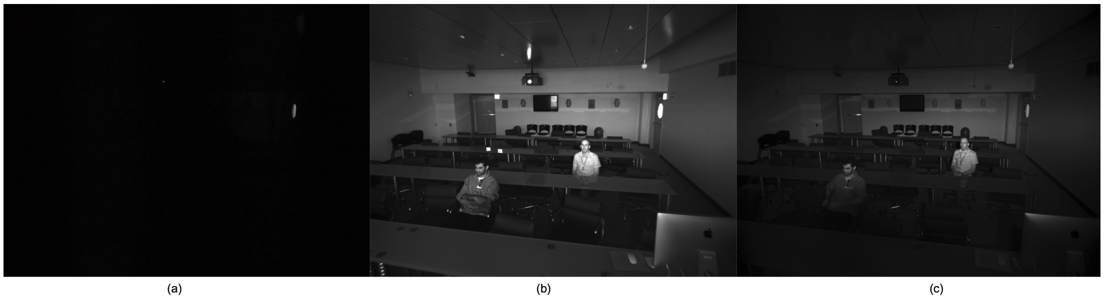


Figure 6.2: Capturing data in low light conditions. **(a) IR illuminators are OFF:** Nothing is visible. **(b) IR illuminators are ON and Lens is OFF:** Audience is visible but reflections from the screen vary the illumination conditions. **(c) IR illuminators are ON and Lens is ON:** The lens acts as a bandpass filter which results in a uniform lighting environment. Note: better uniform signal was obtained using two IR illuminators.

ognizing emotions induced by videos has also been studied in the affective computing community [86, 93, 105, 175, 183]. Emotion recognition has also been used in applications such as for detecting topical relevance, or summarizing videos [6, 86, 87].

Most of the previous work has been: a) limited to stimuli of short duration (i.e. $10-60$ seconds), not applied continually over large periods of time (e.g. up to 2 hours), and b) applied only to individuals, not simultaneously on groups of people. Having a large window of time to monitor human behavior introduces a broad gamut of additional gestures/activities associated with boredom or disengagement, which include fidgeting, doodling, yawning, looking around and pose change as indicators of inattentiveness [191]. These behaviors may be influenced by the fact the person is sharing the environment with other people, and the amount of motion can be used to determine the satisfaction level of audience. This Chapter, describes a method which can measure audience, or group-behavior in a collective manner.

6.3 Audience Analysis Test-bed

6.3.1 Test-bed

Observing people watching visual stimuli from a screen is difficult because : 1) the environment is very dark, and 2) the reflections from the visual stimuli causes nonuniform illumination. The main interesting questions comeing with this scenario are:

- How can we capture the audience face and body moments in these lighting conditions ?
- How do we remove the reflection from the screen to obtain a smooth video ?
- Can we obtain facial expressions from an audience in these lighting conditions ?

Larger lens and sensor size are two important features to consider when selecting a good camera to capture the objects in low-light. In order to capture a smooth signal, we instrumented and proposed a testbed with a low-light camera, infra-red (IR) illuminators and an IR band-pass filter. In order to analyze the collective/synchronized behavior and uninterested/unsynchronized behavior of the audience, capturing a uniform signal is the first and most important task. To counter these issues, the Chapter employs a new hardware solution which gives us a uniform smooth signal.

Infra-red camera:

We used a GX 1920 Infra-red (IR) camera which has 2 Gigabyte ethernet ports each with 240 Mb/s and ICX674 CCD sony sensor with 2/3”. In order to capture more light, we used shorter exposure time with f/1.4 and 9 mm lens with a wider angle. The resolution of the images are 1936 × 1456.

Infra-red Illuminator:

Having only an IR camera is not useful. As shown in Figure 6.2 (a), the outcome from the camera is not visible. To make the environment visible we used two Bosch UFLED95-8BD AEGIS Illuminators (wavelength of 850 nm and 95° wider beam pattern). This illuminator has 18 high efficiency surface mounted LED array and can spread to around 50 m. The constant light technology automatically controls and adjusts the lightings.

During the installation process, we powered down the illuminators and keep next to the camera for safety reasons. In order to prevent the damage to the IR camera, we set up the illuminators beneath the camera. During data capture we make sure the test-screening audience was at least 1.64 m away from the illuminators for health issues.

Infra-red Bandpass Filter:

To guard against reflections from the viewing screen, we employed the use of an IR bandpass filter with wavelength 850 nm. The IR bandpass filter provides a mechanism to pass certain wavelength ranges while rejecting unwanted radiation. This final set-up was used to capture a uniform smooth video with 15 fps as shown in Figures 6.2 (c).

6.3.2 Labelled Data

We captured volunteer audiences of various sizes to watch movies. A summary of the captured data is listed in Table 6.1 while corresponding audience ratings for those movies from *rottentomatoes.com* are shown in Table 6.2. As we were interested in both facial expressions as well as body movements, we manually annotated the following gestures:

Movies	**Movie Type**	**#People**	**Time (s)**	**#Frames**
Movie 1	Comedy	12	3605	54075
Movie 2	2D Animation	09	3921	58815
Movie 3	Horror	07	4200	63000

Table 6.1: The footage of analysed movies.

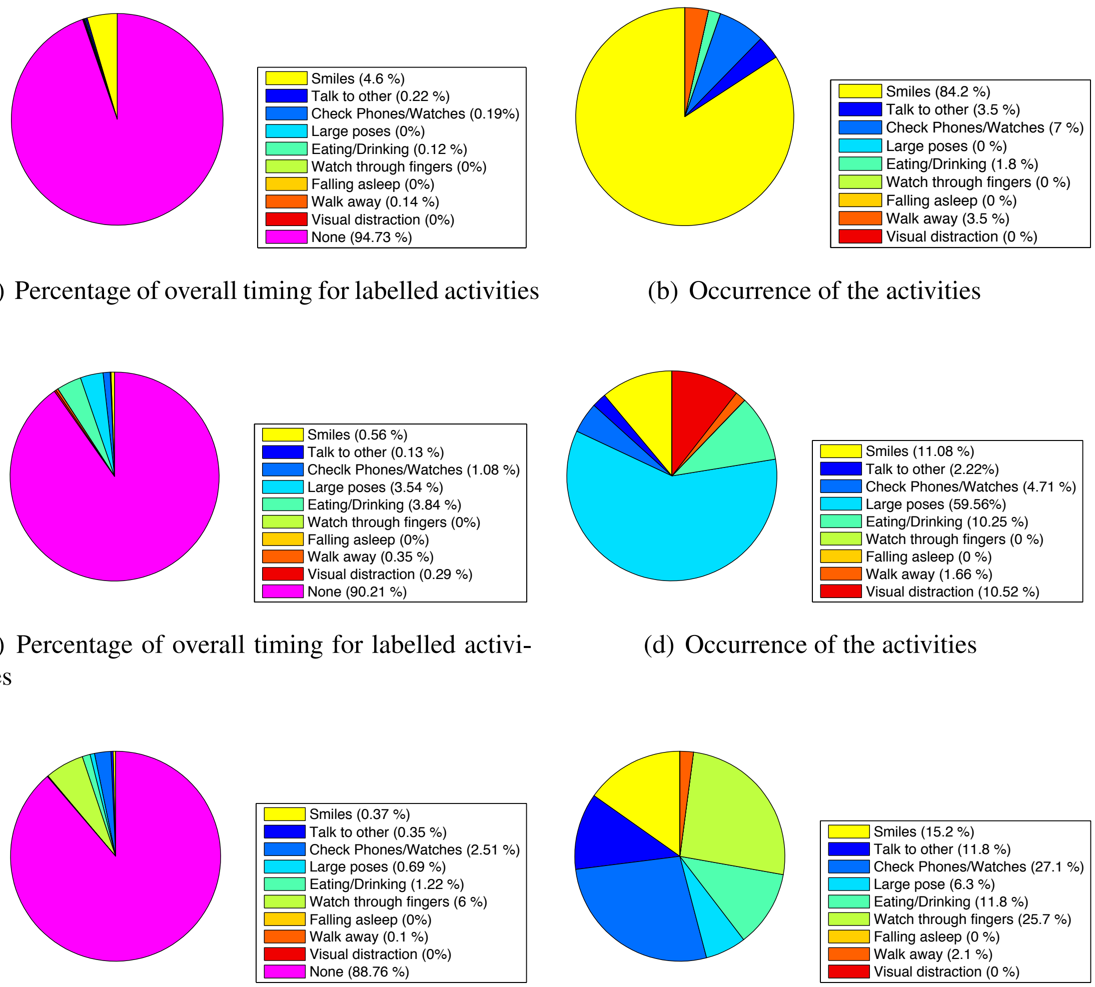
Movie	**Tomatometer (%)**	**Audience (%)**
Movie 1	🍅 75	🍿 75
Movie 2	🍅 98	🍿 34
Movie 3	🍅 65	🍿 55

Table 6.2: Audience ratings for the movies from *rottentomatoes.com*. The red tomato illustrates the movie received good reviews. The popcorn indicates the audience like the movie while splitter popcorn shows audience did not find the movie appealing.

Smiles/laughter: Using FACS [67], we annotated smiles and laughter. The onset of smiles/laughter were labelled as the onset of AU12 and the offset was labelled at the end of that occurrence.

Body movements: We annotated the following common actions: talking to another person, raising arm, moving hand to head/table, moving within chair, eating/drinking, watching through fingers, using laptop/iPad, checking phone/watch.

In terms of activity, approximately 90% of the time no activity was observed, as can be seen in the examples shown in Figure 6.3. However, as the stimulus movies are of different genres, the activities that did occur varied greatly. Figure 6.4, shows the time-series plot of the annotated gestures for both a comedy (Figure 6.4(a)) and a horror movie (Figure 6.4(b)). For the comedy, the peak activities are associated with smiles and/or laughter. Conversely, the peaks for the horror movie coincide with people moving at the same time in response to an exciting or scary part of the movie. Motivated by this analysis, we require a solution that automatically captures both facial (i.e smiles)



(a) Percentage of overall timing for labelled activities

(b) Occurrence of the activities

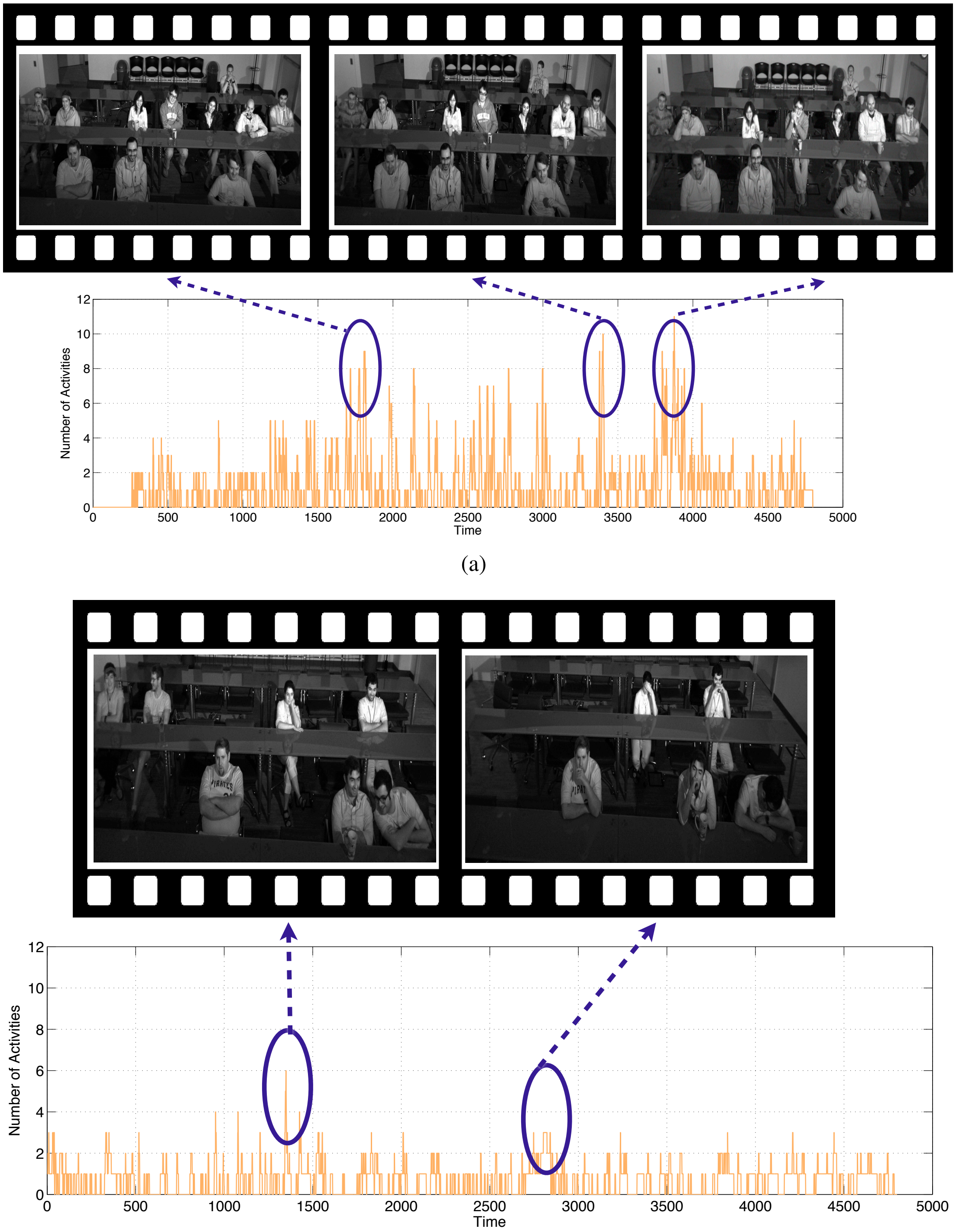
(c) Percentage of overall timing for labelled activities

(d) Occurrence of the activities

(e) Percentage of overall timing for labelled activities

(f) Occurrence of the activities

Figure 6.3: Example of statistical breakdown distribution for movies and a presentation. Top to Bottom: (Top row) a comedy movie, (Second row) a 2D animated movie and (Third row) a horror movie



(a)

(b)

Figure 6.4: The distribution of the number of activities. A sample of peak activities are highlighted: (a) a comedy and (b) a horror movie.

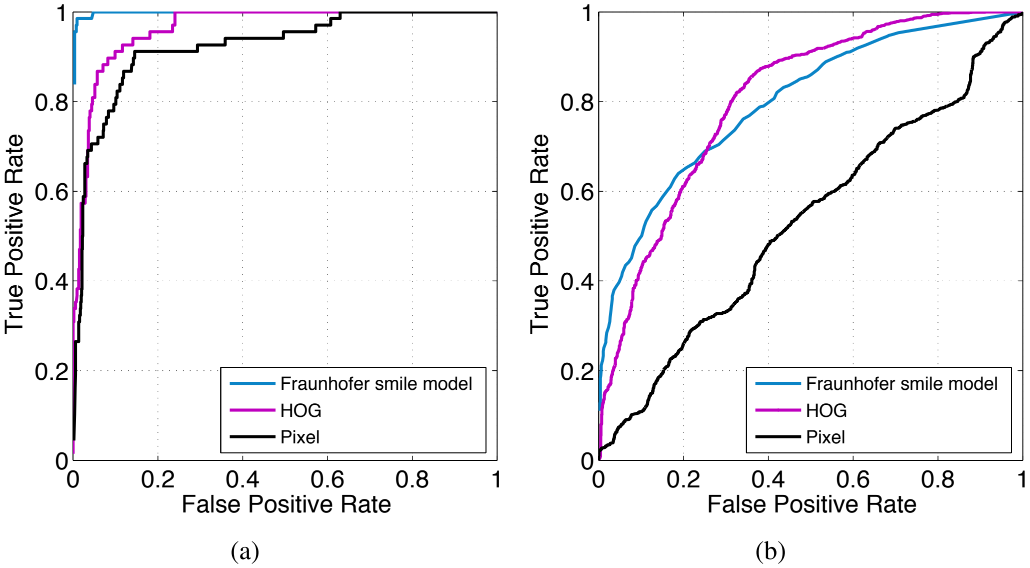


Figure 6.5: ROC curve showing the accuracy of the smile classification using HOG features, pixel intensity values and using the default *Fraunhofer* smile classifier: (a) clean data and (b) audience data.

and body movements.

6.4 Audience Facial Behavior : Smiles

6.4.1 Features for Audience Environment

Generally, the detected face region is post-processed to encode for shift and illumination invariance. Linear filters are often used to extract useful feature representations in computer vision. The Gabor features [73], Histogram of Oriented Gradients (HOG) [50], Scale-invariant feature transform (SIFT) [109] and Local binary patterns (LBP) features [136] are some of the features widely used due to their biological relevance, their ability to encode edges and texture, and their invariance to illumination.

Initially the utility of raw pixels and the HOG features were investigated in clean and

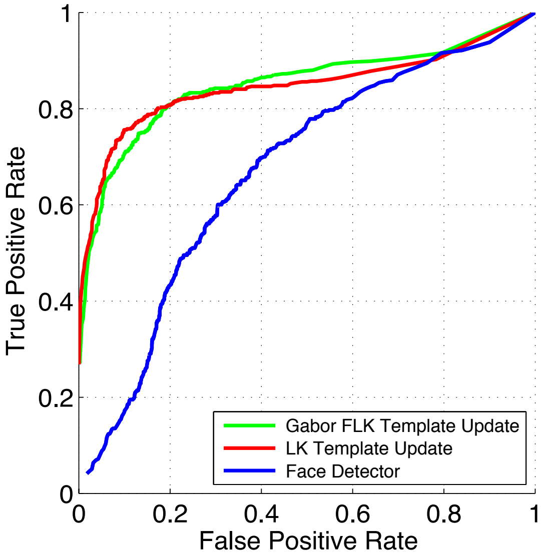


Figure 6.6: Smile recognition performance

audience environments with manually registered 900 face images. In addition to the above features, we investigated the performance with the inbuilt *Fraunhofer* smile classifier. The smile classifier was trained using the examples from CK+ [113] database which extended from CK database [90]. As shown in Figure 6.5 the performance decreased in the audience environment compared with the clean environment. However, HOG features outperformed the other methods with an area-under-the-curve (AUC) of 0.803 which emphasised their biological relevance and their invariance to illumination conditions.

6.4.2 Improved Smile Detection

From the above motivation, the smile detector consists of representing the input face image via HOG descriptors [50] and then training a linear SVM classifier. Our positive instances consisted of 750 labelled smiles from our test-bed and we used approximately the same amount of neutral images for our negative instances. We extracted HOG features using 9 orientation bins with overlapping regions with block size of

2×2, and cell size of 8×8.

We analysed smiles with 2731 labelled frontal face video segments from our test-bed using the detected or fitting faces through off-the-shelf face detector and template update methods (i.e LK and Gabor FLK) respectively. The Gabor FLK and LK showed AUC of 0.783 and 0.779 respectively, while the commercial face-detector had an AUC of 0.661.

However, this task is quite challenging due to fact that the resolution, occlusion of the face and viewing angles for the different participants is quite varied (thick red boxes in Figure 6.7). Examples of successful and unsuccessful detections are shown in Figure 6.7.

6.5 Flow-Based Gesture Analysis

In terms of recognizing human activities at a distance, optical flow has been used as an effective descriptor. Efros et al. [65], used such flow features to recognize actions of people in ballet, soccer and tennis. More recently, Rodriguez et al. [158] used such features to analyze crowds. In terms of recognizing individual and specific actions, there is a plethora of research which has solely focussed on this domain with excellent progress being made in this area [4]. However, we are not interested in specific action of a person but more in the synchrony of action (i.e is everyone doing something (it doesn't matter what) at the same time?). While it would be useful in this context to recognise specific activities such as eating, drinking, looking at mobile phone/wristwatch, looking through fingers etc which each member of the audience may be engaged in, to actually recognise these individual actions would have been an onerous task beyond the scope of this research. The main reason is the difficulty in collecting and annotating ground-truth data for this purpose. Generally it took more than 90 hours

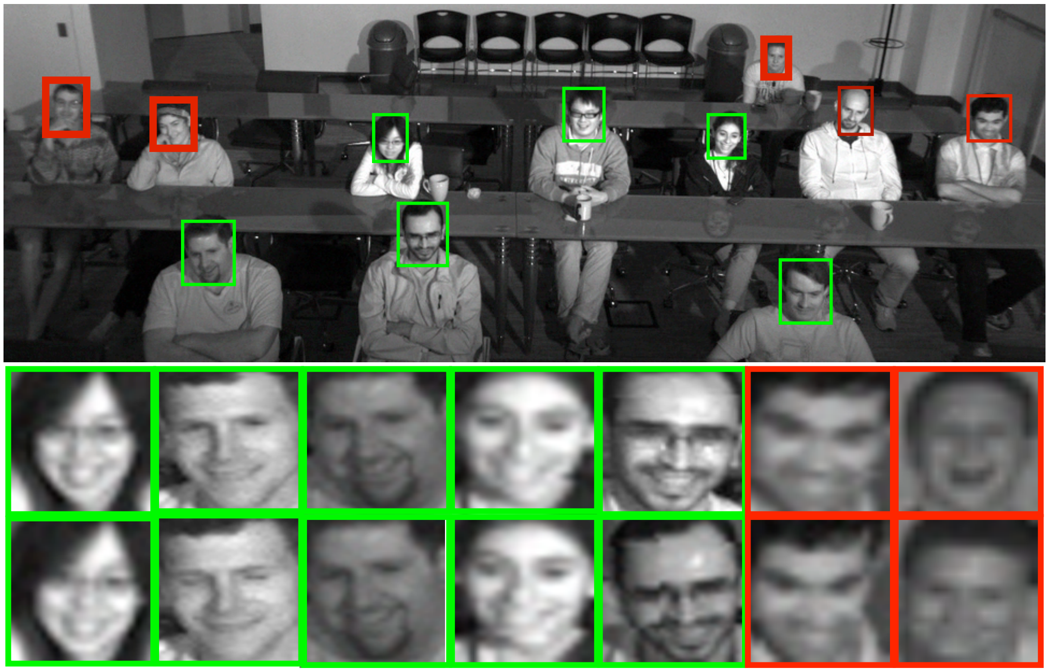


Figure 6.7: An example successful (green boxes) and unsuccessful (red boxes) detections of smiles. Due to the occlusion of the face, viewing angles for the different participants is quite varies and poor resolution (mostly occurs with the volunteers who are sitting back of the theater) recognizing smiles is quite challenging. Visual examples are illustrate with thick red boxes in the top image.

to annotate one session. Even if it is possible to get the level of annotation required it would be expected that the reliability of recognition would poor due to the high level of subjectivity and the poor lighting conditions. Because of that difficulty of recognising individual actions, the focus of the research was on determining the synchrony of common actions among the audience, independent of the action they are engaged in.

This work uses the approach of Efros et al. [65] by obtaining an aggregated spatiotemporal motion descriptor from optical flow. Another benefit was that due to the environment of the screening room, a natural spacing of audience occurred, so each person could watch the movie unoccluded and in comfort, which resulted in each person occupying an uninterrupted 3D volume – meaning that no background subtraction was required to gain the flow features of each person.

6.5.1 Fundamentals of Optical Flow

Optical flow can be defined as physical movement of points in a given image relative to 2D displacement of pixels on the image plane. This assume that the brightness at a point (x, y) does not change over the time. Given an image I, let denote the brightness at the point (x, y) at time t and $t + \Delta t$ as $I(x, y, t)$ and $I(x + \Delta x, y + \Delta y, t + \Delta t)$ respectively. As the brightness of the image at a point (x, y) is constant this can be written as:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{6.1}$$

where the parameters (Δx, Δy) are the horizontal and vertical displacement of point (x, y) and Δt is the small change in time. By assuming the motion at time $t + \Delta t$ is small the image brightness at time $t + \Delta t$ can be written as follows using the first order Taylor series expansion,

$$\begin{aligned} I(x + \Delta x, y + \Delta y, t + \Delta t) \approx I(x, y, t) + &\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y \\ + &\frac{\partial I}{\partial t}\Delta t \end{aligned} \tag{6.2}$$

By combining Equation 6.1 and Equation 6.2 yields to:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{6.3}$$

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0 \tag{6.4}$$

$$I_x v_x + I_y v_y + I_t = 0 \tag{6.5}$$

where $v_x = \frac{\Delta x}{\Delta t}$ and $v_y = \frac{\Delta y}{\Delta t}$ are the velocities in x and y directions and I_x, I_y and I_t are the image derivatives at point (x, y) at time t.

$$I_x = \frac{\partial I}{\partial x}; \quad I_y = \frac{\partial I}{\partial y}; \quad I_t = \frac{\partial I}{\partial t}; \tag{6.6}$$

Equation 6.5 can be rewrite in a more compact form which results the standard optical flow constraint equation as follows

$$\Delta I \mathbf{v} + I_t = 0 \tag{6.7}$$

where $\Delta I = \begin{bmatrix} I_x & I_y \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_x & v_y \end{bmatrix}^T$

The optical flow components $\mathbf{v}$ can be estimate by minimizing the following error term using widely used the LK algorithm [111].

$$\arg\min_{\mathbf{v}} \parallel \mathbf{A}\mathbf{v} - \mathbf{b} \parallel^2 \tag{6.8}$$

where, $\mathbf{A} = \begin{bmatrix} I_x(\mathbf{p}) & I_y(\mathbf{p}) \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} -I_t(\mathbf{p}) \end{bmatrix}^T$ and $\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_i \end{bmatrix}^T$ represent the neighborhood pixels.

6.5.2 Individual Flow-Profile

The synchronous behavior of the audience is hypothesized to give an indication of how *engaged* or *disengaged* the audience is during various segments of the viewing stimuli. The literature [191] has identified that group or individual behaviors between movies/presentations/sports are inconclusive. The positive (+) and negative (-) *engagement* behaviors for movies are shown in Table 6.5.2. It illustrates the higher energy

Type	+ Engagement	- Engagement
Movie/Presentations	smiles watching through fingers	large body poses talk to another person raising arms moving within chair eating/drinking checking phones/watches falling asleep

Table 6.3: The breakdown of the identified behaviors.

gestures as an indication for *disengagement*. Based on these we derived an indication of how *engaged* and *disengaged* profiles during various segments of the viewing stimuli.

To measure the synchronous body movement of an audience we developed an energy-based *flow-profile* measure [180]. Having N audience members, we initialised a local 3D volume for each person in the horizontal and vertical directions x and y over time t as:

$$Q = f(x, y, t) \tag{6.9}$$

We generated a *flow-profile* of each person contained within their 3D temporal volume using horizontal and vertical optical flow components v_x and v_y respectively. Using these flows, we calculated the normalized local 3D energy for a person q as,

$$E_{q,t} = \frac{1}{a_q}\sqrt{v_{q,x,t}^2 + v_{q,y,t}^2} \tag{6.10}$$

where the a_q is the area defined for an individual to move over time. This normalized energy can be vectorized over the duration of the movie time T as,

$$\mathbf{e}_q = [E_{q,1}, E_{q,2}, \cdots, E_{q,T}] \tag{6.11}$$

Finally, we defined a normalized measure of overall audience engagement over the duration of the movie time T which can be used as an indication for *engagement* and

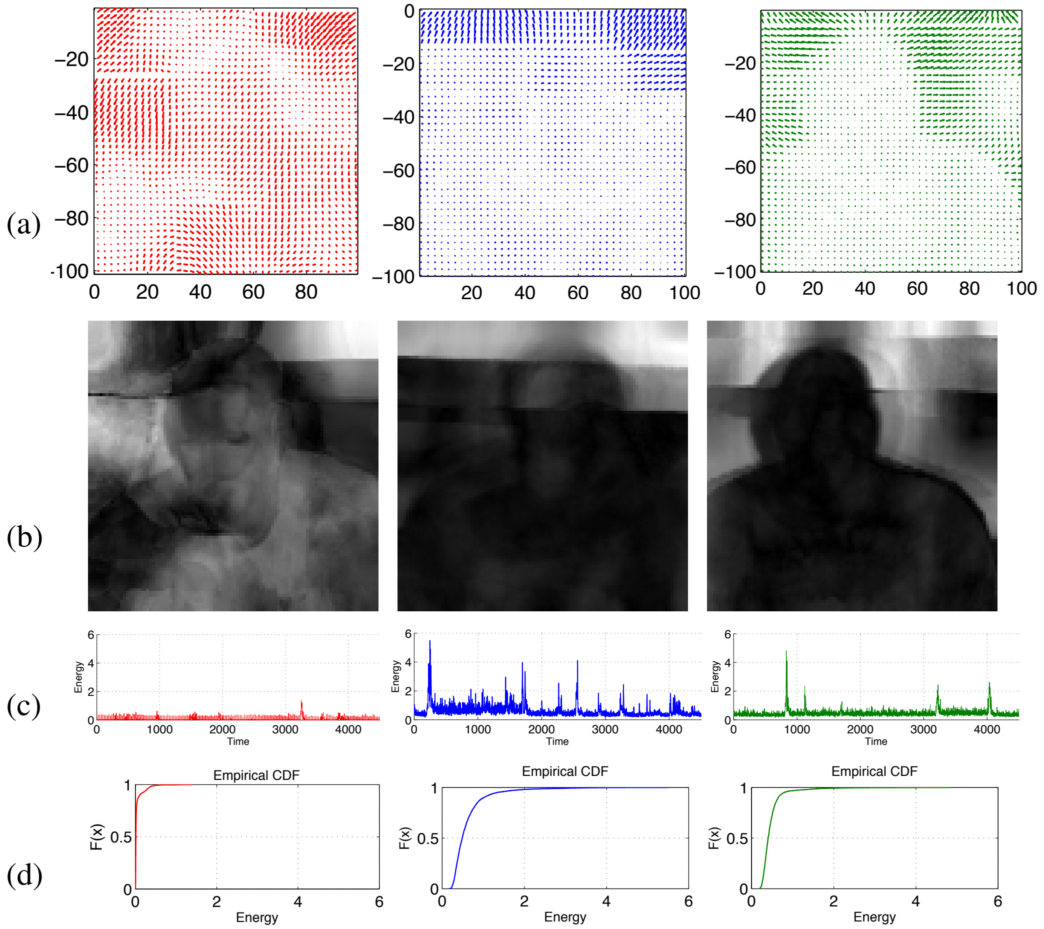


Figure 6.8: Individual flow fields for an audience member who engaged with the movies 1, 2 and 3: (a) average flow field within the 3D volume; (b) average flow magnitude; (c) normalized energy profiles; (d) the cumulative distribution of the normalized energy.

disengagement as follows,

$$\mathbf{e}_{\text{movie}} = \frac{1}{N} \sum_{q=1}^{N} \mathbf{e}_q \tag{6.12}$$

An example of an individual flow-field for an audience member across the three movies is shown in Figure 6.8. We analysed the audience flow profiles and generated an energy profile $\mathbf{e}_i$ for each audience member i which we used to measure audience behavior.

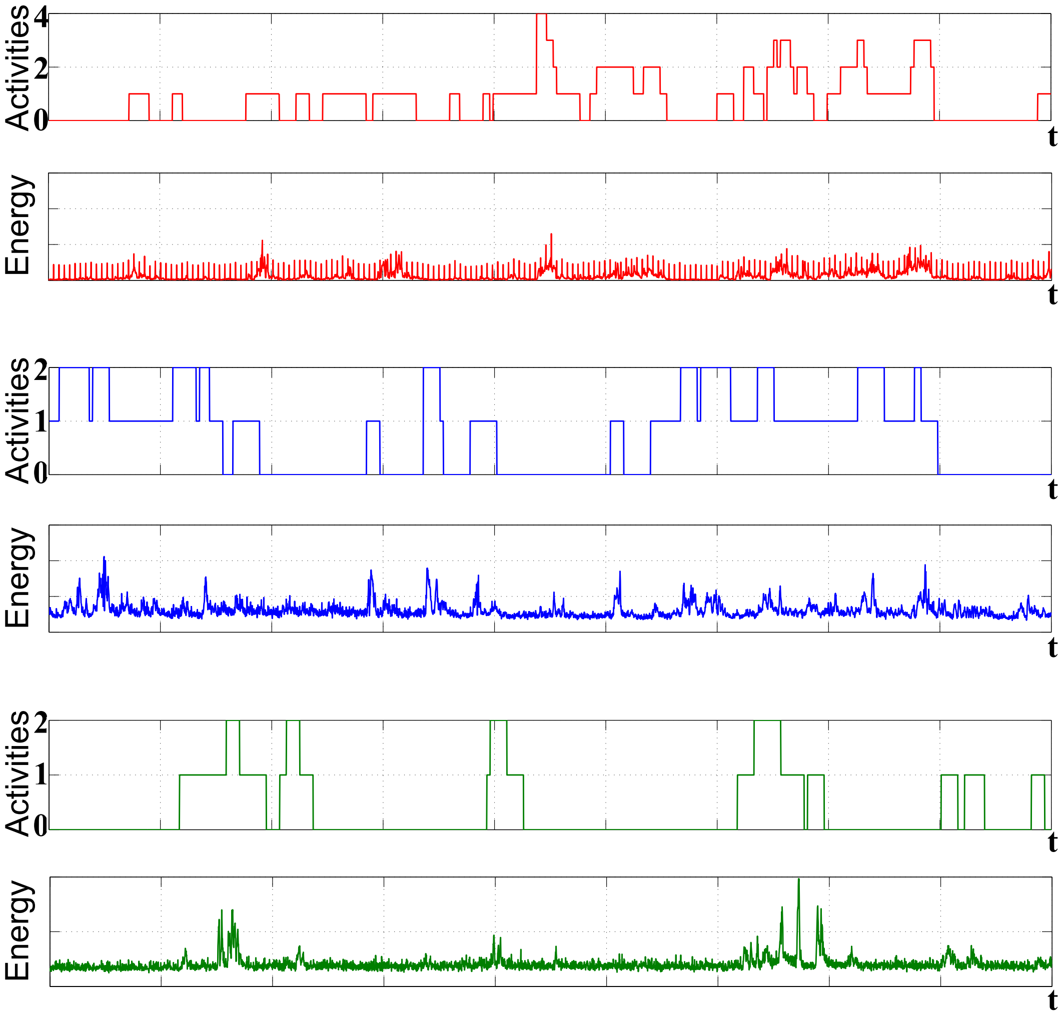


Figure 6.9: The synchrony of audience behavior between (Top) Ground-truth activities and (Bottom) energy profiles for movie 1, 2 and 3.

6.5.3 Face vs Body Features

To see how reliable each feature source was, we analysed the correlation between flow features and ground-truth labels for three movies using body and face features. The overall cross-correlation results are given in Table 6.4. It can be seen that body features are more robust features as we are only interested in the synchrony of movement. An example using body features is shown in Figure 6.9. In this environment, the

Movie	**Max. Correlation** *Body features*	**Max. Correlation** *Face features*
Movie 1	0.69	0.48
Movie 2	0.68	0.24
Movie 3	0.72	0.05

Table 6.4: The cross-correlation results using body and face features. Body features are more reliable than face.

result makes sense as the face is often very small and is sensitive to all types of subtle variations and only contains a very small subset of possible actions that can take place. As our body features subsume the face features, it makes sense to use the body flow feature as our feature representation of each person.

6.6 Temporally Segmenting Audience Behaviors

As shown in Figure 6.3, around 90% of the time people are inactive while watching a movie. This could be due to: i) people not moving at all, ii) intensity or duration of activity being so low or short that it does not warrant labelling, iii) the activity not fitting into the pre-set activities vocabulary. It can be argued that ii) and iii) are due to problems with annotations, but as a result of the long length of input stimuli (approximately 1-2 hours per movie), it is highly impractical and unscalable to get this level of annotation. Even if it is possible to get the level of annotation it would be expected that the reliability of annotation would greatly diminish due to the high level of subjectivity. In terms of automatic analysis, this can be circumvented as the continuous flow features of each person can be used to temporally segment or cluster potentially interesting behaviors.

In terms of temporal segmentation, we used the *change-point* (CP) detection method which is an unsupervised temporal clustering method that has the ability to flag abrupt changes in a stochastic process [84, 85]. Methods can be either *online* (i.e only knowl-

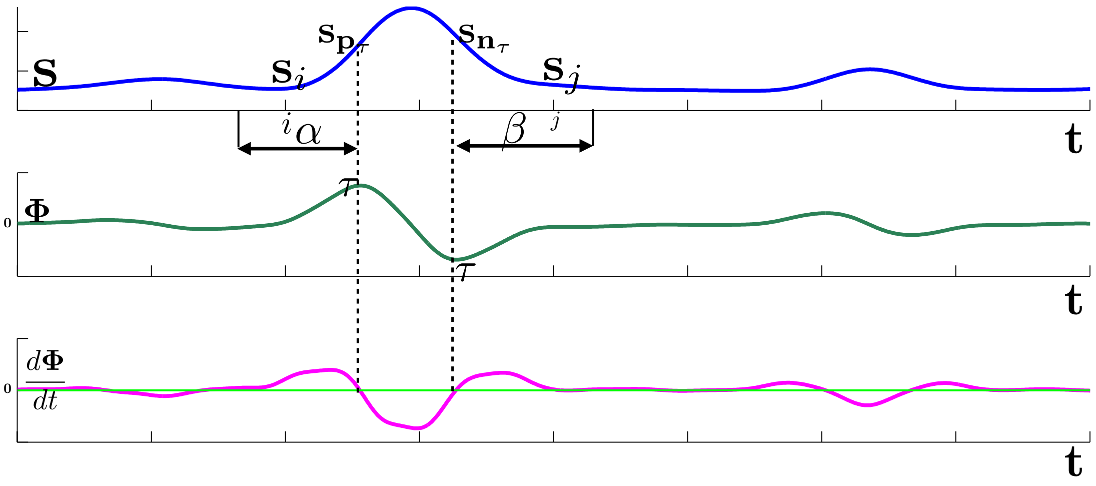


Figure 6.10: CP detection algorithm. (top) Original signal $\mathbf{s}$, (middle) First-order derivative in $\mathbf{s}$ and (bottom) second order derivative in $\mathbf{s}$.

edge of signal up to current time-stamp) such as the *generalized likelihood ratio* [82], or *offline* ((i.e with the full knowledge of the entire signal), such as the *CUSUM* [139]. Due to the computation required to generate the optical flow features and availability of all time stamps, we utilized an *offline* CP detection method. Compared with *CUSUM*, our approach has no gaussian assumption between CPs while *CUSUM* assumes the gaussian distribution between CPs and the parameters are known [18]. The proposed method is able to select an arbitrary number of strongest CPs and can deal with very noisy data.

6.6.1 Proposed Individual Change-Point Detection

Given an audience energy signal $\mathbf{s}$, smoothed over 6 seconds, we first obtain the first-order derivative of the signal $\mathbf{\Phi} = \frac{d\mathbf{s}}{dt}$. We computed positive peaks $\mathbf{p}$ and negative peaks $\mathbf{n}$ by detecting zero crossing values of $\frac{d\mathbf{\Phi}}{dt}$ for a given threshold $\theta = \mu_{\mathbf{s}} + K\sigma_{\mathbf{s}}$. The parameters $\mu_{\mathbf{s}}$ and $\sigma_{\mathbf{s}}$ are mean and standard deviation of signal $\mathbf{s}$ and K is a constant which has to set experimentally on development population. Then the CPs

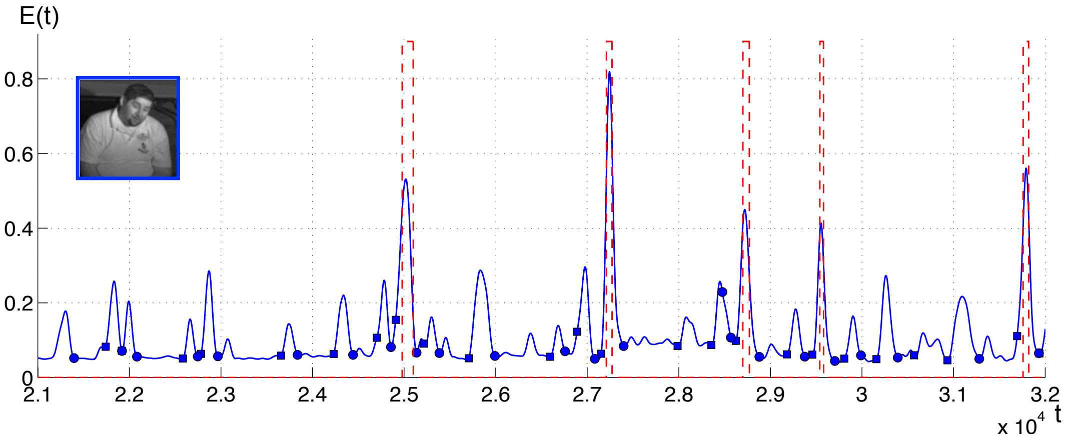


Figure 6.11: An example of our change-point detector compared to the human annotations. Ground-truth annotations are shown with red dotted lines.

can be obtained by maximizing the following objective functions:

$$\forall i, \alpha \leqslant i < \tau : \arg\max_{i} \left\{ \frac{\mathbf{s}_i - \mathbf{s}_{\mathbf{p}_\tau}}{\mathbf{s}_{\mathbf{p}_\tau}} \right\}; \tag{6.13}$$

$$\forall j, \tau < j \leqslant \beta : \arg\max_{j} \left\{ \frac{\mathbf{s}_j - \mathbf{s}_{\mathbf{n}_\tau}}{\mathbf{s}_{\mathbf{n}_\tau}} \right\}; \tag{6.14}$$

where α and β are lower and upper bound around each peak (set it to 15 frames). Equation 6.13 and 6.14 detect CP *i* and *j* where the amplitude of signal highly increases and strongly decreases respectively (Figure 6.10). We compared our change-point approach with manually annotated gestures (Refer to Figure 6.11). It can be seen that the human annotated labels are not able to pick up the subtle movements at the fine granular level (i.e. second) while our automatic approach is able to pickup these movements.

6.6.2 Audience Change-Points

Audiences tend to behave differently in various segments of the movies depending on their interest. These behaviors can be classified as *stillness* or *movement*. Given the normalized smoothed audience energy signal $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ where N is the number of people, first we calculate the first-order derivative $\frac{d\mathbf{x_i}}{dt}$ for $\mathbf{x_i}$ and obtain signal $\mathbf{z} = \sum_{i=1}^{N} \frac{d\mathbf{x_i}}{dt}$. Then we detect audience CPs $\mathbf{t_{syn}}$ which are the zero-crossing values in signal $\frac{d\mathbf{z}}{dt}$ for a given threshold. An example of detected audience CPs for movies 1 and 2 are shown in Figure 6.12. The region between a (+) and a (-) synchrony point (region (b) in Figure 6.12) identified as the indication of *movements* and region between a (-) and a (+) synchrony point can be identified as a indication of *stillness*.

6.7 Summarizing Audience Behavior

To summarize the reaction of the audience to a movie, we segment the movie into one-minute windows and for each window we find the strongest audience change-point. Using that as our index, we use a 1 second window centered at that change-point to summarize the audience behavior over that minute. We piece this together to form a summarization of the audience behavior - which allows someone to get an idea of a 90-minute movie over the course of 90-seconds. Qualitatively, we found that we could find engaged and disengaged segments on the movie using this approach. Visual examples are given in Figure 6.13.

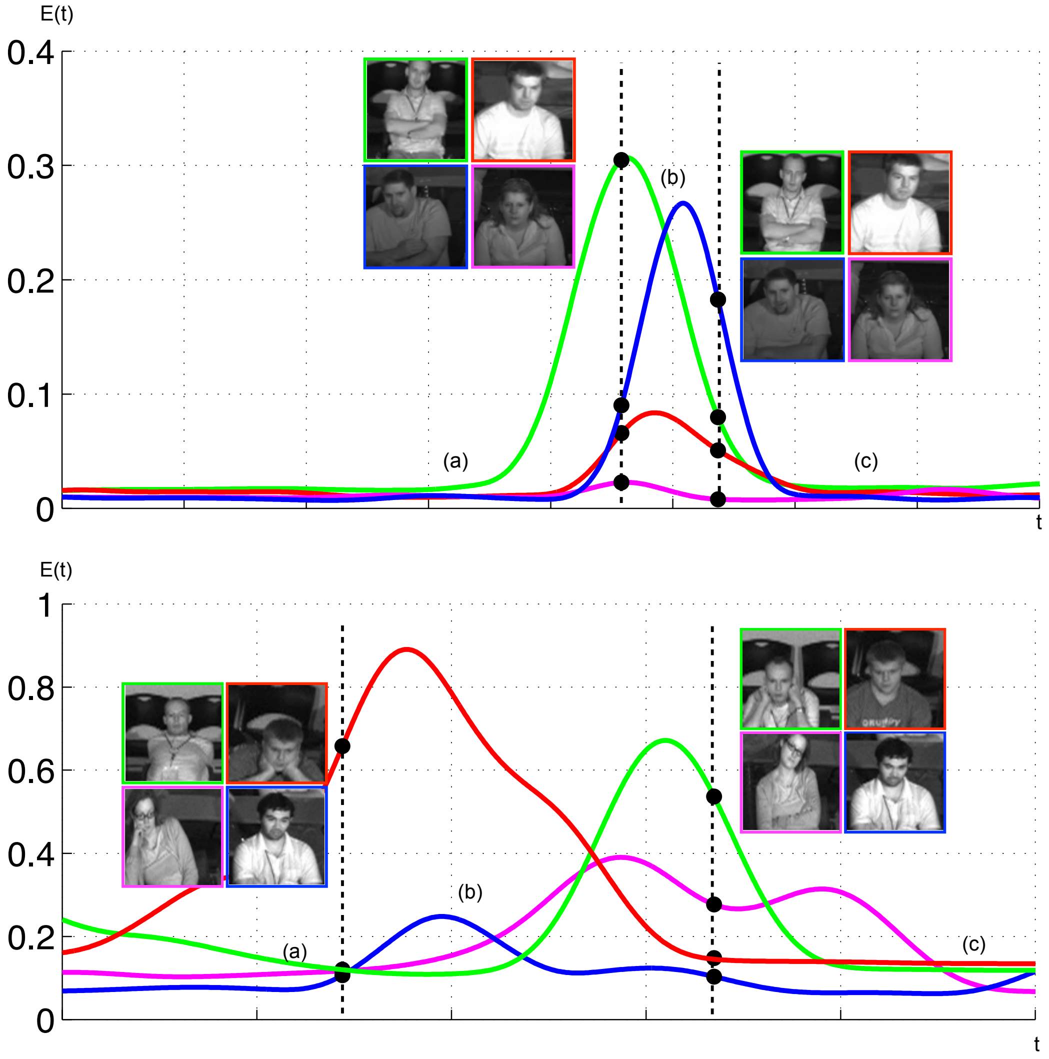


Figure 6.12: Synchrony point detections for a comedy and a 2D animated movie in left and right hand side respectively.

6.8 Entropy of Pair-Wise Correlations

As we are interested in the synchrony of behavior between each audience member at the local-level (i.e pair-wise comparison) as well as the global-level (i.e compared to the whole group), we utilized an *entropy of the pair-wise correlation* approach. In this regard, we first compared the small feature segment (i.e 30 seconds) between two

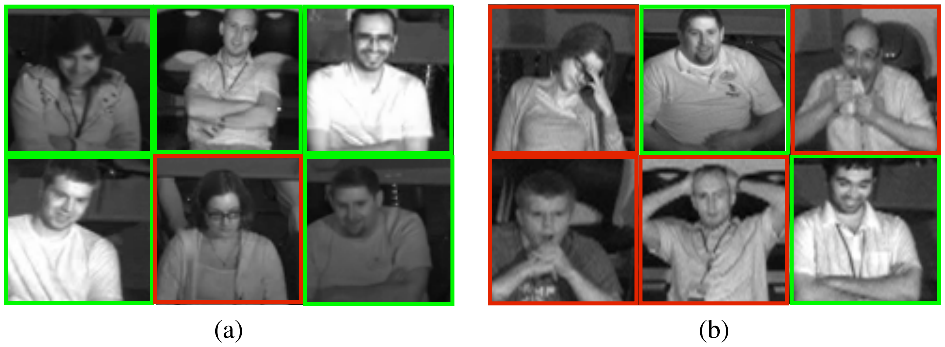


Figure 6.13: An example of movie summarization for: (a) movie 1 and (b) movie 2. The green boxes show examples of similar activities while red boxes illustrates random activities.

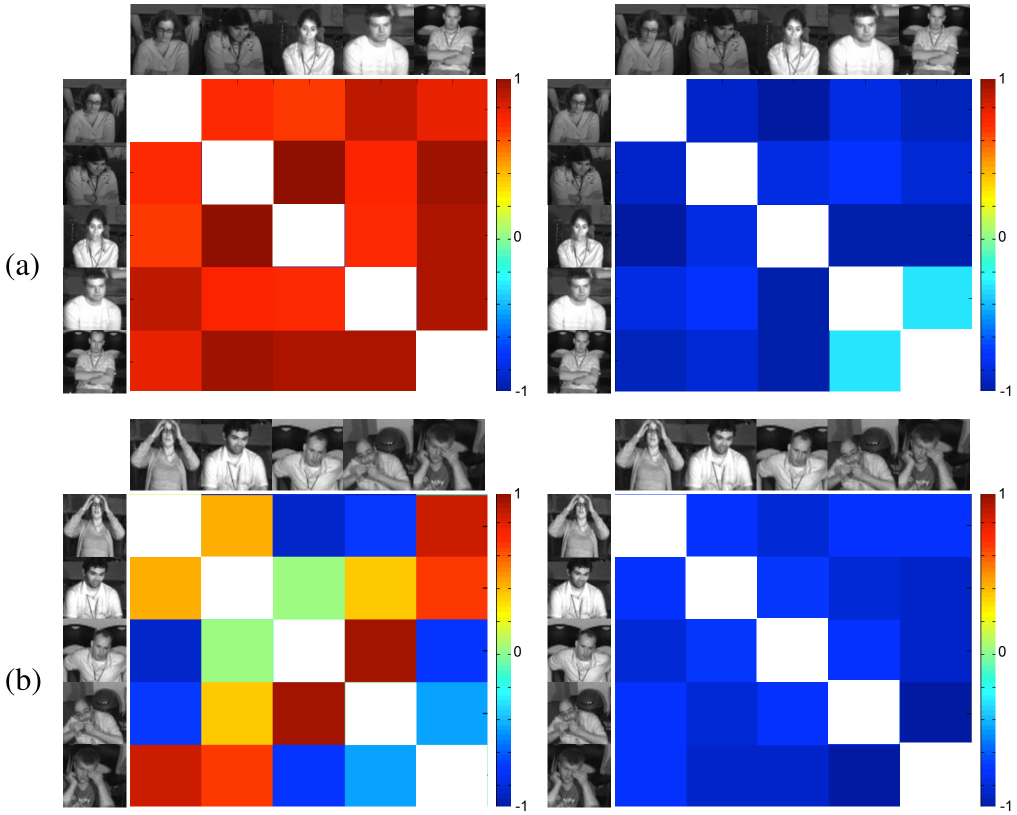


Figure 6.14: A similarity matrix for a small segment of movie for (a) movie 1 and (b) movie 2. The left and right figures are for *movement* and *stillness* regions (no intra-person correlations were conducted (white blocks on the diagonal)).

audience members, s_1 and s_2, and calculate the pair-wise correlation by using,

$$C_{\mathbf{s_1 s_2}} = \exp\left(\frac{-\parallel \mathrm{s}_1 - \mathrm{s}_2 \parallel^2}{2\sigma^2}\right) \tag{6.15}$$

where σ is an adjustable parameter for each similarity matrix (we used $\sigma = 1$). We then exhaustively calculated all the pair-wise correlations between audience members which yielded a similarity matrix which is illustrated in Figure 6.14. As can be seen in (a), when everyone is doing something at the same time (i.e laughing/smiling) the cohesion is high, similarly when everyone is doing nothing the audience cohesion is still high. Given that the similarity matrix of piece-wise correlations can be represented by $\mathbf{X}$, we can generate a probability distribution of $\mathbf{X}$ for that time-segment $p(\mathbf{X})$, which allows us to gain a measure of audience disorder via entropy [173]

$$H(X) = -\sum_{i=0}^{N-1} p(i) \log p(i) \tag{6.16}$$

A high value of entropy means that there is great disorder (i.e random) behavior, while a low value of entropy means that there is cohesion or predictability of behavior.

6.8.1 Predicting Movie Ratings

To gauge how much the general public likes a particular movie, *rottentomatoes.com*, has an interactive feature which allows people to go online and give a rating. Over time the number of ratings aggregate (100k's) and based on these crowd-sourced ratings, they generate an "audience measure". Based on these scores, an average audience measure is obtained, with a movie rating of 75% or higher being deemed a good movie, a movie rating between 50-75% being mediocre and below 50% denoteing a bad movie. Out of our three movies, we had one good, one bad and one satisfactory movie. Using our audience measure, we wanted to see if it could predict the rating of a movie based solely on the audience reaction. To do this, we developed a normalized measure which used the mean of the similarity matrix at each time step M_t, and the entropy of the similarity matrix E_t. We then found the expectation of the ratio using

	M_t	E_t	**Automatic Rating**	**rottentomatoes.com**
Movie 1	0.527	0.750	0.70	0.75
Movie 2	0.265	0.866	0.31	0.34
Movie 3	0.338	0.924	0.37	0.55

Table 6.5: Results of our automatic audience rating measure compared to the crowd-sourced ones from *rottentomatoes.com*.

M_t and E_t to gain a normalized rating measure R across the whole movie as follows:

$$\text{Rating} = \frac{1}{T}\sum_{t=1}^{T}\frac{M_t}{E_t} \qquad (6.17)$$

where T is the number of time segments analyzed within the movie. Using this measure, we show our results in Table 6.5. As can be seen from this result we get a reasonable approximation to the *rottentomatoes.com* crowd-sourced ratings. This shows a proof-of-concept that a possible measure can be obtained, although large amounts of footage of audiences are needed to get an indication of significance.

6.9 Chapter Summary

The feedback from the test-screening audience can be used to better understand the movie. However the current self-report methods are subjective, biased (e.g. loyalty to the brand) and do not contain specific time information. This Chapter seeks to gain an automatic real-time objective measure of audience by analyzing the collective facial and body movements. Due to the complexity and difficulty of this task, no one has previously looked at this problem. In addition to introducing a new problem to the field of face and gesture analysis, as well as a solution on how to capture such data, there are numerous technical challenges which are highlighted in this chapter and solutions to them presented.

The Chapter proposes an automatic real-time objective measure of audience engagement through automatically analyzing the collective/uninterested synchronized behavior in a very dark environment through detecting facial expressions and body gestures. The key contributions of this chapter: (i) Audience environments and test-screenings are very dark and suffer from reflections from the screening. To counter these issues, we employ a hardware solution which gives us a uniform smooth signal. (ii) Chapter 6.3.2 introduces a labelled dataset of audiences of varying sizes watching movies. A key insight from this data collection effort is the lack of movement/actions, which highlights the sensitivity of this task. (iii) Even though the movie viewing environment is very dark and contains views of people at different scales and viewpoints, we can measure audience behavior by improving smile detection by using the FLK algorithm to register audience members' faces. This overcomes instances when there is abrupt change in illumination caused by sudden movement. (iv) Proposal of an *offline* CP detection algorithm to temporally cluster audience behaviors into a series of "interest" segments and (iv) proposal of a method to learn behaviors using crowd-sourced audience ratings from *rottentomatoes.com*.

At a coarse level, nearly all work in face and gesture analysis can be broken down into face and body movements - however, there are many different factors, such as lab vs. real-world, individuals vs. crowd/audience, viewpoints, resolutions, illumination and temporal windows. The best approach of course varies depending on the combination of factors and the target application. This chapter looks at the novel problem of analysing an audience in a dark environment over a long period of attention, i.e. movies.

We demonstrated our approach on three full-length movies and showed that our unsupervised approach can pick up fine motions which human annotators cannot, which allows us to summarize audience behaviors as well as predict them. We showed that we can give a reasonable approximation of audience behavior and in future work we will be collecting a large volume of data to further test out this approach.

Chapter 7

Conclusions and Future Research

The Chapter summaries the dissertation by noting the original contributions made in the fields of computer vision and AVASR. A summary of future revenues is also highlighted in Chapter 7.2.

7.1 Summary of Contributions

Today, computer vision applications such as face detection in cameras/IPhones work reasonably well in controlled environments where the illumination does not change. However, use of the computer vision in many practical applications is still far away, mainly due to the constant change of illumination and very low-light/dark lighting conditions. The key problem is obviously sensing the face in these illumination conditions as the face contains a rich amount of behavioral information. Prototype systems have been built to register human face using either coarse-type of alignment or dense alignment methods to obtain a rich amount of visual information to recognize human behavior.

This thesis has attempted to remove the expenses associated with the lighting variations, which limit the use of computer vision theory for many applications. The normal pipeline would be to first sense the face and then extract out features which contain meaningful information and finally do the classification task. In an attempt to remedy this situation, the work in this thesis has concentrated on researching and developing methods to recognize human behavior through visual information by focusing on face sensing, feature extraction and classification tasks. In addition to that, this dissertation has highly detailed two real-world applications namely, (i) recognizing human speech through lip reading within an automotive environment and (ii) automatically recognizing audience collective/uninterested synchronized behavior in very dark environment while engaging with movies, taking into account the problems associated with lighting variations and providing solutions to overcome those problems with computer vision. The work contained in this dissertation was performed with the intention of addressing these novel and previously unsolved problems in the fields of computer vision and AVASR. Within this spectrum of broad problems, the major contributions stemming from this work are summarized as follows:

(i) Behaviors of coarse-type of face alignments and fine registration methods are investigated. The performance of face alignment methods is also compared with different lighting conditions and demonstrates that it degrades with illumination, especially in low-light conditions when employing the use of coarse-type of face alignment methods (Refer to Chapter 3). In order to overcome the problem with illumination with the fine registration method, Chapter 4.5 demonstrates a novel method showing how LK-inspired AAM fitting gives identical performance in the spatial and Fourier domains. Further, we demonstrate how the effect of multiple filter responses can be re-interpreted as a diagonal weighting matrix in the Fourier domain leading to substantial computational savings, when performing inverse compositional simultaneous fitting across multiple filter responses.

(ii) We demonstrate the process of applying the inverse compositional project-out algorithm in the Fourier domain by showing how: (i) Fourier transforms to the current image, and (ii) the application of multiple filter responses can be completely pre-computed offline. This contribution is of key importance to person-specific AAM face fitting, as it provides an extremely computationally efficient method that affords both invariance to both expression and environmental variations. The method was tested with MPIE database and presents empirically the substantial improvement in person-specific AAM fitting performance over canonical LK inspired fitting algorithms (i.e. simultaneous and project out), when using our proposed Fourier variants. For all our experiments we employed biologically motivated Gabor filter banks (Refer to Chapter 4).

(iii) We provide a comparison of the recognition performance of single channel and multi-channel enhanced speech (in Chapter 5.4.2) with the performance of audio-visual speech using data from a challenging automotive environment (AVICAR [98]), which introduces a number of visual challenges, including changes in illumination and speaker pose as well as severe audio impairment arising from car engine, wind and road noise. Chapter 5.7 shows that visual speech recognition results within a vehicle-environment obviously diminished from what is obtained in ideal laboratory conditions. We extend this study to also demonstrate that the complementary nature of visual information and enhanced audio observed in [45] still holds true when using multi-channel speech enhancement algorithms and state-of-the-art middle integration techniques (i.e synchronous hidden Markov model (SHMM)) for audio-visual fusion.

(iv) Chapter 5.8.2 presents further improvements in speech recognition accuracy over traditional single-camera AVASR approaches that can be obtained when multiple frontal or near-frontal views of speakers' faces are available in a real-world automotive environment. The combination of the four visual streams with a single acoustic stream in a five-stream audio-visual SHMM demonstrated a relative improvement of between 6% and 17% word-level accuracy over traditional single-camera AVASR, and between 9% and 56% relative improvement in word-level accuracy when compared to the acoustic-only approach.

(v) At a coarse level, nearly all work in face and gesture analysis can be broken down into face and body movements - however, there are many different factors, such as lab vs. real-world, individuals vs. crowd/ audience, viewpoints, resolutions, illumination and temporal windows. The best approach of course varies depending on the combination of factors and the target application. In this thesis, we look at the novel problem of analysing an audience in a dark environment over a long period of attention, i.e. movies. Due to the complexity, size and difficulty of this task, no one has previously looked at this problem. This dissertation introduces a new problem to the field of face and gesture analysis with numerous technical challenges.

Audience environments and test-screenings are very dark and suffer from reflections from the screening. To counter these issues, we employ a hardware solution which gives us a uniform smooth signal in Chapter 6.3. Chapter 6.3.2 introduces a labelled dataset of audiences of varying sizes watching movies. A key insight from this data collection effort is the lack of movement/actions, which highlights the sensitivity of this task.

(vi) Even though the movie viewing environment is very dark and contains views of people at different scales and viewpoints, we can measure audience behavior by improving smile detection by using the FLK algorithm to register audience members faces. This overcomes instances when there is abrupt change in illumination caused by sudden movement. Recognizing the synchrony of smiles was challenging due to the occlusion of the face and various pose angles of the participants. However, we were able to improve the synchrony of smiles by 12.6% with proposed Gabor FLK in an audience (dark) environment.

(vii) This thesis introduces a method to obtain an indicator for audience *engagement* or *disengagement* using standard optical flow. We generate a *flow-profile* of each person contained within their local 3D temporal volume via optical flow, which is aggregated into a collective *stillness* measure. It shows that this approach can pick up on the different genres and interest points in movies and can be used to monitor the engagement of the audience over the time using a battery of experiments.

(viii) In addition to that, Chapter 6 proposes an *entropy of pair-wise correlations* measure to give an indication of audience *coherency*. Additionally, it proposed an off-line *change-point* detection algorithm to temporally cluster and summarize audience behaviors into a series of interest segments. We show that the proposed unsupervised approach outperforms human-annotated labels, which do not pick-up these fine details. Using the audience ratings from *rottentomatoes.com*, we are able to learn to differentiate between good and bad movies based on these interest segments. The introduced method showed that we can give a reasonable approximation of audience behavior compared to *rottentomatoes.com* ratings.

7.2 Future Research

This dissertation detailed the problems and solutions in recognizing human behaviour through visual information in noisy environments. The variation of illumination over time or low-light conditions is one of the biggest challenges to recognize human behaviour via visual modality for many applications and this thesis provides solutions towards overcoming this challenge. As a result of this work, a number of different avenues of further work have been identified and can be listed as follows:

(a) Chapter 4.5 introduces a novel method to show how LK-inspired AAM fitting gives identical performance in the spatial and Fourier domains. Even though the introduced method is able to handle the variation of illumination, there is a need for more robust methods to handle the variation of the scale of face.

(b) Throughout this thesis AVASR experiments were conducted using audio modality and two dimensional (2D) lip images as visual modality. 2D images give only the height and the width of the ROI. On the other hand, three dimensional (3D) representation supplies more information that is not available in a 2D image, such as the depth of an object. Therefore having a system that can represent the lip region in 3D will give more visual information about the speaker and it will help to increase the robustness and the effectiveness of the AVASR system. This is an important future direction to be solved, especially in an automotive environment, for the design of an efficient human-vehicle computer interface.

(c) This thesis proposes a novel problem of an automatic real-time objective measure of audience engagement through automatically analysing the collective/uninterested synchronized behavior in very dark environment through detecting facial expressions and body gestures. This thesis used the state-of-the-art smile detection method using HOG features and SVM classifier. However, this task is quite challenging, because the resolution, occlusion of the face and viewing angles for the different participants is quite varied. In the future, there will be a need for analysing strong features through an ability to tackle the above-mentioned challenges in an audience (dark) environment.

(d) To obtain the body features to calculate the audience *engagement/disengagement* and audience ratings in Chapter 6, we used optical flow. However, calculation time for optical flow features is expensive which limits the usage of our method in a real-time audience application. Methods such as image subtraction is need to be investigated to calculate body features.

(e) Chapter 6 proposes a method to obtain movie ratings automatically. The proposed method showed that it gives a reasonable approximation of audience behavior. This shows a proof-of-concept, that a possible measure can be obtained, although large amounts of footage of audiences are needed to get an indication of significance.

(f) The detection or tracking faces in very dark lighting is very challenging over a long period of time (i.e up to 2 hours). Long-term face monitoring in dark conditions will be another research avenue.

(g) Although illumination variation is a major problem for many real-world applications, other variabilities, such as appearance, speaking style and speaker emotion and expression, need to be investigated as well.

Bibliography

[1] *Kitt Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/KITT.

[2] *Open Source Computer Vision Library*, http://www.intel.com/research/mrl /research/opencv.

[3] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," *Speechreading by Humans and Machines*, pp. 461–471, 1996.

[4] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, 2011.

[5] N. Ambady and R. Rosenthal, "Thin Slices of Expressive behavior as Predictors of Interpersonal Consequences : a Meta-Analysis," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[6] I. Arapakis, I. Konstas, and J. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," *in Proceedings of the seventeen ACM international conference on Multimedia*, pp. 461–470, 2009.

[7] E. Aronson and S. Rosenblum, "Space perception in early infancy: perception within a common auditory-visual space," *Science*, vol. 172, pp. 1161–1163, 1971.

[8] A. B. Ashraf, S. Lucey, and T. Chen, “Fast image alignment in the fourier domain,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*, pp. 2480 –2487, June 2010.

[9] A. Ashraf, S. Lucey, T. Chen, K. Prkachin, P. Solomon, Z. Ambadar, and J. Cohn, “The painful face: Pain expression recognition using active appearance models,” *In ICMI*, pp. 9–14, 2007.

[10] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, “Evaluating aam fitting methods for facial expression recognition,” *International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1 –8, 2009.

[11] D. Ayers and M. Shah, “Monitoring human behavior from video taken in an office environment,” *Image and Vision Computing*, pp. 833–846, 2001.

[12] S. Baker and I. Matthews, “Equivalence and efficiency of image alignment algorithms,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[13] S. Baker and I. Matthews, “Lucas-kanade 20 years on: A unifying framework: Part 1: The quantity approximated, the warp update rule, and the gradient descent approximation.,” *International Journal of Computer Vision*, vol. 56, pp. 221–255, February 2004.

[14] R. Bales, “Social inteaction system: Theory and measurement.,” *New Brunswick, NJ:Transaction Publishers*, 1999.

[15] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscesk, I. Fasel, and J. Movellan, “Recognizing facial expression: Machine learning and application to spontaneous behavior,” *IEEE Conference on Computer Vision and Pattern Recogntion*, vol. 2, pp. 568–573, June 2005.

[16] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Fully automatic facial action recognition in spontaneous behavior,” *Interna-*

tional Conference on Automatic Face and Gesture Recognition, pp. 223–230, 2006.

[17] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, M. Frank, and J. Movellan, “Fully automatic facial action recognition in spontaneous behavior,” *7th International Conference on Automatic Face and Gesture Recognition*, pp. 223–230, 2006.

[18] M. Basseville and V. Nikiforov, “Detection of Abrupt Changes - Theory and Application,” 1993.

[19] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer Topics in Signal Processing, Berlin: Springer-Verlag, 2008.

[20] C. Benoit, T. Guiard-Martigny, B. Go, and A. Adjoudani, “Which components of the face do humans and machines best speechread?,” *in Speechreading by Humans and Machines,*, D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer, 1996.

[21] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Washington, DC, USA), pp. 208–211, 1979.

[22] C. Binnie, A. Montgomery, and P. Jackson, “Auditory and visual contributions to the perception of consonants,” *Speech & Hearing Research*, vol. 17, pp. 619–630, 1974.

[23] A. Bobick and J. Davis, “Real-time recognition of activity using temporal templates,” *Applications of Computer Vision (WACV)*, pp. 39–42, 1996.

[24] A. Bobick and J. Davis, “The recognition of human movement using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 257–267, 2001.

[25] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[26] K. W. Bowyer, K. Chang, and P. Flynn, “A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition,” *Comput. Vis. Image Underst.*, vol. 101, pp. 1–15, Jan. 2006.

[27] C. Bregler and Y. Konig, “Eigenlips for robust speech recognition,” *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 669–672, 1994.

[28] H. Buxton, “Learning and understanding dynamic scene activity: a review,” *Image and Vision Computing*, vol. 21, no. 1, 2003.

[29] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, no. 1, pp. 206–224, 2010.

[30] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 679–698, 1986.

[31] J. Carletta, S. Ashby, S. Bourban, M. Flynn, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, and M. W. P. D. Reidsma, “The AMI Meeting corpus: A pre-announcement,” *In Proceedings of Machine Learning for Multimodal Interaction*, pp. 28–39, 2005.

[32] E. Castillo and . H. A. S. Gutiferrez, J. M., “Expert systems and probabilistic network models,” *Springer.*, 1997.

[33] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, “Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions,” *In CVPR*, pp. 1932 –1939, june 2009.

[34] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, 1995.

[35] C. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, pp. 9–31, 2001.

[36] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, T. X. Han, J. Tu, Z. Huang, M. Harper, Y. Xiong, D. Mcneill, R. Tuttle, and T. Huang, "VACE multimodal meeting corpus," *In Proceedings of Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[37] G. Chiou and J. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, pp. 1192–1195, August 1991.

[38] S. Chu and T. Huang, "Bimodal speech recognition using couple hidden markov models," *In Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 747–750, 2000.

[39] J. F. Cohn, "Advances in behavioral science using automated facial image analysis and synthesis," *IEEE Signal Processing Magazine*, vol. 27, November 2010.

[40] A. Colmenarez and T. Huang, "Face detection with information-based maximum discrimination," pp. 782–787, 1997.

[41] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *In Proceedings European Conference on Computer Vision*, vol. 2, pp. 484–498, 1998.

[42] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[43] T. Cootes, A. Hill, C. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image and Vision Computing*, vol. 12, pp. 355–365, July/August 1994.

[44] T. F. Cootes and C. J. Taylor, “On representing edge structure for model matching,” in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 1114–1119, 2001.

[45] S. Cox, I. Matthews, and J. Bangham, “Combining noise compensation with visual information in speech recognition,” *Proceedingsof the ESCA Workshop on Audio-Visual Speech Processing*, pp. 53–56, Rhodes, Sept. 1997.

[46] I. Craw, H. Ellis, and J. Lishman, “Automatic extraction of face features,” *In Pattern Recognition*, pp. 183–187, 1987.

[47] D. Cristinacce and T. F. Cootes, “Feature detection and tracking with constrained local models,” in *British Machine Vision Conference (BMVC)*, pp. 929–938, 2006.

[48] R. Cutler and L. Davis, “Robust real-time periodic motion detection, analysis, and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 781–796, 2000.

[49] N. Daisuke, M. Toshiki, K. Takayuki, M. Kensaku, S. Yasuhito, M. Chiyomi, I. Fumitada, A. Masami, and N. Yoshiki, “Construction of bimodal database for evaluating in-car speech recognition,” *IPSJ SIG Technical Reports*, vol. 12, pp. 35–40, 2005.

[50] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *In CVPR*, pp. 886 –893, 2005.

[51] T. Darrell, G. Gordon, J. Woodfill, H. Baker, and M. Harville, “Robust, real-time people tracking in open environments using integrated stereo, color, and face detection,” *Proceedings of the 1998 IEEE Workshop on Visual Surveillance*, pp. 26–, 1998.

[52] C. Darwin, “The expression of the emotions in man and animals,”

[53] J. G. Daugman, “Two-dimensional spectral analysis of cortical receptive field profiles,” *Vision Research*, vol. 20, no. 10, pp. 847–856, 1980.

[54] J. Daugman, “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional cortical filters,” *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1160–1169, 1985.

[55] J. G. Daugman, “Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression,” *IEEE Trans. PAMI*, vol. 36, pp. 1169–1179, July 1988.

[56] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357 – 366, aug 1980.

[57] D. Dean, *Synchronous HMMs for Audio-Visual Speech Processing*. PhD thesis, Queensland University of Technology, Brisbane, Australia,2008.

[58] T. Dean and K. Kanazawa, “A model for reasoning about persistence and causation,” *Computational Intelligence*, vol. 5, pp. 142–150, Dec. 1989.

[59] B. Dodd, “The acquisition of lip-reading skills by normally hearing children,” *Hearing by Eye: The Psychology of Lipreading*, pp. 163–175, B. Dodd and R. Campbell, Eds. London, England: Lawerence Erlbaum Associates Ltd, 1987.

[60] G. . Duchenne, “Mcanisme de la physionomie humaine,” 1862.

[61] P. Duchnowski, U. Meier, and A. Waibel, “See me, hear me: Integrating automatic speech recognition and lip reading,” *In Proceedings of the International Conference on Spoken Language and Processing (ICSLP)*, pp. 547–550, 1994.

[62] N. Eagle and A. Pentland, “Social network computing,” *Ubicomp 2003: Ubiquitous Computing, Springer-Verlag Lecture Notes in Computer Science*, pp. 289–296, 2003.

[63] G. Edwards, C. Taylor, and T. Cootes, “Learning to identify and track faces in image sequences,” *In International Conference on Computer Vision*, pp. 317–322, 1995.

[64] A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” *In IEEE International Conference on Computer Vision*, pp. 726–733, 2003.

[65] A. Efros, C. Berg, G. Mori, and J. Malik, “Recognizing Action at a Distance,” in *ICCV*, 2003.

[66] P. Ekman and W. Friesen, “The repertorie of noverbal behavior: Catergories,origins, usage and coding,” *Semiotica*, pp. 49–98, 1969.

[67] P. Ekman and W. Friesen, “Manual for the facial action coding system,” *Consulting Psychologists Press*, 1977.

[68] B. Fasel and J. Luttin, “Automatic facial expression analysis: a survey,” *In Pattern Recognition*, pp. 259–275, 2003.

[69] D. J. Field, “Relations between the statistics of natural images and the response properties of cortical cells,” *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2393, 1987.

[70] Y. Freund and R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *in Computational Learning Theory: Eurocolt ’95*, pp. 23–27, Springer-Verlag, 1995.

[71] Y. Fu, X. Zhou, M. Liu, and T. Hasegawa-Johnson, Hunag, “Lipreading by locality discriminant graph,” *In International Conference on Image Processing (ICIP)*, pp. 325–328, 2007.

[72] D. Gabor, “Theory of communication,” *Journal of the Institution of Electrical Engineers (London)*, vol. 93, no. III, pp. 429–457, 1946.

[73] D. Gabor, “Theory of communication,” *In Journal of the Institution of Electrical Engineers*, pp. 429–457, 1946.

[74] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi, “The NIST meeting room pilot corpus,” *In Proceedings of the Language Resource and Evaluation Conference (LREC)*, 2004.

[75] D. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82 – 98, 1999.

[76] A. Goldschen, O. Garcia, and E. Petajan, “Rationale for phoneme-viseme mapping and feature selection in visual speech recognition,” *Speechreading by Humans and Machines*, pp. 505–515, 1996.

[77] H. Graf, T. Chen, E. Petajan, and E. Cosatto, “Locating faces and facial parts,” *International Workshop Automatic Face and Gesture Recognition*, pp. 41–46, 1995.

[78] R. Gross and V. Brajovic, “An image pre-processing algorithm for illumination invariant face recognition,” in *4th International Conference on Audio-and Video Based Biometric Person Authentication (AVBPA)*, pp. 10–18, Springer-Verlag, 2003.

[79] R. Gross, I. Matthews, and S. Baker, “Generic vs. person specific active appearance models,” *Image and Vision Computing*, vol. 23, pp. 1080–1093, November 2005.

[80] R. Gross, J. S. Baker, I. Matthews, and T. Kanade, “Multi-PIE,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[81] S. Gurbuz, Z. Tufekci, E. Patterson, and J. Gowdy, “Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition,” pp. 177–180, Salt Lake City, UT, USA,2001.

[82] F. Gustafsson, “The marginalized likelihood ratio test for detecting abrupt changes,” *Transactions on Automatic Control*, no. 1, 1996.

[83] A. Hanjalic and L. Xu, “Affective video content representation and modeling,” *IEEE Transactions on Multimedia*, vol. 7, pp. 143–154, 2005.

[84] Z. Harchaoui, F. Bach, and E. Moulines, “Kernel Change-Point Analysis,” in *NIPS*, 2009.

[85] M. Hoai and F. De la Torre, “Maximum Margin Temporal Clustering,” in *AISTATS*, 2012.

[86] H. Joho, J. Jose, R. Valenti, and N. Sebe, “Exploiting facial expressions for affective video summarisation,” *In Proceeding of the ACM International Conference on Image and Video Retrieval*, 2009.

[87] H. Joho, J. Staiano, N. Sebe, and J. Jose, “Looking at the viewer: analysing facial activity to detect personal highlights of multime- dia contents,” *In Multimedia Tools and Applications*, pp. 1–19, 2011.

[88] T. Jordan and P. Sergeant, “Effects of facail image size on visual and audio-visual speech processing,” *In Hearing by Eye II (R.Campbell, B.Dodd, and D.Burnham, eds)*, pp. 155–176, 1998.

[89] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Orlando, FL, USA), pp. 4160–4163, 2002.

[90] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *In FG*, pp. 46–53, March 2000.

[91] J. Kim and E. Andre, "Emotion recognition based on physiolog- ical changes in music listening," *In TPAMI*, pp. 2067–2083, 2008.

[92] T. Kleinschmidt, D. Dean, S. Sridharan, and M. Mason, "A continuous speech recognition protocol for the avicar database," *in Proceedings of the International Conference on Signal Processing and Communication Systems*, pp. 339–344, Australia,2007.

[93] S. Koelstra, A. Yazdani, M. Soleymani, M. Christian, J. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos," *Brain Informatics, ser. Lecture Notes in Computer Science*, vol. 6334, pp. 89–100, 2010.

[94] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *IEEE Computer Society*, 2009.

[95] E. Land and J. McCann, "Lightness and retinex theory," *Journal of Optical Society of America*, vol. 61, 1971.

[96] F. Lavagetto, "Converting speech into lip movements: a multimedia telephone for hard of hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3 ,no,1, pp. 90–102, 1995.

[97] B. Lee, *Robust Speech Recognition in a Car Using a Microphone Array*. PhD thesis, University of Illinois at Urbana-Champaign, 2006.

[98] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: An audiovisual speech corpus in a car environment," *In Proc. Interspeech 2004*, pp. 2489–2492, Jeju Island, Korea.

[99] T. Leung, M. Burl, and P. Perona, “Finding faces in cluttered scenes using random labeled graph matching,” *IEEE Intl Conf. Computer Vision*, pp. 637–644, 1995.

[100] M. Lew, “Information theoretic view-based and modular face detection,” pp. 198–203, 1996.

[101] Z. Li, D. Lin, and X. Tang, “Nonparametric discriminant analysis for face recognition,” *IEEE Trans. PAMI*, vol. 31, pp. 755–761, April 2009.

[102] S. Li, J. Sherrah, and H. Liddell, “Multi-view face detection using support vector machines and eigenspace modelling,” *Proceedings of the Interna- tional Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pp. 241–244, 2000.

[103] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. Nefian, “Speaker independent audio-visual continuous speech recognition,” vol. 2, pp. 25–28, August 2002.

[104] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, pp. I–900–I–903 vol.1, 2002.

[105] C. Lisetti and F. Nasoz, “Using noninvasive wearable computers to recognize human emotions from physiological signals,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1672–1687, 2004.

[106] X. Liu, “Generic face alignment using boosted appearance model,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[107] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition,” *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[108] P. Lockwood, C. Baillargeat, J. M. Gillot, J. Boudy, and G. Faucon, “Noise reduction for speech enhancement in cars: non-linear spectral subtraction / kalman

filtering.," *In European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 83–86, 1991.

[109] D. Lowe, "Distinctive image features from scale-invariant keypoints," *In IJCV*, pp. 91–110, 2004.

[110] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision.," in *International Joint Conference on Artificial Intelligence*, pp. 674 – 679, 1981.

[111] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *In Proceeding of the International Joint Conference on Artifical Intelligence*, 1981.

[112] P. Lucey, *Lipreading across multiple views*. PhD thesis, Queensland University of Technology, Brisbane, Australia,2007.

[113] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kande dataset (ck+): A complete facial expression dataset for action unit and emotion-specified expression," *In CVPR4HB*, 2010.

[114] P. Lucey, J. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*, 2011.

[115] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," *In FG*, 2011.

[116] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 24–28, 2006.

[117] A. Madan, R. Caneel, and A. Pentland, “Groupmedia: Distributed multimodal interfaces,” *In International Conference on Multimodal Interfaces*, 2004.

[118] U. Mahbub and M. Ahad, “Action recognition algorithm based on optical flow and ransac in frequency domain,” *SICE Annual Conference (SICE), 2011 Proceedings of*, pp. 1627–1631, 2011.

[119] U. Mahbub, H. Imtiaz, and A. R. Ahad., “An optical flow-based action recognition algorithm,” *CVPRW on Gesture Recognition*, 2010.

[120] I. Matthews and S. Baker, “Active appearance models revisited,” *International Journal of Computer Vision*, vol. 60, pp. 135 – 164, November 2004.

[121] I. Matthews, J. Bangham, and S. Cox, “Audio-visual speech recognition using multiscale nonlinear image decomposition,” *International Conference on Spoken Language Processing*, pp. 38–41, Philadelphia, PA, USA, 1996.

[122] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, “A comapison of model and transform-based visual features for audio-visual lvcsr,” *International Conference on Multimedia and Expo*, pp. 22–25, 2001.

[123] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 3, pp. 305 –317, 2005.

[124] M. McGrath and Q. Summerfield, “Intermodal timing relations and audio-visual speech recognition,” *Journal of the Acoustical Society of America*, vol. 77,no. 2, pp. 678–685, February 1985.

[125] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The extended M2VTS database,” *International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)*, 1999.

[126] J. Miao, B. Yin, K. Wang, L. Shen, and X. Chen, “A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template,” *In Pattern Recognition*, pp. 1237–1248, 1999.

[127] M. Mooij, “Gloabal marketing and advertising: Understanding cultural paradoxes,” 2009.

[128] J. Movellan and G. Chadder, “Channel separability in the audio visual integration of speech: A bayesian approach,” *Speechreading by Humans and Machines*, pp. 473–487, D. G. Stork and M. E. Hennecke, Eds. Berlin: Springer, 1996.

[129] W. Murch, *In the Blink of an Eye: A Perspective on Film Editing*. Silman-James Press, 2001.

[130] R. Navarathna, D. Dean, P. Lucey, C. Fookes, and S. Sridharan, “Recognizing audio-visual speech in vehicles using the AVICAR database,” *In Australasian International Conference on Speech Science and Technology (SST)*, pp. 110–113, 2010.

[131] A. Nefian and M. Hayes, “Face detection and recognition using hidden Markov models,” pp. 141–145, 1998.

[132] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, “Large-vocabulary aduio-visual speech recognition: A summary of the johns hopkins summer 2000 workshop,” *In Proceedings of the Workshop on Multimedia Signal Processing, Special Section on Joint Audio-Visual Processing*, 2001.

[133] C. Neti, G. Potamianos, J. Luettin, H. Matthews, I.and Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, “Audio-visual speech recognition, final workshop 2000 report,” 2000.

[134] R. Nevatia, J. Hobbs, and B. Bolles, “An ontology for video event representation,” *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’04)*, 2004.

[135] R. Nevatia, T. Zhao, and S. Hongeng, “Hierarchical language-based representation of events in video streams,” *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW*, vol. 4, pp. 39–39, 2003.

[136] T. Ojala, M. Pietikainen, and D. Harwood, “A comparative study of texture measures with classification based on feature distributions,” *In Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[137] A. V. Oppenheim and A. S. Willsky, *Signals & Systems*. Prentice Hall, 2nd ed., 1996.

[138] A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera, and E. Zacur, “Av@car: A spanish multichannel multimodel corpus for in-vehicle automatic audio-visual speech recognition,” *IV Inernational Converence on Language Resources and Evaluation*, vol. 3, pp. 763–767, 2004, Lisbon, Portugal.

[139] E. Page, “Continuous Inspection Schemes,” *Biometrika*, 1954.

[140] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, “Human computing and machine understanding of human behavior: a survey,” *Proceedings of the 8th International conference on Multimodal interfaces*, pp. 239–248, 2006.

[141] M. Pantic and L. Rothkrantz, “Automatic analysis of facial expressions: the state of the art,” *In TPAMI*, pp. 1424 –1445, 2000.

[142] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “CUAVE: a new audio-visual database for multimodal human-computer interface research,” *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 217–2020, 2002.

[143] J. Pearl, “Probabilitic reasoning in intelligent systems: Networks of plausible inference,” *Morgan Kaufmann, San Mateo, CA*, 1988.

[144] A. Pentland, “Socially aware computation and communication,” *IEEE Computer*, pp. 63–70, 2005.

[145] E. Petajan, “Automatic lipreading to enhance speech recognition,” *IEEE Global Telecommunications Conference*, pp. 265–272, 1984.

[146] S. Pigeon and L. Vandendorpe, “The M2VTS multimodal face database,” *International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)*, pp. 403–409, 1997.

[147] R. Polana and R. Nelson, “Low level recognition of human motion,” *In Proceedings of the Motion of Non-Rigid and Articulated Objects*, pp. 77–82, 1994.

[148] G. Potamianos and P. Lucey, “Audio-visual asr from multiple views inside smart rooms,” *International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 35–40, 2006.

[149] G. Potamianos and P. Lucey, “Audio-visual ASR from multiple views inside smart rooms,” *International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 35–40, 2006.

[150] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1293–1296, 2003.

[151] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, “Recent advances in the automatic recognition of audio-visual speech,” *Proceedings of the IEEE*, vol. 9, 2003.

[152] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audio-visual automatic speech recognition: An overview,” *In Issues in Visual and Audio-Visual Speech Processing (AVSP)*, 2004.

[153] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, “A cascade image transform for speaker independent automatic speechreading,” *International Conference on Multimedia and Expo (ICME)*, vol. 2, pp. 1097–1100, 2000.

[154] . Psarrou, S. Gong, and M. Walter, “Recognition of human gestures and behaviour based on motion trajectories,” *Image and Vision Computing*, pp. 349–358, 2002.

[155] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition.,” *Proceedings of the IEEE*, 1989.

[156] T. A. Ranney, W. Garrott, and M. Goodman, “NHTSA driver distraction research: past, present and future,” *International Technical Conference on the Enhanced Safety of Vehicles*, 2001.

[157] J. Robert-Ribes, J. Schwartz, and P. Lallouache, T.and Escudier, “Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise,” *Journal of the Acoustical Society of America*, vol. 103, no.6, pp. 3677–3689, 1998.

[158] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, “Data-Driven Crowd Analysis in Videos,” in *ICCV*, 2011.

[159] J. Rottenberg, R. Ray, and J. Gros, “Emotion elicitation using films,” *Series in affective science. Oxford University*, pp. 9–28, 2007.

[160] H. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 23–38, 1998.

[161] T. Ruf, Ernst.A., and C. Kublbeck, “Face detection with the sophisticated high-speed object recognition engine (shore),” *Microelectronic Systems*, pp. 243–252, 2011.

[162] A. Ryan, J. Cohn, S. Lucey, J. Saragih, P. Lucey, F. la Torre, and A. Rossi, “Automated facial expression recognition system.,” *In International Carnahan Conference on Security Technology*, 2009.

[163] T. Sakai, M. Nango, and S. Fujibayashi, “Line extraction and pattern detection in a photograph,” *In Pattern Recognition*, pp. 233–248, 1969.

[164] A. Samal and P. Iyengar, “Automatic recognition and analysis of human faces and facial expressions: A survey,” *Pattern Recognition*, pp. 65–77, 1992.

[165] A. Samal and P. lyengar, “Human face detection using silhouettes,” *In Pattern Recognition and Artificial Intelligence*, pp. 845–867, 1995.

[166] J. Saragih, S. Lucey, and J. F. Cohn, “Deformable model fitting with a mixture of local experts,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2248 – 2255, 2009.

[167] J. Saragih, S. Lucey, and J. Cohn, “Real-time avatar animation from a single image,” *IEEE Face and Gesture Recognition*, 2011.

[168] J. Schlenzig, E. Hunter, and R. Jain, “Recursive identification of gesture inputs using hidden markov models,” *In Applications of Computer Vision*, pp. 187–194, 1994.

[169] H. Schneiderman and T. Kanade, “A histogram-based method for detection of faces and cars,” pp. 504–507, 2000.

[170] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” *In International Conference on Pattern Recognition*, pp. 32–36, 2004.

[171] N. Schwarz and F. Strack, “Reports of subjective well-being:judgmental processes and their methodological implications,” *Well-being: The foundations of hedonic psychology*, 1999.

[172] M. Shan, F. Kuo, M. Chiang, and Y. Lee, “Emotion-based music recommendation by affinity discovery from film music,” *An International Journal Expert Systems with Applications*, 2009.

[173] C. Shannon, “A Mathematical Theory of Communication,” *The Bell System Technical Journal*, 1948.

[174] S. Sirohey, “Human face segmentation and identification,” *Technical Report CS-TR-3176, Univ. of Maryland*, 1993.

[175] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, “Affective characterization of movie scenes based on content analysis and physiological changes,” *International Journal of Semantic Computing*, vol. 3, 2009.

[176] Q. Su and P. Silsbee, “Robust audiovisual integration using semicontinuous hidden markov models,” *International Conference on Spoken Language Processing*, Philadelphia, PA, USA, 1996.

[177] W. Sumby and I. Pollack, “Visual contribution to speech intelligibility,” *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[178] A. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” *Hearing by Eye: The Psychology of Lip-Reading*, pp. 3–51, 1987.

[179] A. Summerfield, A. MacLeod, M. McGrath, and M. Brooke, “Lips, teeth, and the benefits of lipreading,” *Handbook of Research on Face Processing*, pp. 223–233, A. Young and H. Ellis, Eds. Amsterdam, The Netherlands: Elsevier Science Publishers, 1989.

[180] D. Sun, S. Roth, and M. Black, “Secrets of optical flow estimation and their principles,” *In CVPR*, pp. 2432 –2439, 2010.

[181] K. Sung and T. Poggio, “Example-based learning for view-based human face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 39–51, 1998.

[182] W. Suwa, N. Sugie, and K. Fujimora, “A preliminary note on pattern recognition of human emotional expression,” *International Joint Confernece on Pattern Recognition*, pp. 408–410.

[183] K. Takahashi, “Remarks on emotion recognition from multi-modal bio-potential signals,” *IEEE International Conference on Industrial Technology*, vol. 3, pp. 1138 – 1143, 2004.

[184] K. Teddy, “A survey on behavior analysis in video surveillance for homeland security applications,” *IEEE Applied Imagery Pattern Recognition Workshop*, pp. 1–8, 2008.

[185] T. Teixerira, M. Wedel, and R. Pieters, “Emotion-induced engagement in internet video advertisements,” *Journal of Marketing Research*, 2011.

[186] P. Tessier, J. Robert-Ribes, J. Schwartz, and A. Gurin-Dugu, “Comparing models for audiovisual fusion in a noisy-vowel recognition task,” *Speech Communication*, vol. 7,no. 6, pp. 629–642, 1999.

[187] B. Theobald, “Evaluating error functions for robust active appearance models,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 149–154, 2006.

[188] M. Tkalcic, U. Burnik, and A. Kosir, “Using affective parameters in a content-based recommender system for images,” *User Modeling and User-Adapted Interaction*, vol. 20, pp. 279–311, 2010.

[189] M. Tomlinson, M. Russell, and N. Brooke, “Integrating audio and visual information to provide highly robust speech recognition,” *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 821–824, 1996.

[190] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[191] C. Varner and G. Dickinson, "The lecture, an analysis and review of research," *Adult Education Quarterly*, vol. 17, pp. 85–100, 1967.

[192] J. Vidit and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," Tech. Rep. UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[193] A. Vinciarelli, M. Pantic, and H. Bourland, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, 2009.

[194] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *In CVPR*, pp. 511–518, 2001.

[195] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movel-lan., "Automated drowsiness detection for improved driver saftey," *In International Conference on Automotive Technologies*, 2008.

[196] S. Wang, K. Huang, and T. Tan, "A compact optical flow based motion representation for real-time action recognition in surveillance scenes," *In ICIP*, pp. 1117–1120, 2009.

[197] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Towards practical smile detection," *In TPAMI*, 2009.

[198] C. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. PAMI*, vol. 19, pp. 775–779, July 1997.

[199] H.-W. H. Xuedong Huang, Alex Acero, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* 2001.

[200] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden markov model,” *In Computer Vision and Pattern Recognition (CVPR)*, pp. 379–385, 1992.

[201] M. Yang, N. Abuja, and D. Kriegman, “Mixtures of linear subspaces for face detection,” *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pp. 70–76, 2000.

[202] G. Yang and S. Huang, “Human face detection in complex background”,” *Pattern Recognition*, pp. 53–63, 1994.

[203] M. Yang, D. Kriegman, and Ahuja.N., “Detecting faces in images: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 34–58, 2002.

[204] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, “Audio-visual speech recognition using lip movement extracted from side-face images,” *In International Conference on Auditory Visual Speech Processing (AVSP)*, pp. 117–120, 2003.

[205] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland *The HTK Book (for HTK Version 3.2.1)*, 2002.

[206] A. Yuille, P. Hallinan, and D. Cohen, “Feature extraction from faces using deformable templates,” *Internation Journal of Computer Vision*, pp. 99–111, 1992.

[207] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *In TPAMI*, 2009.