Visual Observation of Human Emotions

Varun Jain

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 19 octobre 2011

Présentée par

Varun JAIN

Thèse dirigée par **James L. CROWLEY**
codirigée par **Augustin LUX**

préparée au sein du **Laboratoire d'Informatique de Grenoble à l'INRIA Grenoble Rhône-Alpes**
dans **l'École Doctorale de Mathématiques, Sciences et Technologies de l'Information, Informatique**

Visual Observation of Human Emotions

Thèse soutenue publiquement le **30 mars 2015**,
devant le jury composé de :

Mme. Elisabeth ANDRÉ
Professeur, Universität Augsburg, Rapporteur
Mme. Laurence DEVILLERS
Professeur, LIMSI-CNRS/Université Paris-Sorbonne, Rapporteur
M. Mohamed CHETOUANI
Professeur, Université Pierre et Marie Curie, Examinateur
Mme. Sylvie PESTY
Professeur, Université Pierre Mendès-France, Président
M. James L. CROWLEY
Professeur, Institut Polytechnique de Grenoble, Directeur de thèse
M. Augustin LUX
Professeur, Institut Polytechnique de Grenoble, Co-Directeur de thèse

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

Dedicated to Mummy, Papa and Nishant

ABSTRACT

In this thesis we focus on the development of methods and techniques to infer affect from visual information. We focus on facial expression analysis since the face is one of the least occluded parts of the body and facial expressions are one of the most visible manifestations of affect. We explore the different psychological theories on affect and emotion, different ways to represent and classify emotions and the relationship between facial expressions and underlying emotions.

We present the use of multiscale Gaussian derivatives as an image descriptor for head pose estimation, smile detection before using it for affect sensing. Principal Component Analysis is used for dimensionality reduction while Support Vector Machines are used for classification and regression. We are able to employ the same, simple and effective architecture for head pose estimation, smile detection and affect sensing. We also demonstrate that not only do multiscale Gaussian derivatives perform better than the popular Gabor Filters but are also computationally less expensive to compute.

While performing these experiments we discovered that multiscale Gaussian derivatives do not provide an appropriately discriminative image description when the face is only partly illuminated. We overcome this problem by combining Gaussian derivatives with Local Binary Pattern (LBP) histograms. This combination helps us achieve state-of-the-art results for smile detection on the benchmark GENKI database which contains images of people in the "wild" collected from the internet. We use the same description method for face recognition on the CMU-PIE database and the challenging extended YaleB database and our results compare well with the state-of-the-art. In the case of face recognition we use metric learning for classification, adopting the Minkowski distance as the similarity measure. We find that L_1 and L_2 norms are not always the optimum distance metrics and the optimum is often an L_p norm where p is not an integer.

Lastly we develop a multi-modal system for depression estimation with audio and video information as input. We use Local Binary Patterns - Three Orthogonal Planes (LBP-TOP) features to capture intra-facial movements in the videos and dense trajectories for macro movements such as the movement of the head and shoulders. These video features along with Low Level Descriptor (LLD) audio features are encoded using Fisher Vectors and finally a Support Vector Machine is used for regression. We discover that the LBP-TOP features encoded with Fisher Vectors alone are enough to outperform the baseline method on the Audio Visual Emotion

Challenge (AVEC) 2014 database. We thereby present an effective technique for depression estimation which can be easily extended for other slowly varying aspects of emotions such as mood.

Keywords: Affect Sensing, Facial Image Analysis, Automated Facial Expression Analysis.

RÉSUME

Cette thèse a pour sujet le développement de méthodes et de techniques permettant d'inférer l'état affectif d'une personne à partir d'informations visuelles. Plus précisement, nous nous intéressons à l'analyse d'expressions du visage, puisque le visage est la partie la mieux visible du corps, et que l'expression du visage est la manifestation la plus évidente de l'affect. Nous étudions différentes théories psychologiques concernant affect et émotions, et différentes facons de représenter et de classifier les émotions d'une part et la relation entre expression du visage et émotion sousjacente d'autre part.

Nous présentons les dérivées Gaussiennes multi-échelle en tant que descripteur d'images pour l'estimation de la pose de la tête, pour la détection de sourire, puis aussi pour la mesure de l'affect. Nous utilisons l'analyse en composantes principales pour la réduction de la dimensionalité, et les machines à support de vecteur pour la classification et la regression. Nous appliquons cette même architecture, simple et efficace, aux différents problèmes que sont l'estimation de la pose de tête, la détection de sourire, et la mesure d'affect. Nous montrons que non seulement les dérivées Gaussiennes multi-échelle ont une performance supérieure aux populaires filtres de Gabor, mais qu'elles sont également moins coûteuses en calculs.

Lors de nos expérimentations nous avons constaté que dans le cas d'un éclairage partiel du visage les dérivées Gaussiennes multi-échelle ne fournissent pas une description d'image suffisamment discriminante. Pour résoudre ce problème nous combinons des dérivées Gaussiennes avec des histogrammes locaux de type LBP (Local Binary Pattern). Avec cette combinaison nous obtenons des résultats à la hauteur de l'état de l'art pour la détection de sourire dans le base d'images GENKI qui comporte des images de personnes trouvées "dans la nature" sur internet, et avec la difficile "extended YaleB database". Pour la classification dans la reconnaissance de visage nous utilisons un apprentissage métrique avec comme mesure de similarité une distance de Minkowski. Nous obtenons le résultat que les normes L_1 and L_2 ne fournissent pas toujours la distance optimale; cet optimum est souvent obtenu avec une norme L_p où P n'est pas entier.

Finalement, nous développons un système multi-modal pour la détection de dépressions nerveuses, avec en entrée des informations audio et vidéo. Pour la détection de mouvements intra-faciaux dans les données vidéo nous utilisons de descripteurs de type LBP-TOP (Local Binary Patterns -Three Orthogonal Planes), alors que nous utilisons des trajectoires

denses pour les mouvements plus globaux, par exemple de la tête ou des épaules. Nous avons trouvé que les descripteurs LBP-TOP encodés avec des vecteurs de Fisher suffisent pour dépasser la performance de la méthode de reférence dans la compétition "Audio Visual Emotion Challenge (AVEC) 2014". Nous disposons donc d'une technique effective pour l'evaluation de l'état dépressif, technique qui peut aisement être étendue à d'autres formes d'émotions qui varient lentement, comme l'humeur (mood an Anglais).

Mots-clés: Perception de l'état affectif, Analyse d'image du visage, Reconnaissance d'expression du visage.

PUBLICATIONS

Related Publications

[1] **V. Jain**, J.L. Crowley. Smile Detection Using Multi-scale Gaussian Derivatives. 12th WSEAS International Conference on Signal Processing, Robotics and Automation, 2013.

[2] **V. Jain**, J.L. Crowley. Head Pose Estimation Using Multi-scale Gaussian Derivatives. Scandinavian Conference On Image Analysis 2013.

[3] **V. Jain**, J.L. Crowley, A. Lux. Facial Expression Analysis And The PAD Space. 11th Pattern Recognition and Image Analysis, 2013.

[4] **V. Jain**, J.L. Crowley, A. Lux. Local Binary Patterns Calculated Over Gaussian Derivative Images. International Conference on Pattern Recognition, 2014.

[5] **V. Jain**, J.L. Crowley, A.K. Dey, A. Lux. Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. 4th Audio Visual Emotion Challenge Workshop, 2014.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisors Prof. Crowley and Prof. Lux for holding my hand through this long and rather arduous journey. Apart from the inspiration and support they provided in the course of this research, they also provided me with generous resources for conducting this research. I am also grateful to the members of the jury: Prof. André, Prof. Devillers, Prof. Pesty and Prof. Chetouani for taking time out from their busy schedules and participating in the defense of this thesis.

I take this opportunity to thank the wonderful people I met at Carnegie Mellon University, where I was fortunate enough to spend six months as a visiting researcher. I had the chance to collaborate with some of the smartest people I have ever met including Prof. Dey, Nikola, Julian, Jin-Hyuk and Chandrayee.

I should also thank all the people in PRIMA for the cordial atmosphere in the team. I especially appreciate all the help and company provided by Catherine, Etienne, Claudine, Doms, Remi, Sergi and Thierry. It was a pleasure to be in the company of such warm and talented people. There was never a dearth of ideas for research and never a problem that they could not solve.

Finally, I must thank the people without which none of this would have been possible, my mother, little brother Nishant, my dad, grandparents, my friends from Dehradun and the little witty bear from Finland.

Varun Jain

CONTENTS

1 INTRODUCTION 1
- 1.1 Problem addressed in the thesis 1
- 1.2 Approach adopted 2
- 1.3 Methodology adopted 3
- 1.4 Insights 5
- 1.5 Thesis Summary 6

2 HUMAN EMOTIONS 11
- 2.1 Feedback Theories on Emotion 11
- 2.2 Basic Emotions vs. dimensional theories of affect 12
  - 2.2.1 Basic Emotions 13
  - 2.2.2 Dimensional Theories of Emotions 14
  - 2.2.3 Plutchik's theory 15
- 2.3 Facial Expressions and the Facial Action Coding System 17
- 2.4 Mood 18
- 2.5 Depression 20
- 2.6 Conclusion 21

3 STATIC IMAGE ANALYSIS FOR SENSING AFFECT 23
- 3.1 Image description methods for still images 23
  - 3.1.1 Steerable filter methods 23
  - 3.1.2 Local Binary Patterns 26
- 3.2 Machine Learning methods 28
  - 3.2.1 Dimensionality reduction and Principal Component Analysis 28
  - 3.2.2 Kernel Methods and the Support Vector Machine 29
  - 3.2.3 K-Nearest Neighbors method 30
- 3.3 Experiments 31
  - 3.3.1 Head Pose Estimation 31
  - 3.3.2 Smile Detection 37
  - 3.3.3 Affect Sensing 46
  - 3.3.4 Face Recognition 55
- 3.4 Conclusion and summary 59

4 VIDEO ANALYSIS FOR SENSING AFFECT 61
- 4.1 Methods for video description 61
  - 4.1.1 Local Binary Patterns-Three Orthogonal Planes 61
  - 4.1.2 Dense Trajectories 62
  - 4.1.3 Space Time Interest Points 64
- 4.2 Encoding Techniques 64
  - 4.2.1 Bag of Visual Words 65
  - 4.2.2 Sparse Coding 66

4.2.3 Fisher Vector Encoding 67
4.3 Experiments 67
4.3.1 Facial Expression Recognition 67
4.3.2 Audio Visual Emotion Challenge 2014 Depression sub-challenge 70
4.4 Conclusion 76
5 CONCLUSION 79
5.1 Lessons learned 81
5.2 Impact of the present work 82
5.3 Future scope of this work 83
A APPENDIX 85
A.1 Summary of databases 85
BIBLIOGRAPHY 87

LIST OF FIGURES

Figure 1 General architecture adopted 2
Figure 2 Circumplex model of emotions [115] 14
Figure 3 Plutchik's wheel of emotions [3] 16
Figure 4 Plutchik's cone of emotions [4] 16
Figure 5 Shift in emotional response due to mood 19
Figure 6 Computing LBP response from a pixel's local neighbourhood. 27
Figure 7 LBP Histogram computation [2] 27
Figure 8 SVM classifying linearly separable data [5] 29
Figure 9 SVM with data mapped to a higher dimensional space using kernel [6] 30
Figure 10 A small sequence from the Pointing04 dataset. 32
Figure 11 How the Pointing04 database was collected 32
Figure 12 (a) Graph of Correlation Coeff. vs. C-parameter and $1/\gamma$ for pan and (b) Graph of Correlation Coeff. vs. C-parameter at $1/\gamma = 11$ for pan 33
Figure 13 (a) Graph of Correlation Coeff. vs. C-parameter and $1/\gamma$ for tilt and (b) Graph of Correlation Coeff. vs. C-parameter at $1/\gamma = 6$ for tilt 34
Figure 14 Schematic of our approach. 36
Figure 15 (a) Pan=-15, Tilt=-15 and (b) Pan=0, Tilt=0 were predicted using our approach 37
Figure 16 Examples of (top two rows) real-life smile faces and (bottom two rows) nonsmile faces, from the GENKI-4K database. 38
Figure 17 Imagette divided into cells of 4 X 4 pixels 39
Figure 18 (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 81$ 40
Figure 19 ROC curve for our smile detector 41
Figure 20 Smile intensity using probability estimates 42
Figure 21 Schematic of our approach 42
Figure 22 Examples from the GENKI-4K database illustrating the illumination problems 42
Figure 23 Creating the features: a) original image, b) Gaussian derivative images, and c) concatenation of resulting histograms after applying LBP. 43

Figure 24 Accuracy(%) of different descriptors over the GENKI-4K database 44
Figure 25 BER of different descriptors over the GENKI-4K database 44
Figure 26 Proportion of images with different pose 45
Figure 27 Accuracy with different poses 45
Figure 28 Example Images from the FEED dataset 48
Figure 29 Example Images from the CK database 49
Figure 30 Basic emotions in the affect space 50
Figure 31 (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ for pleasure and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 190$ for pleasure 51
Figure 32 (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ for arousal and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 280$ for arousal 52
Figure 33 Schematic of our approach 52
Figure 34 ROC of the classifier for Pleasure 53
Figure 35 ROC of the classifier for Arousal 53
Figure 36 Comparison of results 54
Figure 37 Results on the FEED database 55
Figure 38 Sample images from the CMU-PIE database showing the different lighting conditions 57
Figure 39 Recognition rate with different reference images 57
Figure 40 Sample images from the extended YaleB database showing the different lighting conditions 58
Figure 41 LBP-TOP histogram computation 62
Figure 42 LBP-TOP computation for facial images 62
Figure 43 LBP-TOP histogram concatenation for facial images 63
Figure 44 Dense Trajectory computation 64
Figure 45 K-means and codebook generation [1] 65
Figure 46 LBP-TOP computed over Gaussian derivative images 69
Figure 47 System Architecture 74
Figure 48 Root Mean Square Error (RMSE) vs. number of clusters 75

LIST OF TABLES

Table 1 Basic Emotions in the PAD space 15
Table 2 AU's associated with basic emotions 18
Table 3 Our MAE as compared with the state-of-the-art 35
Table 4 Our accuracy over discrete poses as compared with the state-of-the-art 35
Table 5 Confusion Matrix for Pan, true values are in the first column, predicted values in the first row 35
Table 6 Confusion Matrix for Tilt, true values are in the first column, predicted values in the first row 36
Table 7 Comparison of prediction time with and without using PCA 36
Table 8 Our accuracy using Multi-scale Gaussian Derivatives (MGD) compared with the accuracy obtained using Gabor Energy Filters (GEF) 41
Table 9 Comparison of prediction time with and without using PCA 41
Table 10 Confusion matrix for 2 class classification 44
Table 11 Labels for the 6 basic emotions 50
Table 12 Comparison of time required for calculating the two types of features 54
Table 13 Comparison of prediction time with and without using PCA 54
Table 14 Comparison of prediction time 55
Table 15 Maximum and average accuracy attained by different methods 58
Table 16 Accuracy(%) over the 4 subsets using different methods 59
Table 17 Accuracy(%) over the 4 subsets using Gaussian derivatives, LBP and their proposed combination 59
Table 18 Our accuracy compared to the accuracy obtained using conventional LBP-TOP features 70
Table 19 Errors for different combinations of descriptors on the development set 75
Table 20 Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for different encoding techniques 75
Table 21 Errors for different sizes of time slice 75
Table 22 Comparison of errors with baseline 76

1 INTRODUCTION

The transition of computing environments from being computer-centered to being human-centered requires a major transformation in how humans interact with computers. The human-centered computing paradigm entails that the interaction between humans and computers should be natural and similar to human social interaction. Effective human social interactions pivots on the successful interpretation of a variety of nonverbal communicative cues such as facial expressions, body language, gestures, postures among other things. The ability to correctly perceive and interpret these nonverbal cues is often associated with emotional intelligence. In the future the "naturalness" and ease of human-machine interaction will depend on how well machines would be able to observe and interpret these previously mentioned nonverbal cues. A context-aware system which could adapt to its user's needs by inferring the affective state of a person would make human-machine interaction more convenient for the user, reduce the cognitive load and make the interaction more intuitive.

A logical step towards the development of an efficient, naturalistic and context-aware human-machine interaction system would be to build technologies to automate the inference of the affective state of a person.

1.1 PROBLEM ADDRESSED IN THE THESIS

In this thesis we investigate the use of computer vision and machine learning techniques for inferring the affective state of a person. An affective state is psycho-physiological construct i.e. it has both psychological and physiological components to it and is a result of a humans' interaction with stimuli. The affective state is manifested in a variety of physiological signals: facial expressions, vocal prosody, heart rate variability, blood volume pulse, galvanic skin response among others.

Facial expressions and gestures play a vital role in social interaction. The face is one of the most visible and least occluded parts of the body. For effective human-computer interaction it is important that computers be equipped with the ability to recognize facial expressions and possibly infer the underlying affective state. Therefore our aim has been to develop techniques for the extraction of visual information from images and videos and for the inference of the affective state using this visual information.

1.2 APPROACH ADOPTED

We adopt a global appearance-based approach for capturing visual information from the facial region. Alternative approaches for facial image analysis include facial keypoint tracking methods [73, 72] and facial model fitting techniques such as Active Appearance Models (AAM) [23] and Active Shape Models (ASM) [22]. A major advantage that appearance based approaches have over model fitting based approaches is that they are computationally less expensive. Keypoint detectors and trackers on the other hand are prone to detection failures and initialization problems. Architectures employing global image appearance techniques can easily be adapted for versatile applications as we demonstrate in this thesis. An architecture developed for a task such as a head pose estimation can easily adapted for another task such as smile detection keeping the general structure intact.

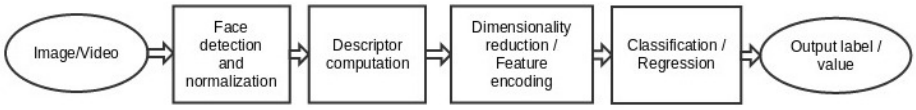


Figure 1: General architecture adopted

A simplified architecture for an appearance based system for facial image analysis is shown in figure 1. The input is in the form of an image or a video. In the next step the face is located, this is commonly done using the Viola-Jones face detector [120]. Additionally, facial landmarks such as the eyes and the nose are located. The region of interest containing the face is then normalized followed by descriptor calculation over the normalized image. Image features are often concatenated to produce a feature vector which may be reduced in dimensions using feature selection techniques or dimensionality reduction techniques such as principal component analysis (PCA) [61]. PCA is a statistical technique for projecting possibly correlated data to a new coordinate system such that the first coordinate accounts for the highest variance, the second coordinate has the second highest variance and so on.

The final step involves the use of a discrimination or regression method to produce the predicted label or values. Support Vector Machine (SVM) [118, 24] is a popular maximum margin classifier that handles the problem of linearly non-separable classes by projecting the data into a higher dimensional space using the kernel trick which entails the implicit projection of inner products onto a higher dimensional space. The radial basis kernel is the most widely used kernel in the machine learning community.

1.3 METHODOLOGY ADOPTED

In our experiments we use the OpenCV implementation of the Viola-Jones face detector for locating the face in images and videos, an alternative to the Viola-Jones face detector is presented in [95] by Ruiz-Hernandez *et al.* who use a combination of multi-scale Gaussian derivatives and adaboost [41] for face detection. Classical approaches such as skin-color detection are still widely in use for face detection [25]. We do not use any form of facial landmark detection and the normalization step only consists of transforming the facial region into an image of fixed size using translation and scaling.

Gaussian derivatives have long been used for a wide variety of applications such as object recognition [132, 101], age estimation [50], image tracking [114, 129]. We explore the use of multiscale Gaussian derivatives [26, 75] for scale invariant image description in our experiments on head pose estimation, smile detection and affect sensing. Multiscale Gaussian derivatives have been shown to efficiently describe image neighborhood appearance and techniques such as the half-octave Gaussian pyramid [27] provide a cost effective way to compute the derivatives.

We test our approach for head pose estimation [55] on the Pointing04 [45] and CMU-PIE [109] databases. The Pointing04 database contains images of 15 subjects in 93 different head poses with images having been collected by asking the subjects to look at markers placed in a room. Using PCA for dimensionality reduction and SVMs for classification we compare our results to the state-of-the-art. The SVM classifier is replaced by a SVM regressor and the approach is tested for continuous poses.

For smile detection we test our approach [54] on the GENKI-4K [127] database containing 4000 images of people collected from the web. Nearly half the images contain a smiling person while the rest contain in a person with any other expression except for smile. As with our experiments on head pose estimation, we use PCA for dimensionality reduction and SVM for classification. The classifier trained on the GENKI data is also used for smile intensity estimation on image sequences from the Cohn-Kanade database [65].

The Cohn-Kanade and FEED [121] databases contain image sequences of subjects displaying Ekman's six basic emotions. The Cohn-Kanade database contains posed data while the FEED database contains spontaneous expressions. Instead of trying to recognize the six basic emotions, we map the six basic emotions to a 2D affect space and relabel the data. Using the exact same architecture as the one used for smile detection, we compare our results with results produced using Gabor Filters [56].

We also formulate a new descriptor using a combination of Gaussian derivatives and Local Binary Patterns (LBP) [86, 87]. Local Binary Pat-

tern features are powerful yet simple to compute descriptors for image texture analysis. They have been used in a rich variety of applications such as face and gender recognition [112, 93, 105], human action recognition [66], lip reading [137] and moving-object detection [48, 49]. In our experiments on smile detection and face recognition we calculate Gaussian derivatives of five orders : I_x, I_y, I_{xx}, I_{yy} and I_{xy} and then calculate Local Binary Patterns over the derivative images.

We re-perform our experiments on smile detection using the new descriptor and also investigate the effect of head-pose in smile detection [57]. Apart from accuracy we use the balanced error rate (BER) metric to compare our results with other methods such as Gabor filters and LBP features calculated over Gabor derivative images.

Using the same combination of Gaussian derivatives and LBP features, we conduct experiments for face recognition. We test our approach on the CMU-PIE and the YaleB [43] extended database. Both these databases contain facial images under a variety of lighting conditions and poses. Employing very simple 1-nearest neighbor method for assigning identity to test images we compare our results with the state-of-the-art which includes descriptors which are explicitly developed for face recognition and handle illumination explicitly.

We combine Local Binary Patterns-Three Orthogonal Planes (LBP-TOP) features [136] with Gaussian derivatives by computing LBP-TOP features over Gaussian derivative image sequences. We use the same five derivatives that we use in our experiments on smile detection and face recognition. The new descriptor is tested on the image sequences from the Cohn-Kanade dataset. The dimensionality of the feature vectors is very high in comparison to the number of observations, therefore a linear SVM is used for classification. The recognition accuracy is compared with the accuracy obtained using standard LBP-TOP features.

We use dense trajectory features developed by [124] for capturing macro level movements in the videos of the Audio Visual Emotion Challenge (AVEC) 2014 database while LBP-TOP features are used to capture the dynamic texture of the facial region [58]. Since the videos are of different durations and for the Depression Sub-challenge a complete video has one Beck-II inventory score [15] for depression, the descriptor information has to be encoded to a feature vector of set size. We perform feature encoding over the visual features extracted using dense trajectories and LBP-TOP features and the precomputed audio LLD features using Fisher Vectors [99]. The depression severity is estimated using a linear Support Vector Machine regressor.

1.4 INSIGHTS

Head pose estimation

We find that even though the Pointing04 database contains discrete head poses, if the SVM classifiers are replaced by SVM regressors, we attain high correlation between the predicted values and the ground truth. These findings indicate that continuous head pose estimation can be performed even when the training data contains purely discrete head poses.

Smile detection

Support Vector Machines produce a probability estimate of the prediction made. The SVM classifier trained on the GENKI-4K smile database is cross-tested on the image sequences of "happiness" in the Cohn-Kanade database [65] and it is found that the probability estimates represent smile intensity.

Gaussian derivatives and Illumination

It has been discovered that images in the GENKI-4K database with extreme lighting conditions are often misclassified. The combination of LBP features and Gaussian derivatives reduces this problem leading to a higher accuracy over the database. The use of Gaussian derivatives with LBP features also provides more invariance to pose as compared to standard LBP features.

Minkowski Distance

In the experiments on face recognition we observe that varying the distance metric in 1-nearest neighbor classification improves the accuracy i.e. instead of using a fixed metric such as L_1 or L_2, if a generalized metric L_p is used with p chosen through cross validation, we can achieve higher accuracy.

LBP-TOP features calculated over Gaussian derivatives

Using a combination of Gaussian derivatives and LBP-TOP features we obtain a slightly better accuracy at expression recognition than the accuracy with standard LBP-TOP features, however the feature vector produced by our method is five times the size of the feature vector obtained

using LBP-TOP features. The trade-off for a marginal increase in accuracy is longer prediction time and higher cost of feature computation.

Early fusion of features in depression estimation

In our experiments on the AVEC 2014 database, we use an early fusion scheme for combining features and it is found that LBP-TOP features contribute most to depression estimation. The predicted values using just LBP-TOP features outperform the baseline. If the challenge database had more observations, late fusion might have been feasible which could have changed the contribution of the three descriptors, LBP-TOP, dense trajectories and audio LLD features, to depression estimation.

1.5 THESIS SUMMARY

Chapter 2

Chapter 2 presents theories on emotion and emotion classification along with a discussion on mood and depression. The first theories on emotion presented in this chapter fall under the category of phenomena based theories. These theories focus on the physiological side to emotions and posit that emotional experience is secondary to physiological changes produced when an organism is presented with stimuli. James-Lange's feedback theory[59, 69] and theories in opposition to his theory are presented. This initial discussion provides a glimpse of the evolution of human understanding of emotions.

The concepts following the discussion on early theories of emotion in chapter 2 are more relevant to the domain of affective computing and have been put to use in later chapters. Paul Ekman, inspired by Darwin's belief that some emotions may be more primitive and basic than others, conducted a series of studies in the 1970's which led him to postulate that six emotions [34, 33] : happiness, sadness, disgust, fear, anger and surprise are not only basic and primary but are also expressed universally across different cultures with the same facial display rules.

A large section of psychologists working on emotions do not agree with the concept of some emotions being more basic or primitive than others. They argue that all emotions are produced by the same neurological system and processes and therefore all emotions are on a level playing field. Russell and Mehrabian presented a circumplex model of emotions [97] where most human emotions can be represented in a 2D space with one dimension representing the positivity and negativity of the emotion while the other dimension stands for how activated or sleepy the person

is. There is not enough consensus on the third dimension but there are emotions which need a third dimension to differentiate themselves from other emotions [98].

Plutchik's model of emotions [92], although failed to find favor with the psychological community, draws a parallel between colors and emotions. His "wheel" of emotions contains eight basic emotions which mix with each other to produce emotional compounds. The cone model proposed by him has a vertical axis that stands for intensity of emotion; the hue axis in the HSV (hue-saturation-value) color space has an analogous in his model in the form of a similarity/dissimilarity axis i.e. emotions similar to each other are adjacent while the emotions which are the most dissimilar are placed diametrically opposite.

Chapter 2 also contains a section on mood. Mood is an affective state like emotion but is more longer lasting and has a less intense emotional experience than emotions. Mood can influence emotional response and Picard [91] has hypothesized a model to link emotions with mood and temperament. A persistent bad mood can be a sign of depression.

Depression is a serious mental illness and depression intensity is often measured using the Beck-II [15] self-report inventory. The Beck-II inventory is the most popular scale adopted by psychologists to quantify depression severity. However self-reports may not be the best way to appraise depression severity.

Chapter 3

Chapter 3 presents our experiments on static image data for sensing affect. We develop an architecture that is applicable for head pose estimation, smile detection and affect sensing using multiscale Gaussian derivatives for image description, principal component analysis for dimensionality reduction and Support Vector Machines for classification or regression. Our results on head pose estimation are better than the state-of-the-art results reported for global appearance based methods.

Experiments on smile detection are performed on the GENKI-4K database containing images collected from the "wild" unlike other smile databases which contain images collected in controlled environments. The SVM classifier trained on Gaussian derivative features calculated on the GENKI database is shown to perform better than the SVM classifier trained on Gabor features. Our smile detector is tested on the image sequences in the Cohn-Kanade database and it is found that the probability estimates produced by the SVM classifier reflect smile intensity. It is observed that our smile detector fails to detect smile correctly when the face is partially illuminated. Leveraging the illumination-invariant property of Local Binary

Patterns, we develop a novel descriptor which is constructed by computing LBP features over Gaussian derivative images.

In case of smile detection with the new descriptor we achieve better results than the ones obtained using LBP features and Gaussian derivative features alone. Additionally we compare our results with Gabor features and LBP features computed over Gabor features. Our method performs better than all the other techniques. It is also noted that our descriptor is the most invariant to head pose.

Cohn-Kanade and FEED [121] databases contain image sequences of subjects experiencing or pretending to experience the six basic Ekman emotions. Mapping these emotions to the 2D affect space, we relabel the databases with affect labels. We employ the same architecture that we developed for head pose estimation and smile detection and compare our results to the results obtained using the technique presented in [28]. Not only do we obtain a higher accuracy but our descriptor takes less time to compute and our feature vector is smaller in size as well.

In the final experiment of chapter 3, the descriptor developed by combining Gaussian derivatives and LBP features is employed for face recognition. Simple 1-nearest neighbor classification with Minkowski distance as the distance metric is used for assigning identity to test images. We outperform the state-of-the-art on the CMU-PIE database and the performance on the YaleB database is at par with the state of the art. It is worthwhile to note that the state-of-the-art includes descriptors developed exclusively for face recognition and handle illumination explicitly while our method, using a very simple classification technique, performs better than other methods which employ much more complex discrimination techniques.

Chapter 4

Chapter 4 contains two set of experiments on affect sensing in videos. The first set of experiments is motivated by our success with the combination of Gaussian derivatives and LBP features. In chapter 4, we propose a novel descriptor created by computing LBP-TOP features over image sequences composed of Gaussian derivative images. It can be considered as an extension of our descriptor from chapter 3 to the temporal domain for capturing dynamic texture. The descriptor is tested on the image sequences in the Cohn-Kanade database. We obtain marginally better results than the ones obtained using the conventional LBP-TOP descriptor but because we use image sequences of Gaussian derivative images of five different orders instead of the original image sequence, our concatenated histogram's size is five times the size of the histogram generated using standard LBP-TOP features.

The second set of experiments in chapter 4 involve the development of a new technique for depression intensity estimation. The AVEC 2014 database [116] contains videos of different durations of subjects experiencing symptoms of depression and the ground truth for the Depression Sub-challenge (DSC) consists of Beck-II inventory scores which represent the severity of depression. We capture video features using two descriptors: LBP-TOP features to capture micro-movements such as intra-facial movements and the dynamic texture of the facial region and dense trajectories for capturing macro movements such as head movements. These video features along with the precomputed audio Low level descriptor (LLD) features are encoded using fisher vectors. The encoding is necessary since videos can be of different durations and whole videos are associated with a unique score. The encoded features are passed to a Support Vector Regressor to obtain the predicted depression intensity. This technique developed for estimating depression intensity should also be applicable for other slowly varying affective states.

Chapter 5

Chapter 5 is the final chapter of the thesis. The chapter contains a brief summary of chapters 2-4 followed by a section highlighting the lessons learned from our experiments. A section is dedicated to the presentation of the probable contributions of the present work to the domain of affective computing. The chapter ends with some ideas on how the current work could be extended.

2

HUMAN EMOTIONS

This chapter summarizes different conceptions associated with emotions. A large part of the discussion is dedicated to the different theories and concepts of emotion representation.

Early psychologists studying emotions concentrated on understanding the processes behind emotion elicitation. Focusing on the physiological aspect of emotions, these psychologists believed that the emotional experience occurs after the physiological reaction to external stimulus. Such theories of emotion fall under the category of phenomena based theories [110]. A brief discussion on these theories is provided in section 2.1.

Models of emotion that are more relevant to affective computing and human-computer interaction described in section 2.2 can be broadly divided in two categories: one conception suggests that some emotions are more fundamental as compared to others and this primary set of emotions is common throughout cultures and races. The other category of theories posit that emotions can be represented in a continuous dimensional space. A major point of disagreement between the two conceptions is that proponents of basic (primary) emotions believe that different emotions arise from separate neural systems while the advocates of dimensional theories are of the opinion that a common and interconnected neurophysiological system is responsible for all affective states. The most popular taxonomy for describing human facial movements according to their appearance is discussed in section 2.3.

In sections 2.4 and 2.5 we describe mood as an emotional phenomenon distinct from emotions and plays the role of a background process in emotion elicitation. Apart from the description of depression section 2.5 also provides a brief description of the BDI-II inventory which is the most commonly used scale for quantifying depression in humans.

2.1 FEEDBACK THEORIES ON EMOTION

The first feedback theory on emotion was proposed by the American philosopher and psychologist William James [59] and Carl Lange [69] a Danish physician, almost at the same time in the 19th century. The James-Lange theory posits that emotion is secondary to physiological responses which in turn are caused by a stimulus. In other words the theory postulates that bodily changes are the primary feelings and are not the result but a necessary precursor to felt emotions. The theory was challenged by

Walter Cannon and he proposed an alternative theory known to us as the Cannon-Bard theory [18].

The Cannon-Bard theory is based on Cannon's understanding of thalamic processes. Cannon and Bard conducted studies on animal physiology. Through these experiments they theorized the role of the brain in the synthesis of physiological responses and emotions. According to their theory, a stimulus causes impulses to travel to the cortex which in turn activates the thalamic processes. Once the thalamic processes are activated they are ready to discharge. When this discharge occurs the physiological response and the emotional experience occur simultaneously but independently. This point in Cannon-Bard's theory is in direct opposition with the James-Lange theory which states that physiological responses occur before the emotional experience.

A new form of feedback theory titled "Facial Feedback Hypotheses" was proposed by McIntosh in 1996 [80] presenting a new relation between the face and emotions. McIntosh raises four questions in his discussion on the hypotheses:

1. Do Facial Actions Correspond to Emotions?
2. Does Facial Movement Modulate Emotions?
3. Can Facial Action Initiate Emotions?
4. Are Facial Actions Necessary for Emotions?

McIntosh's theory is at odds with the James-Lang theory. He hypothesized that facial expressions and not visceral changes could be a possible factor in emotion elicitation.

At the moment the scientific community does not agree on whether bodily feedback is sufficient or even necessary in the elicitation processes of emotions. This lack of consensus on feedback theories and their insufficient explanation of emotion elicitation prompts us to investigate alternate theories to conceptualize emotions.

2.2 BASIC EMOTIONS VS. DIMENSIONAL THEORIES OF AFFECT

There are two viewpoints on the classification of emotions. The first is that emotions are discrete and fundamentally different constructs while the second viewpoint asserts that emotions can be represented using continuous dimensional models. Lang [17] has shown that self-reports of emotion across subjects are more reliable with respect to dimensions than with respect to discrete emotion categories. There are also ways to map basic emotions to multi-dimensional affect spaces as we will discuss later.

Among the proponents of the discrete emotion theory, there are two distinguishable beliefs underlying the assumption that emotions can be divided into categories of primary (basic) and secondary emotions [89]. From the psychological perspective basic emotions are held to be the basic building blocks of secondary emotions while from the biological viewpoint there is a belief that there could be neurophysiological and anatomical substrates corresponding to the basic emotions.

Picard [91] believes that the representation of emotions as discrete categories or as continuous dimensions is a matter of choice much like light which can be described using wave and particle theories, the choice of theory depends on what is being tried to explained. Similarly colors can be described in terms of their component RGB values or by their names, the choice depends on the application.

2.2.1 *Basic Emotions*

Charles Darwin in his book "The Expressions of the Emotions in Man and Animals" first postulated the theory that emotions are universal. Darwin believed that emotion had an evolutionary history that could be traced across cultures and species. A number of researchers since then, prominently Ekman and Friesen [34], through their studies found that certain emotions are expressed via the same facial actions across cultures. Ekman's research challenged the popular belief that emotions expressed by the face are cultural specific and governed by behavioral-learning processes.

Although there are several definitions for basic emotions, the most widely used definition is the one by Ekman [33] who supports the view that not only are the six emotions: joy, sadness, anger, disgust, fear and surprise universal but also basic. Basic or fundamental emotions are defined as biologically primitive, measurable, physiologically distinct and "irreducible constituents of other emotions" [89] as opposed to secondary emotions which can be reduced to a sum of basic emotions. In [60] the authors suggest that a large number of words in the English vocabulary describing emotions can be based on one or more of the five basic emotions: fear, anger, sadness, happiness and disgust.

From the affective computing perspective, the debate on the existence of basic and secondary emotions is not as important as the question whether we are better of representing emotions as discrete categories or otherwise?

2.2.2 *Dimensional Theories of Emotions*

Several theories exist for defining emotions in a continuous space. Most theorists agree on three dimensions out of which two dimensions are agreed to be "arousal" (excited vs. sleepy) and "valence" (positive vs. negative). Wundt [94] believed that the third dimension should represent strain vs. relaxation while Schlosberg [102] believes that it should be attention vs. rejection.

The most commonly used continuous model for emotions used in the affective computing community is the circumplex model [97] introduced by Russell. This model represents emotions in a two-dimensional circular space with two dimensions: arousal and valence. The vertical axis is arousal while the horizontal axis is valence. The center of the circle represents a neutral valence and a medium level of arousal.

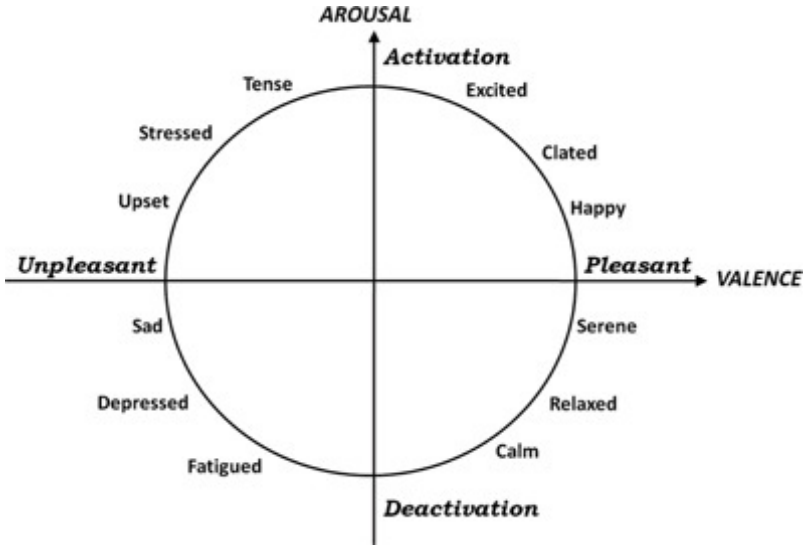


Figure 2: Circumplex model of emotions [115]

The problem with the circumplex model arises when one is asked to examine the difference between "anger" and "fear" in the 2D circumplex space since both the emotion words would be represented by high *unpleasantness* and high levels of arousal. A third dimension is required to discriminate between these two emotions which the authors in [98] call dominance. With the use of this third dimension: dominance, anger and fear become distinguishable because anger ranks high in dominance whereas fear is submissive placing it at the other end of the dominant-submissive spectrum.

In the same paper [98] the authors present a table of emotion words along with their corresponding co-ordinates in the 3D PAD (Pleasure, Arousal and Dominance) space. In this table of emotion words, one can find the six basic emotions and their corresponding values in the 3D space.

However using these co-ordinates is not very easy since they are only mean values of a distribution pattern and have a high standard deviation associated with them essentially making each emotional word a multivariate Gaussian distribution in the PAD space. Also, the authors of [98] have noted that the dimension of dominance has a high correlation with the other two dimensions, raising the question: "Could we find a third dimension that is more *orthogonal* to the other dimensions?".

In this thesis we only used the polarity of each of the six basic emotions in the circumplex space to re-label the data, which was labeled for basic emotions, with pleasure-arousal labels. For example, the emotional word "Happy" has a mean pleasure value of 0.81 and a mean arousal value of 0.51, we re-label the data, which previously had the happy, with the new labels $\mathrm{P} = +$, $\mathrm{A} = +$ and similarly for sadness with the new labels $\mathrm{P} = -$, $\mathrm{A} = -$.

Term	**Pleasure**		**Arousal**		**Dominance**	
	Mean	SD	Mean	SD	Mean	SD
Happiness	0.81	0.21	0.51	0.26	0.46	0.38
Sadness	-0.63	0.23	-0.27	0.34	-0.33	0.22
Surprise	0.40	0.30	0.67	0.27	-0.13	0.38
Fear	-0.64	0.20	0.60	0.32	-0.43	0.30
Anger	-0.51	0.20	0.59	0.33	0.25	0.39
Disgust	-0.60	0.20	0.35	0.41	0.11	0.34

Table 1: Basic Emotions in the PAD space

2.2.3 *Plutchik's theory*

Plutchik [92] devised a wheel of emotions with eight basic bipolar emotions in the form of dimensions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Additionally, he compared his concept of emotions with the three dimensional Hue-Saturation-Value (HSV) color space. Plutchik's HSV-space like cone shaped model has the vertical axis representing intensity while the similarity/dissimilarity between the emotions is represented by the angle. His model does not have a well defined analogue to the saturation dimension in the color space.

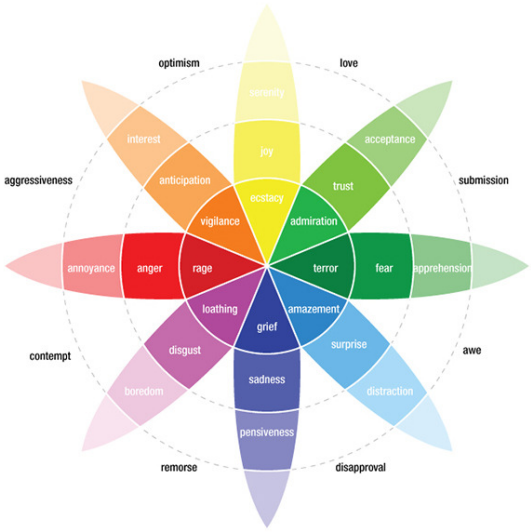


Figure 3: Plutchik's wheel of emotions [3]

Plutchik postulated that just like colors, emotions can have different intensities and emotions can mix with one another to form new emotions. In the exploded model (figure 4), Plutchik placed the eight basic emotions on the basis of bipolarity and similarity. Two adjacent basic emotions can combine to produce a "dyad" which is analogous to the mixing of primary colors.

Figure 4: Plutchik's cone of emotions [4]

Plutchik's model has not found much favor with the affective computing community. He introduced the idea of basic dimensions, however these dimensions do not correlate with any of the fundamental emotions. Most dimensional theories of affect though, have bipolar dimensions just like the ones proposed by Plutchik.

2.3 FACIAL EXPRESSIONS AND THE FACIAL ACTION CODING SYSTEM

Picard [91] defines sentic modulation to be the physical means by which an emotional state is typically expressed. Facial expressions are one of the most easily controllable of the sentic modulations. Since the face is one of the most visible and least occluded parts of the body, we allot a lot of importance to facial expressions when we communicate. For a machine to be able to communicate like humans, machines would have to be endowed with the ability to recognize facial expressions even though facial expressions may not always represent the underlying emotion that generated it.

It has been shown that automated facial expression recognition is more accurate and robust on videos as compared to still images, this could be because images cannot capture the dynamics of facial motion as compared to videos. Psychological studies have also suggested that facial motion is essential to facial expression recognition since facial expressions are dynamic processes. Bassili in [13] has shown that humans are better at recognizing facial expressions from videos in comparison to images.

Most researchers in the field of automated facial expression analysis have treated emotions as discrete and subscribe to Ekman's theory of six basic emotions and their link to facial expressions. Ekman and his colleagues developed the Facial Action Coding System (FACS) [34] as a taxonomy of human facial movements by their appearance on the face. FACS is a convenient system to deconstruct almost any anatomically possible facial expression into component muscle movements which are referred to as Action Units (AU's). An AU can be defined as the relaxation or contraction of one or more muscles. The FACS manual [35] enumerates 46 Action Units, their location on the face, their appearance corresponding to their intensity and Ekman's interpretation on their possible meaning. Apart from the AU's the manual also lists Action Descriptors which are, like AU's based on muscle movements, but their muscular basis has not been as precisely defined as for the AU's. The relation between the six basic emotions and the AU's is presented in the table 2 as proposed by Ekman.

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

Table 2: AU's associated with basic emotions

A wide variety of image and video databases available today are coded for the presence of AU's and the six primary emotions. Most of these databases are collected in controlled lab conditions. In most of these databases the facial expressions have been enacted by people without feeling the underlying emotion linked to those expressions. Some recent databases such as the SEMAINE database [82] contain spontaneous expressions and it is hard to put these expressions in any of the discrete categories. Not only are these expressions representative of "non-pure" emotions but it is often difficult to assign an emotion word to them strengthening the case for dimensional theories of affect being used for representing emotions associated with facial expressions. In this thesis we do not claim to recognize affect rather we attempt to infer affect by observing sentic modulations in the form of facial expressions.

2.4 MOOD

Mood is an emotional state which often lasts longer than emotions and has a more diffuse emotional experience. Picard in [91] states that mood can be considered as a process that is always running in the background. The concept of mood differs from emotions in the sense that it is lower in intensity and changes more slowly with time. Mood differs from other psychological aspects such as temperament and disposition in that mood is more transient in nature just as emotions are more transient in nature as compared to mood.

Certain dispositions and temperaments do however influence mood. People with a cheerful disposition are more likely to have a good mood at a given instant as compared to someone who, say, for example, has an anxious disposition. Similarly mood can bias people towards certain emotions. A bad mood could make it easier for negative-valenced emotions to be activated and vice-versa. The circumplex model of emotions with the two dimensions of pleasure and arousal can be used for an effective description of mood. We often think of mood as being positive, negative or

neutral but it can have more nuances for example a good peaceful mood would be low in arousal as compared to a good mood associated with something exciting. A bad mood due to anger would have a high level of arousal; a bad mood due to immense sadness marked by depression would have low arousal.

Picard presents an abstract and simplified function for emotion response and the role mood plays in it:

$$y = \frac{g}{1 + e^{-(x - x_0)/s}} + y_0 \tag{1}$$

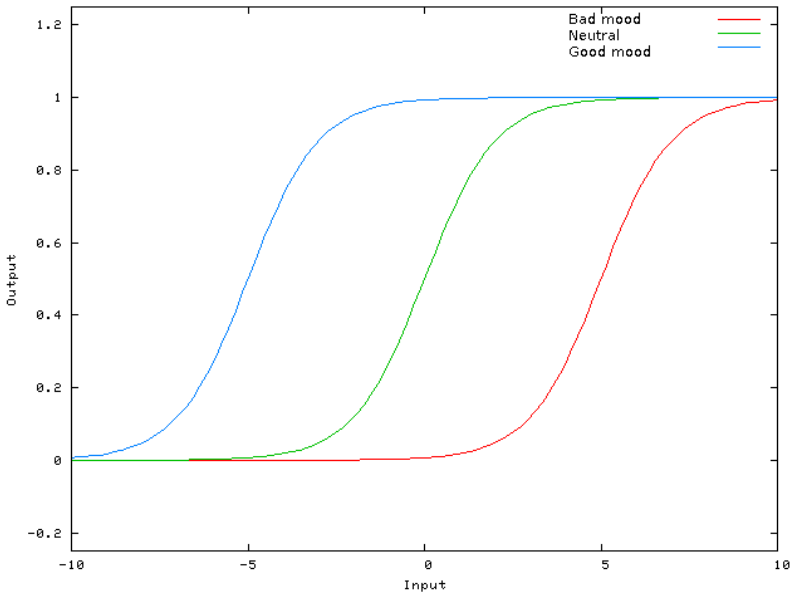


Figure 5: Shift in emotional response due to mood

In this "sigmoidal" like function, x is the input, which may represent multiple possible stimuli, internal or external to the human being. The output y is the emotional response. The parameter s controls the gradient of the curve which is directly tied to the personality of the individual. The parameter g refers to the gain of the sigmoid, this parameter controls the increase or decrease in the amplitude of the sigmoid. This parameter could be related to the activation levels of a person, a person with a higher level of activation might be likely to experience a greater magnitude of emotion. The parameter x_0 is the offset i.e. it shifts the response to the left or right. This parameter is used to represent mood. If the function is used to represent the emotional response to a positive stimuli, a good mood will allow small inputs to trigger a response effectively shifting the sigmoid to the left. Finally the parameter y_0 shifts the sigmoidal curve up or down. This parameter could be used to model the cognitive expectation.

This representation for emotional response is able to take into account influences such as mood, temperament and cognitive expectation and the phenomenon of activation and saturation of emotional responses. It shows how mood acts as a threshold for activation of an emotional response. For example bad mood with a high degree of arousal could reduce the value of x_0 such that even a slightly negative event could activate a response. Similarly a good mood could lead to a higher value of x_0 which wouldn't allow trivial negative events to create any response.

2.5 DEPRESSION

Depression is a serious mental disorder involving persistent bad mood, low self-satisfaction and lack of interest in normal pleasurable activities. Currently depression is diagnosed by a patient's self report or through a mental status examination (MSE). A MSE entails the observation of a patient's state of mind by a psychologist to assess aspects such as attitude, mood, affect and speech. An automated system to detect depression can help both the doctors and patients with diagnosis and treatment monitoring. Such a system will also help to overcome the problem of subjective bias associated with self-reports and MSE.

The Beck Depression Inventory (BDI, BDI-1A, BDI-II), designed by Aaron T. Beck [15], is one of the most widely used self-report questionnaires for measuring the severity of depression. It is a 21-item self-administered rating inventory that measures characteristic attitudes and symptoms of depression. Each item is a list of four statements arranged in increasing severity about a particular symptom of depression. Items are rated on a 4 point scale and the final score is obtained by adding up the ratings for all 21 items. Higher the score, higher the severity of depression.

The current version of the questionnaire, BDI-II, is designed for individuals aged 13 and over, and contains all major content dimensions of depression including sadness, suicidal ideation, experiences of crying, concentration difficulties, changes in appetite or weight, anhedonia, pessimism, self-dislike, self-criticalness, feelings of punishment, belief of being a failure, agitation, feeling of guilt, indecisiveness, lack of energy, change in sleep patters, irritability, feeling of worthlessness, fatigue and change in sex drive.

The BDI-II scales and its predecessors have been translated into a large number of languages and have been extensively used in research and practice. BDI test is perhaps the most used psychological test to date.

In chapter 4 we present a multi-modal system for depression estimation using audiovisual features. The approach is tested on the Audio Visual

Emotion Challenge Dataset which contains videos of people with varying levels of depression.The ground truth is in the form of BDI-II scores.

2.6 CONCLUSION

Feedback theories described in this chapter posit that emotions are a product of physiological response to external stimuli. These theories are not further followed in this thesis but they do provide a glimpse into how our understanding of emotions has evolved over time.

Plutchik's theory in section 2.2.3 describes how basic emotions are analogous to basic colors and just like colors, basic emotions combine to form more complex emotions. Although the theory has not found much favour among the scientific community, most dimensional theories of affect use bipolar dimensions of affect just as in Plutchik's theory of emotion.

In section 3.3.3 of chapter 3, we describe our experiments on inferring affect and its representation in a two-dimensional space using visual information from images. Ekman's theory of basic emotions and their facial representations using Action Units is widely used by the affective computing community and we describe our experiments for recognizing these basic emotions using visual information from videos in section 4.3.1 of chapter 4. Section 4.3.2 of chapter 4 contains description of our experiments on depression assessment using audiovisual information from videos.

Questions on automated facial expression recognition faced by the affective computing community:

1. Some people are more expressive than others, would it be possible for future automated facial expression analysis systems to adapt to individuals according to their temperament?

2. Could a system be developed that would be able to recognize the underlying emotion when the facial expressions belie them and would such a system be able to tell the genuine expressions apart from the forced ones.

For the first question, current machine learning techniques allow for a high degree of generalization. It is possible that this property could allow affect recognition systems to adapt to new subjects.

If facial expressions are the only input to an affect recognition system, correct recognition would not be possible if the facial expressions do not correctly represent the underlying emotions. Even a system with multiple sentic modulations as inputs might not be able to recognize the affective state correctly because a computer can only have limited access to the human mind and body. It could perform better than people especially if the computer has access to biosignals such as heart rate and electromyo-

graphic signals but it would never have access to all the information that the human who experiences those emotions has.

3

STATIC IMAGE ANALYSIS FOR SENSING AFFECT

This chapter presents the techniques developed for head pose estimation, smile detection, affect sensing and face recognition on static images.

Firstly we describe the image description methods used in section 3.1. The machine learning and statistical techniques used are described in section 3.2, while section 3.3 is dedicated to experiments using the methods and techniques described in the preceding two sections. We end this chapter with the summary and insights in section 3.4.

3.1 IMAGE DESCRIPTION METHODS FOR STILL IMAGES

We have used two major categories of descriptors for our experiments on static images: steerable filters and local binary patterns. There are advantages associated with each of the description techniques.

3.1.1 *Steerable filter methods*

The term steerable filters is used to describe a class of filters in which a filter of arbitrary orientation is synthesized as a linear combination of a set of "basis filters" [40]. The process by which the oriented filter is synthesized at any given angle is known as steering.

For our experiments we used two types of steerable filters: Multi-scale Gaussian Derivatives and Gabor Filters. We have compared the performance of these two descriptors for smile detection and head pose estimation. Apart from accuracy we have looked at the calculation and prediction times for these descriptors when a Support Vector Machine is used for classification or regression.

3.1.1.1 *Multi-scale Gaussian Derivatives*

Gaussian derivatives can efficiently describe the neighborhood appearance of an image for recognition and matching[67]. This can be done by calculating several orders of Gaussian derivatives normalized in scale and orientation at every pixel. The basic Gaussian function is defined as:

$$G(x, y; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

Here σ is the variance or scale and defines the spatial support. The basic Gaussian function in equation 2 measures the intensity of the neighbor-

hood and does not contribute to the identification of the neighborhood and can be omitted. The first order derivatives are of the form:

$$G_x(x,y;\sigma) = \frac{\partial G(x,y;\sigma)}{\partial x} = -\frac{x}{\sigma^2}G(x,y;\sigma) \tag{3}$$

$$G_y(x,y;\sigma) = \frac{\partial G(x,y;\sigma)}{\partial y} = -\frac{y}{\sigma^2}G(x,y;\sigma) \tag{4}$$

First order derivatives give information about the gradient (intensity and direction). The second order derivatives are given by:

$$G_{xx}(x,y;\sigma) = \frac{\partial^2 G(x,y;\sigma)}{\partial x^2} = (\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2})G(x,y;\sigma) \tag{5}$$

$$G_{yy}(x,y;\sigma) = \frac{\partial^2 G(x,y;\sigma)}{\partial y^2} = (\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2})G(x,y;\sigma) \tag{6}$$

$$G_{xy}(x,y;\sigma) = \frac{\partial^2 G(x,y;\sigma)}{\partial x \partial y} = \frac{xy}{\sigma^4}G(x,y;\sigma) \tag{7}$$

Second order derivatives provide us with information about image features such as bars, blobs and corners. Higher order derivatives are only useful if the second order derivatives are strong and they are also sensitive to high spatial-frequency noise such as sampling noise.

The inverse-tangent of the ratio of first order derivatives at any image point can be used to determine the direction of maximum gradient.

$$\theta_{max} = \arctan(\frac{G_y(x,y,\sigma)}{G_x(x,y,\sigma)}) \tag{8}$$

It has been shown that Gaussian derivatives are steerable [40] i.e. the filter response can be calculated at any arbitrary orientation and by using appropriate trigonometric ratios the Gaussian derivatives can be rotated in the desired direction.

$$G_\theta = G_x cos(\theta) + G_y sin(\theta) \tag{9}$$

Normalizing Gaussian derivatives in scale is not a trivial task. Several methods have come up in the past addressing this problem. It was shown by Crowley in [26] that Gaussian derivatives be calculated across scales to obtain scale invariant features and then Lowe in [75] adopted Crowley's method to define the intrinsic scale at a point (x,y) as the value of the

scale parameter at which the Laplacian provides a local maximum. The computational cost of directly searching the scale axis for this characteristic scale can be relatively expensive. A cost-effective method for computing Multi-scale Gaussian derivatives is the Half Octave Gaussian Pyramid described in detail in [27] and an integer coefficient version of the same can constructed using repeated convolutions of the binomial kernel (1, 2, 1). The algorithm involves repeated convolutions with a Gaussian kernel in a cascaded configuration. An essential feature of this pyramid is it's invariance to impulse response provided by keeping the sampling rate equal to the Gaussian support at each level of the pyramid.

A key feature of this algorithm is that for different levels of the pyramid the difference of adjacent image pixels in the row and column directions are equivalent to convolution with Gaussian derivatives.

The pyramid is very easy to access, derivative values can be determined for every image position by using bilinear interpolation and derivatives between scale values can be computed using quadratic interpolation between adjacent levels of the pyramid.

The following sets of equations explain how different order of derivatives can be calculated using difference of adjacent image pixels in the row and column directions:

$$\frac{\partial p(x,y,k)}{\partial x} = p * G_x(x,y;2^k\sigma_0) \approx p(x+1,y,k) - p(x-1,y,k) \quad (10)$$

$$\frac{\partial p(x,y,k)}{\partial y} = p * G_y(x,y;2^k\sigma_0) \approx p(x,y+1,k) - p(x,y-1,k) \quad (11)$$

$$\begin{aligned}\frac{\partial^2 p(x,y,k)}{\partial x^2} &= p * G_{xx}(x,y;2^k\sigma_0) \\ &\approx p(x+1,y,k) - 2p(x,y,k) + p(x-1,y,k)\end{aligned} \quad (12)$$

$$\begin{aligned}\frac{\partial^2 p(x,y,k)}{\partial y^2} &= p * G_{yy}(x,y;2^k\sigma_0) \\ &\approx p(x,y+1,k) - 2p(x,y,k) + p(x,y-1,k)\end{aligned} \quad (13)$$

$$\begin{aligned}\frac{\partial^2 p(x,y,k)}{\partial x \partial y} &= p * G_{xy}(x,y;2^k\sigma_0) \\ &\approx p(x+1,y+1,k) - p(x+1,y-1,k) \\ &\quad - p(x-1,y+1,k) + p(x-1,y-1,k)\end{aligned} \quad (14)$$

In the above equations at the k_{th} level of the pyramid the support is defined by $\sigma_k = 2^k\sigma_0$ and the image at the same level is defined by $p(x,y,k)$.

3.1.1.2 *Gabor Filters*

Gabor filters are a category of bandpass filters which can be used for feature extraction and image description. The $Gabor(x, y)$ function can be interpreted as a Gaussian filter modulated by an oriented complex sinusoid:

$$Gabor(x, y) = s(x, y)G(x, y) \tag{15}$$

where $s(x, y)$ is a complex sinusoid, known as the carrier, and $G(x, y)$ is a 2-D Gaussian function [83], known as the Gaussian envelope. The complex sinusoid shifts the Gaussian envelope in the frequency domain resulting in a band-pass filter. Such filters are used for image processing to enhance image structures at specific orientations and scales.

The $Gabor(x, y)$ function is characterized by three parameters: a) the σ and b) the orientation θ of the Gaussian kernel $G(x, y)$ and c) the frequency f of the complex sinusoid $s(x, y)$. The σ sets the resolution of the filter, the θ the orientation of the filter and f the length of the edges that will be captured. The size of σ, so consequently the size of $G(x, y)$, depend on the size of f.

The response of the real and imaginary components can be combined with the L_2-norm to give a Gabor Energy filters (GEF). Such filters respond to an edge or a corner with the local maxima centerd exactly at the edge or the corner. While Gabor wavelets and Gabor Energy functions are now widely known to be effective for facial image analysis, their implementation cost has proven to be a barrier for real time applications on mobile computing platforms. Also, the belief that Gabor filters are a convenient way to compute localised spacial frequency filters has gradually transformed into an widely repeated claim that the Gabor function is the exact function computed in the mammalian visual cortex [31], leading the community to ignore other less expensive mechanisms for computing localised band-pass spatial frequency filters.

3.1.2 *Local Binary Patterns*

Local binary patterns (LBP) operator is a simple yet powerful descriptor for texture analysis [86]. It requires limited computational power and is ideal for real-time applications. Its robustness to monotonic gray-scale changes make it suitable for applications such as facial image analysis where variations in illumination can have major effects on appearance.

Many variations over the original LBP operator have been proposed. One extension allows LBP operator to use neighborhoods of different sizes [87]. Another modification introduced in [87] is the Uniform LBP

which which can be used to reduce the length of the feature vector by using a smaller number of bins and is also invariant to rotation.

A general algorithm for computing LBP features over a facial image is as follows:

Step 1: Divide the image into blocks of pixels.

Step 2: For each pixel in a block, compare the pixel value to each of its neighbors, the neighborhood size is controllable. The comparison is performed either clockwise or counter-clockwise.

Step 3: If the pixel value of the neighbor is greater than the value of the center pixel a "1" is generated otherwise a "0" is generated. For a 3 X 3 neighborhood, an eight bit pattern will be generated, this pattern can be converted to a decimal value for convenience.

Step 4: A histogram is calculated for the block which contains the frequency of occurrence of the patterns generated for each pixel. This histogram may or may not be normalized.

Step 5: Histograms for all the blocks in the image are concatenated to obtain the final descriptor for the image.

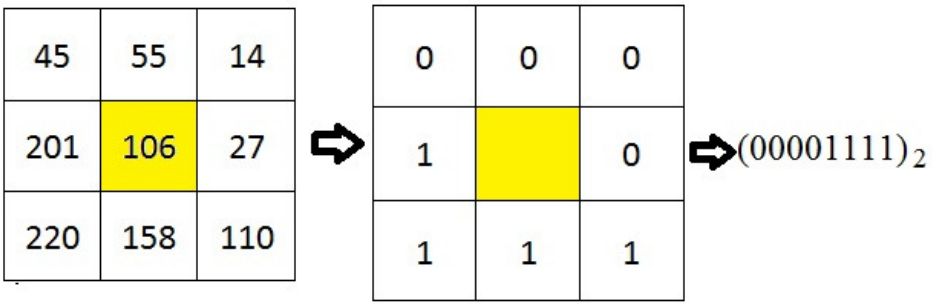


Figure 6: Computing LBP response from a pixel's local neighbourhood.

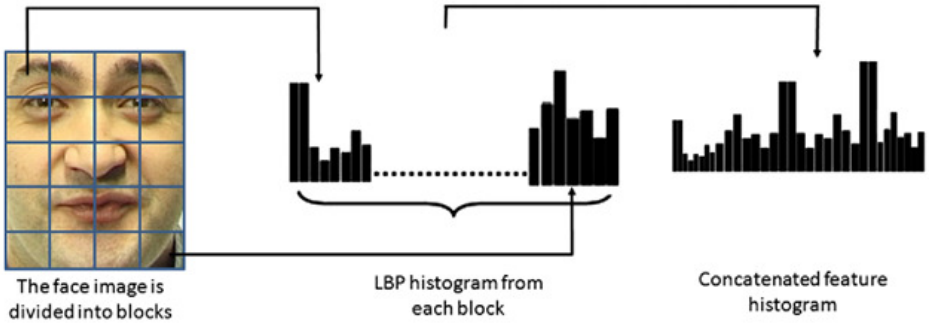


Figure 7: LBP Histogram computation [2]

3.2 MACHINE LEARNING METHODS

This section provides a brief description of the machine learning methods employed in our experiments. In our experiments Principal Component Analysis was used for dimensionality reduction which is discussed next. For classification and regression Support Vector Machines were widely used, K-Nearst Neighbor classification is used for our experiments on face recognition.

3.2.1 *Dimensionality reduction and Principal Component Analysis*

Principal component analysis (PCA) is often used to omit correlated dimensions by transforming the original dimensions into new dimensions which are a linear sum of the original dimensions but are linearly uncorrelated. Then these new dimensions are ranked according to the variance i.e. the dimension which accounts for the most variability in the data gets the first rank and so on [61].

PCA is performed by eigenvalue decomposition of the data correlation matrix after normalizing the data for each dimension. PCA provides scores and loadings. The scores are the transformed values corresponding to each data point and loadings are the coefficients the original variable should be multiplied with to get the score.

Supposing that the data is in the form of a $m x n$ matrix X where m is the number of observations and n is their dimensionality. The mean is calculated along each of the n dimensions(i=1...n)) producing a row vector of means u.

$$u[i] = \frac{1}{m}\sum_{j=1}^{m} X[j,i] \tag{16}$$

The means are subtracted from every row of the observation matrix X to produce a matrix X_o. The covariance matrix C is computed using the equation:

$$C = \frac{1}{m-1} X_o^T * X_o \tag{17}$$

Finding the eigenvectors of the covariance matrix C, they are arranged in ascending order of corresponding eigen values. The first K eigenvectors form the set of basis vectors over which the original observations are projected. The value K is usually chosen such that the variance in the projected data is not below a certain fraction of the variance of the original data.

3.2.2 *Kernel Methods and the Support Vector Machine*

Support Vector Machines (SVM) belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron [118]. The basic idea behind the SVM algorithm is the search for an optimal hyperplane for separating points belonging to different classes, the optimal hyperplane is one which has the maximum distance from the data points. If the points are not linearly separable then the points are projected to a new space using a kernel function. An important feature of kernel functions and kernel space is that the inner product between two vectors in a higher dimensional space can be computed without explicitly projecting the the vectors in the higher dimensional space.

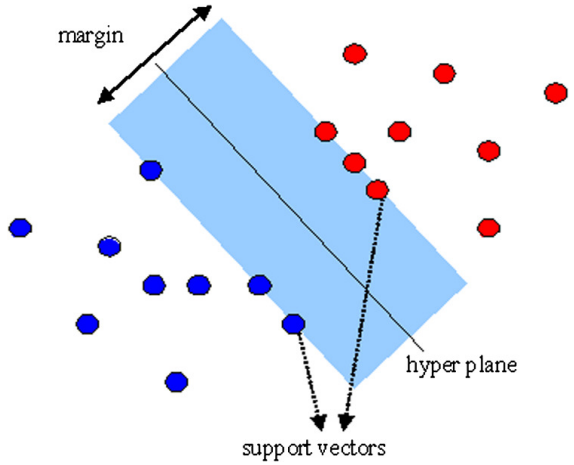


Figure 8: SVM classifying linearly separable data [5]

The most popular kernel used with SVMs is the radial basis kernel, represented by the following equation:

$$K(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} \tag{18}$$

The SVM employed in our experiments was a soft margin SVM, soft margin SVMs are used when the classes are not separable even after transforming the data to a higher dimension. The condition for the optimal hyper-plane can be relaxed by including an extra term ξ, [24]:

$$y_i(X_i^T W + b) \geqslant 1 - \xi_i, (i = 1, ..., m) \tag{19}$$

For minimum error, ξ_i should be minimized as well as $||W||$, and the objective function becomes:

$$\begin{aligned} &\text{minimize} \quad W^T W + C \sum_{i=1}^{m} \xi_i^k \\ &\text{subject to } y_i(X_i^T W + b) \geqslant 1 - \xi_i, \text{ and } \xi_i \geqslant 0; (i = 1, ..., m) \end{aligned} \tag{20}$$

Here C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. $1/\gamma$ or σ is the width of the radial basis kernel. The optimum C and γ are chosen through exhaustive grid search over reasonable limits.

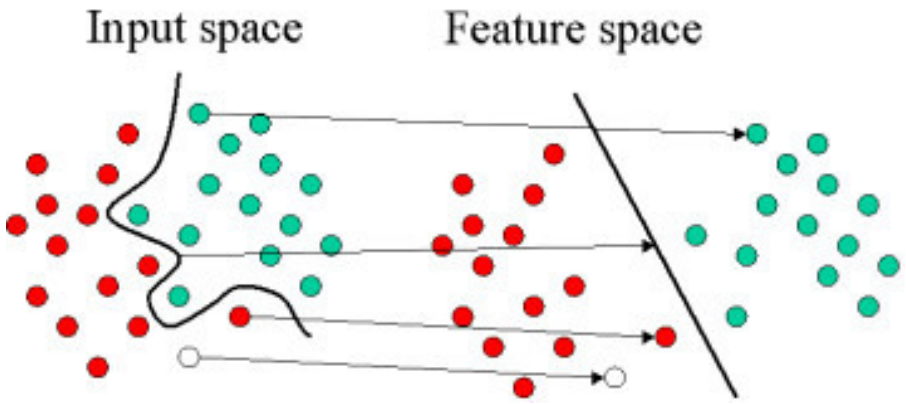


Figure 9: SVM with data mapped to a higher dimensional space using kernel [6]

3.2.3 *K-Nearest Neighbors method*

K-Nearest Neighbor method (k-NN) is a non-parametric technique for clustering that can be used for classification and regression [16]. In the case of classification, the output is a class label. An instance is classified using a majority vote over it's neighborhood, the instance being assigned the label of the majority class in the k nearest neighbors (k is a positive user defined integer, typically small). If $k = 1$, then the instance is simply assigned the label of the nearest neighbor.

Most often L_1 and L_2 metrics are used for distance measures but in section 3.3.4.2 we have shown that the generalized L_p metric also known as the Minkowski distance, where p may or may not be a integer, helps us achieve better face recognition accuracy than the accuracy obtained with L_1 or L_2 metrics.

3.3 EXPERIMENTS

3.3.1 *Head Pose Estimation*

We explore the use of Multi-scale Gaussian derivatives combined with Principal Component Analysis and Support Vector Machines for Head Pose Estimation. The approach is evaluated on the Pointing04 [45] and CMU-PIE [109] data sets.

3.3.1.1 *Head Pose Estimation and the Pointing04 data set*

The problem of head pose estimation involves inferring the orientation of the head from static images or video. It is assumed that the human head has three degrees of freedom, however we estimate only two degrees of freedom namely pan and tilt and the problem is treated as a multi-class classification problem.

The problem of head pose estimation has been approached by the computer vision community in broadly two ways: keypoint-tracking based approaches and appearance based approaches. In keypoint-based approaches facial fiducial points such as eyes, eyebrows, nose, lips etc. have to be located and tracked and then the pose is estimated according to the relative position of these key points [42, 51, 126]. In holistic or appearance based approaches an image descriptor is used to represent the image and a feature vector is assembled using the descriptor values. Then a suitable machine learning technique is used for discrimination between different poses [85, 111].

Stiefelhagen in [111] used horizontal and vertical image derivatives of the first order and used neural networks for discrimination between different poses and applied this approach on the Pointing04 data set. A comprehensive survey on head pose estimation methods [84] shows that Stiefelhagen achieved the best results so far on the Pointing04 data set.

The approach that we present in the following sections was also tested on the Pointing04 data set [52]. This data was collected by Gourier *et al.* [45] at INRIA Grenoble Research Center where 15 people were asked to gaze successively at 93 markers that cover a half-sphere in front of the person. The head pose database consists of 15 sets of images. Each set contains of 2 series of 93 images of the same person at different poses. There are 15 people in the database, wearing glasses or not and having various skin colors. The pose, or head orientation is determined by 2 angles (pan,tilt), which vary from -90 degrees to +90 degrees.

To solve the problem of head pose estimation by an appearance based method, one needs to select an appropriate descriptor to extract features from the image and then a pattern recognition algorithm is required to discriminate between the different poses. We employed Multi-scale Gaus-

sian Derivatives (MGD) and Support Vector Machines (SVM) for head pose estimation on the Pointing04 dataset [45] and show that our choice of descriptor gives better results than those obtained so far.

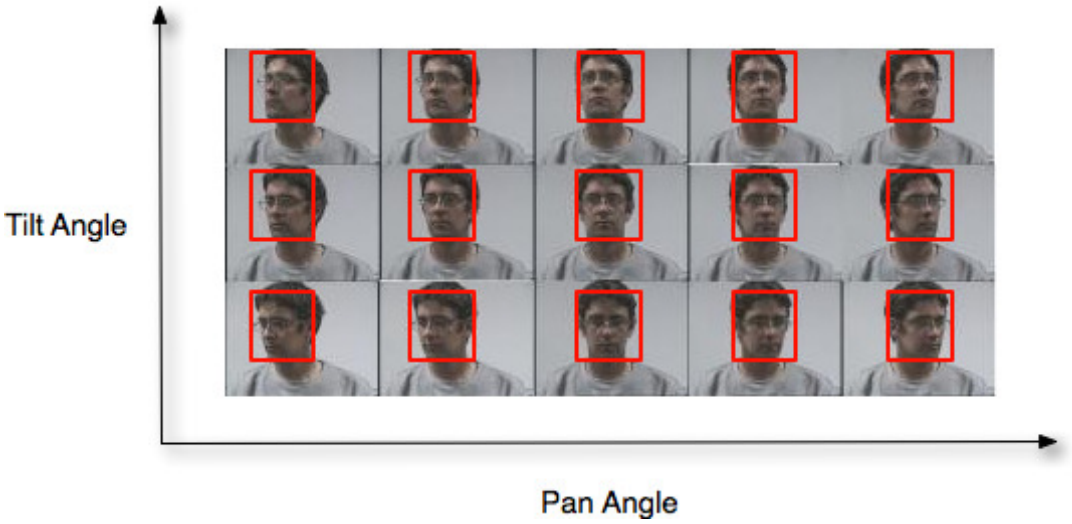


Figure 10: A small sequence from the Pointing04 dataset.

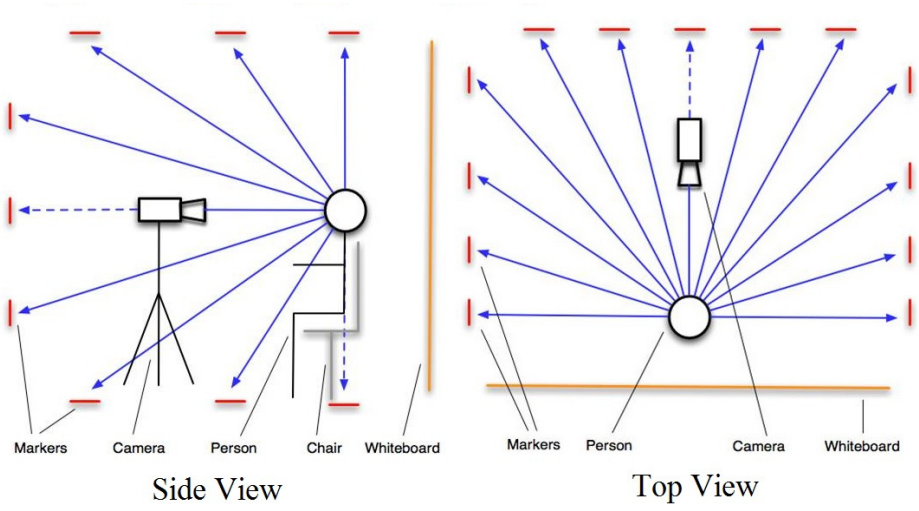


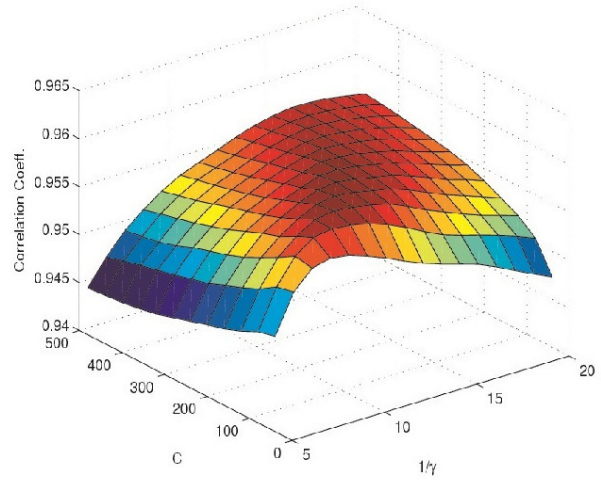
Figure 11: How the Pointing04 database was collected

We use two support vector machines for discriminating between different poses; one is trained for pan and the other for tilt.

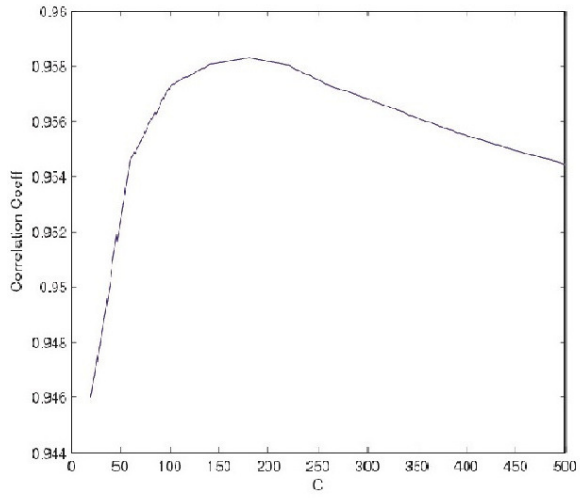
3.3.1.2 *Experimental Procedure and Results*

We used 80 percent of the data for training, 10 percent for validation and the rest for testing. Face detection was then performed on the images in

the dataset using the OpenCV face detector [120]. Following that a Half-octave Gaussian pyramid was constructed over a normalized imagette of the face which is of the size 24 X 36 pixels. The feature vector is generated using first and second order Gaussian derivatives obtained from the pyramid. It was discovered, using the hill climbing algorithm, that Gaussian derivatives of higher orders do not lead to any improvement of accuracy. SVMs were employed for both classification and regression. Cross-validation was used to choose the best hyperparameters for the SVMs.



(a)



(b)

Figure 12: (a) Graph of Correlation Coeff. vs. C-parameter and $1/\gamma$ for pan and (b) Graph of Correlation Coeff. vs. C-parameter at $1/\gamma = 11$ for pan

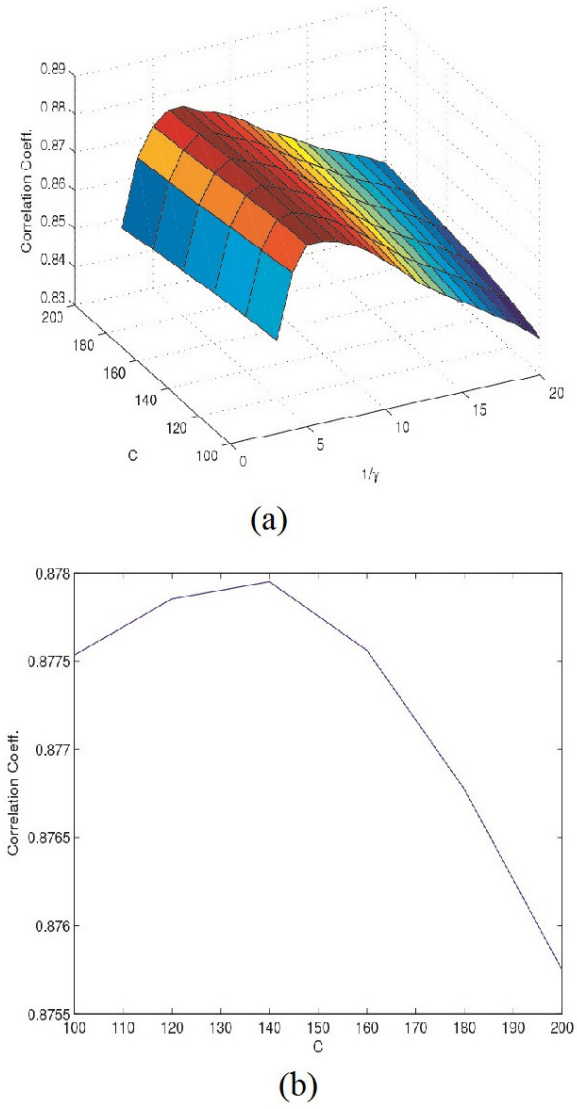


Figure 13: (a) Graph of Correlation Coeff. vs. C-parameter and $1/\gamma$ for tilt and (b) Graph of Correlation Coeff. vs. C-parameter at $1/\gamma = 6$ for tilt

The data was split several times and the accuracy calculated for every split and finally the average was calculated. The results of the Mean absolute error (MAE) are shown in the table 3 and they are better than the state of the art reported in [84].

Our mean absolute errors of 6.9, 8.0 degrees for pan and tilt respectively are much lower than the best error achieved so far by Stiefelhagen [111] which was 9.5, 9.7 degrees for pan and tilt respectively.

The accuracy achieved for the discrete poses: 64.51, 62.72 for pan is much higher than the accuracy reported by Stiefelhagen: 52, 66.3. Our accuracy for tilt is less than the accuracy achieved in [111] because the

authors of that paper considered only 7 out of the 9 poses for tilt in the Pointing04 data set.

MEA	pan	tilt
our approach	6.9	8.0
state-of-the-art	9.5	9.7

Table 3: Our MAE as compared with the state-of-the-art

Accuracy%	pan	tilt
our approach	64.51	62.72
state-of-the-art	52	66.3

Table 4: Our accuracy over discrete poses as compared with the state-of-the-art

For continuous poses the correlation coefficients for pan and tilt were found to be 0.95, 0.87 for pan and tilt respectively showing that the proposed system can work well even for continuous poses even though it is trained on a dataset containing only discrete poses. Table 5 and 6 show the confusion matrices for pan and tilt respectively.

	-90	-75	-60	-45	-30	-15	0	15	30	45	60	75	90
-90	23	1	0	1	0	0	0	0	0	0	0	0	0
-75	5	17	3	0	0	0	0	0	0	0	0	0	0
-60	0	3	10	1	0	0	0	0	0	0	0	0	0
-45	0	0	0	16	1	1	0	0	0	0	0	0	0
-30	0	0	0	4	11	2	0	0	0	0	0	0	0
-15	0	0	0	0	1	10	6	0	0	0	0	0	0
0	0	0	0	0	0	4	24	2	1	0	0	0	0
15	0	0	0	0	0	0	5	11	3	0	1	0	0
30	0	0	0	0	0	0	0	3	23	2	0	0	0
45	0	0	0	0	0	0	0	0	4	9	4	3	0
60	0	0	0	0	0	0	0	0	2	6	7	10	1
75	0	0	0	0	0	0	0	0	0	1	4	5	4
90	0	0	0	0	0	0	0	0	0	0	2	8	14

Table 5: Confusion Matrix for Pan, true values are in the first column, predicted values in the first row

	-90	-60	-30	-15	0	15	30	60	90
-90	3	0	0	0	0	0	0	0	0
-60	0	38	7	0	0	0	0	0	0
-30	0	9	13	4	0	1	0	1	0
-15	0	0	8	17	9	1	0	0	0
0	0	0	2	8	21	11	1	0	0
15	0	0	1	0	7	19	12	0	0
30	0	0	0	0	0	10	23	5	0
60	0	0	0	1	0	0	5	37	0
90	0	0	0	0	0	0	1	0	4

Table 6: Confusion Matrix for Tilt, true values are in the first column, predicted values in the first row

Table 7 shows the prediction times with and without using PCA and we can see that the PCA speeds up the prediction time by a factor of around 200.

	SVM with PCA	SVM without PCA
Prediction time(sec)	0.108	20.17

Table 7: Comparison of prediction time with and without using PCA

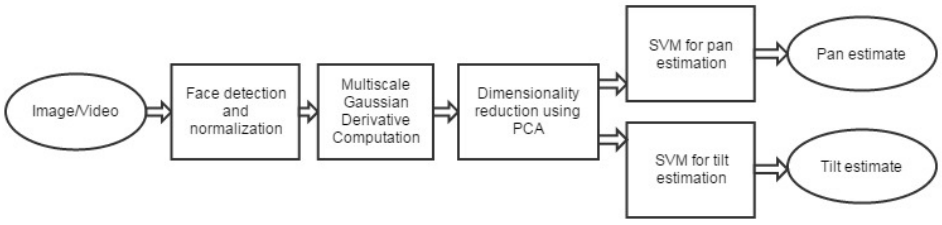
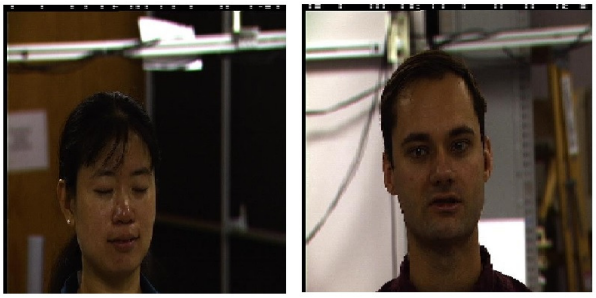


Figure 14: Schematic of our approach.

After training the SVM's on the Pointing04 dataset, they were tested on the CMU-PIE database [109]. Although the CMU-PIE database is not labeled for pose and hence does not let us perform a mathematical analysis, we could see that the predicted values of our SVM's were in agreement with the general orientation of the head in the database.

Two representative images form the PIE dataset are given below along with the results obtained from our SVM's.



(a) (b)

Figure 15: (a) Pan=-15, Tilt=-15 and (b) Pan=0, Tilt=0 were predicted using our approach

3.3.2 *Smile Detection*

Smiles, like other facial expressions, play an important role in human-human interaction. In terms of affect measurements, smiles indicate positive valence which may consequently imply happiness, appreciation or satisfaction. Since smiles can be indicative of a positive mental state, smile detection has various applications ranging from patient monitoring to product rating. We propose two methods for smile detection based on global image appearance. The first experiment on smile detection uses Multi-scale Gaussian derivative features while the second experiment uses a combination of Multi-scale Gaussian derivatives and Local Binary Patterns.

3.3.2.1 *Related Work*

Even though a lot of research has been done on automated facial expression analysis, very few published works have explicitly dealt with smile detection. In [79] McDuff *et al.* used smile intensity to predict how much a viewer likes a particular video. Smile detection is an integral part of emotional state estimation in humans. It also has a variety of applications in consumer surveys, gaming and user interfaces.

A major issue in facial expression analysis is that most of the research is validated on posed databases. In [117] authors have argued that spontaneous expressions are different from posed expressions in both appearance and timing therefore systems developed for recognizing posed expressions might not work well on real world expressions. Spontaneous expressions are much more subtle and complex than posed expressions.

Most smile detection systems in the past have been trained on these posed databases. Deniz *et al.* in [32] presented a smile detection system based on finding keypoints on the face and tested their method on the

DaFex [14] and JAFFE [76] databases which both contain posed smiles. Others in [108, 53, 68] have all experimented on posed databases.

The GENKI-4K database presented by Whitehill and others in [127] contains 4000 images with a wide range of subjects, facial appearance, illumination, geographical locations, imaging conditions and camera models. The images are annotated for smile content(1 = smile, 0 = non-smile). The difference between this database and other facial expression databases is that this database was compiled from images on the internet rather than being captured in a controlled environment.

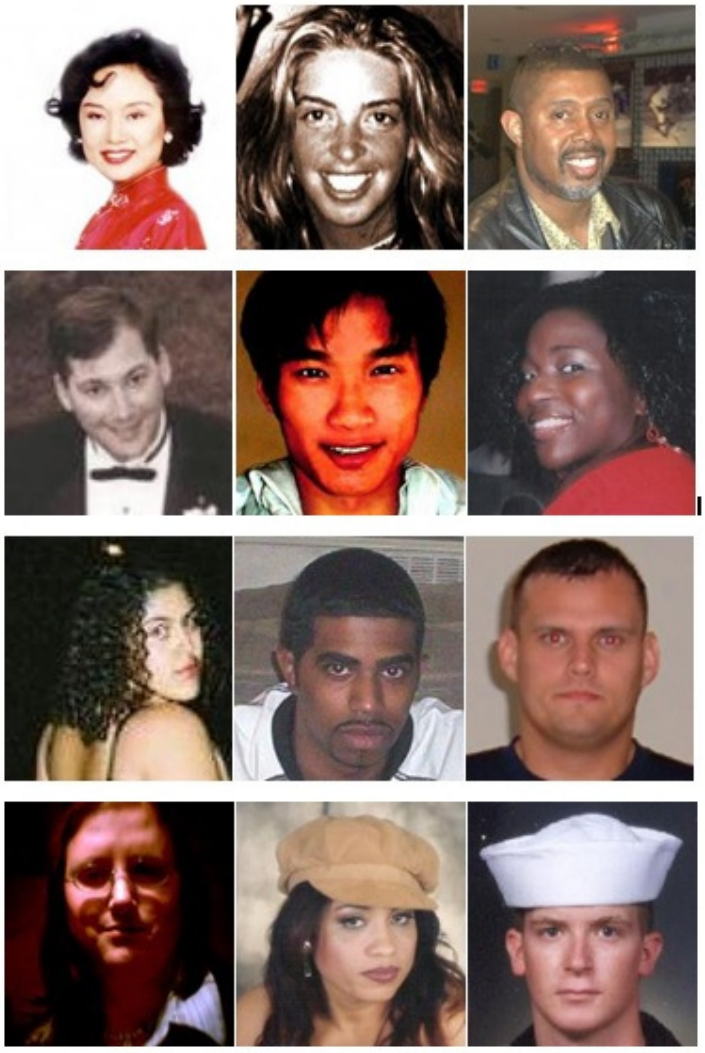


Figure 16: Examples of (top two rows) real-life smile faces and (bottom two rows) nonsmile faces, from the GENKI-4K database.

Shan in [106] presented a comprehensive study on smile detection and proposed his own method which was faster than the state-of-the-art but not more accurate than Gabor filters combined with SVM's.

3.3.2.2 *Experiments using Multi-scale Gaussian Derivatives*

We used 3577 out of the 4000 images in the GENKI-4K dataset removing ambiguous cases and images with serious illumination problems like partial lighting of the face, 60 percent of the data for training and the rest for testing.

Face detection was then performed on these images using the OpenCV face detector [120]. Following that a Half-octave Gaussian pyramid was constructed over a normalized imagette of the face which is of the size 64 X 64 pixels, this size of 64 X 64 pixels for the normalized region was chosen after extensive experimentation where normalized images of 64 X 64 pixels gave better results at smile detection as compared to other sizes. The imagette was divided into cells of 4 X 4 pixels and the feature vector contained the mean and standard deviation of the descriptor values (first and second order Gaussian derivatives obtained from the pyramid) for each cell of 4 X 4 pixels.

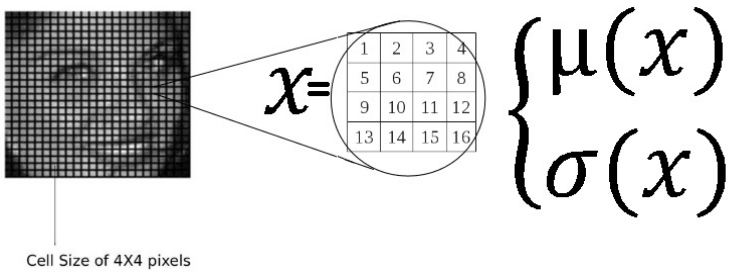


Figure 17: Imagette divided into cells of 4 X 4 pixels

Using PCA the dimensionality of the feature vector was reduced to 61 from the original dimensionality of 3920. A SVM was used for classification, the optimum hyperparameters of which were found using cross validation.

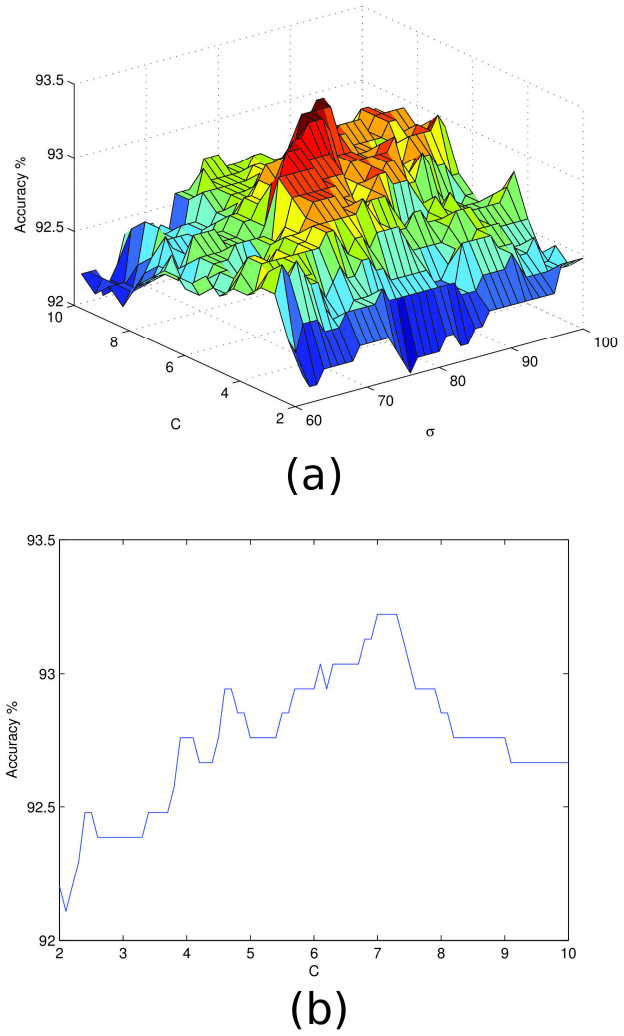


Figure 18: (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 81$

The data was split several times and the accuracy calculated for every split and finally the average was calculated. We achieved a classification accuracy of 92.97% using PCA for dimensionality reduction. Our results are superior to the state of the art Gabor Energy Filters as shown in the table 8. The following settings were used for the Gabor filter: the sinusoidal spatial frequency used three values $(\pi/2, \pi/4, \pi/8)$ while the orientation used six values $(k\pi/6, k \in \{0...5\})$. So effectively we have a filter bank with $3X6 = 18$ filters.

	GEF	MGD with PCA
Accuracy%	90.78	92.97

Table 8: Our accuracy using Multi-scale Gaussian Derivatives (MGD) compared with the accuracy obtained using Gabor Energy Filters (GEF)

Figure 19 shows the ROC for the SVM trained on the GENKI-4k dataset.

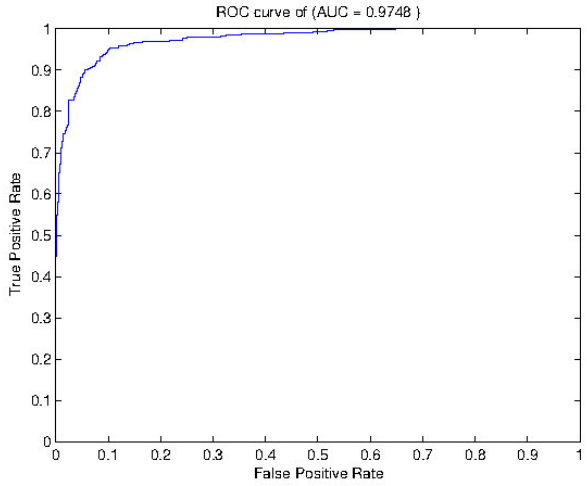


Figure 19: ROC curve for our smile detector

Table 9 shows the prediction times with and with using PCA, as we can observe the use of PCA speeds up the prediction time by a factor of over 60.

	SVM with PCA	SVM without PCA
Prediction time(sec)	0.254	17.133

Table 9: Comparison of prediction time with and without using PCA

The SVM trained over the GENKI-4K database is used for smile detection on images sequences of the Cohn-Kanade database. It was found that the probability estimates of smile detection for the frames in the image sequence are representative of the smile intensity as seen in image 20.

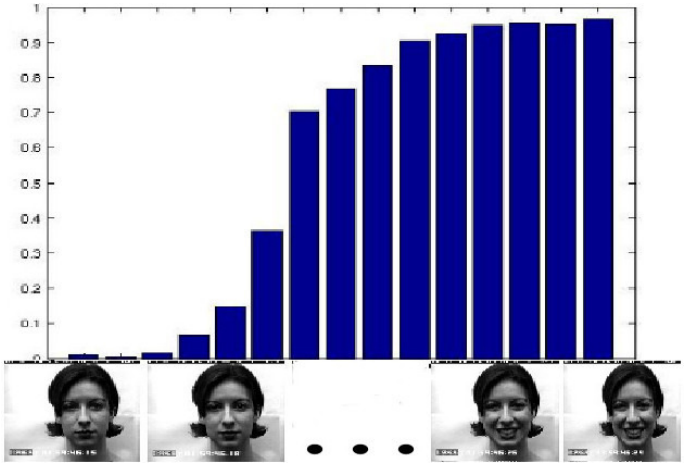


Figure 20: Smile intensity using probability estimates

Figure 21: Schematic of our approach

During our experiments on smile detection we discovered that Gaussian derivatives are not invariant to extreme lighting variations. Some examples of these extreme illumination issues from the GENKI-4K database are shown in figure 22. The next section describes our proposed solution to these illumination problems using a combination of Gaussian derivatives and LBP. The classification accuracy obtained in the experiments in this section and the next cannot be compared because of different training set vs. testing set split ratios.

Figure 22: Examples from the GENKI-4K database illustrating the illumination problems

3.3.2.3 *Experiments with Local Binary Patterns Calculated Over Gaussian Derivatives*

In order to combine Gaussian derivatives with LBP, Gaussian derivative images were produced from the normalized input images by using the

Half-octave Gaussian pyramid [27] which allows for the fast calculation of Gaussian derivatives.

First and second order derivative images of the following order were used: $I_x, I_y, I_{xx}, I_{yy}, I_{xy}$ from the base of the pyramid($\sigma = 1$). Next these derivative images were divided into grids of 4 X 4 local regions with 43.75% overlapping areas from which uniform LBP features were calculated. The local histograms were concatenated to obtain the final feature vector. Since we had 5 derivative images and 16 grids with each grid producing a uniform LBP histogram, we obtained a feature vector of: 5*16*59 = 4720 dimensions.

Unlike the method introduced by the authors of [96], we did not use tensor mathematics and the number of Gaussian features used by us is much lower than the number used in [96]. We followed the hill-climbing algorithm and started adding Gaussian derivatives starting from the base of the pyramid and stopped there because adding derivatives from the level above did not lead to an improvement in accuracy. On the other hand in [96] the authors used 6 levels of the pyramid.

The grid size of 4 X 4 and 43.75% of overlap area was chosen by means of cross-validation. Images were normalized to 66 X 66 pixels, this size was also chosen through cross-validation.

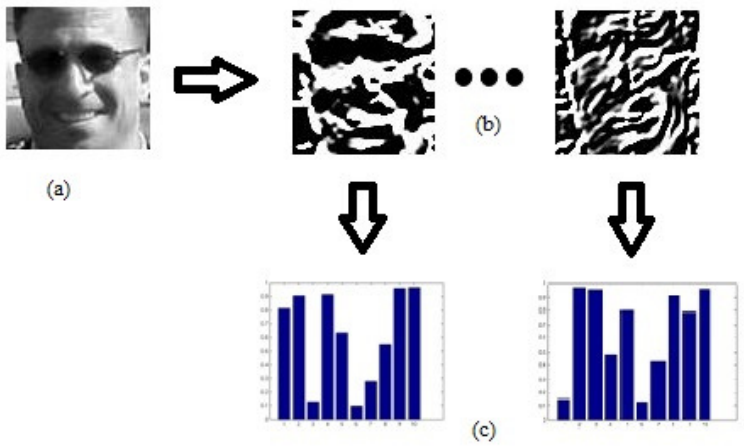


Figure 23: Creating the features: a) original image, b) Gaussian derivative images, and c) concatenation of resulting histograms after applying LBP.

We performed our experiments again on the GENKI-4K dataset using 80 percent of the images for training, 10 percent for cross-validation and the remaining 10 percent for testing. We used SVMs with a radial basis kernel to compare the accuracies obtained by different descriptors. The images were not aligned using facial features such as eyes and the location of the nose. A ten fold cross validation procedure was adopted to obtain

the final results. Apart from measuring the accuracy we also measured the Balanced Error Rate (BER).

		Prediction	
		Class-1	Class+1
Truth	Class-1	a	b
	Class+1	c	d

Table 10: Confusion matrix for 2 class classification

The balanced error rate is the average of the errors on each class: $BER = 0.5 * (b/(a+b) + c/(c+d))$. Where a, b, c, d stand for: true negatives, false positives, false negatives and true positives respectively.

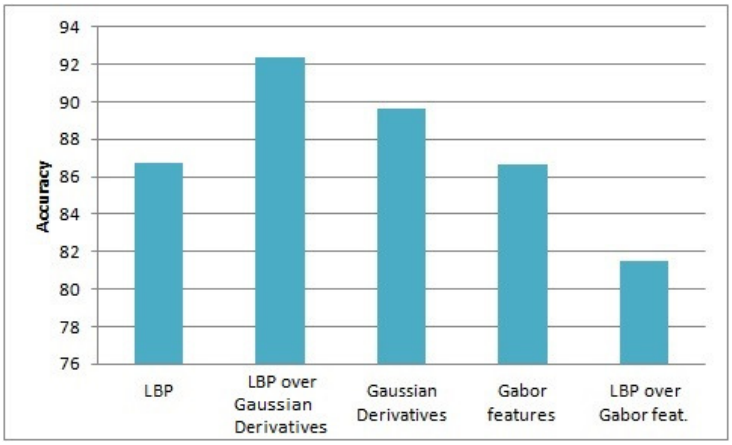


Figure 24: Accuracy(%) of different descriptors over the GENKI-4K database

The proposed technique achieved the highest accuracy of 92.3602% with the lowest BER of 0.0702.

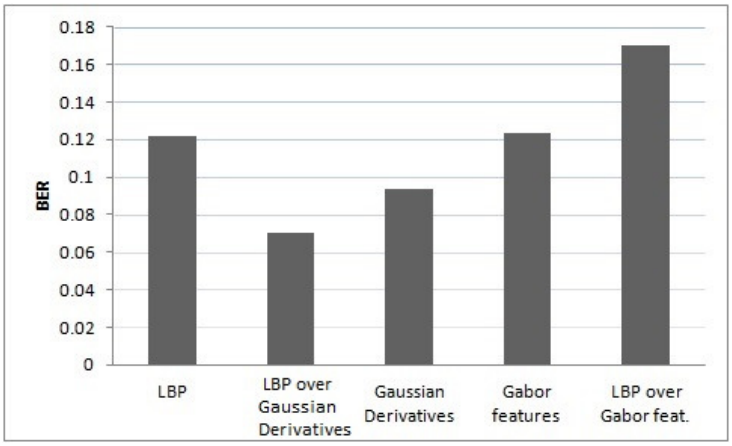


Figure 25: BER of different descriptors over the GENKI-4K database

It is surprising to see that LBP calculated over Gabor features actually perform worse than both Gabor features and LBP alone. This could be because of the curse of dimensionality since we are using Support Vector Machines with a radial basis kernel and the feature vector of LBP calculated over Gabor features has a dimensionality of nearly 17000 whereas the number of training instances is less than 3600.

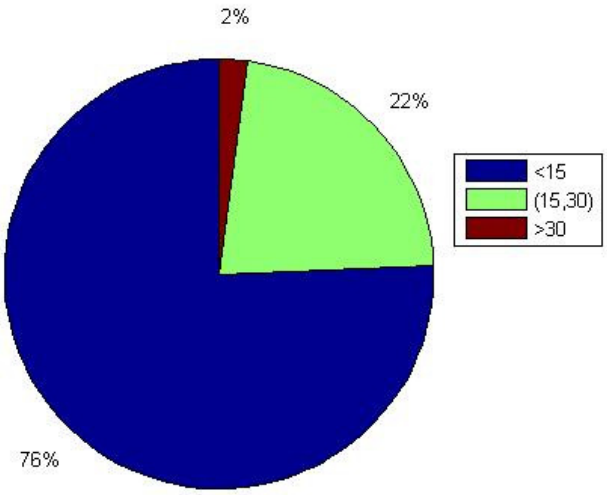


Figure 26: Proportion of images with different pose

We divided the GENKI-4K database into 3 subsets according to the head pose(only yaw). The proportion of images that fall into the three sets is shown in figure 26.

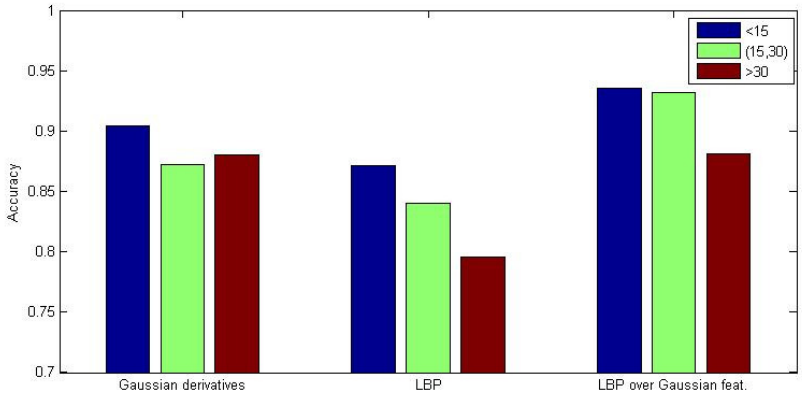


Figure 27: Accuracy with different poses

As expected all the three techniques produce the best results for head pose close to frontal i.e. < 15 degrees, slightly lower for 15 to 30 degrees and lowest for yaw above 30 degrees except for Gaussian derivatives which perform slightly better at yaw above 30 degrees than they perform at yaw between 15 and 30 degrees. The most interesting aspect of the results is that Gaussian derivatives and LBP features calculated over Gaussian features are less susceptible to pose variation than LBP alone explaining why our method performs better than traditional LBP.

3.3.3 *Affect Sensing*

The appearance based technique we developed for head pose estimation and smile detection is extended to affect sensing. The emotional state inferred from facial expressions is represented in the affect space.

3.3.3.1 *Related Work*

Facial expressions are a mirror to human emotions and an important component of human to human interaction. Human computer interaction requires the same ability to read emotions from facial expressions.

In [33] Ekman presented the Facial Action Coding System (FACS), a taxonomy to describe facial expressions in terms of individual muscle movements which we have described in detail in section 2.3. FACS based approaches have been adopted in a variety of vision systems such as the Computer Expression Recognition Toolbox (CERT) [74]. Such systems are trained to estimate the Action Unit (AU) intensities which can then be used to assign one of the six basic emotion labels to that image or frame. The problem arises when the expression in the image is not associated with any of the six basic emotions.

An alternative to such a structured approach is to represent the underlying emotions in a multidimensional emotion space. Shin in [107] used component analysis techniques to recognize emotions and map them to the affect space. Another method was presented by Dahmane and Meunier in [28]. The authors used Gabor wavelets and Support Vector Machines on the Semaine database [82] and they use 4 dimensions (Activation, Expectation, Power and Valence) to represent the emotions that underlie the facial expressions.

In [98] the authors argue that three dimensions are enough to represent any emotion. In our experiments we use the affect space model developed by Russell and Mehrabian and compare our results for Pleasure and Arousal with the results from the technique presented in [28].

Two common ways to describe image features are: appearance based methods and geometric feature based methods. The latter involves detec-

tion and tracking of facial keypoints such as the lip corners, nostrils and eyes. This detection and tracking is done with the help of computationally expensive vision techniques and are not very robust.

The approach we present here does not involve identification of any landmarks on the face and just like the appearance based technique discussed in [28], the image filters are applied to the entire face to obtain the feature vector.

3.3.3.2 *Datasets Used in Experimentation*

Our approach was tested on the Cohn-Kanade [65] and FEED [121] databases. The FEED database from the Technical University of Munich is an image database containing facial images of subjects experiencing the six basic emotions as defined by Ekman [33]. The database has been generated as part of the European Union project FG-NET (Face and Gesture Recognition Research Network).

One of the major problems in working with facial expression databases is that most of the databases contain posed expressions. Thus one of the underlying motivations for the FEED database was to let the observed people react as natural as possible. The emotions were elicited by playing video clips and recording the reactions of the subjects. The recordings were made using a camera mounted on top of the computer screen used to display the video clips.

The database contains image sequences from 18 subjects. Each emotion is elicited three times from every subject. The images were acquired using a Sony XC-999P camera equipped with a 8mm COSMICAR 1:1.4 television lens. A BTTV 878 frame-grabber card was used to grab the images with a size of 640x480 pixels, a color depth of 24 bits and a frame-rate of 25 frames per second. However the images were down-sampled to 320X240 pixels and the color depth reduced to 8 bits before the database was uploaded.

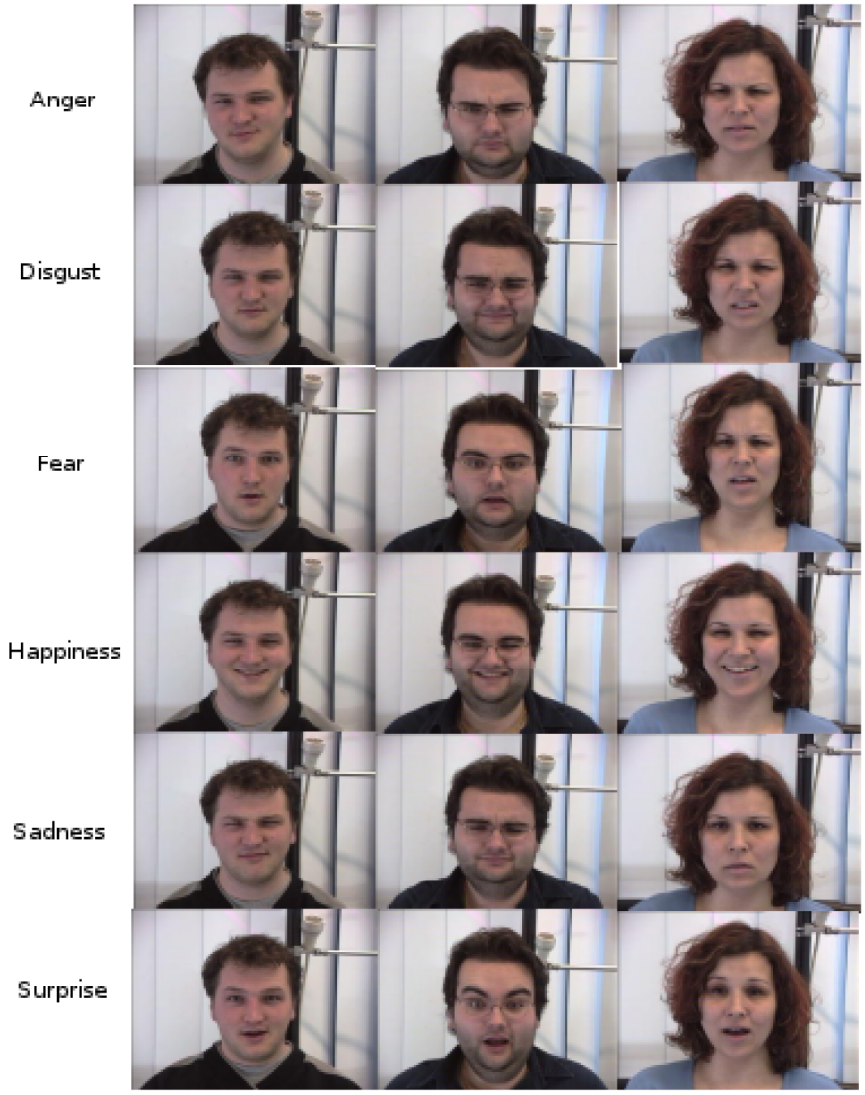


Figure 28: Example Images from the FEED dataset

The Cohn-Kanade (CK) database contains 486 sequences from 97 subjects. A sequence begins with a neutral expression progressing to a peak expression. The peak expression for each sequence is FACS coded and given a basic emotion label. The emotion label refers to what expression was requested rather than what may actually have been performed.

The subjects were 18-30 years old. Sixty-five percent of which were female; 15 percent were African-American and three percent Asian or Hispanic. The subject sat on a chair while two Panasonic WV3230 cameras, each connected to a Panasonic S-VHS AG-7500 video recorder with a Horita synchronized time-code generator recorded the facial expressions. One camera captured the frontal pose of the subject while the other was located at an angle of thirty degrees from the front. Only the image data from the frontal camera is available in the database.

Subjects were instructed by an experimenter to perform a series of facial displays each beginning with a neutral or near neutral face. Six were based on descriptions of prototypical emotions (i.e., joy, surprise, anger, fear, disgust, and sadness). These six tasks were annotated by certified FACS coders. Seventeen percent of the data was comparison annotated. Inter-observer agreement was quantified with kappa coefficient, which is the proportion of agreement above what would be expected to occur by chance. The mean kappa for inter-observer agreement was 0.86. Image sequences from neutral to target display were digitized into 640 by 480 or 490 pixel arrays with 8-bit precision for grayscale values.

Figure 29: Example Images from the CK database

We mapped the basic emotions to the Pleasure-Arousal space as shown in table 11 in accordance with the Pleasure-Arousal values provided by Mehrabian [98]. Instead of using numerical values we assigned class labels (+P -P, +A -A) to perform binary classification.

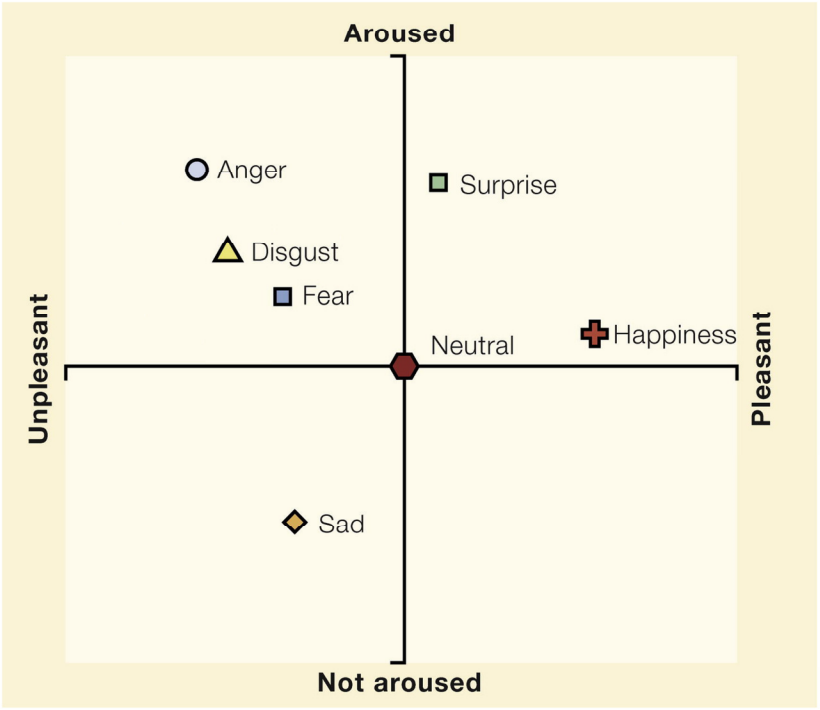


Figure 30: Basic emotions in the affect space

Emotion	P Label	A Label
Joy	+	+
Sadness	-	-
Surprise	+	+
Anger	-	+
Disgust	-	+
Fear	-	+

Table 11: Labels for the 6 basic emotions

The Cohn-Kanade and FEED databases were re-annotated with these class labels. The Cohn-Kanade database was used for training and validation while the FEED database was used for testing.

3.3.3.3 *Experiments and Results*

Employing the same procedure for feature vector generation and dimensionality reduction as described in 3.3.2.2 we used SVMs for classification.

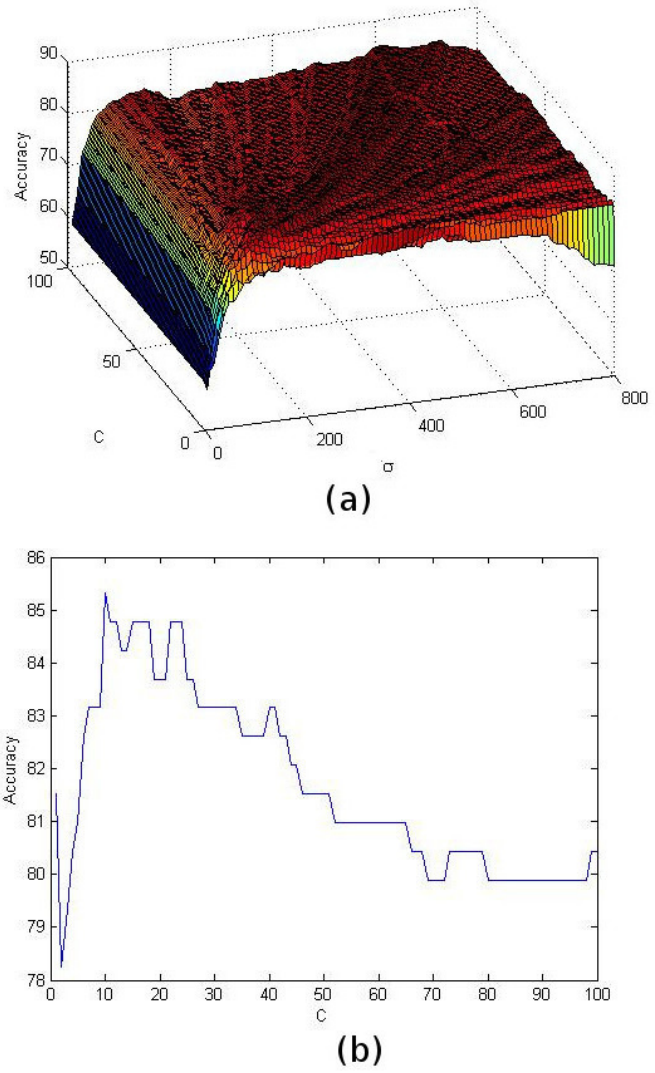
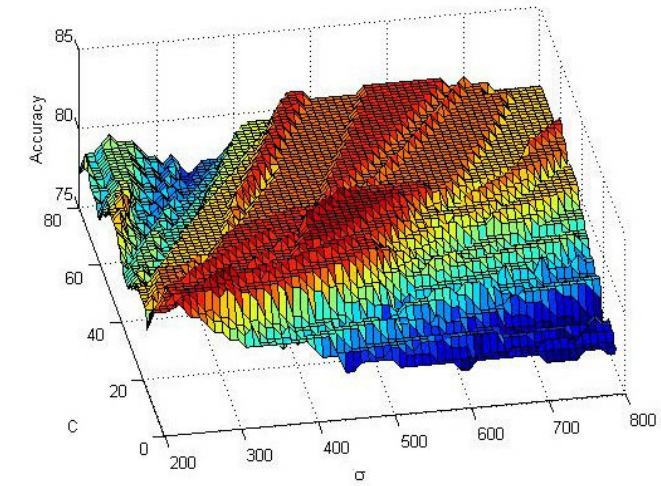
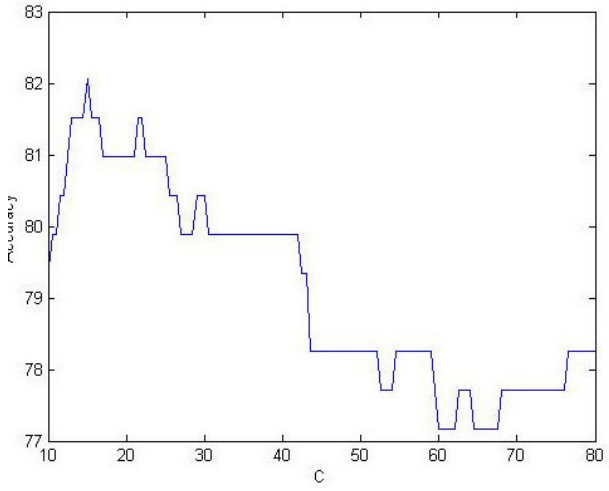


Figure 31: (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ for pleasure and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 190$ for pleasure



(a)



(b)

Figure 32: (a) Graph of classification accuracy vs. C-parameter and $1/\gamma$ for arousal and (b) Graph of accuracy vs. C-parameter at $1/\gamma = 280$ for arousal

Figure 33 reiterates the process.

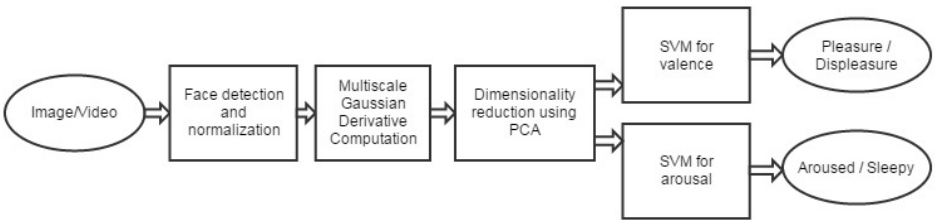


Figure 33: Schematic of our approach

We divided the Cohn-Kanade database into two, 70 percent of the images were used for training and the rest for validation. The database was split several times and the accuracy is calculated for every split and the average is calculated. The ROC for the two SVM's used are shown in the figures below. The first ROC is for the SVM trained for detecting Pleasure and the second one for Arousal.

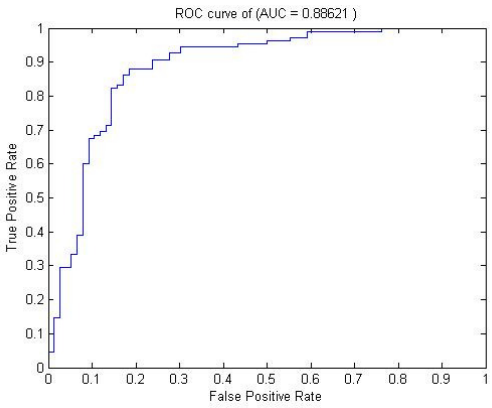


Figure 34: ROC of the classifier for Pleasure

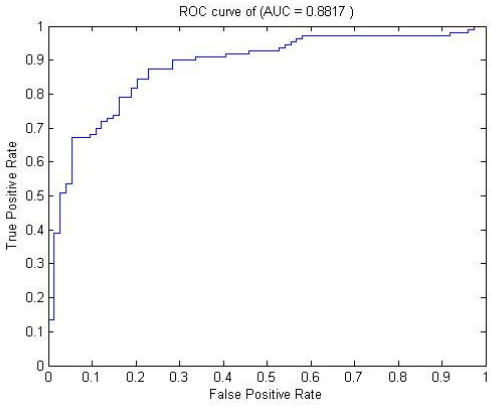


Figure 35: ROC of the classifier for Arousal

The accuracy of our approach over the Cohn-Kanade set is 85.32,82.06 percent for pleasure and arousal respectively. On the other hand the approach developed by Dahmane and Meunier [28] using Gabor filters with three spatial frequencies and six orientations i.e. a total of eighteen filters as discussed in section 3.3.2.2, achieves an accuracy of only 71.80,74.94 percent for pleasure and arousal respectively. Dahmane and Meunier's approach was replicated to obtain the results against which our results were compared.

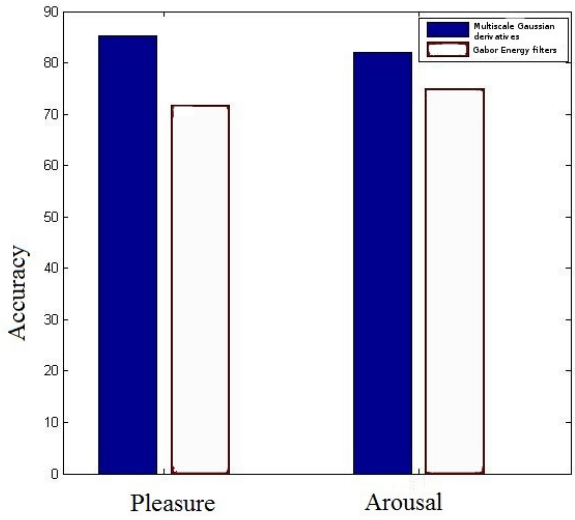


Figure 36: Comparison of results

We also see that it takes much less time to computer Gaussian derivatives using the Half-octave pyramid as compared to Gabor features because of the integer coefficient Half-Octave Pyramid used. Table 12 shows the time to calculate the features for the complete Cohn-Kanade database using the two techniques on the same machine(Intel Xeon Quad-Core 3GHz, 4GB RAM).

	Multi-scale Gaussian Derivatives	Gabor Energy Filters
Calculation Time(sec)	5.36	20.37

Table 12: Comparison of time required for calculating the two types of features

PCA reduces the prediction time by a factor of over 60, table 13 compares the prediction time with and without using PCA.

	SVM with PCA	SVM without PCA
Prediction time(sec)	0.0155	0.8495

Table 13: Comparison of prediction time with and without using PCA

Table 14 shows the prediction time of our technique versus the state of the art because our feature vector is much smaller.

	Multi-scale Gaussian Derivatives	Gabor Energy Filters
Prediction Time(sec)	0.0155	1.06

Table 14: Comparison of prediction time

Our approach is then tested on the FEED database and the accuracy for Pleasure-Displeasure is 70.73% while it is 70.08% for Arousal-Nonarousal.

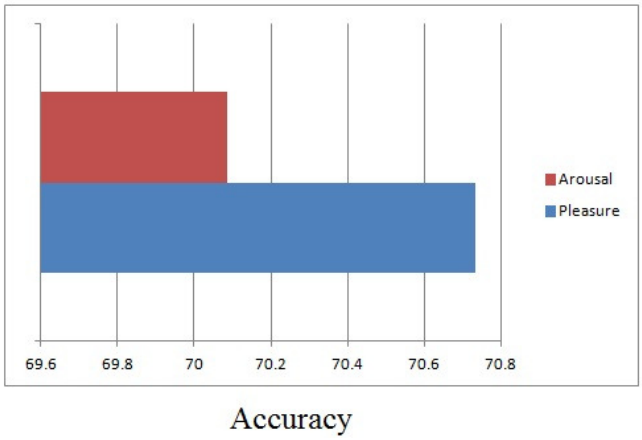


Figure 37: Results on the FEED database

3.3.4 *Face Recognition*

The descriptor combination devised in 3.3.2.3 was employed for face recognition on the CMU-PIE [109] and YaleB [43] data sets. The results were produced using simple 1-Nearest Neighbor method. The experiments on face recognition discussed in the following sections are an illustration of the versatility of the descriptor developed for smile detection using Gaussian features and Local Binary Patterns.

3.3.4.1 *Related Work*

Face recognition involves the identification of individuals from an image or video frame. A major challenge in face recognition is to make the system invariant to illumination since the appearance of the face can change dramatically with changes in lighting conditions. Other problems include aging, occlusions, pose and facial expressions.

People have experimented with feature based techniques for face recognition using methods such as elastic graphs in [130] where the authors generated a graph using fiducial points labeled with Gabor filter responses

and in [9] where Gabor filters were replaced by Histogram of Oriented Gradient features [29]. Holistic approaches are more popular involving descriptor calculation over the entire image rather than on local features of the face.

Li and Yin used the wavelet transform in conjunction with neural networks for face recognition[71]. In [8] the authors used the versatile descriptor LBP for face recognition over the FERET dataset. Ruiz-Hernandez *et al.* [96] combined LBP with Gaussian features maps and then generated a tensor which was then reduced in dimensions using Multilinear Principal Component Analysis and finally recognition was done with a Kernel Discriminative Common Vector method.

In [131] the authors dealt with the problem of illumination by dealing with the effects of illumination on large scale and small scale features explicitly, they achieve the best results by combining their illumination normalization technique with quotient images[124]. Meanwhile in [138] an illumination invariant descriptor was presented for face recognition which also claims to solve the bottleneck associated with heterogeneous lighting.

3.3.4.2 *Experiments with Face Recognition*

Using the protocol we used for smile detection described in 3.3.2.3. We tested our method on the CMU-PIE and the extended YaleB databases. The performance of our method was compared to the state of the art methods such as SQI [123], LTV [20], WF [122], Gradient Face [135], LGH [138] for both the databases and with the method used by Ruiz-Hernandez *et al.* in [96] on the YaleB database. Additionally we compared our performance to Gabor filters, LBP, Gaussian derivatives and LBP calculated over Gabor images for the CMU-PIE dataset. We did not need to replicate the experiments for any of the competing techniques: SQI, LTV, WF, Gradient Face or LGH as the accuracies using all these techniques have been already been reported for both the CMU-PIE database and the Yale B database by Zhu *et al.* in [138].

We used a subset of the original CMU-PIE database, 1428 frontal images from 68 people under 21 variations of illumination conditions were selected. No feature alignment method was used.

Figure 38: Sample images from the CMU-PIE database showing the different lighting conditions

The results are presented in figure 39. Only one image per individual is used as the reference image. All the 21 images taken under different lighting conditions are chosen as the reference images one at a time. We used the L1 distance as the similarity measure. The reference image closest to the test image decides the identity of the person in the test image.

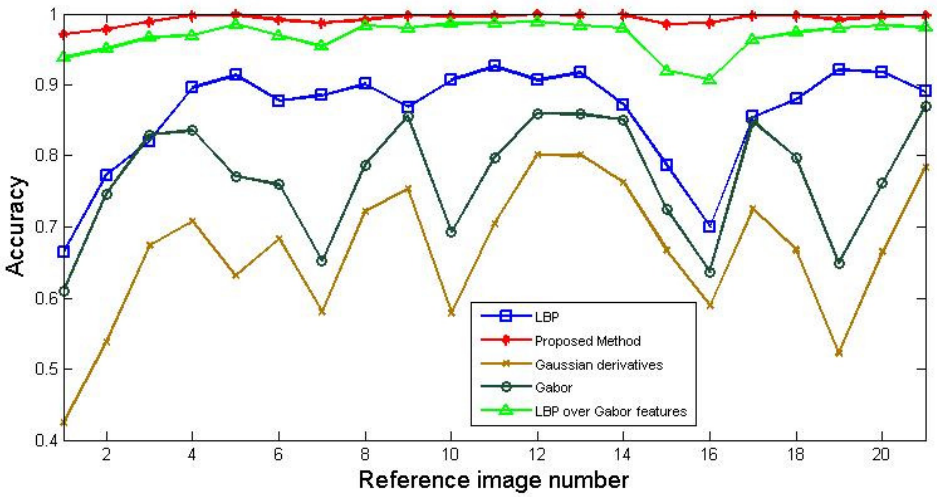


Figure 39: Recognition rate with different reference images

Our approach achieved the highest accuracy of 100% with the image with frontal lighting. Only the performance of LBP calculated over Gabor images comes close to our method. Gaussian derivatives and LBP alone do not achieve very high recognition rates making the case for our method which combines the two.

The maximum and average recognition rates for the different descriptors are given in table 15.

	SQI	LTV	WF	GF	LGH	Ours
Max. Acc%	98.82	95.81	99.71	99.93	**100**	**100**
Avg. Acc%	89.77	80.78	89.52	96.93.93	98.19	**99.26**

Table 15: Maximum and average accuracy attained by different methods

The extended YaleB dataset contains images from 28 individuals captured under 64 different lighting conditions with 9 pose views. We only used the images with frontal views in our experiments.

Figure 40: Sample images from the extended YaleB database showing the different lighting conditions

Researchers have divided the database into 5 subsets in increasing order of complexity of lighting conditions. We used the image number A+000E+00 with the simplest lighting scenario as the reference image. Minkowski distance metric was used as the similarity measure. The p-Minkowski metric between two points a=(x1,y1) and b=(x2,y2) can be given as:

$$d^p(a,b) = [|x1-x2|^p + |y1-y2|^p]^{\frac{1}{p}} \tag{21}$$

The optimum value of p varies from 0.75 to 1.25 on the YaleB database. The reference image closest to the test image decides the identity of the person in the test image. We compare the results of our technique with the state of the art in table 16.

	Set 1	Set 2	Set 3	Set 4
SQI	88.60	**100**	85.75	87.97
LTV	87.28	99.78	66.67	45.49
WF	79.39	99.78	75.88	77.07
GF	94.74	**100**	83.33	75.94
LGH	94.74	**100**	92.54	**96.43**
Ruiz-Hernandez[96]	**100**	**100**	94.7	60.1
Ours	**100**	**100**	**97.22**	79.10

Table 16: Accuracy(%) over the 4 subsets using different methods

Our technique achieved the highest accuracy for the first 3 subsets, on the 4th subset it is beaten by SQI and LGH, which are both techniques that handle the problem of illumination explicitly.

It is interesting to see how the two components of our approach namely Gaussian derivatives and LBP alone match up against their proposed combination.

	Set 1	Set 2	Set 3	Set 4
Gaussian Derivatives	97.53	93.52	62.65	25.13
LBP	99.38	99.38	55.86	33.07
Ours	**100**	**100**	**97.22**	**79.10**

Table 17: Accuracy(%) over the 4 subsets using Gaussian derivatives, LBP and their proposed combination

3.4 CONCLUSION AND SUMMARY

We have presented simple yet effective techniques for static image analysis using the global appearance of images without using any form of key-point detection. We have managed to maintain a high degree of commonality between the architectures used for the different experiments described in section 3.3.

Through these experiments it is evident that not only do Multi-scale Gaussian Derivatives give better accuracy than Gabor filters but they are also computationally cheaper to calculate. In these experiments, we have only used three orientations for Gaussian derivatives whereas we used eighteen for Gabor filters, leading to a shorter feature vector size for Gaussian features.

Using a combination of Gaussian derivative features and Local Binary Patterns, we perform new set of experiments on the the GENKI-4k database for smile detection and on the CMU-PIE and YaleB database for face

recognition and find that this union of descriptors gives better performance figures than the two component descriptors alone. Our proposed method is also more invariant to pose as compared to LBP and less sensitive to illumination changes than Gaussian derivative features.

4

VIDEO ANALYSIS FOR SENSING AFFECT

This chapter presents the techniques developed for facial expression recognition and depression estimation.

Video description methods used in our experiments are described in section 4.1. Section 4.2 describes some commonly used encoding techniques for video features while section 4.3 describes our experiments using the methods and techniques described in the preceding two sections. The chapter concludes with the summary and short discussion about the results in section 4.4.

4.1 METHODS FOR VIDEO DESCRIPTION

In sections 4.1.1, 4.1.2 we provide a brief description of methods that were used for describing video features in the experiments. We also briefly describe space-time interest point (STIP) features 4.1.3 which are baseline features for action recognition. Dense trajectories have recently been shown to perform better than STIP features.

4.1.1 *Local Binary Patterns-Three Orthogonal Planes*

Zhao and Pietikainen in [136] cite the following six criteria for effective dynamic texture recognition:
1. Motion information combined with appearance information
2. Local processing to encapsulate transition information in space and time
3. Robustness to image transformations such as rotation
4. Invariance to change in lighting conditions
5. Computational simplicity
6. Multi-scale analysis

They address these criteria with the development of Local Binary Patterns-Three Orthogonal Planes (LBP-TOP) [136], which is an extension to the Local Binary Patterns image descriptor discussed in 3.1.2. LBP-TOP histograms are generated by concatenating the the co-occurance statistics in the XY, YT and XT planes where the YT and XT planes contain space-time transition information.

LBP-TOP features are calculated over small local neighborhoods of the space-time volume which not only makes them invariant to transformations such as translation and rotation but also robust to illumination

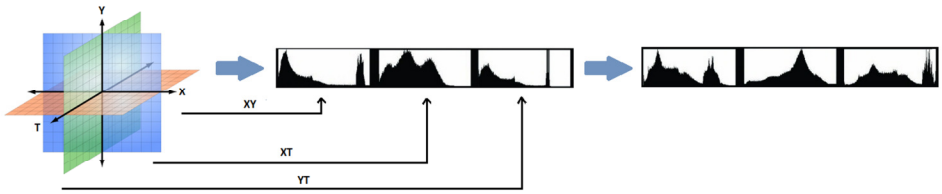


Figure 41: LBP-TOP histogram computation

variations. Since the radii in the three axes are defined over different distance metrics and can have different values the usual circular sampling employed in LBP calculation has been replaced with elliptical sampling.

In [136] the authors also present a block-based approach to apply the LBP-TOP descriptor to videos of the face for applications such as face recognition or facial expression analysis. Results on the Cohn-Kanade dataset for expression recognition suggest that calculating LBP-TOP histograms over the entire facial region only encodes the occurrences of micro-patterns. Information about the location of these micro-patterns is absent. This problem is easily overcome by dividing the video into spatio-temporal volumes. LBP-TOP histograms are calculated over each of these volumes. These histograms contain the appearance and motion information associated with a particular location in space and time. Finally the histograms from all the blocks are concatenated to generate the global descriptor which contains information on micro-pattern occurrences and their relative locations.

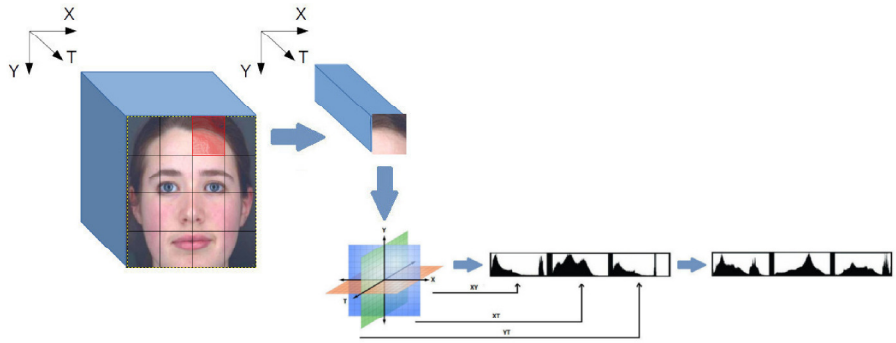


Figure 42: LBP-TOP computation for facial images

4.1.2 *Dense Trajectories*

Trajectories encode the motion information in videos. Dense sampling on all spatial positions and scales assures coverage of foreground motion

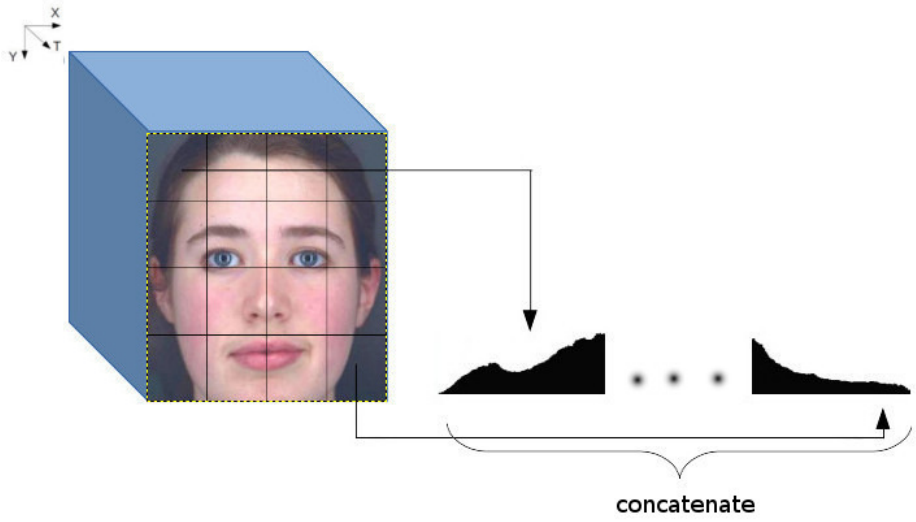


Figure 43: LBP-TOP histogram concatenation for facial images

and of the surrounding context [125]. Robust and efficient optical flow algorithms allow for the extraction of these trajectories.

Densely sampled feature points are tracked through the video except for homogeneous image areas lacking structure. After a dense optical field is computed using adjacent frames, the feature points can be tracked without any extra cost. Feature points from each spatial scale are tracked independently. Static trajectories and trajectories with sudden large displacements, having a large likelihood of being faulty, are eliminated during post processing.

The shape of the trajectory encodes local motion information. Besides this information, the following descriptors are calculated on the 3D space-time volume aligned with the trajectories to extract appearance and motion information.

(a) Motion Boundary Histograms: Optical flow represents motion from various sources including camera motion. It is important that camera motion be removed from the optical flow to perform action recognition. Motion Boundary Histograms (MBH) [30] are calculated by computing the horizontal and vertical derivatives separately on the optical flow. Since these derivatives encode the relative motion between pixels, constant camera motion is eliminated while the information pertaining to motion boundaries is retained.

(b) Optical and Gradient Histograms: Histogram of oriented gradients (HOG) and Histogram of flow (HOF) [29, 30] are computed along the dense trajectories, HOG encapsulates the appearance information whereas HOF encodes the local motion information. For both HOG and HOF, the orientations are quantized into 8 bins plus an additional zero bin for HOF.

Both the feature vectors are L_2 normalized and concatenated to obtain the fused feature vector.

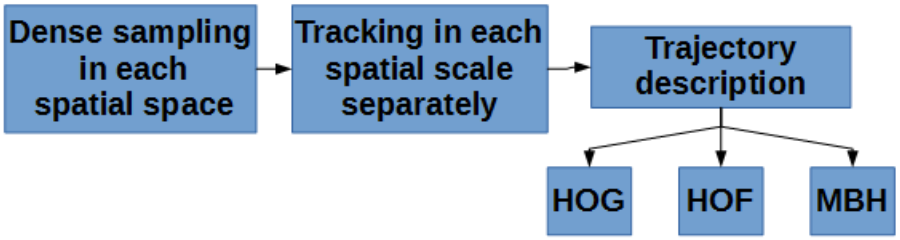


Figure 44: Dense Trajectory computation

4.1.3 *Space Time Interest Points*

Points in an image with notable local variation of image intensities are often referred to as "interest points". They are "interesting" because they are rich in information and at the same time are relatively invariant to perspective image transformations.

Laptev in [70] extended the concept of interest points to the spatio-temporal domain. He demonstrated how the Harris interest point detector can be extended to the spatio-temporal domain by choosing separate scale parameters for time and space so that the image values of interest points have large variations in both spatial and temporal directions. Automatic scale selection was provided by defining a new differential operator that could simultaneously assume a maxima for both the spatial and the temporal scale for a particular space-time event.

For each interest point, a variety of descriptors can be calculated over the space-time neighborhood. HOG and HOF descriptors are the most widely used descriptors for STIP. The HOG and HOF histograms calculated over the space-time volume are concatenated and used for building a visual dictionary.

4.2 ENCODING TECHNIQUES

Visual encoding techniques transform large amount of local image or video information into a compact representation which can be used as a visual signature for the image or video. The baseline method is to compute a spatial histogram of visual words which we describe in the following section while recent methods for visual encoding rely on soft assignment of features to mixture components as in the case of Fisher Vector encoding 4.2.3.

The techniques described in the following sections have been used for applications such as image classification [63], action recognition [38, 88], image retrieval [113], image re-ranking and object detection [7].

4.2.1 *Bag of Visual Words*

The bag-of-words model is an orderless document representation used in natural language processing and information retrieval. Bag-of-words can also be applied to image classification.

In document classification, a bag of words is a sparse vector of occurrence counts of words i.e. a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of "visual words".

To use the bag-of-words (BoW) representation for image classification, images are treated as documents and "visual words" need to be defined. To achieve this, it usually includes following three steps: Feature detection, feature description and codebook generation. When using dense trajectories, the trajectories themselves are the detected features and the HOG/HOF and MBH descriptors calculated on the 3D volumes aligned with the trajectories are the feature descriptors used to generate the codebook and later transformed to a histogram representation.

In the case of STIP, the detected features are the space-time interest points detected using the 3D Harris detector. The HOG/HOF descriptors calculated over the space-time volumes serve as the input for codebook generation and once the codebook is generated, the HOG/HOF feature descriptors are represented as sparse histograms.

Codebook generation is generally performed using unsupervised learning methods such as k-means clustering. The clusters learned using k-means clustering are the "visual words". It is important that the training set used for learning these “visual words”is sufficiently representative so that the codebook learned is comprehensive enough.

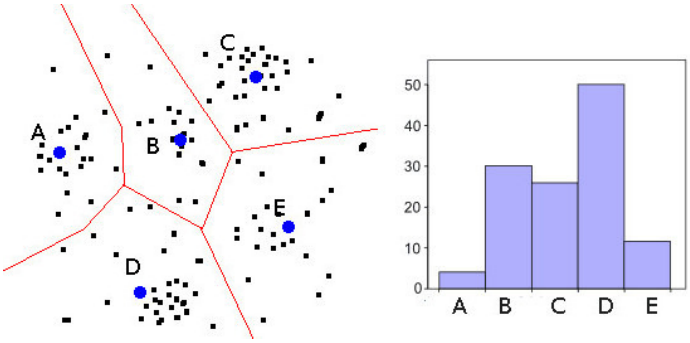


Figure 45: K-means and codebook generation [1]

Once the codebook is generated, the vector quantizer maps the feature vector to the closest "visual word" in the codebook. There is no perfect way to decide the size of the codebook, the optimum size is found empirically and varies from application to application.

4.2.2 *Sparse Coding*

Sparse coding involves the decomposition of a signal into a linear combination of basis signals. This set of basis functions is called a dictionary. Dictionary learning traces its roots back to wavelet based signal processing. Images were often decomposed using predefined dictionaries with the wavelets serving as basis functions. Recently, it has been shown that learning the dictionary instead of using a ready-made dictionary leads to much better signal reconstruction. Say we have a training set of signals $X = [x_1, x_2 ... x_n]$ in $\mathbb{R}^{m \times n}$, we optimize the cost function:

$$f_n(\mathbf{D}) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, \mathbf{D}) \tag{22}$$

where $\mathbf{D}$ in $\mathbb{R}^{m \times k}$ is the dictionary and each column represents a basis vector, l is the loss function and should be small for a $\mathbf{D}$ good at representing an input signal x. The signal dimension m is usually small compared to the number of samples n for typical image retrieval operations. Also the number of samples n is much larger than the number of dimensions in the basis vector i.e. k. It is worth noting that in sparse coding, overcomplete dictionaries with $k > m$ are allowed.

The cost function $l(x_i, \mathbf{D})$ is defined as the optimal solution of the l_1 sparse coding problem:

$$l(x, \mathbf{D}) = \min_{\alpha \epsilon \mathbb{R}^k} \frac{1}{2} \|x - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{23}$$

In equation 23 λ is the regularization parameter, $\alpha = [\alpha_1 ... \alpha_n]$ in $\mathbb{R}^{k \times n}$ is the coefficient of sparse decompositions and l_1 regularization over α yields a sparse solution.

Most of the contemporary dictionary learning algorithms are second-order iterative batch processes, using the complete training set at each iteration in order to minimize a constrained cost function. These methods are faster than first order gradient descent methods but even they cannot handle "Big-data". Only online dictionary learning methods such as the one described by Mairal *et al.* in [77] can address the problems associated with "Big data" and dynamic training.

Once the dictionary has been learnt the next step involves encoding each local descriptor into an N-dimensional vector by fitting a linear model with a l_1 sparsity constraint. Finally for a given image or video, the sparse vectors are pooled. It has been empirically found that max-pooling works better than average pooling. The pooled vector is finally normalized to generate the single signature sparse vector for the image or video.

4.2.3 *Fisher Vector Encoding*

An alternative to the Bag of Words approach is the Fisher Vector Encoding (FV) framework. Unlike Bag of Words and Sparse Coding where features are expressed as combinations of "visual word", Fisher Vector encoding records the difference between the features and "visual words" [99].

A generative model in the form of a Gaussian Mixture Model (GMM) is built over a subset of the samples, it can also be considered as a "probabilistic visual vocabulary", now each sample can be characterized by its deviation from the GMM. The deviation is quantified in the form of gradients of the log-likelihood of the sample with respect to the GMM parameters.

In a mathematical sense, BoV coding is a special case of Fisher Vector Encoding where the gradient calculation has been limited to mixture weight parameters of the GMM. The improvement in accuracy of FV over BoV can be attributed to the additional gradient information incorporated in FV. Chatfield *et al.* in [19] reported that Fisher Vector encoding works better than a variety of encoding techniques at image classification on the PASCAL VOC challenge[37].

Although Fisher Vectors are not as sparse as BoV histograms, they perform well with linear classifiers whereas BoV histograms usually require specialized kernels such as chi-square kernels to perform efficient classification. In our experiments we used the VLfeat implementation [119] of Fisher Vector encoding which is available for free download from www.vlfeat.org.

4.3 EXPERIMENTS

4.3.1 *Facial Expression Recognition*

Most appearance based methods for automatic facial expression analysis use static descriptors such as Gabor filters, Gaussian derivatives and Local Binary Patterns. However as Bassili in [13] pointed out, the knowledge of facial movement is integral to accurate facial expression recognition. It follows that we look at spatio-temporal texture and not just static spatial

textures. Extending our work from 3.3.2.3 to the temporal domain, we present a new descriptor for videos that provides us with spatio-temporal texture using a combination of LBP-TOP features and Gaussian derivatives. The descriptor is tested on the CK database and the results presented in 4.3.1.2.

4.3.1.1 *Related Work*

A number of systems have been developed for recognizing the six prototypical emotional expressions. Two recent surveys [39, 134] provide a comprehensive list of facial expression recognition systems. Here we will only provide an overview of systems that use facial motion and dynamic texture information for expression recognition.

Dense optical flow has been used by researchers on local regions of the face as well as holistically for extracting facial motion. Lien *et al.* in [73] tracked facial feature points on the face using a pixel-wise optical flow algorithm, reduced correlation in the data using PCA and used Hidden Markov Models to recognize the facial expressions. Mase and Pentland in [78], Otsuka and Ohya in [90] and Yoneyama *et al.* in [133] used optical flow to track specific regions of the face to estimate facial motion.

Other authors have used motion models for facial motion extraction. These too have been applied both holistically and locally. In [36] Essa and Pentland presented 3D motion and muscle models which they stabilised using a Kalman filter. Apart from tracking low-level features using optical flow, methods have been developed for tracking higher level facial features [72]. One major issue with high-level feature tracking is that initialization is not always easy and prone to failure.

Local Binary Patterns provide a simple and effective way to describe texture and extending them to the temporal domain in the form of LBP-TOP allows for the extraction of dynamic texture. Baltrušaitis *et al.* in [12] presented a multi-modal system for affect recognition using LBP-TOP features and audio features. The authors combined Continuous Conditional Random Fields (CCRF) with Support Vector Machines for Regression (SVR) for modeling continuous emotion in dimensional space helping them attain the best results on the Audio Visual Emotion Challenge (AVEC) 2012 database. The successful use of LBP-TOP features for affect recognition and our own success with LBP features combined with Gaussian derivatives motivated us to try to further improve LBP-TOP features by combining them with Gaussian derivatives.

4.3.1.2 *Procedure and Results*

We extended the technique presented in the previous chapter for combining Gaussian derivative features with Local Binary Patterns to the tem-

poral domain. The Half-Octave Gaussian pyramid was used to produce Gaussian derivative images for each frame in the video. First and second order derivative images were extracted from the base of the pyramid ($\sigma = 1$): $I_x, I_y, I_{xx}, I_{yy}, I_{xy}$. Subsequently derivative images of the same order were used to produce space-time volumes over which the LBP-TOP operator is applied. Each of these space-time volumes were divided into overlapping spatial regions as presented in 3.3.2.3. LBP-TOP histograms were calculated for each local spatio-temporal volume which were then concatenated to obtain the joint histogram. The final histogram for a video was obtained by concatenating the histograms for the space-time volume of each derivative order.

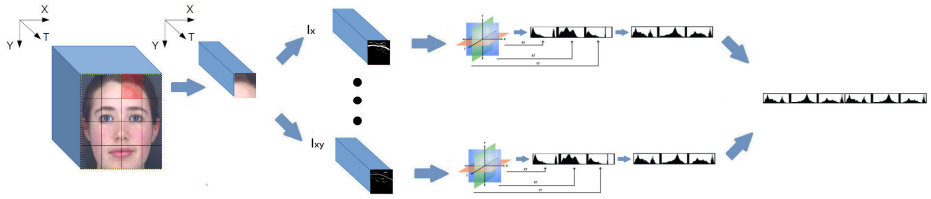


Figure 46: LBP-TOP computed over Gaussian derivative images

For the Cohn-Kanade database the optimum grid size was found to be 4 X 4 with an overlap of 43.75% with images containing the facial region being normalized to 66 X 66 pixels. Since we had 5 order of derivatives and 16 grids with each grid producing three LBP histogram (one for xy plane, one for yt plane and one for xt), we had a feature vector of: 5*16*3*59=14160 dimensions.

In terms of pre-processing of the database, only the first frame of the video was used to localize the facial region using the Viola-Jones face detector [120]. The same location of the face was used for subsequent frames in the video. No other pre-processing was performed in the experiments. Gaussian derivatives provide a degree of invariance to image rotation whereas LBP features are moderately invariant to illumination variations.

Emotion	**LBP-TOP calculated over Gaussian derivatives**	**LBP-TOP**
Happiness	98.14	96.69
Sadness	95.01	93.21
Surprise	100	100
Fear	86.33	85.03
Anger	84.91	83.77
Disgust	96.54	97.67
Total	94.56	92.72

Table 18: Our accuracy compared to the accuracy obtained using conventional LBP-TOP features

Note that the results obtained using standard LBP-TOP features are different from the reported results in [136] because of different experimental protocols adopted.

Given the high dimensionality of data a linear SVM was used for classification and the results (% accuracy) shown in table 18 are at par with the state-of-the-art. The results are however just marginally better than the results produced with LBP-TOP features. The increase in computational complexity and the increase in dimensionality of features by a factor of five while giving a marginal better accuracy does not justify the use of our descriptor.

4.3.2 *Audio Visual Emotion Challenge 2014 Depression sub-challenge*

Facial expressions, eye gaze and head motion are important visual features used by psychologists to gauge depression in patients. Advances in the field of computer vision allow us to automatically observe these visual features. However most research has focused on static images and posed facial expressions. Techniques that work well for posed emotions may not work well for spontaneous expressions [103]. In [11], the authors underline the importance of spatio-temporal information for affect sensing. In the work presented here, we extract spatio-temporal information using two different visual descriptors from videos of people with depression and use linear support vector machines to quantify the level of depression.

4.3.2.1 *AVEC 2014 data*

The AVEC 2014 database is a subset of the AVEC 2013 database. The AVEC 2013 database contains 150 videos of subjects interacting with a computer. The total number of subjects is 84 with their age ranging between 18 to 63 years. The mean age is 31.5 and the standard deviation is 12.3. The subjects were recorded between 1 to 4 times and the period between two recordings of the same subject was 2 weeks. One subject was recorded 4 times, 18 subjects were recorded three times, 31 subjects twice while the remaining 34 were only recorded once. The length of the recording varied from 20 minutes to 50 minutes (mean = 25 minutes). The total duration of the recordings was is approximately 240 hours.

For the AVEC 2013 database, human-computer interaction involved 14 tasks which were presented to the subject using MS PowerPoint. The AVEC 2014 subset only includes 2 of these tasks. The two tasks are provided as separate videos and thus the database contains 300 videos in total. These two tasks were selected as the tasks performed by the largest number of subjects. Although the AVEC 2014 is a subset of the AVEC 2013 database, 5 new pairs of videos were added to replace 5 previous pairs which were deemed unfit for the challenge.

The two tasks selected are:

(a) Northwind: The participants read aloud a section of the fable "Die Sonne und der Wind" in German.

(b) Freefrom: Participants answered one question out of a set of questions of the form: "What is your favorite dish?"; "Discuss a sad childhood memory" again in German.

Although the original recordings were made at a variable sampling rate, the final audio was produced by resampling the originals to a 128 Kbps bitrate using the AAC codec. Similarly the videos were also recorded at different sampling rates but were finally resampled to 30 fps with a resolution of 640 X 480 pixels using the H.264 codec. Finally the Audio- Video recordings were packaged in a mp4 container.

The database is split into three partitions: training, development and test; each split containing 50 Northwind-Freeform video pairs. The data was split assuring that the partitions have similar distributions of age, gender and depression levels. There are no session overlaps between partitions i.e. multiple recordings from the same original clip will stay in the same partition. Labels are available only for the training and development sets.

4.3.2.2 *Related Work*

In [44], the authors looked at the change over time in severity of depression and facial expressions. Facial actions units were coded both manu-

ally and automatically and high agreement rates were achieved between the two approaches. It was found that facial expressions were consistent with the "social risk hypothesis" which states that patients with depression tend to withdraw from society. When the depression is severe, patients' facial expressions are more likely to be associated with contempt while the frequency of smiling is reduced. When the patients feel better, they display social signals indicating their disposition to associate. This work validated the use of automated facial expression analysis for behavioral science and heralded the use of automatic facial expression analysis in clinical science.

Scherer *et al.* [100] recognized vertical head gaze, vertical eye gaze, smile intensity and smile duration as important nonverbal behaviors for sensing psychological disorders such as social anxiety and depression. They employed a multimodal sensor framework called Multisense which included a face tracker, a head tracker, a system for observing eye gaze and a Microsoft kinect sensor for skeleton tracking and audio capture. They discovered that people with depression generally have a downward angle of gaze as compared to non-depressed people. It was also found that depressed people have lower intensity smiles and have shorter duration smiles, on average. The authors identified statistical differences in the nonverbal behavior of patients with depression and anxiety. These findings suggest that head pose and facial expressions are important visual cues for depression and related psychological disorders. However the authors were not able to find a suitable visual descriptor for capturing information from the fidgeting motion of patients.

In [21], the authors used FACS and Active Appearance Models (AAM) to distinguish between depressed and non-depressed subjects but they did not extend their framework to assess the level of depression. Without using subject specific AAM as in [21] and [81], Joshi *et al.* in [62] used LBP-TOP and STIP features in conjunction with Bag of words (BoW) encoding to detect depression. They experimented with a variety of feature fusion techniques and combine audio features such as loudness, pitch, intensity and Mel-frequency cepstral coefficients (MFCC) to develop a multimodal depression sensing system.

The best performing system for depression estimation at AVEC 2014 was developed by Williamson *et al.* [128]. They developed a multimodal system where high-level features are extracted from low-level features which on the other hand are extracted from different modalities such as speech prosody and facial action unit activations. In [104] the authors use a two-model regression framework for depression level estimation using baseline video features and short-term acoustic features mapped to an i-vector space. Kächele et al. in [64] infer depression intensity using only meta-knowledge. This meta-knowledge is in the form of features such as

length of the video sequence, gender of the subject, semantic content of the video among others.

4.3.2.3 *Experiments and results*

The labels for the training and development set of AVEC 2014 are available to participants. To obtain the errors over the test set results are mailed to the organisers in order . Participants get 5 attempts to test their results on the test set. Each partition contains 50 Beck Depression Index-II labels. Each label corresponds to a pair of videos, Freeform and Northwind.

To extract the visual information from the videos we computed LBP-TOP features to capture the intra-face movement and dense trajectories to capture macro movements such as those of the head and the shoulders.

We split the videos into individual frames and performed face detection and alignment using Openimaj[47]. Openimaj normalized the detected face into an imagette of 80 X 80 pixels. Zhao and Pietikäinen [136] demonstrated that it is best to divide the imagette into overlapping spatial regions and calculate the LBP-TOP features separately for each spatial region over a time slice and finally concatenate the results from the different regions and time slices. This technique helps encode the occurrence of micro-patterns and their relative locations in the image.

Using cross-validation, it was seen that for an imagette of our size, spatial regions of 10 X 10 pixels work best with a 50% overlap. We computed LBP-TOP features for 2 different sizes of the temporal slices, t = 3s and t = 1s.

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the LBP-TOP feature vector and decorrelate the features. This reduces the computation time and also reduces the size of the Fisher Vectors as this is linearly dependent on the feature vector size [88]. We chose a dimensionality of D = 64, assuring that the variance in the projected data is at least 95% of the original data. A Gaussian mixture model was fit over a subset of reduced-dimensionality training features which was used to create one Fisher vector per video. The optimum number of clusters was chosen using cross-validation over the development set.

For dense trajectories we used the following settings: length of the trajectory was set to 15 frames, the stride for dense sampling feature points was set to 5 pixels and the neighborhood size for computing the descriptor was set to 32 pixels. A set of features (HOG, HOF and MBH) over the dense trajectories was generated for each video. Just as with the LBP-TOP features, PCA was performed followed by fitting of a GMM over a subset of projected trajectory features from the training set. Millions of trajectories were generated for the training set alone, a subset of $3.6X10^5$ was used to fit the GMM. The fitted model was used to generate a Fisher vector for each video.

The low level descriptors (LLD) audio features provided with the AVEC 2014 database were reduced in dimensionality using PCA (D = 64) and a GMM was fit over the projected features followed by Fisher vector generation.

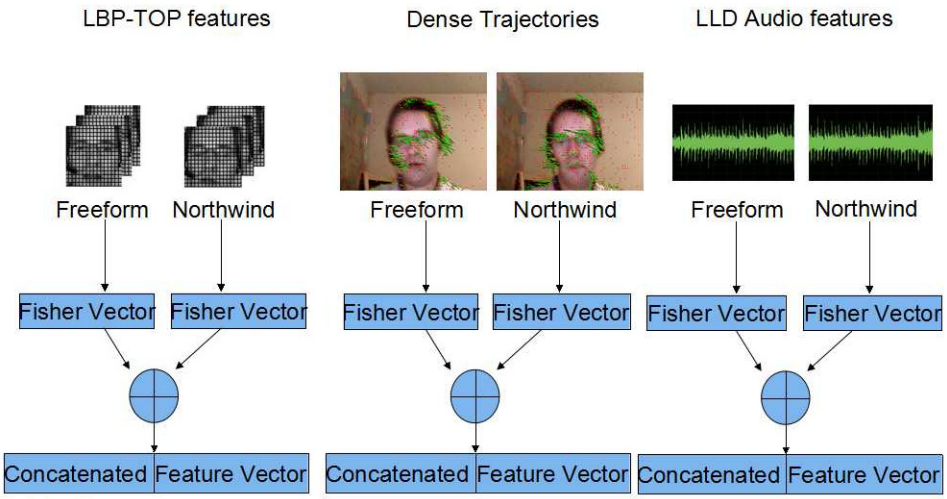


Figure 47: System Architecture

The LBP-TOP, dense trajectory and audio features having been transformed into Fisher vectors were concatenated for each pair of videos, Freeform and Northwind, and fed as input to a linear support vector machine (SVM). Linear SVM was chosen because in the feature matrix, the number of columns, was much more than the number of rows. A feature vector of D = 64 produces a FV of 4096 columns, for each pair of videos there will therefore be 2*4096=8192 columns whereas the number of samples in the training set is just 50. Figure 48 shows how the optimum number of clusters was chosen for fitting the GMM on LBP-TOP features. The minimum development error was achieved at 35 clusters; a similar analysis gave us a minima at 40 for dense trajectories and at 50 for LLD audio features.

	LBP-TOP	LLD	Dense Trajectories	LBP-TOP+Dense Trajectories	LBP-TOP+LLD	LLD+Dense Trajectories	LBP-TOP+LLD+Dense Trajectories
MAE	6.9697	9.7457	9.8668	6.9679	6.9662	9.5229	**6.9643**
RMSE	8.1674	11.514	11.7985	8.1647	8.1645	11.2538	**8.1618**

Table 19: Errors for different combinations of descriptors on the development set

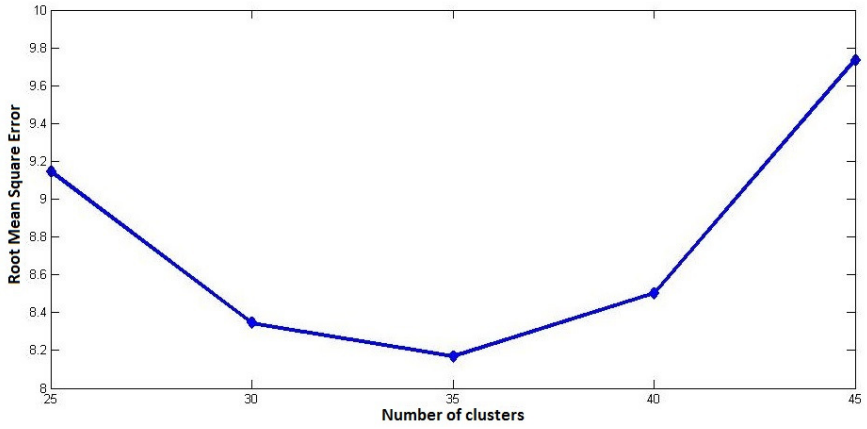


Figure 48: Root Mean Square Error (RMSE) vs. number of clusters

We compare our results obtained using Fisher vector encoding with results produced using sparse coding in table 20 for LBP-TOP features on the development set.

Encoding Technique	Fisher Vector Encoding	Sparse Coding
MAE	6.9697	10.1785
RMSE	8.1674	11.9858

Table 20: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for different encoding techniques

For sparse coding, we varied the dictionary size from 250-750 and the minima was attained at 550. We used max pooling in the final encoding step which has been shown to perform better than average pooling.

Window Size	1 second	3 seconds
MAE	7.5520	6.9697
RMSE	8.9025	8.1674

Table 21: Errors for different sizes of time slice

It can be seen in table 21 that a time slice of 3 seconds works better than a slice of 1 second on the development set.

In table 19 we see that the minimum errors are obtained by combining all the three descriptors: LBP-TOP, dense trajectories and LLD audio features. However using LBP-TOP features alone and encoding them using Fisher Vectors, we achieved error values very close to the error values attained by the combination of all three features. Given the computational effort required to generate and encode dense trajectories and LLD features, we opted to use LBP-TOP features alone for our results on the test set.

We only tested the early fusion technique because we only have 50 samples for training and another 50 for development. In case late fusion is performed, we would need two layers of regressors with the output of one layer forming the input for the second and the training data of just fifty samples getting split between the two layers.

Finally we compare our errors on the development and testing set with the baseline in table 22.

	Development Set		Test Set	
	Our Method	Baseline	Our Method	Baseline
MAE	6.9697	-	**8.3988**	8.857
RMSE	**8.1674**	9.26	**10.2491**	10.859

Table 22: Comparison of errors with baseline

Our method performs better than the baseline method. It is worth noting that these results are produced using LBP-TOP features alone combined with Fisher Vector encoding. The baseline [116] uses LGBP-TOP features [10] which are LBP-TOP features calculated over several orders of Gabor images; LBP-TOP features are therefore computationally simpler to compute. In the baseline draft paper it is not mentioned how they encode the visual information to obtain a unique feature vector corresponding to each label hence we cannot compare the computational efficiency of our system with the baseline.

4.4 CONCLUSION

A dynamic appearance based descriptor is presented which performs at par with the state-of-the-art but arguably costs more to compute than the standard LBP-TOP feature while giving just a meagre improvement in accuracy.

We then present a multimodal system for automated depression evaluation. It allows to quantitatively estimate the likelihood of depression using visual features.

Our experiments show that dense trajectories and LLD features do not significantly improve the results when combined with LBP-TOP features. It is seen that LBP-TOP features alone combined with Fisher Vector encoding are enough to beat the baseline.

We believe that this novel framework for depression assessment can be easily extended for predicting other slowly changing labels such as mood.

5

CONCLUSION

In this thesis we have presented methods for inferring affect from visual information. In most of our experiments this information has been extracted from the facial region in images and videos. Computer vision techniques for capturing the global appearance of image regions have been employed. A common architecture has been presented that can be used for affect recognition and ancillary tasks such as head pose estimation and face recognition.

Chapter 2 provides a discussion on the different notions and concepts related to emotion. The chapter begins with a brief discussion on the early theories on emotion. Early psychologists working on emotions reckoned that emotional experience is secondary to physiological changes. Most modern researchers today do not agree with these early theories on emotions. Two popular representations for emotions that are currently popular among the affective community have been discussed in this chapter. The belief that emotions are discrete and some emotions are more "basic" than others has been propounded by one school of thought. Another group of psychologists believes that emotions can be represented in a multi-dimensional space and that all emotions are generated by the same neurological system and processes and therefore some emotions cannot be more primary than others. Plutchik proposed an emotion representation which is analogous to the HSV (hue-saturation-value) color model. He suggested that emotions can mix like colors and some basic emotions mix to provide complex emotions. We also discuss a popular taxonomy of facial movements and how Ekman's basic emotions are manifested in facial expressions. We believe that as far as the human-computer interaction and affective computing communities are concerned, the question that "which form emotion representation is more correct ?" is not as relevant as the question that "which form of emotion representation is more appropriate for our application ?".

Chapter 2 also includes a discussion on mood. We define mood as an affective state more diffuse and longer lasting that emotions. Picard's model that establishes a link between mood and emotional response is also described in the section on mood. The final section is dedicated to depression and the inventory currently employed by psychologists to quantify depression.

In chapter 3 we present the techniques developed for affect sensing from static image data. Our focus has been to develop methods which

utilize global image appearance without using any techniques for tracking keypoints on the face or fitting a model over it. We present a common architecture for head pose estimation, smile detection and affect sensing. We employ multiscale Gaussian derivatives for capturing the global image appearance. Using a combination of Gaussian derivatives, principal component analysis for dimensionality reduction and support vector machines for discrimination we obtain state-of-the-art results. We also devise a new image description method which involves the calculation of Local Binary Pattern (LBP) features over Gaussian derivative images. This descriptor is tested for two different applications: smile detection and face recognition. In the case of face recognition we use a simple 1-nearest neighbor method for classification. We outperform the state of the art at smile detection and our results are at par with the state-of-the-art for face recognition.

In chapter 4 we continue with our philosophy of using global image appearance for affect sensing. We present the techniques developed for sensing affect from video data. As with the experiments described in chapter 3, we do not use any facial keypoint detection or tracking methods and neither do we use any model fitting techniques such as Active Appearance Models (AAM) or Active Shape Models (ASM) in the experiments described in chapter 4. Local Binary Patterns calculated on Three Orthogonal Planes (LBP-TOP) are an efficient and computationally simple way to capture texture from video data. Motivated by our success in face recognition and smile detection using a combination of Gaussian derivatives and Local Binary Patterns, we combined Gaussian derivatives with LBP-TOP features to produce a new descriptor for capturing dynamic texture. The descriptor developed is tested on video sequences from the Cohn-Kanade database to recognize the six basic emotions. Although our results are marginally better than the results obtained using the standard LBP-TOP features, the dimensionality of the feature vector of our descriptor is five times the size of the feature vector obtained using LBP-TOP features. In chapter 4 we also develop a technique to estimate severity of depression in human subjects using audio-visual information. We use two descriptors to extract information from the visual component: dense trajectories to capture macro level movements and LBP-TOP features to capture intra-facial movements while precomputed Low Level Descriptors (LLD) features capture the audio information. Since the videos in the Audio Visual Emotion Challenge (AVEC) 2014 database are of different sizes, the audio-visual information extracted is also of different sizes. We employed state-of-the-art encoding techniques to transform the audio-visual information extracted from the videos into a global signature for the video. This signature allows us to use support vector regression for estimating the severity of depression. This multi-modal system developed outperforms the base-

line and can be used for the estimation of other slowly changing affective states such as mood.

5.1 LESSONS LEARNED

Head pose estimation

The Pointing04 database contains discrete head poses. Support Vector Machine (SVM) classifiers were used to discriminate between the different poses however when we used SVM regressors trained over the same data, a high correlation between the predicted values and ground truth was obtained. These results suggest that head pose estimation can be performed on continuous poses even with a training data exclusively containing discrete poses.

Smile detection

The SVM classifier trained on the GENKI-4K database produced probability estimates for each prediction. When this classifier was used on image sequences from the Cohn-Kanade database, it turned out that the probability estimates were indicators of smile intensity.

Gaussian derivatives and Illumination

While working on smile detection it was discovered that faces which are partially illuminated were often misclassified. Calculating Local Binary Pattern features over Gaussian derivative images leads to higher invariance to illumination issues. It was also found that the combination of LBP features and Gaussian derivatives is more invariant to pose than LBP features alone.

Minkowski Distance

While performing the experiments on face recognition it was discovered that the L_1 and L_2 norms may not always be the optimum distance metrics and sometimes the optimum metric is L_p with p as a variable. In the experiments on the YaleB database p was a value between 0.25 and 2.

LBP-TOP features calculated over Gaussian derivatives

The combination of LBP-TOP features and Gaussian derivatives provides marginally better results than LBP-TOP features but produce a feature

vector five times the size of the feature vector obtained with LBP-TOP. This effectively leads to a reduction in prediction speed by a factor of five.

Early fusion of features in depression estimation

In our experiments on depression estimation, we concatenated the fisher vector encodings of LBP-TOP features, dense trajectories and LLD audio features. It was found that the accuracy obtained using the LBP-TOP features alone is only marginally lower than the accuracy obtained using all three features. Since our database was of a limited size and late fusion of features would have entailed splitting the training data into two smaller sections for training two layers of regressors, we could only test the early fusion approach.

5.2 IMPACT OF THE PRESENT WORK

We have presented an architecture than can be, without much change, used for diverse applications such as head pose estimation, smile detection and affect sensing from static image data. Since we only use global image appearance, this architecture is easily adaptable to mobile systems and other devices with limited computing power.

The combination of Gaussian derivatives and LBP features outperforms state of the art for smile detection without using any form of keypoint tracking or facial model fitting methods and the results are are at par with the state of the art for face recognition. It is worthwhile to note that our descriptor competes with descriptors developed explicitly for face recognition which can handle complex lighting conditions.

We have shown that LBP-TOP features encoded with Fisher Vectors provide a unique signature to videos of uniform length even if the videos are of different durations. This has allowed us to work with audio-visual data where a single label is assigned to a whole video sequence. For example, if we have samples of a physiological signal, correlated with a slowly varying affective state such as mood, of different time durations and features are extracted from these samples, the quantity of features obtained from these two samples may be different. In order to compare the two samples one could employ Fisher Vector encoding to obtain a sample signature of uniform length as presented in our experiments.

5.3 FUTURE SCOPE OF THIS WORK

The AVEC 2014 database used for depression estimation only contains 150 samples, 50 each for training, validation and testing. These limited number of samples did not allow us to try different feature fusion techniques. A larger database would allow the testing of the late fusion scheme.

The descriptor developed in chapter 4 by the combination of Gaussian derivatives and LBP-TOP features performed better than standard LBP-TOP features but the feature vectors suffered from high dimensionality. Feature selection techniques could be explored to achieve results better than standard LBP-TOP features while keeping the dimensionality low.

Although we presented a novel technique for head-pose estimation in the present work, we have not used the information for inferring the affective state. There is evidence to suggest that head-pose and head movements are important modalities for affect sensing [46].

A logical extension to the work presented in this thesis is the development of a multimodal system for affect sensing. A single modality may not represent the underlying emotional state. However if a multimodal system has access to biosignals such as heart rate and glavanic skin response among others, it could be able to *recognize* the true affective state of the person instead of just *inferring* it.

A

APPENDIX

A.1 SUMMARY OF DATABASES

Database	Summary
Pointing04 [45]	Consists of 15 sets of images. Each set contains of 2 series of 93 images of the same person at different poses. There are 15 people in the database, with/without glasses and varying skin colors. The pose, or head orientation is determined by 2 angles which vary from -90 degrees to +90 degrees.
CMU-PIE [109]	Consists of 41,368 images of 68 people in 13 different poses, 43 different illumination conditions, and with 4 different expressions.
Genki-4K [127]	Images contain faces spanning a wide range of illumination conditions, geographical locations, personal identity, and ethnicity. Contains 4000 face images labeled as either âsmilingâ or ânon-smilingâ by human coders. The pose of the faces is approximately frontal as determined by an automatic face detector.
Cohn-Kanade [65]	Includes 486 sequences from 97 posers. Each sequence begins with a neutral expression and progresses to a peak expression. The peak expression for each sequence is fully FACS [35] coded and given an emotion label. The emotion label refers to what expression was requested from the subject rather than what may have been performed.
FEED [121]	Contains facial images of subjects experiencing the six basic emotions as defined by Ekman [33]. The database contains image sequences from 18 subjects. Each emotion is elicited three times from every subject.
Yale B [43]	The extended YaleB dataset contains images from 28 individuals captured under 64 different lighting conditions with 9 pose views.

AVEC-2014 [116]	The AVEC 2014 database is a subset of the AVEC 2013 database. The AVEC 2013 database contains 150 videos of 84 subjects interacting with a computer. The subjects were recorded between 1 to 4 times and the period between two recordings of the same subject was 2 weeks. The AVEC 2014 subset includes two tasks. The two tasks are provided as separate videos and thus the database contains 300 videos in total. Although the AVEC 2014 is a subset of the AVEC 2013 database, 5 new pairs of videos were added to replace 5 previous pairs which were deemed unfit for the challenge. The database is split into three partitions: training, development and test; each split containing 50 video pairs. Labels are available only for the training and development sets.

BIBLIOGRAPHY

[1] A clustered scatter plot, . URL http://mnemstudio.org/clustering-k-means-introduction.htm. Accessed: 17/01/2015.

[2] Example of an lbp based facial representation, . URL http://goo.gl/EnzPYh. Accessed: 3/10/2014.

[3] Plutchik's wheel of emotions, . URL http://ryanwm.com/thesis/lit_review/emotion.php. Accessed: 12/12/2014.

[4] Plutchik's cone of emotions, . URL http://www.talentdifferent.com/emotions-2146.html. Accessed: 12/12/2014.

[5] SVM classifying linearly separable data, . URL http://opticalengineering.spiedigitallibrary.org/article.aspx?articleid=1653966. Accessed: 15/01/2015.

[6] SVM with data mapped to a higher dimensional space using kernel, . URL http://www.statsoft.com/textbook/support-vector-machines. Accessed: 15/01/2015.

[7] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, pages 113–128, 2002. URL http://cogcomp.cs.illinois.edu/papers/AgarwalRo02.pdf.

[8] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *ECCV (1)*, pages 469–481, 2004.

[9] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol. Face recognition using hog-ebgm. *Pattern Recognition Letters*, 29(10): 1537–1543, 2008.

[10] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *ACII*, pages 356–361, 2013.

[11] Z. Ambadar, J. Schooler, and J. F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting

subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[12] T. Baltrušaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[13] J. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37(11):2049–2058, 1979.

[14] A. Battocchi, F. Pianesi, and D. Goren-Bar. Dafex: Database of facial expressions. In *Intelligent Technologies for Interactive Entertainment, First International Conference, INTETAIN 2005, Madonna di Campiglio, Italy, November 30 - December 2, 2005, Proceedings*, pages 303–306, 2005.

[15] A. Beck. *Depression: Causes and Treatment*. University of Pennsylvania Press, Philadelphia, 1972.

[16] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[17] M. M. Bradley and P. Lang. Measuring Emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[18] W. Cannon. The James-Lange theory of emotion: A critical examination and an alternative theory. *American Journal of Psychology*, 39:106–124, 1927.

[19] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, pages 1–12, 2011.

[20] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang. Total variation models for variable lighting face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1519–1524, 2006.

[21] J. F. Cohn, T. S. Kruez, I. A. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre. Detecting depression from facial actions and vocal prosody. In *ACII*, pages 1–7, 2009.

[22] T. F. Cootes and C. J. Taylor. Active shape models - 'smart snakes'. In *BMVC92*, pages 266–275. Springer London, 1992.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Computer Vision - ECCV'98*, pages 484–498. Springer Berlin Heidelberg, 1998.

[24] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[25] J. L. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 640–645. IEEE, 1997.

[26] J. L. Crowley and A. C. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):156–170, 1984.

[27] J. L. Crowley, O. Riff, and J. H. Piater. Fast computation of characteristic scale using a half-octave pyramid. In *Scale Space 03: 4th International Conference on Scale-Space theories in Computer Vision, Isle of Skye*, 2002.

[28] M. Dahmane and J. Meunier. Continuous emotion recognition using gabor energy filters. In *Affective Computing and Intelligent Interaction 2011*, pages 351–358, 2011.

[29] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[30] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer Berlin Heidelberg, 2006.

[31] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A (JOSA A)*, 2:1160–1169, July 1985.

[32] O. Déniz, M. C. Santana, J. Lorenzo-Navarro, L. Antón-Canalís, and G. Bueno. Smile detection for user interfaces. In *ISVC (2)*, pages 602–611, 2008.

[33] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-â4):169–200, 1992.

[34] P. Ekman and W. Friesen. *Facial Action Coding System:A Technique for the Measurement of Facial Movement*. Palo Alto, 1978.

[35] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System: The Manual on CD ROM*.

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

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[38] S. R. Fanello, I. Gori, G. Metta, and F. Odone. Keep it simple and sparse: Real-time action recognition. *J. Mach. Learn. Res.*, 14(1): 2617–2640, Jan. 2013. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2567709.2567745.

[39] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *PATTERN RECOGNITION*, 36(1):259–275, 1999.

[40] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[41] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[42] A. Gee and R. Cipolla. Fast visual tracking by temporal census. *Image and Vision Computing*, 14(2):105–114, 1996.

[43] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, June 2001. ISSN 0162-8828. doi: 10.1109/34.927464. URL http://dx.doi.org/10.1109/34.927464.

[44] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, pages 1–8, 2013.

[45] N. Gourier, D. Hall, and J. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proceedings of POINTING'04 International Workshop on Visual Observation of Deictic Gestures*, 2004.

[46] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In J. Or, editor, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218. I-Tech Education and Publishing, Vienna, Austria, 2008.

[47] J. Hare, S. Samangooei, and D. Dupplaw. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *ACM Multimedia 2011*, pages 691–694. ACM, November 2011. URL http://eprints.soton.ac.uk/273040/. Event Dates: 28/11/2011 until 1/12/2011.

[48] M. Heikkilä and M. Pietikäinen. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):657–662, 2006. doi: 10.1109/TPAMI.2006.68. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.68.

[49] M. Heikkilä, M. Pietikäinen, and J. Heikkilä. A texture-based method for detecting moving objects. In *British Machine Vision Conference, BMVC 2004, Kingston, UK, September 7-9, 2004. Proceedings*, pages 1–10, 2004. doi: 10.5244/C.18.21. URL http://dx.doi.org/10.5244/C.18.21.

[50] J. A. R. Hernandez, J. Crowley, and A. Lux. " how old are you?" a possible answer using tensors of binary gaussian receptive maps. In *British Machine Vision Conference*, 2010.

[51] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 242–247, 1996.

[52] http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html. Head Pose Image Database.

[53] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *CW*, pages 437–444, 2005.

[54] V. Jain and J. Crowley. Smile detection using multi-scale gaussian derivatives. In *12th WSEAS International Conference on Signal Processing, Robotics and Automation*, 2013.

[55] V. Jain and J. L. Crowley. Head pose estimation using multi-scale gaussian derivatives. In *Image Analysis*, pages 319–328. Springer, 2013.

[56] V. Jain, J. Crowley, and A. Lux. Facial expression analysis and the pad space. In *PRIA-11-2013 Pattern Recognition and Image Analysis*, pages 579–582, 2013.

[57] V. Jain, J. Crowley, and A. Lux. Local binary patterns calculated over gaussian derivative images. In *22nd International Conference on Pattern Recognition*, pages 3987–3992. IEEE Computer Society, 2014.

[58] V. Jain, J. L. Crowley, A. Dey, and A. Lux. Depression estimation using audiovisual features and fisher vector encoding. In *4th International Workshop on Audio/Visual Emotion Challenge*, pages 87–91, 2014.

[59] W. James. What is an emotion. *Mind*, 9:188–205, 1884.

[60] P. N. Johnson-Laird and K. Oatley. The language of emotions: An analysis of a semantic field. *Cognition and emotion*, 3(2):81–123, 1989.

[61] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2002.

[62] J. Joshi, R. Goecke, A. Dhall, S. Alghowinem, M. Wagner, M. Breakspear, J. Epps, and G. Parker. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on MultiModal User Interfaces*, 7(3):217–228, 2013.

[63] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610 Vol. 1, Oct 2005. doi: 10.1109/ICCV.2005.66.

[64] M. Kächele, M. Schels, and F. Schwenker. Inferring depression and affect from application dependent meta knowledge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 41–48, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3119-7. doi: 10.1145/2661806.2661813. URL http://doi.acm.org/10.1145/2661806.2661813.

[65] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, March 2000.

[66] V. Kellokumpu, G. Zhao, and M. Pietikäinen. Recognition of human actions using texture descriptors. *Mach. Vis. Appl.*, 22(5): 767–780, 2011. doi: 10.1007/s00138-009-0233-8. URL http://dx.doi.org/10.1007/s00138-009-0233-8.

[67] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological cybernetics*, 55(6): 367–375, 1987.

[68] U. Kowalik, T. Aoki, and H. Yasuda. Broaference - a next generation multimedia terminal providing direct feedback on audience's satisfaction level. In *INTERACT*, pages 974–977, 2005.

[69] C. Lange. Om Sindsbevoegelser: Et psykofysiologiske Studie. *Kopenhagen: Kronar*, 1885.

[70] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[71] B. Li and H. Yin. Face recognition using rbf neural networks and wavelet transform. In *ISNN (2)*, pages 105–111, 2005.

[72] Y. li Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):97–115, 2001.

[73] J. Lien. *Automatic recognition of facial expression using hidden Markov models and estimation of expression intensity*. PhD thesis, The Robotics Institute, CMU, 1998.

[74] G. Littlewort, J. Whitehill, T. W. I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition*, pages 298–305, 2011.

[75] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, 1999.

[76] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *3rd International*

Conference on Face & Gesture Recognition (FG '98), April 14-16, 1998, Nara, Japan, pages 200–205, 1998.

[77] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, page 87, 2009.

[78] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions on Information and Systems*, 74(10): 3474–3483, 1991.

[79] D. McDuff, R. E. Kaliouby, D. Demirdjian, and R. W. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, pages 1–7, 2013.

[80] D. N. McIntosh. Facial feedback hypotheses: Evidence, implications, and directions. *Motivation and Emotion*, 20(2): 121–147, 1996.

[81] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear. An approach for automatically measuring facial activity in depressed subjects. In *ACII*, pages 1–8, 2009.

[82] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[83] J. R. Movellan. Tutorial on Gabor Filters. Available at `mplab.ucsd.edu/wordpress/tutorials/gabor.pdf`.

[84] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[85] S. Niyogi and W. Freeman. Example-based head tracking. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 374–378, 1996.

[86] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[87] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7): 971–987, 2002.

[88] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, pages 1817–1824, 2013.

[89] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–331, 1990.

[90] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using hmm. In *FG*, pages 442–447. IEEE Computer Society, 1998.

[91] R. Picard. *Affective Computing*. The MIT Press, Cambridge, MA, 1997.

[92] R. Plutchik. *Emotion: a psychoevolutionary synthesis*. Harper & Row, New York, 1980.

[93] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. 2009.

[94] R. Reisenzein. Wundt's three-dimensional theory of emotion. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 75:219–250, 2000.

[95] J. Ruiz-Hernandez, A. Lux, and J. Crowley. Face detection by cascade of gaussian derivates classifiers calculated with a half-octave pyramid. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–6, Sept 2008. doi: 10.1109/AFGR.2008.4813457.

[96] J. A. Ruiz-Hernandez, J. L. Crowley, A. Méler, and A. Lux. Face recognition using tensors of census transform histograms from gaussian features maps. In *BMVC*, pages 1–11, 2009.

[97] J. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[98] J. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(11):273–294, 1977.

[99] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[100] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *FG*, pages 1–8, 2013.

[101] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[102] H. Schlosberg. Three dimensions of emotion. 1954.

[103] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image Vision Comput.*, pages 1856–1863, 2007.

[104] M. Senoussaoui, M. Sarria-Paja, J. a. F. Santos, and T. H. Falk. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 57–63, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3119-7. doi: 10.1145/2661806.2661819. URL http://doi.acm.org/10.1145/2661806.2661819.

[105] C. Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4): 431–437, 2012.

[106] C. Shan. Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing*, 21(1):431–436, 2012.

[107] Y.-s. Shin. Recognizing facial expressions with pca and ica onto dimension of the emotion. In *Proceedings of the 2006 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR'06/SPR'06, pages 916–922, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-37236-9, 978-3-540-37236-3.

[108] Y. Shinohara and N. Otsu. Facial expression recognition using fisher weight maps. In *FGR*, pages 499–504, 2004.

[109] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–51, 2002.

[110] A. Sloman. Prolegomena to a theory of communication and affect. In A. Ortony, J. Slack, and O. Stock, editors, *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, pages 229–260. Springer, Berlin, Heidelberg, 1992.

[111] R. Stiefelhagen. Estimating head pose with neural networks - Results on the pointing04 icpr workshop evaluation data. In

Proceedings of ICPR Workshop Visual Observation of Deictic Gestures, 2004.

[112] X. Tan, S. Chen, Z. Zhou, and F. Zhang. Face recognition from a single image per person: a survey. 2006.

[113] J. S. Tiezheng Ge (USTC), Qifa Ke (Microsoft). Sparse-coded features for image retrieval. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.

[114] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010.

[115] G. Valenza, P. Allegrini, A. Lanatà, and E. P. Scilingo. Dominant lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation. *Frontiers in Neuroengineering*, 5(3), 2012. ISSN 1662-6443. doi: 10.3389/fneng.2012.00003. URL http://www.frontiersin.org/neuroengineering/10.3389/fneng.2012.00003/abstract.

[116] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 - 3D dimensional affect and depression recognition challenge. In *4th ACM international workshop on Audio/visual emotion challenge*, 2014.

[117] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, pages 162–170, New York, NY, USA, 2006. ACM. ISBN 1-59593-541-X.

[118] V. Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998.

[119] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874249. URL http://doi.acm.org/10.1145/1873951.1874249.

[120] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 20(17):137–154, 2004.

[121] F. Wallhoff. Facial expressions and emotion database. http://www.mmk.ei.tum.de/w̃af/fgnet/feedtum.html, Technical University of Munich, 2006.

[122] B. Wang, W. Li, W. Yang, and Q. Liao. Illumination normalization based on weber's law with application to face recognition. *IEEE Signal Process. Lett.*, 18(8):462–465, 2011.

[123] H. Wang, S. Z. Li, and Y. Wang. Generalized quotient image. In *CVPR (2)*, pages 498–505, 2004.

[124] H. Wang, S. Z. Li, Y. Wang, and J. Zhang. Self quotient image for face recognition. In *ICIP*, pages 1397–1400, 2004.

[125] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011. URL http://hal.inria.fr/inria-00583818/en.

[126] J.-G. Wang and E. Sung. Em enhancement of 3d head pose estimated by point at infinity. *Image and Vision Computing*, 25 (12):1864–1874, 2007.

[127] J. Whitehill, G. Littlewort, I. R. Fasel, M. S. Bartlett, and J. R. Movellan. Toward practical smile detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):2106–2111, 2009.

[128] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 65–72, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3119-7. doi: 10.1145/2661806.2661809. URL http://doi.acm.org/10.1145/2661806.2661809.

[129] S. A. Winder and M. Brown. Learning local image descriptors. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[130] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In L. C. Jain, U. Halici, I. Hayashi, and S. B. Lee, editors, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, chapter 11, pages 355–396. CRC Press, 1999. ISBN 0-8493-2055-0.

[131] X. Xie, W.-S. Zheng, J.-H. Lai, P. C. Yuen, and C. Y. Suen. Normalization of face illumination based on large-and small-scale features. *IEEE Transactions on Image Processing*, 20(7): 1807–1821, 2011.

[132] J. J. Yokono and T. Poggio. Oriented filters for object recognition: an empirical study. In *FGR*, pages 755–760, 2004.

[133] M. Yoneyama, Y. Iwano, A. Ohtake, and K. Shirai. Facial expressions recognition using discrete hopfield neural network. In *ICIP (1)*, pages 117–120, 1997.

[134] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[135] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu. Face recognition under varying illumination using gradientfaces. *IEEE Transactions on Image Processing*, 18(11):2599–2606, 2009.

[136] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007.

[137] G. Zhao, M. Barnard, and M. Pietikäinen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11 (7):1254–1265, 2009. doi: 10.1109/TMM.2009.2030637. URL http://dx.doi.org/10.1109/TMM.2009.2030637.

[138] J.-Y. Zhu, W.-S. Zheng, and J.-H. Lai. Logarithm gradient histogram: A general illumination invariant descriptor for face recognition. In *FG*, pages 1–8, 2013.