

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

What Can Spontaneous Facial Expression Tell Us?

Permalink

<https://escholarship.org/uc/item/7jh3926h>

Author

Yang, Songfan

Publication Date

2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

What Can Spontaneous Facial Expressions Tell Us?

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Songfan Yang

December 2014

Dissertation Committee:

Dr. Bir Bhanu , Chairperson
Dr. Yingbo Hua
Dr. Subir Ghosh

The Dissertation of Songfan Yang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This thesis would not have been possible without the help of a great number of people. I would like to first thank my advisor, Dr. Bir Bhanu, for his support throughout my entire Ph.D. era. I am fortunate to have him as my advisor in a special time of my life. I would also like to thank my thesis committee member, Dr. Yingbo Hua and Dr. Subir Ghosh, for their help in improving the quality of the thesis. I am thankful for Dr. Anastasios Mourikis for his advice and help.

I am very grateful to my colleagues and friends at UC Riverside. Many thanks to my labmates, Le An, Mehran Kafai, Zhixing Jin, Yiming Li, Xiaojing Chen, Yu Sun, Ninad Thakoor, and etc. for all the great experience we have shared together. I would like to express my deepest memory for a dear friend and colleague of mine, Suresh Kumar, who had influenced me with positives. May him rest in peace in heaven. Many thanks to my friends at UCR, Jun Wang, Amy Zhang, Mingyang Li, Yiming Chen, Shengyang Xu, for all the happiness they have brought to my life.

I am eternally grateful to my family. I am thankful for my parents, Jian Yang and Xiaohua Yu, for their life-long love and continue support. Every single part of me is a masterpiece of them. I am exceptionally grateful to my wife, QQ, and my daughter, Brandy. I wish to share every heartbeat of my life with you. I would like to express my gratitude for uncle John and auntie Wen for the nest they have weaved for me and my family in the America. I would like to thank my parent-in-law, Jianrong Mo and Gaohua Chen, for the numerous care they have provided for my family, especially my daughter.

Portion of this dissertation has been published in “Understanding discrete facial expressions in video using an emotion avatar image” by Yang et al. ©The IEEE.

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

What Can Spontaneous Facial Expressions Tell Us?

by

Songfan Yang

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2014
Dr. Bir Bhanu , Chairperson

Facial expression plays a significant role in human communication. It is considered the single most important cue in the psychology of emotion [93]. Facial expression is taken as a universally understood signal, which triggers a discrete categorical basic emotion [93], including joy, sadness, fear, surprise, anger, and disgust. Thus, automatic analysis of emotion from images of human facial expression has been an interesting and challenging problem for the past 30 years [83]. Aiming towards the applications of human behavior analysis, human-human interaction and human-computer interaction, this topic has recently drawn even more attention.

Automatic analysis of facial expression in a realistic scenario is a much more difficult problem due to that the 2-D imagery of human facial expression consists of rigid head motion and non-rigid muscle motion. We are tasked to solve this “coupled-motion” problem and analyze facial expression in a meaningful manner. We first proposed an image-based representation, Emotion Avatar Image, to help person-independent expression recognition. This method allows us to analyze facial expression in a canonical space, which makes the comparison of corresponding features more accurate and reasonable. Second, an real-time registration technique is designed to improve frame-based streaming facial action unit (AU) recognition. We do not always have the luxury of obtaining the temporal segmented discrete facial expressions,

e.g., joy or surprise. This chapter describes a frame-based method for registration. It not only aligns faces (objects in general) to a reference, but also guarantees temporal smoothness, both of which are essential for spontaneous expression analysis. Third, the proposed accurate expression recognition techniques are then applied to the field of advertising, where facial expression is demonstrated to be closely correlated with the commercial viewing behavior of audiences.

Contents

List of Figures	x
List of Tables	xiv
1 Introduction	1
2 Emotion Understanding from Emotion Avatar Image	4
2.1 Motivation	4
2.2 Related Work, Motivation, and Contributions	10
2.2.1 Related Work	10
2.2.2 Motivation	11
2.2.3 Contributions	13
2.3 Technical Approach	14
2.3.1 Face Detection	14
2.3.2 Emotion Avatar Image Representation	15
2.3.2.1 SIFT Flow Alignment	15
2.3.2.2 Avatar Reference and EAI	18
2.3.2.3 Characteristics of EAI	20
2.3.3 Feature Extraction	22
2.3.3.1 LBP	23
2.3.3.2 LPQ	23
2.3.4 Classification	24
2.4 Experimental Results	25
2.4.1 System Implementation	25
2.4.2 Challenge Evaluation Protocol	27
2.4.3 Challenge Results and Discussions	29
2.4.4 Evaluation on Cohn-Kanade Dataset	33
3 A Dense Flow-based Framework for Real-time Object Registration in Video	36
3.1 Motivation	36
3.2 Related Work and Contributions	39
3.2.1 Related Work	39
3.2.2 Contributions of This Work	40
3.3 Flow-based Real-time Object Registration	41
3.3.1 The Generalized Model	42
3.3.2 The Flow-based Instantiation	44

3.3.3	Error Propagation and Close-loop Rectification	47
3.3.4	Computational Cost Analysis	48
3.4	Experimental Results	48
3.4.1	Facial Action Unit Recognition	48
3.4.2	Rear-view Vehicle Recognition	55
3.4.3	Multi-frame Image Super-resolution	58
4	Zapping Index:Using Smile to Measure Advertisement Zapping Likelihood	62
4.1	Motivation	62
4.2	Related Work and My Contributions	65
4.2.1	Automatic Facial Expression Recognition	65
4.2.2	Zapping Analysis	65
4.2.3	Contributions	66
4.3	Technical Approach	67
4.3.1	Smile Detection	67
4.3.2	AdEmotion Data Collection	70
4.3.3	Data Characteristics	73
4.3.3.1	The Zapping Distribution	73
4.3.3.2	The Mean Smile Response	75
4.3.3.3	The Maximum Smile Response	76
4.3.3.4	The Volume of Smile Response	79
4.3.4	Zapping Index	80
4.3.4.1	The Smile Histogram Feature	80
4.3.4.2	Zapping Classification	81
4.4	Experimental Results	83
4.4.1	Zapping Index on Individual Sequences	83
4.4.2	Zapping Index vs. Zapping Distribution	84
4.4.3	Zapping Index vs. Smile Response	85
4.4.4	User Preference from Zapping Index	85
4.4.5	Limitations	88
5	Conclusions	93
	Bibliography	95

List of Figures

2.1	Existing face registration techniques cannot handle out-of-plane head rotation. .	5
2.2	Sample sequence for posed data. Very little head motion is involved.	7
2.3	The uncontrolled acted data from GEMEP-FERA dataset. This data is used for testing.	12
2.4	The overview of the proposed approach.	15
2.5	SIFT flow face registration performs well when pose change is small or large. It captures the facial muscle motion in both cases but the results are very noisy. .	17
2.6	Avatar Reference face model and EAI representations for the first three levels. For comparison, level-0 EAIs are the average of every face image from their corresponding videos without alignment. Higher levels of EAI have more facial feature details and a homogenous face model.	20
2.7	The level-2 EAI representation for 7 training subjects and all emotions. The facial features are reasonably aligned, and person-specific information is attenuated.	21
2.8	The boxplot of the 10-fold cross validation result on 155 GEMEP-FERA training data with respect to different values of SVM parameter C	25
2.9	Box plot of 10-fold cross-validation results on 155 training videos using different level of EAIs. The average classification rate is connected for LPQ texture descriptor to show the improvement at each level. This is to demonstrate that I adopt level-2 EAIs because of its potential to good performance and relative computational efficiency.	26
2.10	Avatar Reference from level-0 to level-7. Higher level of Avatar Reference will have excessive facial details due to overfitting. Level-1 is used in my system. .	26
2.11	Comparison of my test results to the baseline method [111]. The value for my method for each category is taken from the “average rate” in its corresponding confusion matrix. The EAI representation possesses the largest improvement in predicting the unseen person-independent data—31% for LPQ, 26% for LBP. . .	31

2.12	Examples of EAI representation for the same facial expression from the same subject. The EAIs are consistent with the given subject's identity and facial expression. Therefore, the results are near perfect for person-specific case. . . .	31
2.13	Comparison of classification result in the <i>primary test</i> for person-specific, person-independent, and overall cases [2]. Teams are ranked based on the overall performance (numbers are labeled). My EAI methods rank the first place and the third place for LPQ and LBP, respectively. UCR: University of California at Riverside; UIUC-UMC: University of Illinois at Urbana-Champaign; University of Missouri; UCSD-CERT: University of California at San Diego; ANU: Australian National University; UCL: University College London; UMont.: University of Montreal; NUS: National University of Singapore; QUT-CMU: Queensland University of Technology; Carnegie Mellon University; MIT-Cambridge: Massachusetts Institute of Technology; University of Cambridge	32
3.1	Comparison of registration results. Row 3 is the absolute difference of frame 1 and frame 2. Column 2 is the point-based affine registration method used in [65] [113], where affine transformation is computed from 68 facial feature points generated by the state-of-the-art detector [140]. Column 3 uses SIFT flow [67] to align with the Avatar Reference face model from [129]. Ideally, I would like the frame difference to show only at locations where the non-rigid motion is present (mouth area in this case). The proposed method, SOFIT, achieves the most plausible result.	37
3.2	Registration Example. The input sequence is registered with respect to the reference frame shown in bottom right. The flow visualization is coded as in [14]. Better viewed in color.	46
3.3	The error accumulation with respect to time. The error is defined as the SIFT flow of the current frame to the canonical reference frame. I use loop-closure (LC) to update the global flow estimation and rectify the error. The LC is carried out every 300 frames in this experiment.	47
3.4	Qualitative result comparison. Row 1 and 2 are the first and fifth frame of a sequence. Row 3 to 5 are the cumulative absolute frame difference of 5 unaligned frames using method SOFIT, SIFT flow [67], point-based affine where points are detected using deformable part-based model (DPM) [140], respectively. Row 6 are without alignment. The proposed alignment technique captures the correct non-rigid motion of face, for example, eyebrows raise for first column and mouth open for second column.	52
3.5	More results using SOFIT registration technique.	53
3.6	Sample registration results. The first row of both (a) and (b) are misclassified examples by SVM while the second row are correctly classified. In (a), SOFIT is able to align detected vehicles to a similar scale. In (b), SOFIT mainly corrects the missing bottom and align similar parts of the vehicle according to their structure. The bottom right shows the avatar reference frame generated by [129] used in the initialization step.	56

3.7	Sample vehicle detection result. Red rectangle is the result of vehicle ROI detection and yellow dash rectangle is the result after enforcing bilateral symmetry.	57
3.8	The boxplot of the 10-fold cross validation comparison. (a) unaligned, (b) aligned with initialization, (c) SOFIT.	58
3.9	Comparison of super-resolution results using different registration methods for 2 subjects. For each row from left to right: one of the LR inputs (enlarged by pixel replication), sub-pixel registration (SP) [58], frequency domain based registration (FR) [116], and the proposed registration method. I use two SR methods to reconstruct the high-resolution outputs: iterated back-projection (BP) [53] and normalized convolution (NC) [91]. The red blocks show the magnified parts from the pink blocks.	59
3.10	Image quality comparison between video-based super-resolution results and the proposed method. A recent non-reference image quality assessment method [126] is used. The higher the score, the better the estimated visual quality is. LR is low-resolution input images. SP denotes the super-resolved results using [58]. FR is the super-resolved results using [116]. These benchmark methods are compared with the proposed SOFIT method.	60
4.1	ROC curve for my person-independent smile detection algorithm.	68
4.2	Sample smile response results. The response value reflects the intensity of smile.	68
4.3	Handling smile detection failure. (a) Smile response failure cases where (1) and (2) exist because the subjects' neutral faces are similar to smiling faces; (3) is due to out-of-plane rotation. (b) The baseline response of each subject is used to eliminate the failure cases automatically.	69
4.4	The data collection environment. A duplicate monitor is used for data synchronization.	71
4.5	The zapping distribution. The data-driven threshold at 90% is used to separate the data into zapping and non-zapping classes.	74
4.6	The average smile response of zapping and non-zapping classes for the first 30 seconds (720 frames). Each sequence is bounded by its standard error of the mean (SEM).	76
4.7	Sample frames of smile response from zapping and non-zapping classes.	77
4.8	The distribution of the zapping / non-zapping data based on their maximum smile response.	77

4.9	The analysis of the smile response volume. Fig. 4.9(a) and Fig. 4.9(b) are interpreted as follows: the probability is F if the smile response of a sequence is <i>above</i> y for x percent of its entire length. Fig. 4.9(c) is generated by subtracting Fig. 4.9(a) from Fig. 4.9(b), which shows that high smile response is more effective than high volume in distinguishing zapping from non-zapping. (Better viewed in color)	78
4.10	Smile histogram of zapping and non-zapping sequences. The selected sequences are the same as in Fig. 4.7.	81
4.11	The ROC plot for zapping/non-zapping classification. The naïve baseline is by assigning test labels based on the class distribution.	82
4.12	The moment-to-moment Zapping Index for individual sequences. The selected sequences are the same as in Fig. 4.7.	83
4.13	The relationship of Zapping Index (ZI) and zapping distribution for each ad. Zapping distribution is the normalized histogram measured with the primary y-axis, while ZI is displayed by the secondary y-axis bounded by its standard error (shown with dotted lines). Generally speaking, the more sharply ZI decreases, the less zapping happens.	86
4.14	The relationship of smile response and Zapping Index (ZI) for each ad. Broadly speaking, they are in inverse relation. But the ZI is less volatile as compared to smile response.	87
4.15	The pair-wise distance of individual ad in zapping distribution and Zapping Index (ZI) groups. ZI is highly correlated with viewer's zapping behavior exhibited in zapping distribution, which demonstrates that ZI is a reasonable measure for zapping. (The distances within each group are normalized between 0 and 1.)	88
4.16	All four possible types of user preferences represented by ZI. Each ZI pattern is the average of one subject on the entire ad category. Flat ZI response such as both cases in (a) shows no smile response as well as no interest to the ad category, and vice versa.	89
4.17	Zapping Index vs. zapping distribution for the ads in <i>running shoe</i> category. ZI is less effective in predicting zapping probability compared to Fig. 4.13 since soliciting smile is not the intention of these ads. Hence, users tend to zap less even if ZI value is high.	90
4.18	Zapping Index vs. smile response for the ad in the <i>running shoe</i> category. Smile response is low but this does not always lead to a decrease in the value of ZI.	91

List of Tables

2.1	Confusion Matrix for EAI+LPQ (person-independent)	27
2.2	Confusion Matrix for EAI+LPQ (person-specific)	28
2.3	Confusion Matrix for EAI+LPQ (overall)	28
2.4	Confusion Matrix for EAI+LBP (person-independent)	29
2.5	Confusion Matrix for EAI+LBP (person-specific)	30
2.6	Confusion Matrix for EAI+LBP (overall)	30
2.7	A comparison of selected facial expression recognition techniques.	34
2.8	A comparison of methods proposed by different teams which participated in the FERA challenge [2]. The ranked results on the same GEMEP-FERA challenge dataset are given by Fig. 2.13.	35
3.1	Person-independent AUC-score result on FERA-GEMEP AU training set. Feature: Local Binary Pattern	50
3.2	Person-independent AUC-score result on FERA-GEMEP AU training set. Feature: Local Phase Quantization	51
4.1	Advertisement Selection	73
4.2	Another Advertisement Category: “Running Shoe”	89

Chapter 1

Introduction

Existing video based facial expression recognition techniques analyze the spatial information in every frame as well as explore the temporal relation among frames. On the contrary, In chapter 2, I present a new image-based representation and associated reference image called Emotion Avatar Image (EAI), and Avatar Reference, respectively. The representation is computed from the videos of facial expression. The approach for facial expression consists of the following steps: (a) face detection; (b) face registration of video frames with the Avatar Reference to form the EAI representation; (c) computation of features from EAIs using both Local Binary Patterns (LBP) and Local Phase Quantization (LPQ); and (d) classifying the feature as one of the emotion type using a linear Support Vector Machine (SVM) classifier. my system is tested on the Facial Expression Recognition and Analysis Challenge (FERA2011) data, GEMEP-FERA dataset. The experimental results demonstrate that the information captured in a EAI for a facial expression is a very strong cue for emotion inference. Moreover, my method suppresses the person-specific information for emotion, and performs well on unseen data.

A recognition task usually begins with aligning objects with similar shape. For example, in the field of facial expression analysis, aligning faces in a meaningful manner is a

challenging problem to solve. The realistic streaming expression contains non-rigid muscle motion and rigid head motion. The goal is to unify the rigid motion while maintaining the non-rigid motion. In chapter 3, I present a video-based alignment framework to tackle this problem. I assume that the object recognition task is carried out in a unified space represented by a canonical object model; individual object undergoes a global transformation (e.g., affine) with respect to this canonical model. This is to rectify the object-independent transformation such as rotation while maintaining the object-specific identifiable information. I define a general model for object alignment in a Bayesian framework, and rigorously show that a special case of the model results in a SIFT-flow- and optical-flow-based least-square problem. I demonstrate that dynamic programming can be used to speed up the computation. This algorithm is designed for streaming data that allows the analysis of facial muscle dynamics under subtle changes in natural spontaneous expressions. I extend my experiment to vehicle recognition and image super-resolution to demonstrated generalization of the method. Visual and quantitative results demonstrate that the proposed method aligns objects in a meaningful manner and leads to superior recognition results.

In marketing and advertising research, “zapping” is defined as the action when a viewer stops watching a commercial. Researchers analyze users’ behavior in order to prevent zapping which helps advertisers to design effective commercials. Since emotions can be used to engage consumers, in chapter 4, I leverage automated facial expression analysis to understand consumers’ zapping behavior. Firstly, I provide an accurate moment-to-moment smile detection algorithm. Secondly, I formulate a binary classification problem (zapping/non-zapping) based on real-world scenarios, and adopt smile response as the feature to predict zapping. Thirdly, to cope with the lack of a metric in advertising evaluation, I propose a new metric called Zapping Index (ZI). ZI is a moment-to-moment measurement of a user’s zapping probability. It gauges not only the reaction of a user, but also the preference of a user to commercials. Finally, ex-

tensive experiments are performed to provide insights and I make recommendations that will be useful to both advertisers and advertisement publishers.

Chapter 2

Emotion Understanding from Emotion Avatar Image

2.1 Motivation

Early stage research on facial expression recognition focused on static images [83]. Both feature-based and template-based approaches were investigated. Recently, researchers have been using image sequences or video data in order to develop automated expression recognition systems. As demonstrated in the fields of computer vision [135] [123] [108] [128] and psychology [4] [17], various types of dynamic information, such as dynamic appearance and dynamic geometry are crucial for the recognition of human expressions.

However, extracting the facial dynamics from an expression sequence is not a trivial problem. It requires near perfect alignment for the head pose and facial features. The inherent challenge for facial expression recognition is the dilemma between compensating the rigid motion of the head pose and extracting the non-rigid motion of facial muscles. Most existing algorithms and real-time computer programs [64] [37] are only capable of analyzing frontal face

with near upright angle. This is not due to the failure to detect a face but due to the failure to register the detected face reasonably in a video.

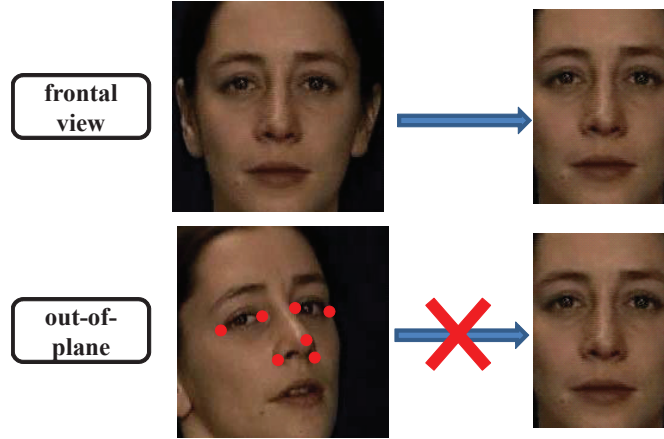


Figure 2.1: Existing face registration techniques cannot handle out-of-plane head rotation.

As shown in Fig. 2.1, when a subject’s face is in frontal view, near frontal view, or has minor in-plane rotation, the alignment can be done easily by in-plane image transformation. We can detect both eye locations, scale the distance of both eyes to a constant value for every subject, and then rotate the image to guarantee both eyes are horizontally aligned. Finally, we can translate the whole image such that the eyes are located at some predefined locations. This registration technique is suitable for some early stage research experiments where facial expression data are acquired under controlled conditions. One restriction is that in the collected data, not much head movement should be involved. To accomplish this, data are collected by cameras mounted on the subject’s head to eliminate the head motion [84].

Three types of data are used in the facial expression recognition community, namely posed data, acted data, and spontaneous data. For datasets that are collected from stationary camera such as MMI dataset [85] and Cohn-Kanade (CK) dataset [57] (Fig. 2.2), the subjects show facial expressions with minimum head movement, and therefore, help researchers to focus on the non-rigid facial muscle movement. Thus, these datasets fall into the category of posed fa-

cial expressions, meaning the subjects are given instructions before showing expressions. Subjects are conscious about controlling their facial muscle movement. All the expressions start from a neutral face, which provides a good reference for computing the non-rigid facial motion. However, experiments demonstrate that in human-to-human interaction such as conversation, people tend to adapt their head movements and facial expressions in response to the stimulus [19]. This is a strong evidence for the fact that facial expression is incorporated with head motion. This fact is also true in MMI and CK dataset, where most of the videos contain minor rigid head motion (both in-plane and out-of-plane). Therefore, the in-plane image transformation technique cannot align the entire dataset.

One technique that state of the art algorithms use to resolve in-plane rotation is 2D affine transformation. A number of facial “anchor points” are defined whose motion is relatively stable during facial expressions. Such anchor points include eye locations, inner and outer eye corners, and the tip of the nose. We could also define a corresponding target location for each anchor point. Once the anchor points are detected, the affine transformation matrix could be computed by minimizing the sum of the least square error of detected location and target location of the anchor points. This affine transform is subsequently applied to the entire face image to complete the registration step.

The affine-based registration performs quite well when in-plane or minor out-of-plane head motion is present. However, it is very sensitive to noise. The anchor points are not entirely stable during a facial expression. The eye corner could be unstable if the subject is blinking; the tip of the nose could also be moving, and so forth. The typical number of anchor point is around 6. If every points is not detected precisely, large error will be generated and the affine transformation for the whole image will be unacceptable.

Moreover, affine-based registration is not temporally smooth. If a minor change oc-



(a) Cohn-Kanade dataset [57]



(b) MMI dataset [85]

Figure 2.2: Sample sequence for posed data. Very little head motion is involved.

curred to an anchor point for 2 consecutive face images; the affine transform matrix will be off by a small amount. After applying this affine transform to the entire face image, every single pixel is affected due to this small change. This will result in a fake motion for the stationary face regions. Therefore, the whole dynamic analysis based on this registration method will be imprecise.

Another registration technique is through the Active Appearance Model (AAM) [25] [73]. At first, it requires manual labeling of a subset of the frames for each data, which is not desired in an automated system [71]. Although person-independent AAM approach is developed later, this technique can be inaccurate due the false feature location. Therefore, it is sensitive to noise.

The other issue besides face registration is the person-independent property (subjects in the test data are not used for training) of the algorithm. As mentioned earlier, basic emotions are universal across culture, gender, age, etc. Thus, the computer algorithms are expected to extract person identity invariant features. This is different from the purpose of face recognition in which person-dependent properties should be retained when extracting facial features from video sequences in the task of face recognition [6, 11]. Computer algorithms cannot be trained with data for all the human beings. The generalization ability allows the system to predict for the unseen people. This property enables the system to carry out facial expression recognition from a person-dependent (or person-specific) environment to a person-independent environment. Unfortunately, few works have addressed the person-independence issue.

The person-specific information can be eliminated at two steps in the system: face registration and feature extraction. In-plane image transformation based and AAM based registration techniques do not change the geometry or appearance of facial features, and therefore, the person-specific information is retained. The affine transformation based registration algorithms are able to change the geometry and the appearance of a person to a limited extent. When

a face is in a near frontal view (where the affine based registration accomplishes the most plausible result) and not much transformation is needed, the identity is mostly unaltered. When faces are not in the frontal view, the affine based algorithm is able to change the person's identity by a large amount, but unfortunately, that is when this approach performs poorly and most of the registration results are unacceptable.

The person specific information can also be eliminated through feature extraction. Features that are extracted could be categorized into geometry-based and appearance-based. Geometry-based approaches track the geometry of landmark points over time and use their dynamics as the feature. If the locations of the facial landmark points are normalized and only the amount of location change is considered to be the feature, it falls into the category of person-independent feature. For instance, emotion "joy" is typically accompanied with smile which results in the expanding and pulling up of the mouth corner. This geometric clue tends to be universal. However, this point-based inference is very sensitive to noise. The location change of a landmark point is typically a few pixels at regular video resolution; meanwhile, a minor head movement results in changes in the order of tens of pixels which is also very difficult to recover. This minor rigid motion of head disguises the true motion of the landmark points and, therefore, generates large error in the extracted feature. On the other hand, the appearance-based approaches such as Local Binary Patterns (LBP) [80], Gabor wavelets [63], Local Phase Quantization (LPQ) [81], concentrate on the dense response of filters to intensity values of a face. These methods are inherently person-dependent unless the identity is eliminated in the face registration process.

The challenges mentioned above encourage us to develop a system which accurately registers face images and at the same time eliminates the person-specific information. To pinpoint the key emotion of an image sequence while circumventing the complex and noisy dynamics, I also seek to summarize the emotion video containing possibly hundreds of frames. If I can

find a single good image representation based upon which to make judgments, I would be able to capture the emotion, that an entire video expresses, in a computationally efficient manner.

In this work, I adopt the recently introduced SIFT flow algorithm [67] to register the facial images. By matching the dense SIFT descriptors across image pairs, this method is able to generate satisfactory alignment results for facial features. Although SIFT flow is originally designed for image alignment at the scene level, it is reasonable to apply it here to facial expression recognition since a human face can be considered as a scene in this case. It is capable of globally aligning the head/face region while maintaining the shape and motion of facial features for consecutive frames. In order to solely extract the facial motion information irrespective of person-specific information, I iteratively build a single “Avatar Reference” face model, onto which I align all the face images. Later, I update the Avatar Reference face model, and also the single good representation, Emotion Avatar Image (EAI), for each video consisting of frames for an emotion. The model is term Avatar because the subjects are morphed towards homogeneity while the emotions are successfully retained. Subsequently, the EAIs are passed through LBP and LPQ texture descriptors for feature extraction individually. Finally, Support Vector Machine (SVM) with linear kernel is used for classification. My approach is able to transform the whole expression recognition problem from an image sequence back to a single image that is suitable for real-time application.

2.2 Related Work, Motivation, and Contributions

2.2.1 Related Work

A large amount of effort has been focused on describing facial expression features. Based on the feature in use, as introduced earlier, I can broadly divide the methods into three categories, i.e., geometry-based approaches, appearance-based approaches, and the combination

of the two. Geometry-based approaches track the facial geometry information based on a set of facial landmark points over time and classify expressions based on their deformation. On the other hand, appearance-based approaches use information from the facial texture described by various types of texture descriptors, such as LBP, Gabor wavelets, and LPQ. The dynamics of the texture deformation can also be included for feature extraction. Table 2.7 provides a thorough comparison of methods from the literature based on the usage of registration techniques, feature types, dynamic feature, classifier, and the dataset.

In this chapter, the methods that are compared with the proposed method are listed in Table 2.8. In this table, I also analyze their registration techniques, features, classifiers similar to the comparison shown in Table 2.7. Also, I include the features and classifiers that are adopted in these papers. Later in Section 2.4, I provide a comparison of the methods on the same data, the GEMEP-FERA challenge dataset [1].

2.2.2 Motivation

Based on how the data is acquired, it can be categorized into three classes: posed data, acted data, and spontaneous data. When posed data is collected, subjects are given a series of instructions such as emphasize on the facial muscle movement and try not to move head. Posed data played an important role in the early stage research, because it provided researchers with more insight about the relation of expression to the muscle movement. CK database [57] and MMI database [85] fall into this category.

The ultimate goal of the research community is to recognize spontaneous facial expressions. However, spontaneous data is very hard to acquire. Facial expressions can be called spontaneous when subjects are not aware that they are being recorded, and express emotions naturally. Since it very difficult to design a fully unaware environment when collecting data, no spontaneous dataset coded with explicit emotions is publically available.

The intermediate stage between the previous two, namely the acted data has less control than the posed data but subjects are fully aware when data is being recorded. The GEMEP-FERA challenge dataset [1] that this work used belongs to this class, and is shown in Fig. 2.3. In the process of data collection, subjects are not asked to control themselves, but just to convey a certain emotion. These experiments have no control about body pose, head pose, gesture, or occlusion and therefore, are very challenging for expression recognition by an automated system.



Figure 2.3: The uncontrolled acted data from GEMEP-FERA dataset. This data is used for testing.

To motivate my method, I analyze the specifications of GEMEP-FERA dataset:

1. 10 subjects (5 male, 5 female) are involved with their upper body visible.
2. Subject's age is approximately between 25 and 60 years by observation.
3. Video resolution is 720×576 and face resolution is around to 200×200 pixels.
4. Average video length is about 2 seconds with 30 fps frame rate.
5. Each video contains one subject displaying expressions corresponding to an emotion.

6. Five emotions are involved: Anger, Fear, Joy, Relief, Sadness. This is different from the typical 6 basic emotions datasets.
7. There are 3 to 5 videos for each subject with the same emotion.
8. Most subjects are uttering meaningless phrases while displaying an expression [111].
9. Videos do not start with the neutral face or end at apex or offset, unlike posed data.
10. Multiple apexes are involved in some videos.
11. The neutral face is not always available.

The above observations provide us the following key facts that inspire my system:

1. Good registration is demanding and previous registration techniques (in-plane image transformation and affine-based transformation) are not suitable for this dataset.
2. Dynamics is hard to recover because the neutral reference face is not always available.
3. Constant lip motion limits the geometry based approaches.

2.2.3 Contributions

Existing work intensely emphasizes on analyzing the sequential change of the facial feature. Nevertheless, since the onset and offset for a realistic data are hard to detect, if a near-apex frame is able to be picked up to represent an entire expression session, I can avoid extracting subtle sequential facial feature deformations, and describe emotions in a reliable manner.

The contributions of this work are: (1) Iteratively build a reference face model, Avatar Reference. This homogenous reference face model can capture the nature of the entire dataset. (2) Condense a video sequence into single image representation, EAI, in facial expression recognition. The EAI representation registers the facial features at a meaningful location and main-

tains the non-rigid facial muscle movement. Being able to suppress the person-specific information, the EAI representation also allows the expression recognition tasks to be carried out in a person-independent manner.

To my best knowledge, until now, little work has been done to condense a video sequence into a tractable image representation for emotion recognition. As the results show later, if the expression is not extremely subtle such that even human visual system is unable to capture, my algorithm can distinguish most of the differences among expressions.

2.3 Technical Approach

Fig. 2.4 outlines my approach in four major steps. After automatic extracting the face from a raw video, I provide some insights about the EAI representation that suppresses the person-specific information while maintaining the shape and texture information of the facial features. Both LBP and LPQ texture descriptors are applied to generate the features, and then, the linear SVM classifier is used for classification.

2.3.1 Face Detection

I first extract the face from the video using the Viola and Jones face detector [118] implemented in OpenCV. This algorithm achieves high quality performance and is suitable for real-time processing. The detection rate is near perfect on GEMEP-FERA [1] dataset. Since the face resolution is around 200×200 pixels, I resize the detected face image exactly to this resolution using bicubic interpolation. This process removes the noise and smoothes the raw images.

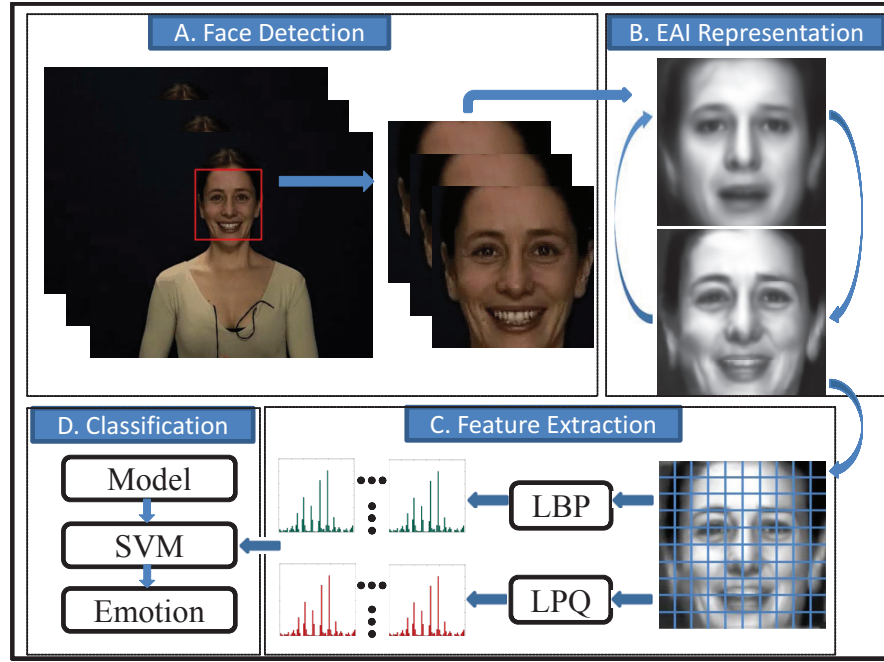


Figure 2.4: The overview of the proposed approach.

2.3.2 Emotion Avatar Image Representation

2.3.2.1 SIFT Flow Alignment

SIFT flow is recently introduced by Liu et al. [67]. It is originally designed to align an image to its plausible nearest neighbor which can have large variations. The SIFT flow algorithm robustly matches dense SIFT features between two images, while maintaining spatial discontinuities.

In [67], the local gradient descriptor, SIFT [69], is used to extract a pixel-wise feature component. For every pixel in an image, the neighborhood (e.g. 16×16) is divided into a 4×4 cell array. The orientation of each cell is quantized into 8 bins, generating a $4 \times 4 \times 8 = 128$ -dimension vector as the SIFT representation for a pixel, or the so called SIFT image. The SIFT image has a high spatial resolution and can characterize the edge information.

After obtaining the per-pixel SIFT descriptors for two images, a dense correspondence

is built to match the two images. Similar to optical flow, the objective energy function that I attempt to minimize is designed as:

$$E(\mathbf{w}) = \sum_p (\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|_1, t) + \quad (2.1)$$

$$\sum_p \eta(|u(\mathbf{p}) + v(\mathbf{p})|) + \quad (2.2)$$

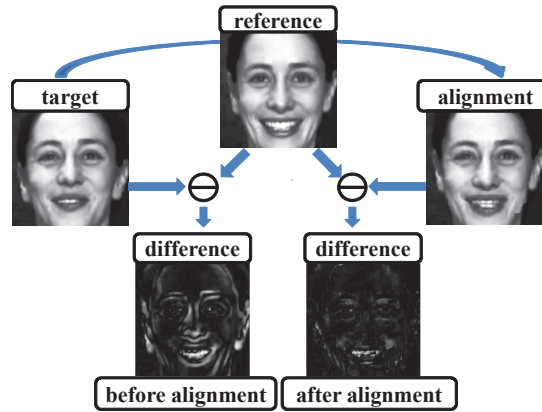
$$\sum_{(\mathbf{p}, \mathbf{q}) \in \varepsilon} (\alpha |u(\mathbf{p}) - u(\mathbf{q})|, d) + \quad (2.3)$$

$$(\alpha |v(\mathbf{p}) - v(\mathbf{q})|, d)$$

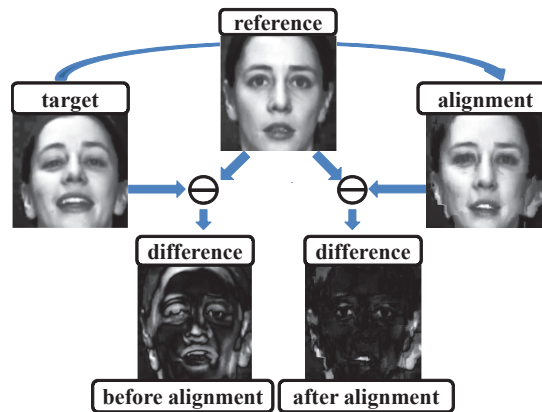
where $\mathbf{p} = (x, y)$ is the grid coordinates of images, and $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at $\mathbf{p}$. s_1 and s_2 are two SIFT images to be matched. ε contains all the spatial neighbors (a four-neighbor system is used). The data term in Eq. 2.1 is a SIFT descriptor match constraint that enforces the match along the flow vector $\mathbf{w}(\mathbf{p})$. The small displacement constraint in Eq. 2.2 allows the flow vector to be as small as possible when no other information is available. The smoothness constraint in Eq. 2.3 takes care of the similarity of flow vectors for adjacent pixels. In this objective function, truncated $L1$ norm is used in both the data term and the smoothness term with t and d as the threshold of matching outliers and flow discontinuities, respectively.

The dual-layer loopy belief propagation is used as the base algorithm to optimize the objective function. Then, a coarse-to-fine SIFT flow matching scheme is adopted to improve the speed and the matching result.

Fig. 2.5(a) shows two frames with minor pose change from GEMEP-FERA training dataset [1]. I align the target frame with respect to a reference frame. For comparison purpose, I take the absolute difference between images before alignment and after alignment with respect to the reference separately. Comparing the two difference images in Fig. 2.5(a), the rigid head motion from minor pose change is eliminated, while the non-rigid facial motion is maintained



(a) Minor difference. Only true facial motions are captured as shown by the corresponding difference image of before alignment and after alignment.



(b) Major difference. The difference image (bottom right) of the reference and the alignment result shows the true facial motions are captured in the mouth and eye areas.

Figure 2.5: SIFT flow face registration performs well when pose change is small or large. It captures the facial muscle motion in both cases but the results are very noisy.

in the areas of mouth, eyes, and lower cheek. The areas of non-rigid motion match with the motion in the original two images from human visual observation. Nevertheless, the difference image also shows that the SIFT flow alignment process is noisy; discontinuities can be observed in the aligned result.

Consider a case with a major pose change in Fig. 2.5(b), the head pose motion is out-of-plane and facial appearance changes significantly. The registration result is in the upright pose, and non-rigid motion in the mouth and eye areas can still be captured. Differences at periphery are due to the lack of correspondences for SIFT flow vectors. However, this information is still useful as it captures the pose change, which is also an important cue in facial expression recognition [19]. Differences at periphery show that the pose change and the true facial feature motion are decomposed. Similar to the minor pose change case, the noise and discontinuity are issues in the aligned result.

2.3.2.2 Avatar Reference and EAI

SIFT flow has the potential to align images with large spatial variation. This is useful in aligning the face image given the possibility of a large head pose change or an occlusion. However, the person-specific information still has to be eliminated. I seek to build a reference face with respect to which each face image can be aligned.

Algorithm 1. *Avatar Reference and EAI*

Given:

$I^{(m,n)}$: face image from sequence m , frame n

M : total number of image sequences

N_m : total number of frames in sequence m

Q : user defined number of levels

A_i^{ref} : Avatar Reference at level- i

EAI_i^m : EAI representation for sequence m based on the level- i Avatar reference A_i^{ref}

$I_{align}^{(m,n)}$: the alignment result for a face image $I^{(m,n)}$ using SIFT flow

Initialization: $A_0^{ref} = \frac{1}{\sum_{m=1}^M N_m} \sum_{m=1}^M \sum_{n=1}^{N_m} I^{(m,n)}$

for $i = 1 \rightarrow Q$ **do**

for $m = 1 \rightarrow M$ **do**

for $n = 1 \rightarrow N_m$ **do**

$I_{align}^{(m,n)} \leftarrow SIFTflow(I^{(m,n)}, A_{i-1}^{ref})$

end for

$EAI_i^m \leftarrow \frac{1}{N_m} \sum_{n=1}^{N_m} I_{align}^{(m,n)}$

end for

$A_i^{ref} \leftarrow \frac{1}{\sum_{m=1}^M} \sum_{m=1}^M EAI_i^m$

end for

In Algorithm 1, I design an iterative averaging method to generate an Avatar Reference face model. To put simply, I initialize the algorithm by averaging all possible face images in the training dataset. Using this average face as the reference initially, I align each face image in the video using SIFT flow. After alignment, user can update the Avatar Reference using all the aligned faces. The iteration number defines the level of the Avatar Reference (level 0 means the average of all the unaligned face images). The Avatar Reference models for the first 3 levels are shown in row 1 of Fig. 2.6. From my observation, the Avatar Reference is not always a neutral face. It captures the most likely facial appearance throughout the whole dataset, and therefore, it has less total variation in registration. The mouth is open for the level-1 and level-2 Avatar Reference face result (as shown in Fig. 2.6, row 1). This is because most of the subjects in the training data are uttering meaningless phrases [1], therefore, have a lot of mouth movement.

In Algorithm 1, once the Avatar Reference face model is obtained, I establish the single representation EAI for the sequence of face images at the current level. As demonstrated earlier, single aligned face image possesses errors and discontinuities. Therefore, I describe an image sequence as the average of all frames within this sequence. The statistical justification of the EAI representation is similar to [47]. Assuming the distribution of every aligned face frame is subject to an addition of a true face and an additive noise. The noise is further assumed to be uncorrelated identical Gaussian distribution. During the averaging process, the noise variance is reduced by a factor of N , where N is the number of face image. Thus, the alignment noise can be removed from the EAI representation.

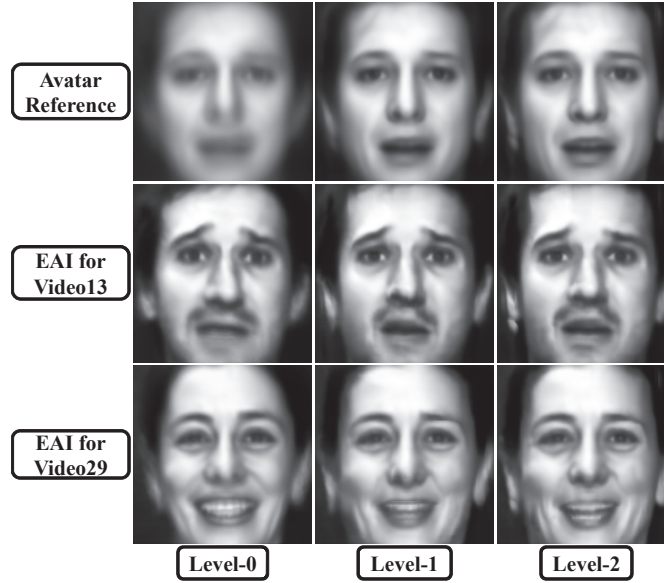


Figure 2.6: Avatar Reference face model and EAI representations for the first three levels. For comparison, level-0 EAIs are the average of every face image from their corresponding videos without alignment. Higher levels of EAI have more facial feature details and a homogenous face model.

2.3.2.3 Characteristics of EAI

In this work, I attempt to test the performance of EAIs at different levels. As shown in Fig. 2.6 (row 2 and row 3), the quality of the EAIs improves as the level of Avatar Reference



Figure 2.7: The level-2 EAI representation for 7 training subjects and all emotions. The facial features are reasonably aligned, and person-specific information is attenuated.

becomes higher. A high level Avatar reference model enhances the facial details, and attenuates the person-specific information. Meanwhile, EAI representation retains the expression information that is recognizable by human visual system. Fig. 2.7 shows the EAI representation for 7 subjects in the training data and for all emotions. Since all the faces are aligned with respect to the same Avatar Reference, the EAI representation can be seen to align facial features, such as nose, eyes, and mouth reasonably. This lays the foundation for extracting meaningful facial feature motion. Besides, aligning every face image with the Avatar Reference allows the elimination of the person specific information to a great extent.

The EAIs in Fig. 2.7 can also be observed to capture the non-rigid facial feature motion and therefore. They retain the corresponding emotion information. This is due to the small constraint intensity parameter η in Equation. 2.2. The larger values of η will penalize more on the large flow vectors, which will result in less morphing for the alignment result. Ideally, if two face images are perfectly aligned, all the facial features should be at exactly the same locations. The facial feature motion will be eliminated in this case. Practically, the real facial feature motions during an expression are larger than the SIFT flow compensation, and subsequently, can be maintained in the noisy alignment results. The accumulation process will smooth the alignment results while capturing the real motion caused by an facial expression.

2.3.3 Feature Extraction

I transform the facial expression recognition problem from a video to a static image. To test the effectiveness of the single image representation EAI, I describe the facial texture from EAI using the well-known texture descriptor LBP and recently proposed blur insensitive LPQ descriptor. I expect to receive similar improvements for both methods.

2.3.3.1 LBP

LBP is a powerful and well known texture descriptor. In this work, I used the extended version of basic LBP by Ojala et al. [80] where the LBP descriptor is gray-scale and rotation invariant. To briefly go over this extended work, the operator, denoted as $LBP_{P,R}^{u2}$, is applied to an circularly symmetric neighborhood with P number of pixel on the circle of radius R . Superscript “ $u2$ ” denotes the uniform property. Uniform LBP is favorable since it reduces the feature dimension. For example, the $LBP_{8,1}^{u2}$ adopted in this work will generate 59 basic patterns while the $LBP_{8,1}$ has 256 possibilities. Since these parameter settings are used in the baseline method [111], I adopt the same settings for better comparison .

After thresholding each pixel in its neighborhood with respect to the center value, the histogram is used to accumulate the occurrence of the various patterns over a region. In my experiment, I resize the face images to 200×200 , and each image is divided into size 20×20 blocks to capture the local texture pattern. Therefore, the LBP feature vector in use is of dimension $59 \times 10 \times 10 = 5900$. As mentioned earlier, the face resolution is close to 200×200 , hence I resize all face images to this uniform value to minimize the information lost.

2.3.3.2 LPQ

The blur insensitive LPQ descriptor is originally proposed by Ojansivu et al. in [81]. The spatial blurring is represented as multiplication of original image and a point spread function (PSF) in frequency domain. The LPQ method is based upon the invariant property of the phase of the original image when the PSF is centrally symmetric.

LPQ method examines a local $M \times N$ neighborhood $N_{\mathbf{x}}$ at each pixel position $\mathbf{x}$ of the image $f(\mathbf{x})$, and extracts the phase information using the short-term Fourier transform defined

by

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in N_{\mathbf{x}}} f(\mathbf{x} - \mathbf{y}) e^{j2\pi \mathbf{u}^T \mathbf{y}} = \omega_{\mathbf{u}}^T \mathbf{f}_{\mathbf{x}} \quad (2.4)$$

where $\omega_{\mathbf{u}}$ is the basis vector of the 2-D DFT at frequency $\mathbf{u}$, and $\mathbf{f}_{\mathbf{x}}$ is another vector containing all M^2 image samples from $N_{\mathbf{x}}$.

The local Fourier coefficients are at four frequency points: $\mathbf{u}_1 = [a, 0]^T$, $\mathbf{u}_2 = [0, a]^T$, $\mathbf{u}_3 = [a, a]^T$, and $\mathbf{u}_4 = [a, -a]^T$, where a is a sufficiently small scalar. I use $a = 1/7$ in my experiment. The vector for each pixel is obtained as

$$\mathbf{F}_{\mathbf{x}} = [F(\mathbf{u}_1, \mathbf{x}), F(\mathbf{u}_2, \mathbf{x}), F(\mathbf{u}_3, \mathbf{x}), F(\mathbf{u}_4, \mathbf{x})] \quad (2.5)$$

The phase information is recovered by a scalar quantizer

$$q_j(\mathbf{x}) = \begin{cases} 1, & \text{if } g_j(\mathbf{x}) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $g_j(\mathbf{x})$ is the j th component of the vector $G_{\mathbf{x}} = [\text{Re}\{\mathbf{F}_{\mathbf{x}}\}, \text{Im}\{\mathbf{F}_{\mathbf{x}}\}]$. The resultant eight binary coefficients $q_j(\mathbf{x})$ are represented as integer values between 0-255 using binary coding

$$f_{LPQ}(\mathbf{x}) = \sum_{j=1}^8 q_j(\mathbf{x}) 2^{j-1} \quad (2.6)$$

Besides, the decorrelation process is added to the original LPQ implementation to eliminate the dependency among the neighboring pixels. Similar to LBP, I divided the 200×200 face image into size 20×20 regions. Therefore, the LPQ feature vector is of dimension $256 \times 10 \times 10 = 25600$.

2.3.4 Classification

I adopt Linear SVM as the classifier [23] for my system. As demonstrated in Fig. 2.8, the 10-fold cross validation accuracy is not affected if C is not extremely small. Therefore, I choose $C = 1$ for my system.

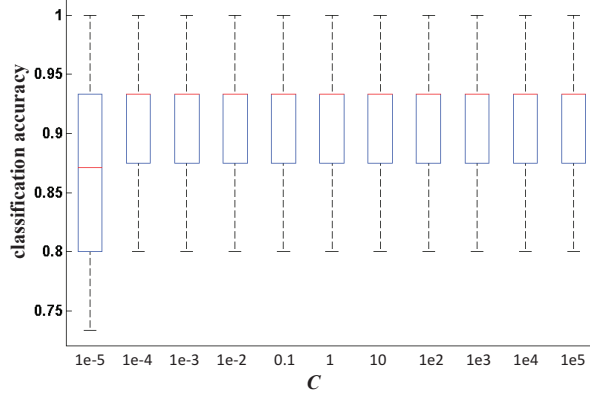


Figure 2.8: The boxplot of the 10-fold cross validation result on 155 GEMEP-FERA training data with respect to different values of SVM parameter C .

2.4 Experimental Results

2.4.1 System Implementation

Similar to my previous work [131], after extracting the faces from the raw data using the Viola and Jones face detector [118], the face images are then aligned to level-1 Avatar Reference face model based on the Algorithm 1, and single representation EAIs are generated. Subsequently, using both LBP and LPQ operators, I extract the feature from all the EAIs separately. Specifically, the $LBP_{8,1}^{u2}$ is used in my experiment. The parameters for the LPQ operator are $M = 9$, $a = 1/7$, $\rho = 0.9$. Lastly, as demonstrated in Section 2.3.4 the classifier I used is the linear SVM [23] classifier with $C = 1$.

The reason why I use level-2 EAI face model is statistically demonstrated in Fig. 2.9. I carry out a series of 10-fold cross-validation experiments on only the training of GEMEP-FERA dataset using first 11 levels of the EAIs and test on LPQ texture descriptor. The cross-validation procedure results in person-specific category because I do not exclude the test subjects from the training. Fig. 2.9 shows that the performance improves as the level of the EAI increases for the first 3 levels. This is consistent with my discussion on Avatar Reference level in Section 2.3.2.

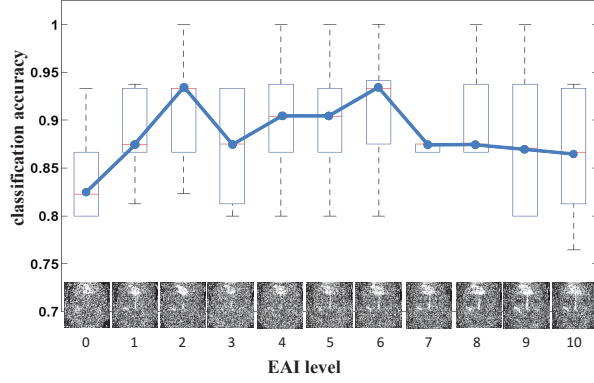


Figure 2.9: Box plot of 10-fold cross-validation results on 155 training videos using different level of EAIs. The average classification rate is connected for LPQ texture descriptor to show the improvement at each level. This is to demonstrate that I adopt level-2 EAIs because of its potential to good performance and relative computational efficiency.

The performance peaks at both level-2 and level-6. After analyzing Avatar Reference and the corresponding EAI representation, the overfitting issue occurs to the Avatar Reference as the level increases as shown in Fig. 2.10. Artifact facial details are excessively displayed through the higher number of iteration in Algorithm 1. The system with level-6 EAI may not have a good generalization to unseen data.



Figure 2.10: Avatar Reference from level-0 to level-7. Higher level of Avatar Reference will have excessive facial details due to overfitting. Level-1 is used in my system.

Table 2.1: Confusion Matrix for EAI+LPQ (person-independent)

	truth					
		Anger	Fear	Joy	Relief	Sadness
prediction	Anger	12	3	0	0	1
	Fear	0	7	0	0	0
	Joy	0	5	19	4	0
	Relief	1	0	1	11	2
	Sadness	1	0	0	1	12
	total rate	0.86	0.47	0.95	0.69	0.8
average rate		0.75				

2.4.2 Challenge Evaluation Protocol

My method and other 9 methods (including the baseline [111]) are compared using the Facial Expression Recognition and Analysis Challenge (FERA2011) data, GEMEP-FERA dataset [1]. As the part of FERA2011 challenge, 155 training videos were given at first containing 7 subjects a month before the deadline. Then, the 134 test videos were released one week before the deadline. I ran the test videos using my system which takes each video session as the input and outputs the emotion label. All predicted labels were then submitted to the organization panel of FERA2011. After evaluation, the results were provided in three different categories: person-independent, person-specific, and overall.

Table 2.2: Confusion Matrix for EAI+LPQ (person-specific)

	truth					
prediction		Anger	Fear	Joy	Relief	Sadness
	Anger	13	0	0	0	0
	Fear	0	10	0	0	0
	Joy	0	0	10	0	0
	Relief	0	0	1	10	0
	Sadness	0	0	0	0	9
total rate		0.92	1	1	0.91	1
average rate		0.96				

Table 2.3: Confusion Matrix for EAI+LPQ (overall)

	truth					
prediction		Anger	Fear	Joy	Relief	Sadness
	Anger	25	3	0	0	2
	Fear	0	17	0	0	0
	Joy	0	5	29	4	0
	Relief	1	0	2	21	2
	Sadness	1	0	0	1	21
total rate		0.93	0.68	0.94	0.81	0.84
average rate		0.84				

Table 2.4: Confusion Matrix for EAI+LBP (person-independent)

	truth					
		Anger	Fear	Joy	Relief	Sadness
prediction	Anger	12	4	4	0	5
	Fear	0	8	0	0	0
	Joy	1	3	16	0	0
	Relief	1	0	0	14	1
	Sadness	0	0	0	2	7
	total rate	0.86	0.53	0.8	0.88	0.47
average rate		0.71				

2.4.3 Challenge Results and Discussions

The confusion matrices for EAI using LPQ operator are shown in Table 2.1, 2.2, and 2.3, with test results on person-independent, person-specific, and overall, respectively. Similarly, the confusion matrices for EAI using LBP operator are presented in Table 2.4, 2.5, and 2.6.

I first compare my two methods to the baseline approach, because all of these approaches adopt similar procedures and parameter settings. In both LBP and LPQ cases, my EAI representation achieves the most significant improvement over the baseline result in the person-independent test (from 0.44 to 0.75 in LPQ; from 0.44 to 0.71 in LBP). This improvement can be visualized in Fig. 2.11. This is a positive evidence that my approach eliminates the person-specific information and captures the facial expression information. Also, this demonstrates the desired ability of EAI for predicting the unseen data in real applications.

In person specific test, my method achieves 96% classification accuracy. This result

Table 2.5: Confusion Matrix for EAI+LBP (person-specific)

	truth					
prediction		Anger	Fear	Joy	Relief	Sadness
	Anger	11	0	1	0	0
	Fear	0	9	1	0	0
	Joy	2	1	8	0	0
	Relief	0	0	1	10	1
	Sadness	0	0	0	0	9
total rate		0.85	0.9	0.73	1	0.9
average rate		0.87				

Table 2.6: Confusion Matrix for EAI+LBP (overall)

	truth					
prediction		Anger	Fear	Joy	Relief	Sadness
	Anger	23	4	5	0	6
	Fear	0	17	1	0	1
	Joy	3	4	24	0	0
	Relief	1	0	1	24	2
	Sadness	0	0	0	2	16
total rate		0.85	0.68	0.77	0.92	0.64
average rate		0.77				

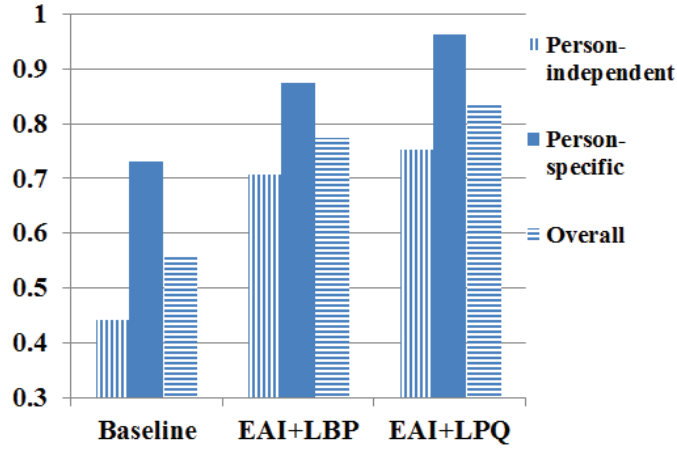


Figure 2.11: Comparison of my test results to the baseline method [111]. The value for my method for each category is taken from the “average rate” in its corresponding confusion matrix. The EAI representation possesses the largest improvement in predicting the unseen person-independent data—31% for LPQ, 26% for LBP.

is illustrated in Fig. 2.12. In the training data, each subject displays the same expression 3 to 5 times. I compare the EAI representation for the same expressions from the same subject. The EAI representations achieves consistency when a subject displayed the same expressions in different videos.



Figure 2.12: Examples of EAI representation for the same facial expression from the same subject. The EAIs are consistent with the given subject’s identity and facial expression. Therefore, the results are near perfect for person-specific case.

Fig. 2.13 shows that my EAI representation combined with LPQ and LBP descriptors rank the first and the third place, respectively in the *primary test*. Since the ground truth label for each emotion video is easy to tell, the FERA2011 organizer required a *secondary test* where no

participant can see the data. I submitted my facial expression recognition system program using EAI+LPQ to the organizer. Secondary test data is approximately half the size of the primary test set. My approach achieves 86% overall classification rate [2], which is consistent with the primary test.

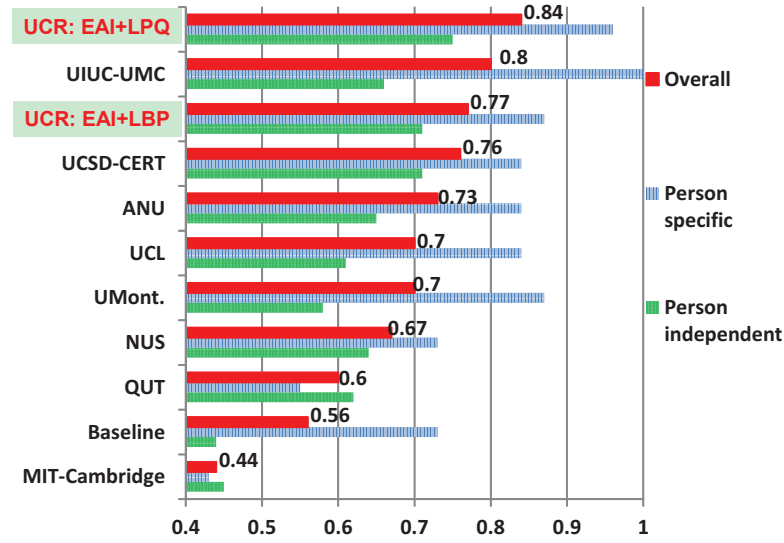


Figure 2.13: Comparison of classification result in the *primary test* for person-specific, person-independent, and overall cases [2]. Teams are ranked based on the overall performance (numbers are labeled). My EAI methods rank the first place and the third place for LPQ and LBP, respectively. UCR: University of California at Riverside; UIUC-UMC: University of Illinois at Urbana-Champaign; University of Missouri; UCSD-CERT: University of California at San Diego; ANU: Australian National University; UCL: University College London; UMont.: University of Montreal; NUS: National University of Singapore; QUT-CMU: Queensland University of Technology; Carnegie Mellon University; MIT-Cambridge: Massachusetts Institute of Technology; University of Cambridge

The inherent characteristic of my approach is to eliminate facial dynamics while maintaining the emotion information. Unlike most of the other approaches [66] [15] which treat each frame as a single training instance (total of 8995 frames from 155 videos if all the images in the training set are used), my method only considers them as 155 EAIs. Given more training videos, my system will most likely be improved since 155 videos of five emotions (approximately 30 videos/emotion on average) may not be sufficiently large to represent a single emotion across a large population.

2.4.4 Evaluation on Cohn-Kanade Dataset

I also evaluated my system using 316 sequences from the CK [57] dataset. Since no subject data of the same facial expression has been collected more than once, the 10-fold cross validation experiment that was carried out belongs to the person-independent category. The average classification accuracy is 71%, which is consistent with my person-independent result in Table 2.1. My algorithm performs not as good as [135] [123] [137] on the CK dataset. This is because the frontal view face images from the CK dataset do not need sophisticated registration techniques. Thus, good dynamic facial features can be easily captured. Therefore, those approaches that use dynamic features outperform my approach that is based on simple features computed from the EAI representation. However, in a more realistic case where a good registration result is difficult to achieve (such as GEMEP-FERA), my approach outperform the approaches using complex dynamic features [28] [77].

Table 2.7: A comparison of selected facial expression recognition techniques.

Authors	Registration	Feature	Dynamic feature	Classifier	Dataset
Yacoob et al. [127]	not mentioned	geometry : optical flow	yes	rule-based classifier	posed data
Essa et al. [39]	3D model fitting	geometric : 3D motion and muscle models	yes	Euclidean norm	posed data
Wang et al. [121]	not mentioned	geometry : B-spline curve	yes	Euclidean norm	posed data
Hu et al. [49]	not mentioned	geometry : variation of active shape model	yes	probabilistic model	posed data
Valtar et al. [114]	not mentioned	geometry : dynamics of 20 facial points	yes	probabilistic SVM	MMI [85] + CK [57]
Pantic et al. [82]	affine transform	geometry : facial points dynamics	yes	rule based	MMI [85]
Donato et al. [31]	in-plane transform	appearance : Gabor wavelets	no	nearest neighbor	posed data
Zhao et al. [135]	in-plane transform	appearance : LBP-TOP	yes	SVM	CK [57]
Bartlett et al. [16]	in-plane transform	appearance : Gabor wavelets	no	SVM and Adaboost	CK [57] + posed data
Wu et al. [123]	in-plane transform	appearance : Gabor motion energy	yes	SVM	CK [57]
Tian et al. [107]	in-plane transform	hybrid : geometric + transient facial features	no	Neural network	CK [57] + posed data
Lucey et al. [71]	AAM	hybrid : 2D shape/appearance + 3D shape	no	SVM	acted data: RU [16]
Zhou et al. [137]	AAM	hybrid : geometry + SIFT [69]	yes	multidimensional assignment algorithm	CK [57] + RU [16]

Table 2.8: A comparison of methods proposed by different teams which participated in the FERA challenge [2]. The ranked results on the same GEMEP-FERA challenge dataset are given by Fig. 2.13.

Paper ID	Registration	Feature	Dynamic feature	Classifier
UCR: EAI [131]	SIFT flow [67]	appearance: LQP [80] or LBP [81]	no	SVM
UIUC-UMC [102]	affine transformation	appearance: SIFT + motion	no	SVM
UCSD-CERT [66]	affine transformation	appearance: Gabor wavelets	no	SVM
ANU [30]	constrained local model	appearance: gradients histogram + LPQ	no	SVM
UCL [77]	in-plane transform	appearance: motion history histogram + LBP variation	yes	SVM_2K [78]
UMont. [28]	in-plane transform	appearance: histograms of oriented gradients	yes	SVM
NUS [98]	in-plane transform	hybrid: accumulated motion image	no	SVM
QUT [24]	constrained local model	appearance: pixel based feature + LBP	no	SVM
Baseline [111]	in-plane transform	appearance: LBP	no	SVM
MIT-Cambridge [15]	not mentioned	hybrid: geometry + Gabor wavelets	no	Hidden Markov Model + Dynamic Bayesian Nets

Chapter 3

A Dense Flow-based Framework for Real-time Object Registration in Video

3.1 Motivation

Video-based image registration is a fundamental problem for computer vision and pattern recognition. It has various applications such as face recognition [119], facial expression recognition [112], image stitching [101], etc. Depending upon different applications, there can be specific requirements for the registration techniques [109] [22]. Broadly speaking, in the process of registration most algorithms overlay objects spatially and temporally to obtain the optimal alignment. When there exists non-rigid motion and this motion is an essential feature for classification, the key issue is to align features associated with the non-rigid motion. As an example, in the human facial expression analysis field, the non-rigid muscle motion is of the central focus. However, accurate facial expression analysis is difficult due to the following aspects:

1. The facial muscle motion is non-rigid. In the real data facial expressions are coupled with

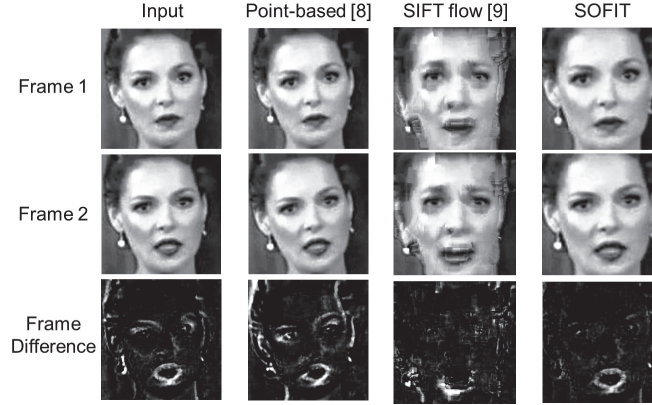


Figure 3.1: Comparison of registration results. Row 3 is the absolute difference of frame 1 and frame 2. Column 2 is the point-based affine registration method used in [65] [113], where affine transformation is computed from 68 facial feature points generated by the state-of-the-art detector [140]. Column 3 uses SIFT flow [67] to align with the Avatar Reference face model from [129]. Ideally, I would like the frame difference to show only at locations where the non-rigid motion is present (mouth area in this case). The proposed method, SOFIT, achieves the most plausible result.

rigid motion of head pose.

2. The head pose comprises of both in-plane rotation and out-of-plane rotation.
3. The muscle motion is subtle in real-world spontaneous expressions.
4. The data are streaming instead of being in a batch form.
5. The consecutive frames should comply with temporal smoothness constraint for micro-expression analysis.
6. The imaging condition is varying, such as the illumination or resolution of the face region.

In this chapter, I propose a video based registration approach, namely SIFT and optical flow image transform (SOFIT)¹, that tackles the aforementioned challenges in real-world datasets for facial expression analysis. Inspired by the philosophy of [131] where the alignment is done with respect to a reference face model, I transform every frame of the streaming

¹supplementary materials available at the project page: <http://ee.ucr.edu/~syang/sofit/index.html>

data onto a reference with canonical pose, expression, and illumination. This philosophy is significant in the facial expression analysis because facial muscle motion is similar for the same expression irrespective of the person [34], but the facial feature location (such as eyes, nose, mouth) of different people varies. Thus, finding a canonical reference feature location for all the faces is favorable for analyzing the dynamics of facial features across population.

As illustrated in Fig. 3.1, I consider registration of frame 2 with respect to frame 1. All methods in this figure are able to account for the in-plane head rotation. However, as seen in the frame difference images (row 3) for the point-based affine transformation (column 2) and the SIFT flow transformation (column 3), there is a motion on most parts of the face. This is similar to the original face image (column 1) where the image is the output of Viola-Jones face detector [118] and is not registered. This suggests us to impose the temporal smoothness constraint so that the frame difference is small for areas with no motion; while for areas with motion (mouth area in this case), the frame difference should capture this change, as demonstrated by the results of the proposed method (column 4).

In this chapter, I model the alignment problem by three parts. First, each frame is aligned with respect to a reference frame in a general distance measure. It is then instantiated to the SIFT flow criterion thereafter. Second, my model enforces a smoothness constraint on adjacent frames. It is realistic for the consecutive frames to comply with the smoothness constraint. I realize this by depending this current transformation estimation on a number of previous frames in an optical flow criterion. Third, large transformation is penalized to prevent overfitting.

3.2 Related Work and Contributions

3.2.1 Related Work

To analyze facial expressions, behavioral scientists have developed facial action coding system (FACS) [34] as an objective standard to describe the muscle motion. According to FACS, human (coders) can decompose every possible facial behavior into action units (AU), which roughly correspond to the muscles that produce them. Automatic AU recognition [135][113], has been quite successful for well-aligned, posed data, such as MMI [85] and CK+ [57] dataset. Unfortunately, AU recognition in an uncontrolled real-world environment remains a difficult problem, as shown in the Facial Expression Recognition and Analysis Challenge (FERA2011) [111] due to the difficulties mentioned in the introduction. Existing face registration approaches attempt to solve different aspects of the aforementioned challenges. In the face recognition and image retrieval communities, researchers attempt to get rid of the non-rigid motion from facial data through registration using an ensemble of images [62][50][89]. These approaches are not suitable for the facial expression recognition domain, where the following two criteria should be met:

1. Non-rigid facial muscle motion should be retained, which carries essential information for expression inference.
2. Facial feature should be aligned under various muscle motion, pose variation.
3. Low intensity facial muscle motion should be captured for spontaneous facial expression analysis

To align faces with expressions, the state-of-the-art systems [113], [65] track a set of anchor points on the face and estimate the affine transformation based on which the entire face is warped. Although the most recent facial point detection techniques [72][125][140] are able to

achieve outstanding detection results, there are two significant issues that needs to be addressed. First, the affine estimation is sensitive to small perturbation of point detection results. Typically in point-based method, a number of facial fiducial points (20 points in [113] and [72]) are detected. Each point carries much more weights in the estimation of affine matrix compared with methods that use corresponding information from the entire image. as demonstrated by Fig. 3.1.

Besides, affine transform parameter estimation by a small set of points can be susceptible to detection errors. In a realistic case where the resolution of the face is not high enough, the accuracy of feature point detection will also degrade. Yang and Bhanu [131] adopt SIFT flow technique [67] to align every frame to a reference face. An et al. [5, 8, 10] extend the work in [131] to warp face images towards a frontal face template, in order to achieve more accurate face matching. As shown in Fig. 3.1 column 3, the outcome of the SIFT flow transform has a large amount of discontinuities and artifacts. Although they solve this issue by generating image-based face representations (Emotion Avatar Image) and a reference model (Avatar Reference), carrying out the double layer loopy-belief propagation for every frame is computationally expensive and not suitable for real-time systems.

3.2.2 Contributions of This Work

The contributions of this work are summarized as follows:

1. I propose a novel real-time video-based object registration framework, SOFIT, that aligns the real-world data with non-rigid local motion, e.g., muscle motion from facial expression (Section 3.3).
2. SOFIT is an holistic approach and no detection of local features (eyes, nose, mouth) is needed. Therefore, it is tolerant to noise and low image resolution. The proposed method

results in temporally smooth image sequences.

3. I demonstrate versatility of my registration method in fields of spontaneous facial expressions analysis and existing AU recognition system performance enhancement, image super-resolution, vehicle recognition.

One of my previous work [130] also uses the idea of SIFT/optical flow approximation for alignment estimation. In this chapter, I propose a general framework to the alignment problem and rigorously show that how the final close-form solution is obtained. Unlike [130] where transformation estimation is propagated by affine matrix multiplication, the solution in this chapter shows that it is superior to propagated the flow field for transformation approximation instead. Due to the robustness of this framework, it is no longer desirable to greedily check the registration result for every frame, such as in [130]. Instead, a loop-closure rectification is used to correct the propagation error, which can be computed in parallel with the propagation computation for further speedup.

3.3 Flow-based Real-time Object Registration

The objective of this work is to align objects in video-based data in an uncontrolled environment. The original inputs to my system are faces detected by the Viola-Jones detector [118] for the analysis and illustrations in expression analysis domain. I first write the generalized model in a Bayesian framework. A flow-based approximation results in an efficient close-form solution. I also point out a dynamic programming implementation that will further optimize the registration algorithm.

3.3.1 The Generalized Model

Let $p = (x, y)$ be the grid coordinate of i -th frame of grayscale image, I_i . Given a sequence of N unregistered frames of a object, my goal is to align individual frame with respect to a canonical representation of this object I_c . Denote a generalized form of transformation function, $\Pi_i(\cdot)$, to register frame i , which I intend to solve. I assume the standard Gaussian measurement model

$$Q_i = \text{Dist}(\Pi_i(I_i), I_c) + w \quad (3.1)$$

where $\text{Dist}(\Pi_i(I_i), I_c)$ in Eq. 3.8 is the distance measure between the transformed frame I_i , i.e., $\Pi_i(I_i)$, and the canonical representation I_c . In this work, I attempt to align every frame w.r.t. the canonical representation such that they share similar structure. However, in general, it is applicable to many other distance measures. w is i.i.d. normally distributed zero-mean measurement noise. I also write a Gaussian transformation regularity model as

$$Y_i = \Pi_i(p) - p + m \quad (3.2)$$

it penalizes excessive transformation due to overfitting. m is also an i.i.d. noise process in zeros-mean Gaussian distribution. The joint probability of all variables can be written as

$$L = P(\Pi_{1:N}, Y_{1:N}, Q_{1:N}, I_c) \quad (3.3)$$

$$= P(Q_{1:N}, Y_{1:N} | \Pi_{1:N}, I_c) P(\Pi_{1:N} | I_c) P(I_c) \quad (3.4)$$

where $\Pi_{1:N}$ is short for $\Pi_1, \dots, \Pi_N$. Dropping the constant term and using the independence of my model definition in Eq. 3.1, 3.2, I obtain

$$L \propto P(Q_{1:N}|\Pi_{1:N}, I_c)P(Y_{1:N}|\Pi_{1:N})P(\Pi_{1:N}) \quad (3.5)$$

$$= \prod_{i=1}^N P(Q_i|\Pi_i, I_c) \prod_{i=1}^N P(Y_i|\Pi_i) \prod_{i=1}^N P(\Pi_i|\Pi_{1:i-1}) \quad (3.6)$$

where $\prod_{i=1}^N P(\Pi_i|\Pi_{1:i-1})$ can be viewed as the smoothness constraint. With the weakly coupled Markov assumption, I only take into account $H = \min(i, h)$ number of frames prior to frame i . It states that the aligned frame I_i should have similar appearance with its previous h neighbors (if $h < i$). Thus, the joint probability can be written as

$$L \propto \prod_{i=1}^N P(Q_i|\Pi_i, I_c) \prod_{i=1}^N P(\Pi_i|\Pi_{i-H:i-1}) \prod_{i=1}^N P(Y_i|\Pi_i) \quad (3.7)$$

$$= \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{1}{2\sigma} \text{Dist}(\Pi_i(I_i), I_c)\right\}. \quad (3.8)$$

$$\prod_{i=1}^N \frac{1}{\epsilon\sqrt{2\pi}} \exp\left\{\frac{1}{2\epsilon H} \sum_{j=1}^H \|\Pi_i(I_i) - \Pi_{i-j}(I_{i-j})\|_2^2\right\}. \quad (3.9)$$

$$\prod_{i=1}^N \frac{1}{\tau\sqrt{2\pi}} \exp\left\{\frac{1}{2\tau} \|\Pi_i(p) - p\|_2^2\right\} \quad (3.10)$$

Here, the smoothness constraint $P(\Pi_i|\Pi_{i-H:i-1})$ obeys zeros-mean Gaussian distribution. σ, ϵ, τ control the variance for the corresponding Gaussian distribution. Maximizing the likelihood, L , is equivalent to minimizing its negative log likelihood, E , where

$$E(\Pi_i) = \sum_{i=1}^N \text{Dist}(\Pi_i(I_i), I_c) \quad (3.11)$$

$$+ \frac{\alpha}{2H} \sum_{i=1}^N \sum_{j=1}^H \|\Pi_i(I_i) - \Pi_{i-j}(I_{i-j})\|_2^2 \quad (3.12)$$

$$+ \frac{\beta}{2} \sum_{i=1}^N \|\Pi_i(p) - p\|_2^2 \quad (3.13)$$

where the constant terms are dropped. $\alpha = \sigma^2/\epsilon^2$ and $\beta = \sigma^2/\tau^2$ can be considered as two scaling parameters on the smoothness term and penalty term, respectively.

3.3.2 The Flow-based Instantiation

Since my objective is structural matching, I opt to use SIFT flow [67] for similarity matching under large variation, non-rigid transformation. SIFT flow [67] was originally designed to align an image to its plausible nearest neighbor which can have large variations. The SIFT flow algorithm robustly matches dense SIFT features [68] between two images, resulting in a structural coherent image pairs. Thus, the data matching term in Eq. 3.11 can be instantiated in the coordinate space as

$$\sum_{i=1}^N Dist(\Pi_i(I_i), I_c) = \sum_{i=1}^N \| \Pi_i(p) - (p + f_s^i) \|_2^2 \quad (3.14)$$

where f_s^i (shorthanded for $f_s(I_i, I_c)$) is the SIFT flow field given by matching I_i to canonical frame I_c .

Regarding the smoothness constraint in Eq. 3.12, I consider the optical flow between frames. Optical flow compute the motion between two frames by matching the corresponding intensity values. In the context of video processing, it is reasonable to assume that the frame rate is high enough to compute accurate optical flow for consecutive frames. Thus, with pixel level correspondence, the smoothness constraint in Eq. 3.12 can be approximated as

$$\begin{aligned} & \frac{\alpha}{2H} \sum_{i=2}^N \sum_{j=1}^H \| \Pi_i(I_i) - \Pi_{i-j}(I_{i-j}) \|_2^2 \\ & \simeq \frac{\alpha}{2H} \sum_{i=2}^N \sum_{j=1}^H \| \Pi_i(p) - (p + f_s^{i-j} + f_o^{i-j,i}) \|_2^2 \end{aligned} \quad (3.15)$$

where $f_o^{i-j,i}$ is the optical flow field from frame $i-j$ to frame i .

Now, the cost function is written as the sum of three L-2 norm terms in Eq. 3.14,3.15,3.13.

I adopt L-2 norm because that the speed of algorithm is the main concern in this work, and I intend to derive close-form solution for the optimization problem. I further assume that the

transformation function is affine. As shown in my previous work [130], the computation of the X, Y -component can be decomposed, which creates space for speedup from parallel computation. Thus, the cost function is instantiated as

$$E(\Pi_i) = \sum_{i=1}^N \| T_i p - (p + f_s^i) \|_2^2 \quad (3.16)$$

$$+ \frac{\alpha}{2H} \sum_{i=2}^N \sum_{j=1}^H \| T_i p - (p + f_s^{i-j} + f_o^{i-j,i}) \|_2^2 \quad (3.17)$$

$$+ \frac{\beta}{2} \sum_{i=1}^N \| T_i p - p \|_2^2 \quad (3.18)$$

where T_i is the 3×3 affine matrix. With minor abuse of notion, p is now a horizontal stacked coordinate location template in size $3 \times m$, where $m = r \times c$, assuming that r and c are the image height and width, respectively. Each column of p is a coordinate point $(x, y, 1)$ in homogeneous coordinates. Taking the derivative of E w.r.t. Π_i and setting it to zero result in

$$T_i = (1 + \alpha + \beta)^{-1} \cdot \left(f_s^i + \frac{\alpha}{H} \sum_{j=1}^H (f_s^{i-j} + f_o^{i-j,i}) + \beta p \right) p^T (p p^T)^{-1} \quad (3.19)$$

It is observed from Eq. 3.19 that for an certain input frame i , I have to compute its SIFT flow w.r.t. the canonical reference frame. Computing SIFT flow for every frame is time-consuming. However, given accurate optical flow estimation between frames, I can approximate the SIFT flow computation of the i -th frame by the sum of the SIFT flow of the $(i-1)$ -th frame and the optical flow of $(i-1)$ -th to i -th frame, i.e.

$$f_s^i = f_s^{i-1} + f_o^{i-1,i} \quad (3.20)$$

This gives rise to the final close-form solution as

$$T_i = (1 + \alpha + \beta)^{-1}.$$

$$(f_s^{i-1} + f_o^{i-1,i} + \frac{\alpha}{H} \sum_{j=1}^H (f_s^{i-j} + f_o^{i-j,i}) + \beta p) p^T (p p^T)^{-1} \quad (3.21)$$

Eq. 3.21 is suitable for dynamic programming implementation where I cache the previous SIFT flow, f_s^{i-1} . For the current frame i , I only carried out several optical flow computations, i.e., $f_o^{i-H,i}, \dots, f_o^{i-1,i}$. When H is small, e.g., 3-5, the optical flow is accurate and the total amount of optical flow computation is small. Besides, individual optical flow can be parallel computed to further speed up the algorithm. What's more, I use iteratively reweighted least squares (IRLS) [51] to robustly estimate the affine transformation matrix.

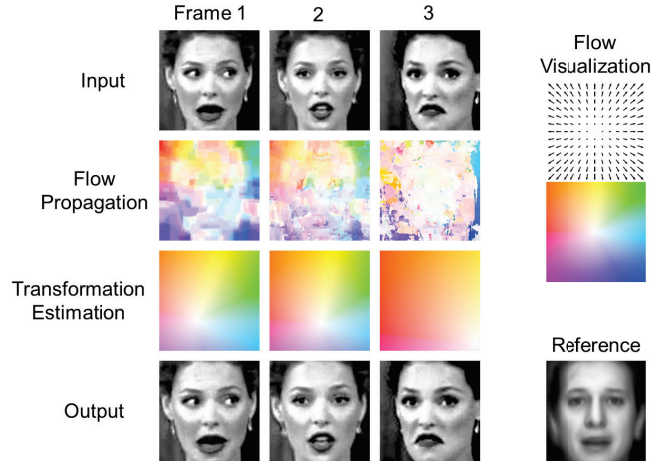


Figure 3.2: Registration Example. The input sequence is registered with respect to the reference frame shown in bottom right. The flow visualization is coded as in [14]. Better viewed in color.

An registration example for face is visualized in Fig. 3.2. The flow propagation in Eq. 3.21, i.e., $(1 + \alpha + \beta)^{-1} (f_s^{i-1} + f_o^{i-1,i} + \frac{\alpha}{H} \sum_{j=1}^H (f_s^{i-j} + f_o^{i-j,i}) + \beta p)$, is visualized in the second row for consecutive frames. The affine transformation is then robustly estimated for each frame. The output sequence are registered with respect to the reference frame and the it complies smoothness constraint.

3.3.3 Error Propagation and Close-loop Rectification

In my model, I sacrifice model optimality to efficient implementation. Therefore, the average registration error accumulates with time. The registration error is defined as the deviation from the canonical reference frame. Since I care about structural similarity, I compute the mean length of the SIFT flow from the current frame to the reference frame. For error analysis, I need videos with length of more than one minute (1800 frames in my case with 30 fps) to observe the noticeable cumulative error. Therefore, I register a long sequence from Youtube and plot the error in Fig.3.3. Although this error measurement consists global rigid head motion and local non-rigid muscle motion, I am still able to observe the error accumulation effect.

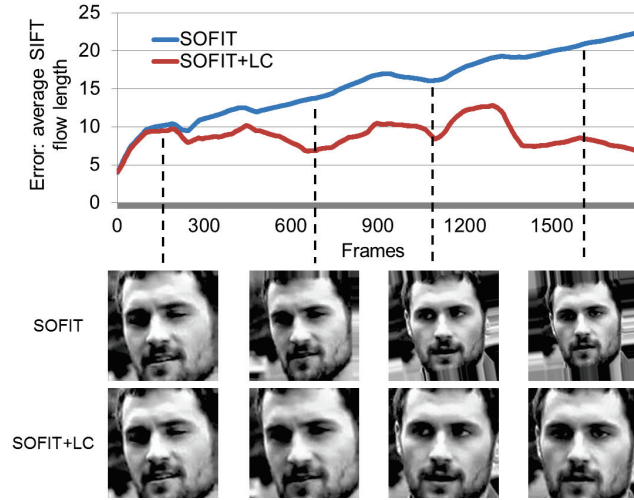


Figure 3.3: The error accumulation with respect to time. The error is defined as the SIFT flow of the current frame to the canonical reference frame. I use loop-closure (LC) to update the global flow estimation and rectify the error. The LC is carried out every 300 frames in this experiment.

To solve this issue, I intend to update the global estimation at a certain rate without affecting the propagation computation. Inspired by the loop-closure strategy in robotics [55], I update fs in Eq. 3.21 by recomputing the SIFT flow every 300 frames. This update frequency is chosen because cumulative error is negligible in the 300 frames, and this also provides enough

time for f_s to be updated in parallel and will not affect the overall flow propagation. As seen in Fig. 3.3, the error is rectified due to this global flow estimation update.

3.3.4 Computational Cost Analysis

I obtain considerable speedup from dynamic programming implementation for Eq. 3.21. In essence, the SIFT flow only needs to be computed for the initialization. The steps to be carried out for every subsequent frame are the following:

1. Compute the dense optical flow with respect to the previous H frame in parallel
2. Estimate affine transformation matrix using Eq. 3.21 based on the update coordinate value.

I adopt the optical flow in OpenCV [20]. The aforementioned steps can be finished in less than 150ms for a 100×100 image on a dual-core Intel i5 2.67GHz laptop with 4GB memory. Since the bottleneck of my method is optical flow computation, one can further speedup the algorithm by GPU implementation. The initialization takes on the average 1.6 second using the aforementioned settings. The computation is SIFT flow and optical flow are also in parallel. The loop-closure re-initialization is also in parallel with the the optical flow computation, and it will not affect the speed of the registration.

3.4 Experimental Results

3.4.1 Facial Action Unit Recognition

I demonstrate SOFIT face registration technique by facial action unit (AU) recognition on FERA Challenge dataset [2]. The goal AU recognition is to detect 12 frequently occurring AUs on a per-frame basis. I use the same protocol as the Facial Expression Recognition and

Analysis Challenge (FERA2011) [111] AU sub-challenge. The data I use for training is the GEMEP-FERA training dataset, which includes 87 sequences and around 5400 frames. The pose and gesture of the subjects in this dataset are uncontrolled, and therefore, this dataset is more realistic and complex compared to MMI [85] and CK+ [57] datasets.

The main concerns here are two folds:

1. Is registration really an important issue in real-world AU recognition in uncontrolled environment?
2. If yes, can I improve the recognition performance just by adopting a better registration algorithm, i.e., SOFIT?

To address these issues, I use the exact same feature as the baseline approach for better comparison. The only variable in my experiment design is using different registration methods. The baseline registration method detects both eye locations of the face, scale, and in-plane rotate the face. This registration belongs to in-plane image transformation category as summarized in [129]. The state-of-the-art registration technique is called Emotion Avatar Image (EAI) [131], [129]. It is another variation of SIFT flow method [67]. I have also included my earlier registration work [130] for comparison.

The feature extraction and classification are conducted the same as the baseline approach. After I extracted the face from Viola-Jones face detector [118], I resize them to 100×100 images and register them using the proposed method. Subsequently, I divide the image into 10×10 blocks, where the same type of Local Binary Pattern (LBP) [80] texture feature for each block is computed and concatenated. I then train 12 linear SVM binary classifiers based on the implementation of [23], each of which is trained independently regardless of the co-occurrence

Table 3.1: Person-independent AUC-score result on FERA-GEMEP AU training set. Feature: Local Binary Pattern

	Baseline	EAI	SOFIT	SOFIT
	[111]	[129]	[130]	
AU1	0.69	0.68	0.73	0.76
AU2	0.69	0.68	0.68	0.70
AU4	0.58	0.68	0.66	0.62
AU6	0.68	0.69	0.79	0.78
AU7	0.61	0.62	0.75	0.67
AU10	0.68	0.61	0.69	0.70
AU12	0.68	0.75	0.75	0.74
AU15	0.52	0.54	0.65	0.69
AU17	0.61	0.66	0.71	0.69
AU18	0.57	0.72	0.72	0.74
AU25	0.53	0.55	0.67	0.57
AU26	0.52	0.56	0.59	0.58
average	0.61	0.64	0.70	0.69

Table 3.2: Person-independent AUC-score result on FERA-GEMEP AU training set. Feature: Local Phase Quantization

	Baseline	EAI	SOFIT	SOFIT
	[111]	[129]	[130]	
AU1	0.67	0.67	0.72	0.77
AU2	0.63	0.69	0.70	0.69
AU4	0.57	0.55	0.65	0.67
AU6	0.78	0.74	0.79	0.81
AU7	0.61	0.58	0.66	0.73
AU10	0.61	0.65	0.70	0.73
AU12	0.76	0.78	0.78	0.81
AU15	0.56	0.65	0.62	0.69
AU17	0.64	0.63	0.66	0.68
AU18	0.72	0.74	0.76	0.73
AU25	0.53	0.54	0.56	0.64
AU26	0.53	0.54	0.63	0.59
average	0.63	0.65	0.69	0.71



Figure 3.4: Qualitative result comparison. Row 1 and 2 are the first and fifth frame of a sequence. Row 3 to 5 are the cumulative absolute frame difference of 5 unaligned frames using method SOFIT, SIFT flow [67], point-based affine where points are detected using deformable part-based model (DPM) [140], respectively. Row 6 are without alignment. The proposed alignment technique captures the correct non-rigid motion of face, for example, eyebrows raise for first column and mouth open for second column.

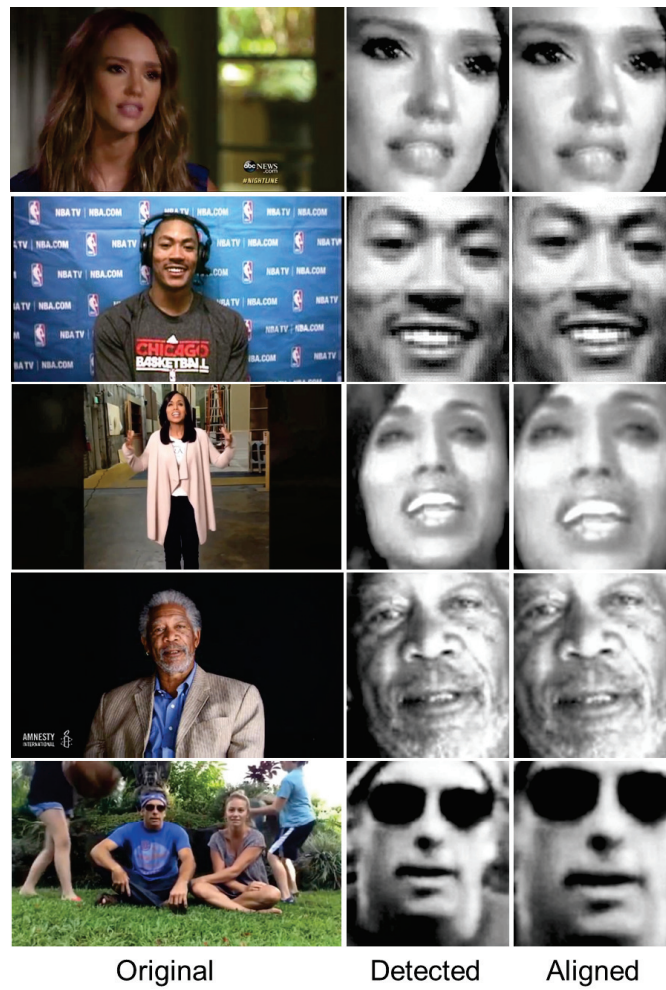


Figure 3.5: More results using SOFIT registration technique.

of AUs.

For registration methods with reference frame, i.e., EAI [129], SOFAIT [130], and SOFIT, I use the level-1 Avatar Reference [129] generated from FERA Challenge training data [2]. In generating a EAI representation, I need to select a temporal length parameter. In the original EAI paper [131], I choose the length of a single video (around 2 seconds) for this parameter for facial expression recognition on a per-video basis. To generalize this registration technique in AU recognition on a per-frame basis, I heuristically determine the best value for the temporal length parameter. I carry out a leave-one-subject-out cross validation on the FERA-AU training data, and determine the parameter value to be 0.56 second for the best F1 score over all AUs. This means for each frame in a video, approximately 14 closest frames will be used to compute EAI representation. For the boundary frames, i.e. the starting and ending 7 frames, I simply assign their values to be the 8th frame from the beginning and the 8th from the end, respectively. Thereafter, the aforementioned features are extracted from the EAI representations.

Since the ground-truth label is only available for the FERA-AU training set, I carried out a person-independent cross validation experiment, where no test subject is used for training, and the average performance is reported. Due to the finite scale of the training exemplars, person-independent test is essential to demonstrate the generalization ability of an approach to unseen subjects. Table 3.1 and 3.2 shows the performance measured by local binary pattern (LBP) and local phase quantization (LPQ). The score is area under curve (AUC) of the receiver operating characteristic (ROC) curve. As seen from Table 3.1 and 3.2, SOFIT and SOFAIT outperform the baseline and EAI registration method. Although EAI performs well in discrete emotion recognition [129], in a per-frame based setting, however, EAI lacks ability to reveal the subtle motion. When using LBP feature (Table 3.1), SOFIT and SOFAIT are comparable and SOFIT performs better in the case of using LPQ (Table 3.2). SOFIT with LPQ feature obtains the highest score in 9 out of 12 categories.

I should point out that the video sequences in FERA-GEMEP is relatively short, i.e. around 2 seconds. Thus, the propagation error is not high and the appearance of registration results are similar using SOFIT and SOFAIT in most of the cases. In a longer video however, e.g., Fig. 3.3, the error of SOFAIT rises faster than SOFIT. Moreover, I also observe similar reference-independent effect as shown in [130], where performance will not degenerate as the reference frame changes.

Fig. 3.4 shows the qualitative evaluation why my registration improves the baseline and EAI approaches. I compute the absolute frame difference of the first 5 frames for both unaligned and aligned faces. As shown in the third and fourth column of Fig. 3.4, the unaligned frame difference reveals motion mainly caused by the edge feature of a face, while after alignment, the non-rigid muscle motion is retained. More visual results are shown in Fig. 3.5.

3.4.2 Rear-view Vehicle Recognition

In the application of surveillance and traffic monitoring, the ability to classify the type of a vehicle is essential for vehicle identification [56], circumventing problems such as license spoofing. Current work on vehicle recognition [90] [79] [133] [87] rely on license plate for alignment. Matching a video of an unknown vehicle with database faces various challenges such as view variation, illumination change, etc. Video data were collected from a camera mounted on top of a freeway lane over several days, during which 1664 vehicles were collected. The collected rear-view vehicle data fit in the context of my dense-flow based registration method, i.e., object shapes are similar and the structure/texture variation is key for recognition.

Three individuals manually labeled four vehicle type, namely sedan, minivan, pickup, and SUV. The ground-truth is set to be the majority label by all three individuals. As some of the vehicles are hard to be identified, and there is a tie between different individuals (three different



(a) First row: first frame without alignment; second row: the average of individual frames aligned by SOFIT



(b) First row: first frame aligned by SIFT-flow affine; second row: the average of individual frames aligned by SOFIT

Figure 3.6: Sample registration results. The first row of both (a) and (b) are misclassified examples by SVM while the second row are correctly classified. In (a), SOFIT is able to align detected vehicles to a similar scale. In (b), SOFIT mainly corrects the missing bottom and align similar parts of the vehicle according to their structure. The bottom right shows the avatar reference frame generated by [129] used in the initialization step.

classes are picked by the three individuals), the data is not included. Finally 1505 vehicle with ground-truth label are used as my data for evaluation. Vehicle ROI were detected by the moving object detection method from [104]. ROIs are then refined by removing shadows and enforcing bilateral symmetry [105]. For each vehicle sequence, there are approximately 10 frames. Fig. 3.7 shows a sample frame illustrating a detection result.

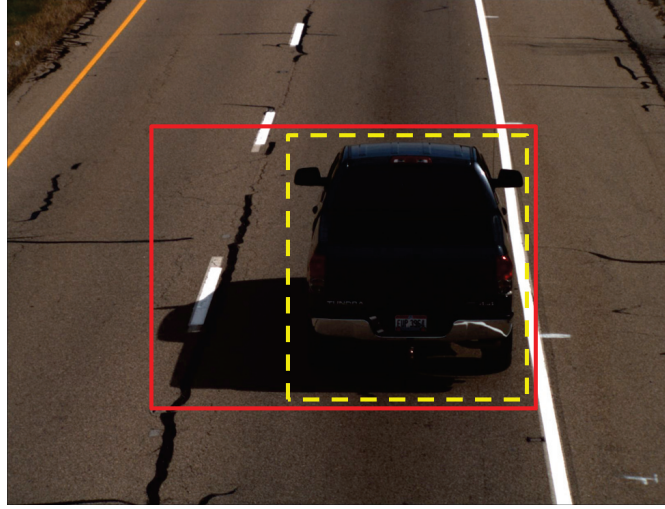


Figure 3.7: Sample vehicle detection result. Red rectangle is the result of vehicle ROI detection and yellow dash rectangle is the result after enforcing bilateral symmetry.

Due to my camera setup, the recorded vehicle is moving away from the scene. Therefore, I consider the first frame of the video as the frame where the vehicle is fully observed. As I can imagine, the first frame is the one with the best quality of the entire sequence. Three methods in comparison are used to generate image representation of ROIs from each video. (a) First frame without alignment, (b) First frame with SIFT-flow affine alignment (namely the initialization step of my approach), (c) The median representation of the entire sequence aligned using SOFIT. I demonstrate that by simply taking the mean of the entire aligned sequence, the representation is superior to just using a single frame. The reference frame in use is the also the level-1 Avatar Reference generated by following [129]. Again, it is essentially an image based summary of the entire vehicle in canonical appearance. Sample alignment results are shown in

Fig. 3.6.

The ROIs are all resized to 200×200 for all three methods. The locally normalized Harris strength (LNHS) features [87] is used as the texture descriptor. I adopt the VLFeat [117] implementation for $\sigma = 2.0$. The Harris strengths are split into 32×32 cells and normalized in 3×3 blocks. Descriptions are generated by shifting block by 1 pixel. I then use linear Support Vector Machine [88] as the classifier without parameter tuning (i.e., $C = 1.0$). A 10-fold cross validation is carried out and the result is reported in Fig.3.8. The F1 score shows that SOFIT alignment is able to leverage image sequences and generate a robust representation for superior classification accuracy.

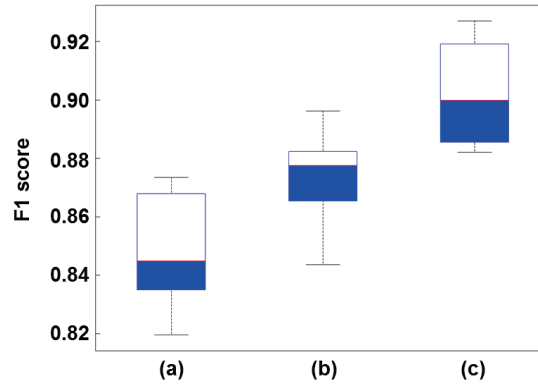


Figure 3.8: The boxplot of the 10-fold cross validation comparison. (a) unaligned, (b) aligned with initialization, (c) SOFIT.

3.4.3 Multi-frame Image Super-resolution

In the imaging process, it is common to acquire images with low-resolution (LR) and/or certain artifacts such as blurriness, noise, etc. Image super-resolution (SR) is the process of generating a high-resolution (HR) image from one or more low-resolution inputs. In the past few decades, there has been extensive work on super-resolution methods. Based on the inputs, the SR algorithms can be classified in two categories : single-image based [100, 7,

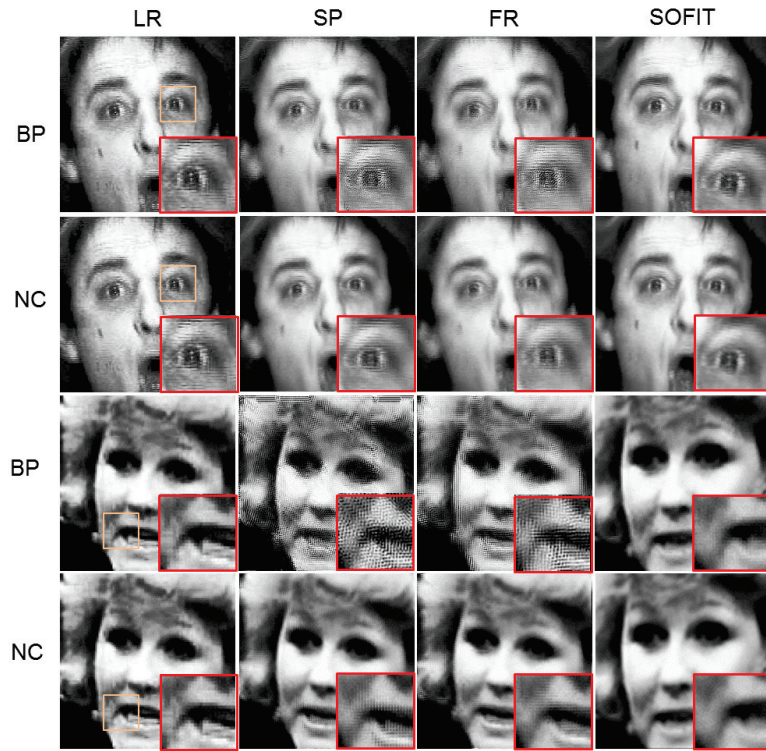


Figure 3.9: Comparison of super-resolution results using different registration methods for 2 subjects. For each row from left to right: one of the LR inputs (enlarged by pixel replication), sub-pixel registration (SP) [58], frequency domain based registration (FR) [116], and the proposed registration method. I use two SR methods to reconstruct the high-resolution outputs: iterated back-projection (BP) [53] and normalized convolution (NC) [91]. The red blocks show the magnified parts from the pink blocks.

12, 9] and multi-image based methods [53]. I apply SOFIT registration algorithm proposed in this paper to generate aligned images as inputs to different multi-image based SR methods. Here I compare my registration method with two registration methods: frequency domain based method (FR) [116], and registration using sub-pixel displacement (SP) [58]. These registration methods are then used in two SR methods: iterated back-projection (BP) [53], and normalized convolution based method (NC) [91]. Figure 3.9 shows the comparison of some sample results using different registration methods in different SR algorithms.

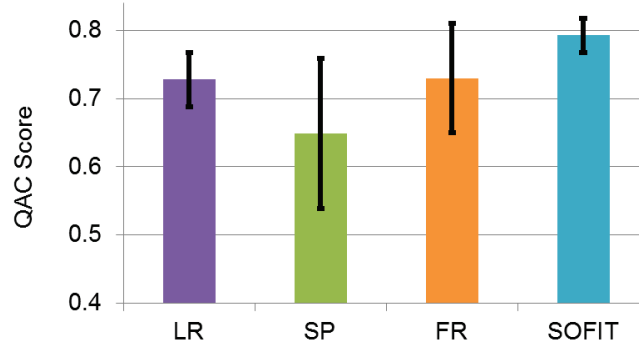


Figure 3.10: Image quality comparison between video-based super-resolution results and the proposed method. A recent non-reference image quality assessment method [126] is used. The higher the score, the better the estimated visual quality is. LR is low-resolution input images. SP denotes the super-resolved results using [58]. FR is the super-resolved results using [116]. These benchmark methods are compared with the proposed SOFIT method.

From Fig. 3.9 I see that with my registration method, the SR results are significantly improved over other SR methods in terms of visual quality. Despite the poor quality of the input frames, the results by my method are smooth with much fewer artifacts (e.g., noise, blockiness). The gain on the performance of SR directly comes from the accuracy of the proposed registration method. The output images by my methods are also well rectified which would be desirable for future processing purposes, such as AU and expression recognition.

To quantitatively evaluate the image quality using my registration method, I compute a recent proposed non-reference image quality index, Quality-aware clustering (QAC) [126],

for output images using my method and competing super-resolution methods. QAC is a general purpose blind image quality assessment method that has high linearity to human perception of image quality. Fig. 3.10 lists the average image quality scores on 87 sequences from GEMEP-FERA database [2]. Compared to the LR input and output using different super-resolution methods, the output of SOFIT achieves the highest scores with lowest standard deviation, which indicates better visual quality through this general quantitative measure.

Chapter 4

Zapping Index:

Using Smile to Measure Advertisement

Zapping Likelihood

4.1 Motivation

In recent years, multimedia data (e.g., images, videos, audios) on the Internet keep on increasing at a phenomenal rate. For example, 72 hours of video data are uploaded to the YouTube every minute [42]. More and more people tend to spend time watching videos on the Internet instead of using the traditional media such as TV. In addition, with the vastly growing popularity of mobile devices (e.g., smart phones, tablets), easy Internet access continues to attract more traffic on mobile networks. As predicted, in 2017 the video contents will account for 66% of all mobile data traffic¹.

The popularity of the Internet videos implies a huge potential for online commercial

¹http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html

advertisement (ad). The marketing expenses for commercial ads on the Internet are growing. For instance, the cost of a 30-second commercial on TV at prime time in the US was around 0.5 million US dollars in the fall of 2012². At some specific venues, the cost of commercials may be much higher. As an example, the cost of a 30-second commercial in the Super Bowl event in US has hit 4 million US dollars in 2013. With the increased advertising cost on TV and decreased audiences, marketers are gradually switching their focus to online advertising, in favor of their large audience base and lower cost to publish.

As a well-known example, the TrueView in-stream advertising [86] is a popular online advertising tool by Google Inc. The ad is shown prior to the video requested by the user. The user has the option to skip the ad and move directly to the desired video after 5 seconds of viewing the ad. The advertisers are billed if a user watches the ad at least for 30 seconds or the complete ad, if it is less than 30 seconds long. In such a case, for the online media provider (e.g., Google) to obtain the maximum profit and for the advertiser to reach the widest audience and achieve the advertising goal, it is their common interest to draw viewers' attention to the online commercials.

In marketing and advertising research, to evaluate the attention to the commercials, zapping is considered as an important topic [38], if not the most important one. Commercial viewers often have the option to 'zap' a commercial by either switching channels or simply turning off the source. The action of zapping indicates that the viewer is no longer interested in the commercial and this behavior means the loss of a consumer for the advertiser. To evaluate the effectiveness of advertising, several methods can be adopted. Self-report, which registers a respondent's subjective feeling, suffers from an important limitation referred as "cognitive bias", and may not always be able to capture lower-order emotions in an accurate way [92].

Facial expression is one of the richest source of communication [35]. Automatic facial

²<http://domainestimations.com/?p=14174>

expression recognition finds its applications in human behavior analysis, human-human interaction and human-computer interaction. Automatic facial expressions analysis is non-intrusive and can be dynamically analyzed as a commercial is playing [103]. Accurate facial expression analysis facilitates the marketing and advertising researchers in understanding a user's emotional state and behavior. This has the potential to improve the effectiveness of advertising or even design interactive commercials to enhance the advertising experience.

Recently, smile has been demonstrated as an useful indicator of a user's preference of commercials [75]. Teixeira et al. [103] develop a statistical approach using facial expressions to study advertising and they find that surprise and joy are effective in retaining a viewer's attention. Furthermore, applying machine learning and data mining techniques to advertising research enables us to exploit the underlining relationships between commercials and users by performing experiments with large data.

In this chapter, I attempt to understand a user's behavior in watching an ad. I make prediction on a user's zapping probability and provide guidance to ad publishers and advertisers. This can benefit ad publishers (such as YouTube) to understand the user's reaction to a certain commercial and, therefore, decide its value. Besides, this can also benefit advertisers so that they have an evaluation tool to analyze the feedback of their ad. Advertisers can leverage this behavior feedback to make better commercials.

I propose a measurement called Zapping Index (ZI), which is a prediction of the moment-to-moment zapping probability when an user is watching a commercial. The motivation for developing ZI is the following:

- The need for marketing metrics is well recognized. A survey of CEOs shows that CEO's top concern about marketing was the lack of performance metrics [52]. ZI creates a new metric for marketer and advertisers.

- ZI helps to study the affective behavior of an audience.
- ZI helps to improve the effectiveness of an ad.

To calculate the ZI, I opt to use smile response and set it apart from other facial expressions. As demonstrated in [38], entertaining information has a strong relation to zapping. Smile is a reflection of joy and happiness triggered by entertainment. Moreover, current computer vision algorithms perform well on automatic smile detection [96] [122].

4.2 Related Work and My Contributions

4.2.1 Automatic Facial Expression Recognition

Facial expression recognition techniques can be broadly divided as geometric-based approaches, appearance-based approaches, and the combination of both.

Geometric-based approaches track the facial geometry over time and infer expression based on the facial geometry deformation. Some exemplar methods include: Active Shape Model (ASM) [48], Active Appearance Model (AAM) [71], particle filter [115], geometric deformation [61]. Appearance-based approaches, on the other hand, emphasize on describing the appearance of facial features and their dynamics. Whitehill et al. [122] use a bank of Gabor energy filters to decompose the facial texture. The volume of local binary patterns (VLBP) is extracted in [135]. Yang et al. [129] aggregate the dynamics of the facial expression into a single image, Emotion Avatar Image (EAI), for high accuracy person-independent expression recognition.

4.2.2 Zapping Analysis

The attention paid to a commercials determines the interests of the audience in that commercial and only those commercials that retain a viewer's attention can produce desired

communication effects [103]. As the consumers have a choice to switch away from either a TV commercial or an online video commercial, it is challenging for the advertisers to retain consumers' attention during the course of a commercial [38]. The term zapping implies that the receiver of a commercial is no longer interested in its content/presentation, thus opt not to continue watching the commercial. In [46] a hierarchical Bayes approach is used to analyze the dynamics of attention to TV commercials. It investigates how the likelihood of a commercial zapping varies with time and shows that across-ad heterogeneity in zapping is related to the underlying characteristics of the commercial. Elpers et al. [38] demonstrate that both the entertainment and the information value of a commercial have a strong multiplicative effect on the probability for a commercial to be watched by viewers. Two experiments with a total number of 190 subjects and 45 commercials were conducted to support this finding. Kooij et al. [59] show that zapping has influence on end-user's Quality of Experience (QoE) for Internet Protocol television (IPTV). Further study on zapping is conducted in [97] and various solutions are proposed to reduce zapping to keep the user staying with the IPTV broadcasting. Teixeira et al. [103] incorporate joy and surprise expression recognition from a Bayesian Neural Network classification system to analyze the user's zapping decision. They conclude that the velocity of the joy response highly impacts the viewer's zapping behavior. But they have not made any prediction on the moment-to-moment zapping probability.

4.2.3 Contributions

The contributions of this work are summarized as follows:

1. I introduce an accurate person-independent video-based smile detection method. The smile response intensity is well transformed from a probability score from 0 to 1.
2. I perform zapping detection/classification in a non-intrusive manner based on facial ex-

pression cues.

3. I propose a novel metric called Zapping Index (ZI) for ad evaluation. ZI is a moment-to-moment prediction of a user’s zapping probability.
4. I collect a database, named AdEmotion, for the analysis of zapping prediction and user preference evaluation. I demonstrate the usefulness of ZI in measuring user preference. This results in advices for both ad publishers as well as providers about the effectiveness of an ad. The AdEmotion dataset will be publicly available in the near future on my website.

4.3 Technical Approach

In this section, I first introduce how moment-to-moment smile detection is carried out. Subsequently, I describe the data collection procedure. I formulate the problem of distinguishing zapping from non-zapping as a binary classification problem. After analyzing the characteristics of the data, I propose a new feature which is a temporal histogram of the smile measurement over time. I adopt SVM classifier to train the zapping classification model, which is then used to generate the Zapping Index.

4.3.1 Smile Detection

The goal is to compute the probability of smile on a per-frame basis. The faces are first extracted using Viola-Jones face detector [118]. I then follow my previous work [130] to align the faces using dense flow-based similarity registration technique. This registration algorithm aligns every frame with a face to a reference face and the registration results are temporally smoothed. Thus, the person-independent spontaneous facial expression recognition can be carried out in a meaningful manner. The aligned faces which are scaled to 200×200 pixels,

are divided into 20×20 pixel regions. The Local Phase Quantization [81] texture descriptor is computed for each of the regions. These outputs are then concatenated to form the feature for smile detection.

The smile detection is formulated as a binary classification problem with the smiling face and neutral face being the two class labels. I adopt the linear Support Vector Machine (SVM) [23] for classification. For accurate person-independent smile detection, the classifier is trained on multiple databases with a large number of subjects from: FEI [106], Multi-PIE [45], CAS-PEAL [43], CK+ [70], and data from Google image search similar to [13]. In total, 1543 subjects (1543 smiling faces and 2035 neutral faces) are included for training.

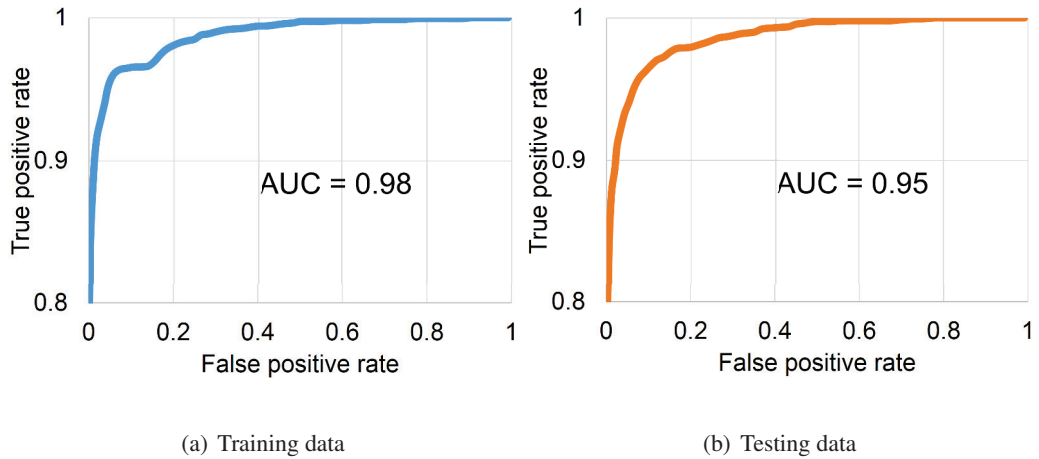


Figure 4.1: ROC curve for my person-independent smile detection algorithm.

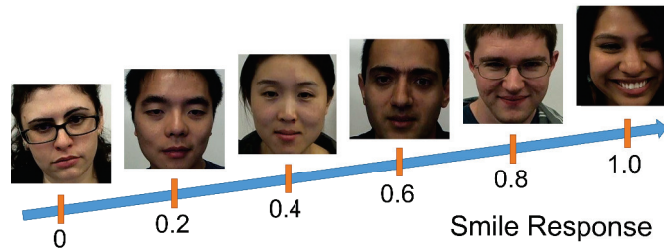


Figure 4.2: Sample smile response results. The response value reflects the intensity of smile.

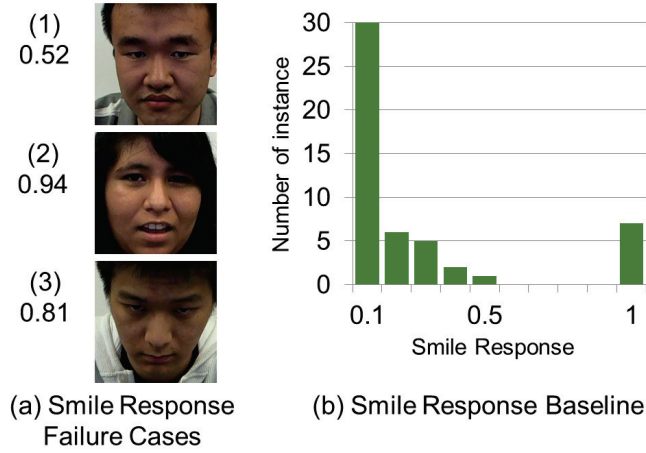


Figure 4.3: Handling smile detection failure. (a) Smile response failure cases where (1) and (2) exist because the subjects’ neutral faces are similar to smiling faces; (3) is due to out-of-plane rotation. (b) The baseline response of each subject is used to eliminate the failure cases automatically.

During the training, the smile detection is carried out in a person-independent manner and no test subject is included during training. The Area Under Curve (AUC) is 0.98 for the 10-fold cross validation (see Fig. 4.1(a)). To demonstrate the generalization of this classifier, I carried out a test on a selection of 10,000 sample frames from my AdEmotion database (see Section 4.3.2) that I collected in this research. The Area Under Curve (AUC) is 0.95 in Fig. 4.1(b), which means that the smile classifier performs well on the new data. The probability output of the SVM smile classifier is then recorded as the smile response. Since I structure *smiling* vs. *neutral* instead of *smiling* vs. *non-smiling* classification, an interesting finding is that, the classification rate of test data is not only superior, but also the probabilistic output of smile detector is able to capture the smile intensity as illustrated in Fig. 4.2. The reason I choose smile vs. neutral expression classification setup in this experiment is that the subjects are concentrated on the viewing experience. Most of the expressions other than smile are of neutral nature, and very few subjects display excessive non-smile expression. In this case, the probabilistic outputs closely correlate to the smile intensity even when it is low. One thing worth noting is that, there are neutral examples with open mouth in the training data, and therefore, the

classifier is not just naively predicting random mouth motion but rather muscle motion caused by smile.

For proof of concept, I have verified the probabilistic outputs with the manually annotated smile intensity results. I have gathered three annotators, and each is given 500 frames sampled from the entire AdEmotion data. The annotators score the smile intensity of each frame by comparing it with the reference figure similar to Fig. 4.2. The median value of all three annotators is selected as the ground-truth smile intensity to mitigate the effect from a large discrepancy among annotators. The resulting absolute mean error intensity is 0.216 between the prediction and ground-truth.

Some failure cases are shown in Fig. 4.3(a). In order to eliminate the subjects whose smile response is inaccurate, I leverage the fact that the expression for a subject is distributed around the neutral as I mentioned earlier. Therefore, for each sequence, I quantize the smile response to 0.1 accuracy, and take the *mode* of the quantization to approximate the baseline expression response for a subject. As a result, all the 7 error cases, whose smile baseline is 1, are able to be separated as shown in Fig. 4.3(b).

4.3.2 AdEmotion Data Collection

Participants were seated in front of a 23 inch monitor, with a Logitech c910 webcam mounted on the top of the monitor. The webcam resolution is set to 960×544 pixels. The average resolution on face is approximately 220×220 pixels. Participants were shown a series of 8 video ads in random order selected from 2 categories shown in Table. 4.1. The length of the ad ranged from 30 to 90 seconds. Participants were instructed that they could watch each ad until the end or zap at any moment by clicking on the skip button. In either situation, participants were given a 30 seconds break to reduce the emotional effect to the subsequent ad watching experience. They were also given a questionnaire during each break that contained the

following questions:

1. Did you like the commercial?
2. Did you skip the ad?
3. Why did you skip? Mark all that applies:
 - The ad is not funny.
 - The ad is not informative.
 - I have seen this ad before.

There has been research [38] that shows that the lack of entertainment and information factors are the two major reasons for zapping. I design these questions in order to analyze different aspects of this dataset. In this work, my focus is on using the response provided by subjects to verify the predictions from the zapping index, which will be discussed later in detail.

The entire data collection procedure lasts 8 minutes on the average for each participant, and no one is interrupted during this procedure. The participants' facial behavior during the entire procedure is recorded by the webcam at 24 fps.

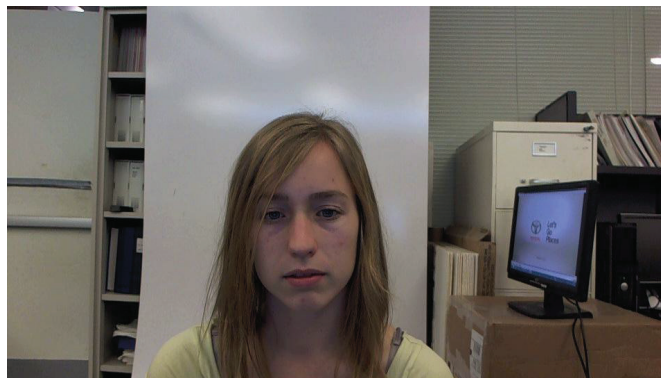


Figure 4.4: The data collection environment. A duplicate monitor is used for data synchronization.

During recording, there is a secondary monitor behind the participant, which displays the same content watched by the participant. The recording camera is able to capture a subject's facial expression as well as the corresponding content that he or she is watching. I designed this setting for data synchronization. In order to analyze the facial expression responses of different participants with respect to certain ad, I manually separated the expression data according to the ad information shown on the duplicate monitor. No data during the interval of two ads is used in this work. The setup of data collection environment is shown in Fig. 4.4.

The ads shown in Table 4.1 are selected by the following criteria:

1. Popular category: The *Fast Food* and *Car* are the two categories that almost everyone is familiar with and well connected to in the United States.
2. Minimum gender bias: I do not consider the gender effect in this research. The selected ad categories have less gender bias compared with categories such as *Beer* or *Makeup*.
3. Recognizable brand: Since online video user is the target market. I select the ad from brands that either have their official YouTube Channel or participated in the YouTube ad campaign. In this way, I have access to the ad for this research.
4. Varying entertainment levels: I have carefully evaluated the entertainment information in each ad. My final ad selection includes both kinds of ads that are very amusing and that are less entertaining.

55 college students have participated in my data collection. There are 31% female, 40% Asian, 25% Euro-American, 16% Afro-American, and 19% other ethnicity groups.

Table 4.1: Advertisement Selection

Category	Brand	Ad Name
Car	Toyota	I Wish
	Honda	I Know You
	Chevy	Wind Test
	Nissan	Enough
Fast Food	Jack In The Box	Hot Mess
	Subway	New Footlong
	Carl's Jr.	Oreo Ice Cream
	Pizza Hut	Make It Great

4.3.3 Data Characteristics

I analyze the characteristics of AdEmotion dataset in terms of zapping distribution and smile response. These characteristics are essential in motivating my zapping classification feature. I could potentially design systems that recognize other facial expressions. However, in this application, subjects concentrated on watching ads, and therefore, the dominant facial expressions are neutral and smile. This is also demonstrated to be true in an “in-the-wild” ad-watching data, namely AMFED [76]. Therefore, in light of the idea of Occam’s razor, I have focused specifically on the smile expression specifically in this work.

4.3.3.1 The Zapping Distribution

Since participants are given the option of zapping at anytime, I show the distribution of fraction of an ad that is being watched in Fig. 4.5. In other words, it is the distribution of the portion of the ad that has been watched. I fit a Gaussian mixture model with two components

to the distribution and find that 90% of the ad fraction is the best value to separate the two components of the mixture. In Fig. 4.5, the probability is dramatically higher in 90% to 100% range. This means that a large portion of the ads have been watched until the end. In the 0% to 90% range, the first half (0% to 45%) has a slightly higher probability than the second half on the average. This informs us that subjects in my experiments tend to zap early if they do not feel like watching an ad.

One interesting fact that is worth noting is that the popular TrueView advertisement publisher only bills the advertiser if an ad has been watched for more than 30 seconds [86]. If I have a better understanding of the zapping behavior, I can create a win-win-win situation: the user receives more desirable video content; the advertiser obtains more attention from users; and the publisher (such as YouTube from Google) gains more revenue.

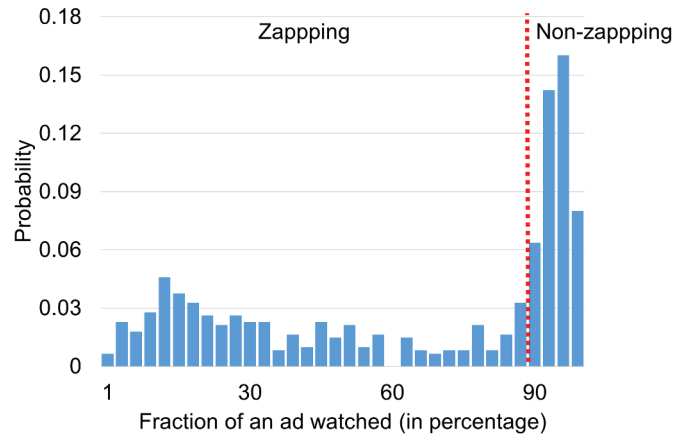


Figure 4.5: The zapping distribution. The data-driven threshold at 90% is used to separate the data into zapping and non-zapping classes.

Thus, based on the zapping distribution (see Fig. 4.5), I define two different classes: zapping and non-zapping, and use 90% of the ad length as the threshold in separating these classes. Given the facial expression response of a user watching an ad, one of my goals is to determine in an automated manner the class of the sequence. This can be formulated as a binary classification problem where zapping is the positive class and non-zapping is the negative class.

The analysis of the class characteristics in Section 4.3.3.2, 4.3.3.3, and 4.3.3.4 provides us the motivation for the feature used in zapping classification.

4.3.3.2 The Mean Smile Response

I conduct person-independent smile detection as described in Section 4.3.1. I present my motivations to the feature selection for zapping classification, and ultimately, establish a strong correlation of the zapping index from my predictor and the viewer's zapping behavior.

I analyze the average smile response in the first 30 seconds (720 frames from 24 fps webcam device) for both zapping and non-zapping classes. In Fig. 4.6, the moment-to-moment mean smile response is bounded by the positive and negative standard error of the mean (SEM). Since the user can zap at any time, the sequence are of various lengths. Therefore, the average smile response is computed as follows:

$$r_m(t) = \frac{\sum_{i=1}^N r_i(t)}{\sum_{i=1}^N I_i(t)}, \quad I_i(t) = \begin{cases} 1, & \text{if } r_i(t) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

where $r_i(t)$ is the smile response of sequence i at time t , N is the number of sequences, $I_i(t)$ is the indicator function to check the existence of the smile response of sequence i at time t . In other words, for each frame, the average smile response is computed based on the available responses. Using the similar idea, I compute the SEM bound by:

$$sem(t) = \frac{std(r(t))}{\sum_{i=1}^N I_i(t)} = \sqrt{\frac{\sum_{i=1}^N (r_i(t) - r_m(t))^2}{\sum_{i=1}^N I_i(t)(\sum_{i=1}^N I_i(t) - 1)}} \quad (4.2)$$

where $r_m(t)$ and $std(r(t))$ are the mean and the standard deviation of available smile responses at time t , respectively.

In Fig. 4.6, the smile response level for the two classes is initially about the same. Thereafter, the response of the non-zapping class increases for the rest of the 30 seconds. On

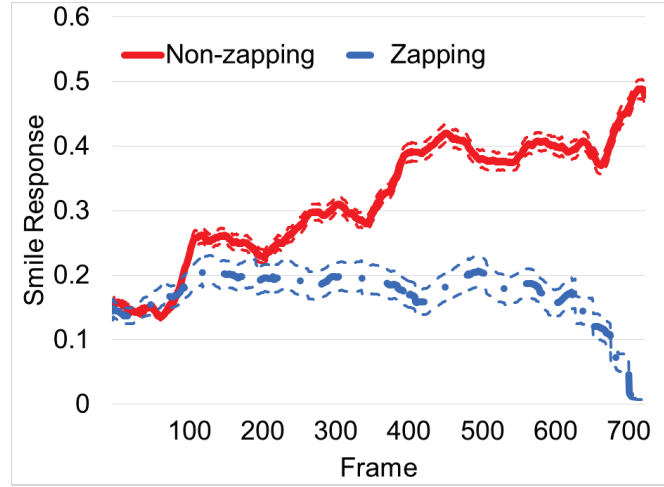


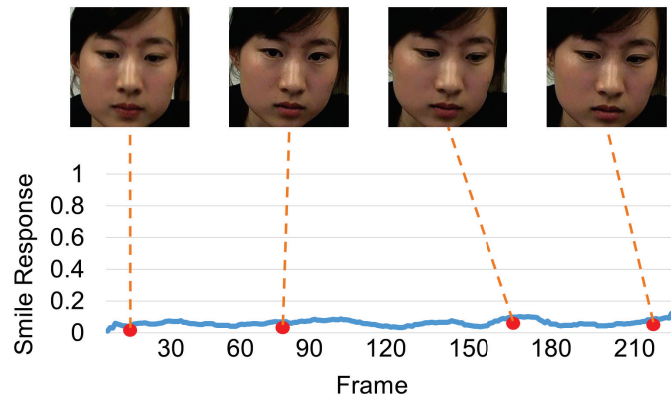
Figure 4.6: The average smile response of zapping and non-zapping classes for the first 30 seconds (720 frames). Each sequence is bounded by its standard error of the mean (SEM).

the contrary, for the zapping class, the response remains around 0.2 and decreases toward the end. *Therefore, the moment-to-moment average smile response is a good feature to separate zapping from non-zapping class.* This observation is also in line with the conclusion in [103] that smile level largely correlates with the zapping behavior.

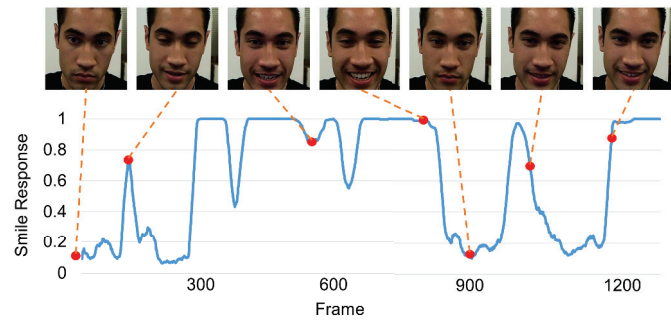
4.3.3.3 The Maximum Smile Response

The maximum smile response of the sequences is also different for zapping and non-zapping classes. Two examples are shown in Fig. 4.7.

I plot the distribution of sequences from the two classes based on their maximum smile response in Fig. 4.8. The total probability of each group sums up to 1. As illustrated in Fig. 4.8, if a sequence's maximum smile response is above 0.5, then the chance is higher that it belongs to the non-zapping class, and vice versa for maximum smile response below 0.5. The probability reaches the highest for the non-zapping class if the maximum smile response is above 0.9. On the contrary for zapping class, majority of the sequences are with the maximum smile response less than 0.1.



(a) zapping class



(b) non-zapping class

Figure 4.7: Sample frames of smile response from zapping and non-zapping classes.

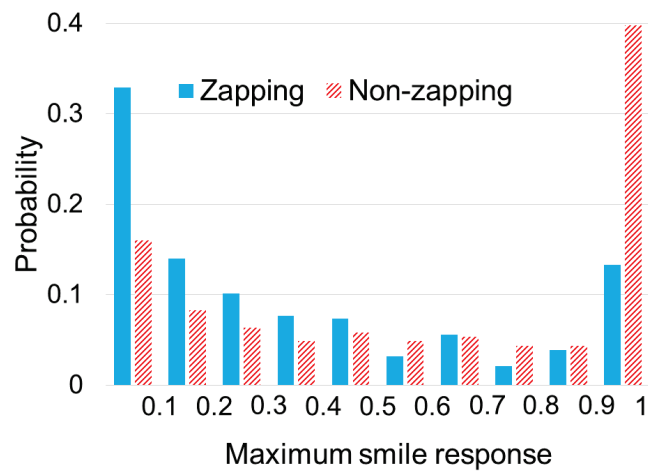


Figure 4.8: The distribution of the zapping / non-zapping data based on their maximum smile response.

For non-zapping class, the probability is the second highest (15.5%) when the smile response is less than 0.1. Observations on my data show that a few participants watch the entire ad but display minor smile expression. This means that entertaining content is not the only reason to keep the user engaged. Besides, the interview of participants also shows that a small group of people enjoyed the ad but prefer not to show their feelings through facial expression.

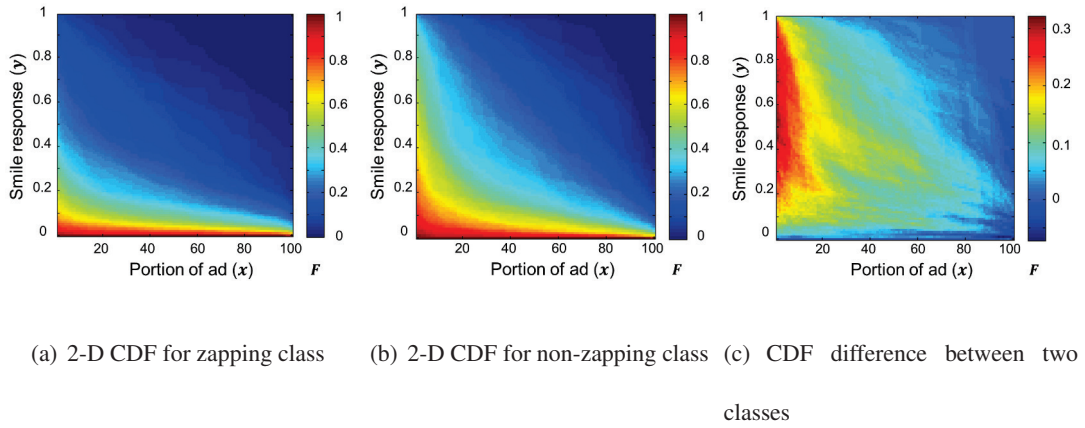


Figure 4.9: The analysis of the smile response volume. Fig. 4.9(a) and Fig. 4.9(b) are interpreted as follows: the probability is F if the smile response of a sequence is *above* y for x percent of its entire length. Fig. 4.9(c) is generated by subtracting Fig. 4.9(a) from Fig. 4.9(b), which shows that high smile response is more effective than high volume in distinguishing zapping from non-zapping. (Better viewed in color)

For zapping class, the probability decreases as maximum smile response increases, and reaches the minimum when smile response is between 0.7 and 0.8. However, the probability increases thereafter. After examining the data, I found that several subjects were engaged by the ad and were smiling with high intensity in the beginning. Unfortunately, they zapped right away when the brand's logo or name showed up at the end of the ad. After interviewing with them, I found that most of the people behaved like this because they thought the ad is about to finish. From the advertiser's point of view, this scenario should be considered as a success. However, from the publisher's point of view, they will not get paid since they consider this scenario as zapping [86].

Remark 2. Note to Advertisement Publisher: *In my analysis, 13.3% of the time, users zap because advertiser's brand is displayed at the end when the ad is not finished. In this case, the advertisers take the benefit since users consume the content of the ad. The publisher (such as YouTube), on the other hand, loses revenue since the ad is neither watched for more than 30 seconds nor completed by the user under this circumstance. If the billing policy is changed from "30 seconds" to "27 seconds" or from "complete the entire advertisement" to "complete 90% of the advertisement", less zapping is likely to happen in my experiment. In light of this observation, a publisher is suggested to change the billing policy in the aforementioned manner while maintaining the effectiveness of a commercial.*

4.3.3.4 The Volume of Smile Response

In addition to maximum smile response, I also analyze how the volume of the smile response distinguishes the zapping and non-zapping classes. The volume of smile response is defined as the portion of the length of the sequence that is above a certain smile response level. Fig. 4.9(a) and Fig. 4.9(b) are a variation of 2-D cumulative distribution function (CDF) defined as:

$$F_{XY}(x, y) = P(X \geq x, Y \geq y) \quad (4.3)$$

where X is a random variable that measures the portion of an advertisement that is watched, and Y represents smile response. It can be interpreted as follows: if the smile response of a sequence is *above* y for x percent of its entire length, the probability of this event is F .

In both Fig. 4.9(a) and Fig. 4.9(b), the CDF is 1 when smile response is close to 0 (bottom edge of the figure). This is because all the data satisfy the criteria that the smile response is always above 0. The reason why the upper right corner is of value 0 is because no sequence has a smile response above 0.9 for 90% of the time.

Compared to the zapping class, the CDF of the non-zapping class in Fig. 4.9(b) is

close to symmetrical along the diagonal line from the bottom left corner to the upper right corner. This shows that, for non-zapping class, both the level of the smile response and its volume play important roles in the data distribution. For example, the CDFs are similar for non-zapping class for the following two cases: (1) smile response is above 0.6 for 20% amount of the time; (2) smile response is above 0.2 for 60% amount of the time.

I show the CDF difference in Fig. 4.9(c) by subtracting Fig. 4.9(a) from Fig. 4.9(b). The major difference exists where the smile volume is low and the smile response level is high. This informs us that, in distinguishing zapping from non-zapping, high smile response level is more important than high volume.

Remark 3. Note to Advertiser: *My statistics shows that users tend to zap less if their smile response level is higher or their smile response is above a certain level for a longer period of time. Moreover, if an advertiser has to choose between “high smile response level + low volume” and “low smile response level + high volume”, the former is more effective in preventing zapping. My experimental result suggests that, practically, it is preferred to design an ad with one or two entertaining scenes that highly impact the user’s engagement than to include several entertaining scenes with mediocre impact, if eliciting laugh is the objective of a commercial.*

4.3.4 Zapping Index

4.3.4.1 The Smile Histogram Feature

Based on the data characteristics, I find that the mean, max, and the volume of smile response of a video sequence are essential for distinguishing zapping from non-zapping. It is natural to use the histogram of the smile response as the key feature.

As shown in Fig. 4.10, the cumulative smile histogram is calculated for the entire sequence. It is then normalized between 0 and 1. In the two typical examples in Fig. 4.10, the

probability is high when the smile response is low for the zapping class. For non-zapping class, on the contrary, the smile response is more evenly distributed.

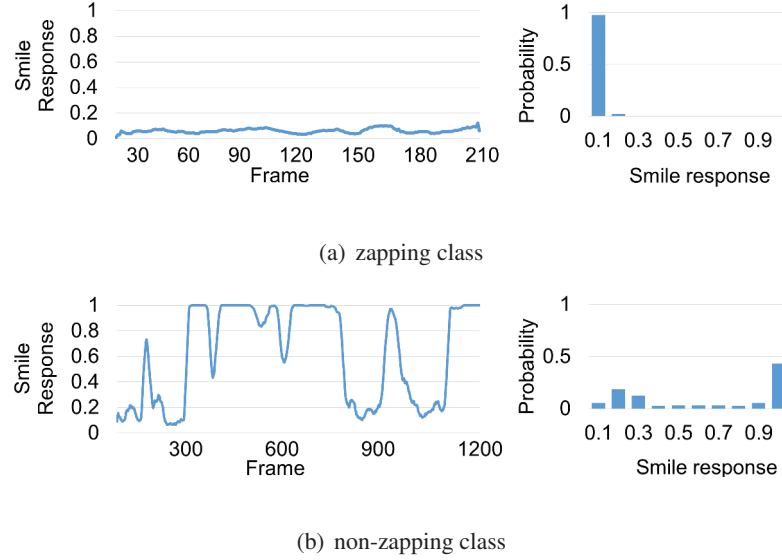


Figure 4.10: Smile histogram of zapping and non-zapping sequences. The selected sequences are the same as in Fig. 4.7.

4.3.4.2 Zapping Classification

In order to distinguish zapping from non-zapping sequences, we formulate it as a binary classification problem. The class labels of the data are assigned based on the 90% threshold shown in Fig. 4.5.

During the training phase, the histograms of all the sequences are computed. I use SVM [23] with the radial basis function as the kernel function to train my classifier. The *double-layer* 10-fold cross validation is then carried out to avoid overfitting. The first layer is for parameter optimization and the second layer uses the optimized parameters for model training. The number of bins is then determined to be 10 by the second layer of cross validation. Then a validation set is constructed by randomly generating 4000 frames from the entire dataset.

In comparison, I provide the baseline result from naively assigning labels based on

class distribution shown in Fig. 4.5. In addition, I include the result of using the normalized cumulative smile response of individual sequences. Fig. 4.11 shows the ROC curve of the aforementioned approaches. As AUC scores illustrated in Fig. 4.11, smile histogram feature (0.83) significantly outperform smile response feature (0.60), which is congruent with my analysis in Section 4.3.3 that the mean, maximum, and volume of smile response are essential in characterizing zapping behavior.

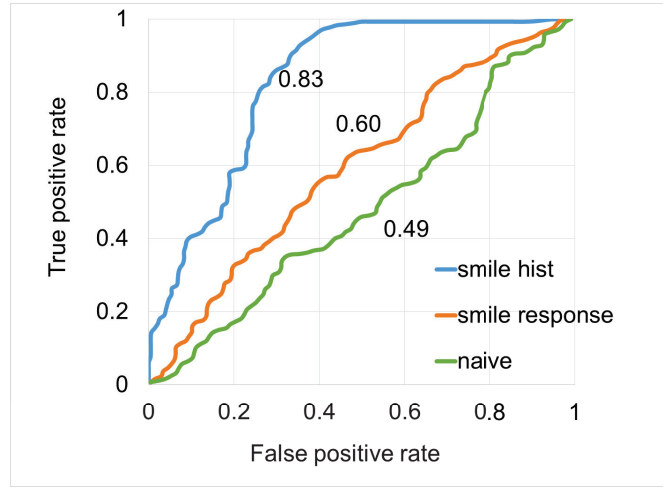
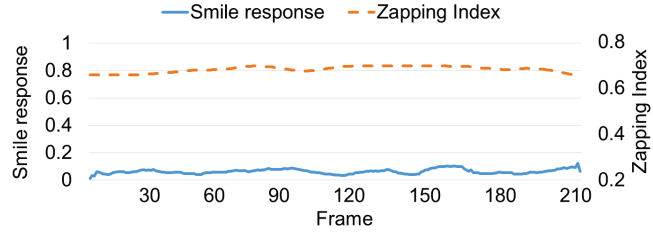


Figure 4.11: The ROC plot for zapping/non-zapping classification. The naïve baseline is by assigning test labels based on the class distribution.

During the testing phase, my goal is to measure the moment-to-moment zapping probability. Thus, at each frame, the smile histogram feature is computed, which is then passed to the classifier, and the probability output is considered as the zapping index. The upcoming discussion in the next section shows why ZI is a valid measurement for zapping behavior and how it is related to zapping prediction and user preference discovery.



(a) zapping class



(b) non-zapping class

Figure 4.12: The moment-to-moment Zapping Index for individual sequences. The selected sequences are the same as in Fig. 4.7.

4.4 Experimental Results

In this section, I explore the characteristics of the Zapping Index (ZI) on individual sequences, across each ad, and across each ad category. I also visualize their relationships with data distribution that show the popularity of each ad. Moreover, I am able to understand the preference of users to different ad categories.

4.4.1 Zapping Index on Individual Sequences

According to my design, a larger value of ZI means higher probability of zapping. Fig. 4.12 provides the Zapping Index for the two running examples. Generally speaking, if a user displays low smile response, the ZI remains around 0.65; the ZI decreases when smile response increases. The ZI value of 0.65 coincides with my discussion related to Fig. 4.8, when maximum smile response of a sequence is between 0 and 0.1, there is a two thirds chance that it

belongs to the zapping class.

As discussed in Section 4.3.3.3, if a user's smile response reaches the maximum, he or she will most unlikely zap. In Fig. 4.12(b), there might be minor increase of ZI if smile response drops from the maximum. However, the ZI value will remain low even after the increase, which illustrates that the user is less likely to zap.

4.4.2 Zapping Index vs. Zapping Distribution

In Fig. 4.13, I compare the relationship between the average ZI value and the zapping distribution for each ad. I compute the mean and SEM bound of ZI similar to the description in Section 4.3.3.2.

For ads which have less zapping as shown by the zapping distribution (e.g., Toyota and Nissan), the decrease rate of the moment-to-moment ZI is high which means small likelihood of zapping by users. Indeed, these two ads are the funniest ads among all of my selections. On the contrary, for ads for which a larger number of participants zapped (e.g. Pizza Hut and Subway), the ZI curve only has a slight decrease. Therefore, my ZI measurement correlates with the zapping distribution.

To further demonstrate that ZI is a quality measurement for zapping, I analyze the correlation between ZI and zapping distribution in Fig. 4.13. I treat a ZI sequence for each ad as a feature and compute the pair-wise Euclidean distance between features. I compute the distance the same way for zapping distribution. The pair-wise distance for both zapping distribution and ZI are plotted in Fig. 4.15(a) and Fig. 4.15(b), respectively. I observe similar patterns in Fig. 4.15(a) and Fig. 4.15(b), which means that ZI preserves distance between ads in zapping distribution. For example, the distance is large for *Toyota* and *Pizza Hut* in Fig. 4.15(a); this means that viewers' behavior is dramatically different in watching these two ads. In Fig. 4.15(b), ZI captures the same difference. Therefore, the measurement of ZI is highly correlated

with the viewer's zapping behavior.

4.4.3 Zapping Index vs. Smile Response

I show the comparison of ZI and smile response in Fig. 4.14. Generally speaking, ZI has an inverse relationship with smile response. One may argue that smile response itself is a good indicator for zapping prediction. However, as I can see from Fig. 4.14, smile response is a measurement of user's smile expression at every moment, and therefore, it is volatile as time changes. ZI, on the other hand, is a better measurement for predicting zapping. It is less volatile by taking into account the maximum smile and smile occurring volume information overtime (see Fig. 4.9). Yet, it is also sensitive enough to capture the noticeable changes in smile response.

One interesting pattern that is worth noting in ads such as Jack In The Box, Nissan, Toyota is that there is a major drop for the smile response at the end. After examining the participant's expression as well as the ads themselves, I found out that most participants smile due to the entertaining scene at the end. They tend to smile less as soon as they saw the brand's logo or name. This phenomenon is congruent with my discussion in Section 4.3.3.3.

4.4.4 User Preference from Zapping Index

An important factor, if not the most important one, in considering the advertising campaign solution is the target market [60]. Advertisement is viewed as useful information for the right target, but viewed as harassment for the wrong target. Advertising publishers explore a large data from users to discover the target market. User information such as age, gender, ethnicity, geography, income, lifestyle, online behavior, etc., are leveraged to infer the user preference. The most recent work [74] has demonstrated that smile response is able to reveal

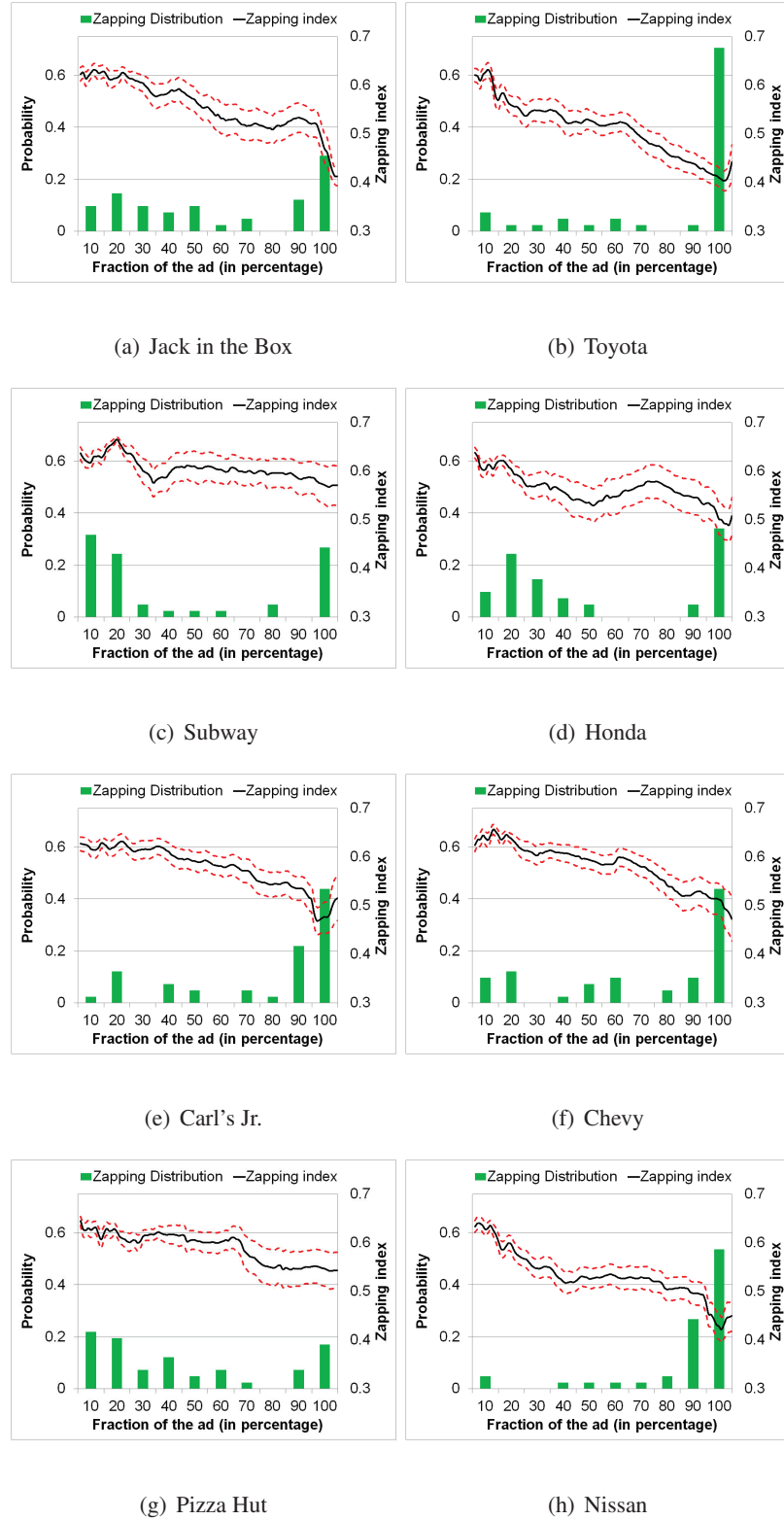
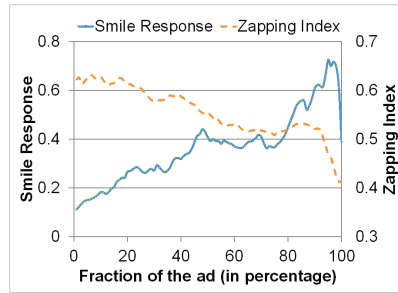
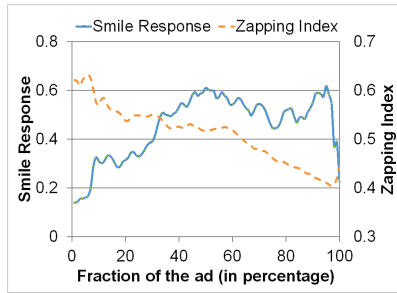


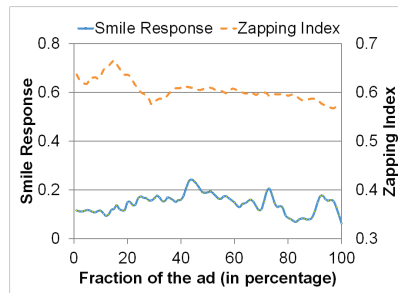
Figure 4.13: The relationship of Zapping Index (ZI) and zapping distribution for each ad. Zapping distribution is the normalized histogram measured with the primary y-axis, while ZI is displayed by the secondary y-axis bounded by its standard error (shown with dotted lines). Generally speaking, the more sharply ZI decreases, the less zapping happens.



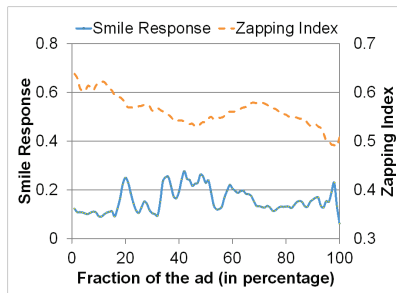
(a) Jack in the Box



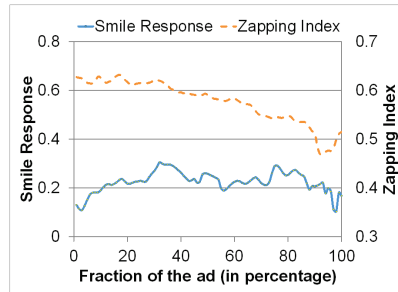
(b) Toyota



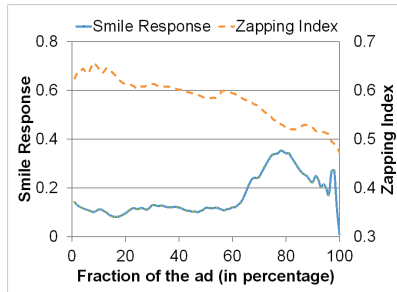
(c) Subway



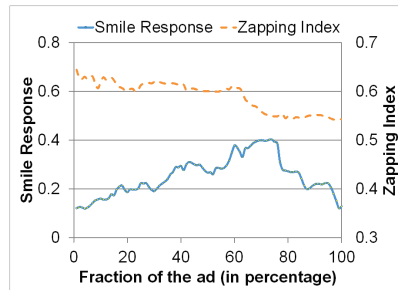
(d) Honda



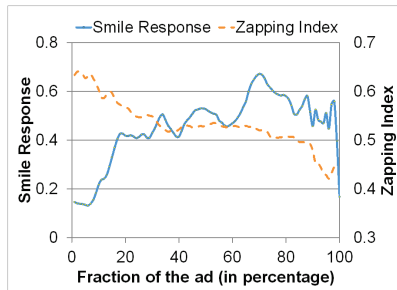
(e) Carl's Jr.



(f) Chevy



(g) Pizza Hut



(h) Nissan

Figure 4.14: The relationship of smile response and Zapping Index (ZI) for each ad. Broadly speaking, they are in inverse relation. But the ZI is less volatile as compared to smile response.

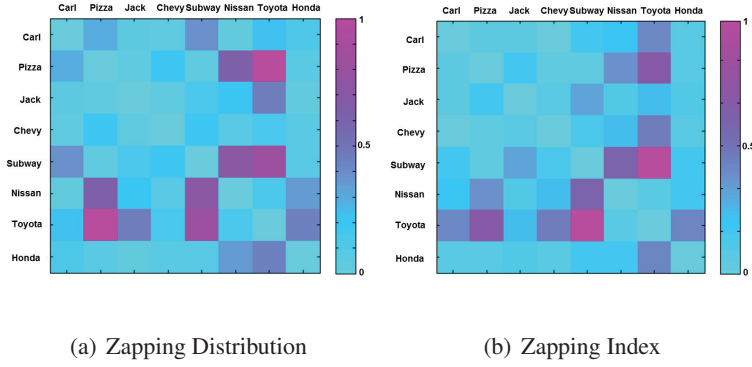


Figure 4.15: The pair-wise distance of individual ad in zapping distribution and Zapping Index (ZI) groups. ZI is highly correlated with viewer’s zapping behavior exhibited in zapping distribution, which demonstrates that ZI is a reasonable measure for zapping. (The distances within each group are normalized between 0 and 1.)

whether an ad is liked by a viewer. In this work, I show that Zapping Index, derived from smile response, is another type of user information which directly shows viewer’s preference. Thus, ZI may have a potential impact on the future of advertising. I analyze user preferences for two different ad categories in my experiment. Fig. 4.16 shows four typical samples of user preferences of two different ad categories expressed by ZI. These four types include: like neither category, like first category but not the second, like the second category but not the first, and like both categories. By classifying the user based on their ZI, it is possible to accurately measure their preferences, which will benefit both the advertising provider and the publisher.

4.4.5 Limitations

In this work, I only consider the entertaining value of an ad and use smile response to compute ZI for the prediction of a user’s zapping probability. However, entertainment is not the only reason that engages a user. For example, information content is another major reason for user engagement [38], in which case, user will not necessarily smile. Under this situation, my ZI measurement may be less effective in predicting zapping probability.

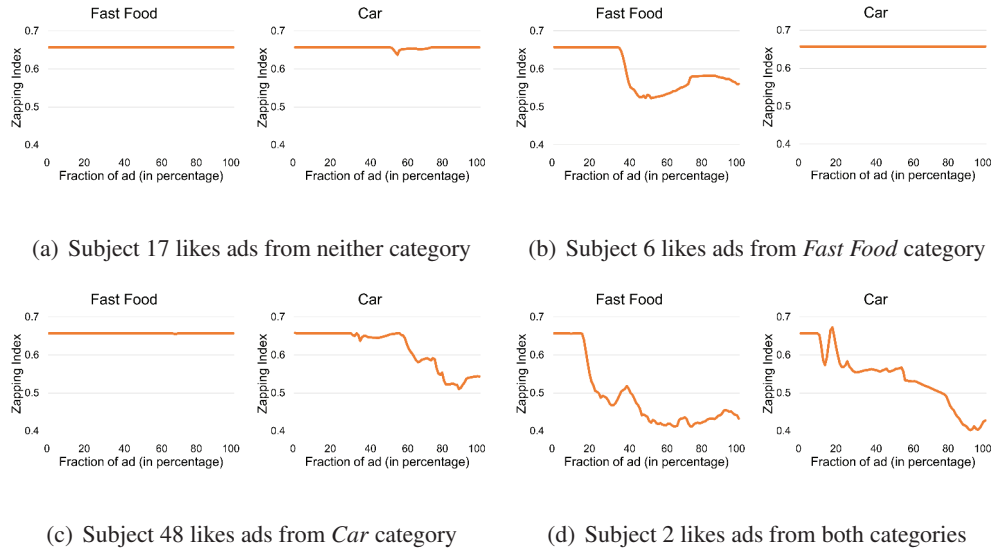


Figure 4.16: All four possible types of user preferences represented by ZI. Each ZI pattern is the average of one subject on the entire ad category. Flat ZI response such as both cases in (a) shows no smile response as well as no interest to the ad category, and vice versa.

Table 4.2: Another Advertisement Category: “Running Shoe”

Category	Brand	Ad Name
Running Shoe	Nike	Flyknit Lunar 1+
	Adidas	Boost
	Puma	Mobium and Adaptive
	Under Armour	I Will Innovation

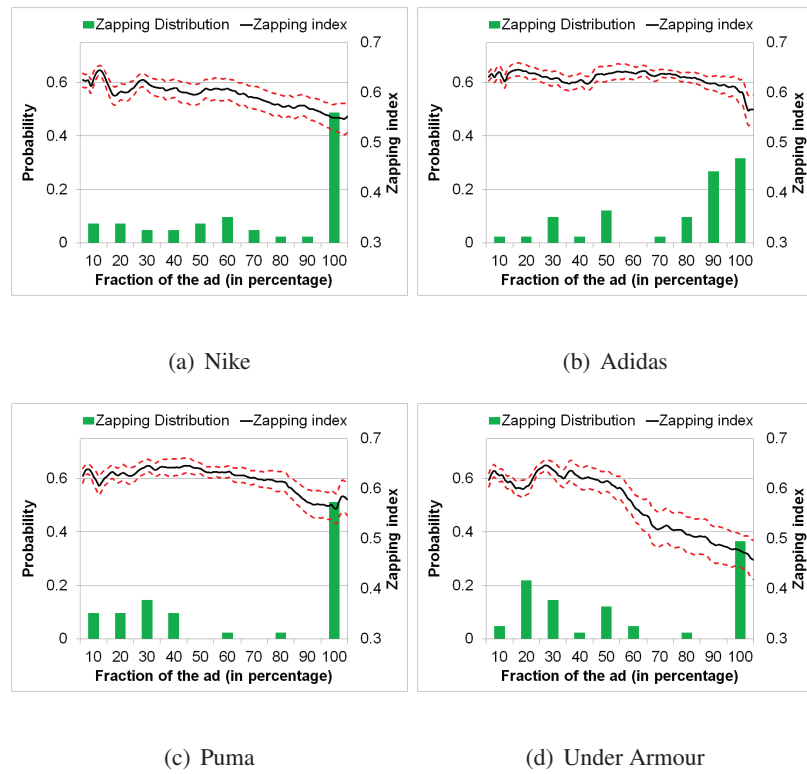


Figure 4.17: Zapping Index vs. zapping distribution for the ads in *running shoe* category. ZI is less effective in predicting zapping probability compared to Fig. 4.13 since soliciting smile is not the intention of these ads. Hence, users tend to zap less even if ZI value is high.

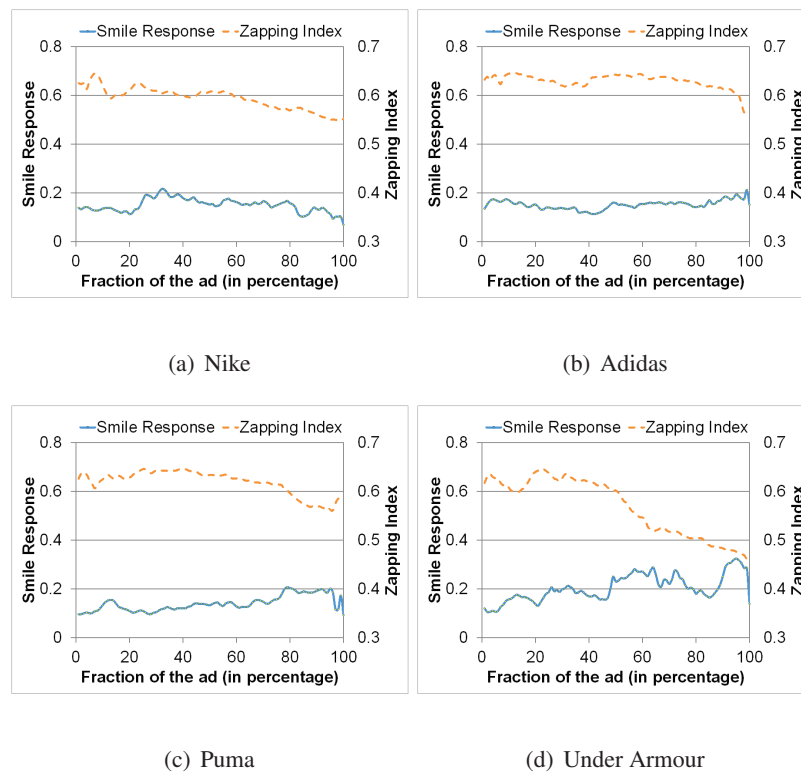


Figure 4.18: Zapping Index vs. smile response for the ad in the *running shoe* category. Smile response is low but this does not always lead to a decrease in the value of ZI.

To test my assumption, I selected another four ads from the *running shoe* category, shown in Table. 4.2. I make the selections by following the criteria described in Section 4.3.2. The only difference is that I selected the high quality ad but not necessarily amusing.

Similar to Fig. 4.13 and Fig. 4.14, I explore the relationship of ZI vs. zapping distribution and ZI vs. smile response in Fig. 4.17 and Fig. 4.18, respectively. The ads from the *Running Shoe* category are not intended to make a user laugh but rather to provide information by demonstrating their technology. Hence, the average smile response is low and the ZI does not necessarily decrease in Fig. 4.18, which shows that users are likely to zap. However, based on the zapping distribution in Fig. 4.17, majority of the participants have watched the ads without zapping. Thus, ZI is less effective as a zapping prediction metric when the intention of the ad is not engaging people through entertaining factors.

Chapter 5

Conclusions

In order to tackle the automatic analysis of facial expression in real-world scenarios, I have present two novel approaches, EAI and SOFIT. EAI condense a video sequence into a single image representation. I adopt SIFT flow for aligning the face images which is able to compensate for large rigid head motion and maintain facial feature motion detail. Then, an iterative algorithm is used to generate an Avatar Reference face model onto which I align every face image. I experimentally demonstrate that level-2 EAI has the potential to generate higher classification rate. My EAI representation combined with LPQ and LBP texture descriptors achieved excellent performance in both person-independent and person-specific cases when tested on the challenging facial expression recognition dataset, GEMEP-FERA dataset. Given the consistency of my EAI representation, the performance of my approach is dramatically improved when compared with the baseline [111] and other approaches [102] [66] [30] [77] [28] [98] [24] [15].

SOFIT is a video-based real-time face registration technique. This approach utilizes holistic dense flow-based information, and therefore, it is robust to detection error and noise. Minor out-of-plane head rotation can also be corrected by employing structural information

from SIFT flow. Besides, this method is able to generate temporally smooth registration results which are essential for spontaneous facial expression analysis and super-resolution. Last but not least, with the dynamic programming implementation, this method is suitable for real-time processing. Experimental results demonstrate its applications in AU recognition, vehicle recognition, and image super-resolution.

Given the effective expression analysis tool, I explored the automated facial expression recognition in the application of online advertising. I demonstrated that users' zapping behavior has a close relationship with their smile response. I created an advertising evaluation metric, Zapping Index, to measure a user's zapping probability. A higher value of ZI reveals that the user has a higher chance of zapping. ZI can also be used to measure a user's preference to different categories of commercials. This is beneficial to advertisers as well as to ad publishers.

Bibliography

- [1] *FERA2011 Challenge data*. <http://sspnet.eu/fera2011/fera2011data/>.
- [2] *FERA2011: Facial Expression Recognition and Analysis Challenge*. <http://sspnet.eu/fera2011/>.
- [3] Efficient smile detection by extreme learning machine. *Neurocomputing*, 2014.
- [4] Zara Ambadar, J. Schooler, and Jeffrey Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 2005.
- [5] L. An, B. Bhanu, and Songfan Yang. Boosting face recognition in real-world surveillance videos. In *IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, 2012.
- [6] L. An, M. Kafai, and B. Bhanu. Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2013.
- [7] Le An and B. Bhanu. Image super-resolution by extreme learning machine. In *19th IEEE International Conference on Image Processing (ICIP)*, 2012.
- [8] Le An, B. Bhanu, and Songfan Yang. Face recognition in multi-camera surveillance videos. In *Int. Conf. on Pattern Recognition (ICPR)*, 2012.
- [9] Le An and Bir Bhanu. Face image super-resolution using 2D CCA. *Signal Processing*, 2014.
- [10] Le An, Bir Bhanu, and Songfan Yang. Unified face representation for individual recognition in surveillance videos. *Augmented Vision and Reality*. Springer Berlin Heidelberg, 2013.
- [11] Le An, M. Kafai, and B. Bhanu. Face recognition in multi-camera surveillance videos using dynamic Bayesian network. In *Sixth International Conference on Distributed Smart Cameras (ICDSC)*, 2012.
- [12] Le An, N. Thakoor, and B. Bhanu. Vehicle logo super-resolution by canonical correlation analysis. In *19th IEEE International Conference on Image Processing (ICIP)*, 2012.
- [13] Le An, Songfan Yang, and Bir Bhanu. Efficient smile detection by extreme learning machine. *Neurocomputing*, 2014.

- [14] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, and R. Szeliski. A Database and Evaluation Methodology for Optical Flow. In *Proc. ICCV*, 2007.
- [15] Tadas Baltrusaitis, Daniel McDuff, Ntombikayise Banda, Marwa Mahmoud, Rana el Kaliouby, Peter Robinson, and Rosalind Picard. Real-time Inference of Mental States from Facial Expressions and Upper Body Gestures. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [16] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *Proc. FG*, 2006.
- [17] John N. Bassili. Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face. *Personality and Social Psychology*, 1979.
- [18] Roger Best. *Market Based Management*. Prentice Hall; 5th edition, 2008.
- [19] Steven M. Boker, Jeffrey F. Cohn, Iain Matthews, and Timothy R. Brick. Effects of Damping Head Movement and Facial Expression in Dyadic Conversation Using Real-time Facial Expression Tracking and Synthesized Avatars. *Philosophical Transactions B of the Royal Society*, 2009.
- [20] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [21] E. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 2011.
- [22] Y. Caspi and M. Irani. Spatio-temporal alignment of sequences. *IEEE Trans. PAMI*, 2002.
- [23] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines*, 2001.
- [24] Sien Wei Chew, Patrick J. Lucey, Simon Lucey, Jason Saragih, Jeffrey Cohn, and Sridha Sridharan. Person-independent Facial Expression Detection Using Constrained Local Models. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [25] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Trans. PAMI*, 2001.
- [26] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least Squares Congealing for Unsupervised Alignment of Images. In *Proc. CVPR*, 2008.
- [27] Albert Cruz, Bir Bhanu, and Songfan Yang. A Psychologically-Inspired Match-Score Fusion Model for Video-Based Facial Expression Recognition. In *Proc. ACII workshop on AVEC*, 2011.
- [28] Mohamed Dahmane and Jean Meunier. Emotion Recognition using Dynamic Grid-based HoG Features. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [29] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. CVPR*, 2005.
- [30] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion Recognition Using PHOG and LPQ features. In *Proc. FG Workshop on FERA Challenge*, 2011.

- [31] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying Facial Actions. *IEEE Trans. PAMI*, 1999.
- [32] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional Speech: Towards A New Generation of Databases. *Speech Communication*, 2003.
- [33] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. A New Emotion Database: Considerations, Sources and Scope. In *Proc. the ISCA Workshop on Speech and Emotion*, 2000.
- [34] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
- [35] Paul Ekman. Facial expression and emotion. *American Psychologist*, 1993.
- [36] Paul Ekman and Erika Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.
- [37] R. El Kaliouby and P. Robinson. Mind Reading Machines: Automated Inference of Cognitive Mental States from Video. In *Proc. SMC*, 2004.
- [38] Josephine L.C.M. Woltman Elpers, Michel Wedel, and Rik G.M. Pieters. Why do consumers stop viewing television commercials? two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research*, 2003.
- [39] I.A. Essa and A.P. Pentland. Coding, Analysis, Interpretation, and Recognition of Facial Expressions. *IEEE Trans. PAMI*, 1997.
- [40] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *JMLR*, 2008.
- [41] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning Low-Level Vision. *IJCV*, 2000.
- [42] Elia Gabarron, Luis Fernandez-Luque, Manuel Armayones, and Annie YS Lau. Identifying measures used for assessing quality of youtube videos with patient health information: A review of current literature. *Interactive Journal of Medical Research*, 2013.
- [43] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Trans. on Systems, Man and Cybernetics, Part A*, 2008.
- [44] T. Gehrig and H.K. Ekenel. A Common Framework for Real-time Emotion Recognition and Facial Action Unit Detection. In *Proc. CVPR Workshops*, 2011.
- [45] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 2010.
- [46] Paul Gustafson and S. Siddarth. Describing the dynamics of attention to tv commercials: A hierarchical Bayes analysis of the time to zap an ad. *Journal of Applied Statistics*, 2007.

- [47] J. Han and B. Bhanu. Individual Recognition Using Gait Energy Image. *IEEE Trans. PAMI*, 2006.
- [48] Changbo Hu, Ya Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2004.
- [49] Changbo Hu, Ya Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *Image and Vision Computing*, 2006.
- [50] G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *Proc. ICCV*, 2007.
- [51] P. J. Huber. *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, 1981.
- [52] Paul Hyde, Edward Landry, and Andrew Tipping. Making the perfect marketer. *Strategy?+?Business*, 2004.
- [53] Michal Irani and Shmuel Peleg. Improving Resolution by Image Registration. *Graph. Models Image Process.*, 1991.
- [54] Rik G.M. Pieters Josephine L.C.M. Woltman Elpers, Michel Wedel. Why do consumers stop viewing television commercials? two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research*, 2003.
- [55] Michael Kaess and Frank Dellaert. A markov chain monte carlo approach to closing the loop in SLAM. In *Proc. ICRA*, 2005.
- [56] M. Kafai and B. Bhanu. Dynamic Bayesian networks for vehicle classification in video. *IEEE Transactions on Industrial Informatics*, 2012.
- [57] T. Kanade, J.F. Cohn, and Yingli Tian. Comprehensive Database for Facial Expression Analysis. In *Proc. FG*, 2000.
- [58] D. Keren, S. Peleg, and R. Brada. Image Sequence Enhancement Using Sub-pixel Displacements. In *Proc. CVPR*, 1988.
- [59] Robert Kooij, Kamal Ahmed, and Kjell Brunnström. Perceived quality of channel zapping. In *Fifth IAESTED Int. Conf. on Communication Systems and Networks (CSN 2006)*, 2006.
- [60] Philip Kotler. *Marketing management*. Pearson Education India, 2009.
- [61] I. Kotsia and I. Pitas. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Trans. IP*, 2007.
- [62] E.G. Learned-Miller. Data Driven Image Models Through Continuous Joint Alignment. *IEEE Trans. PAMI*, 2006.
- [63] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of Facial Expression Extracted Automatically from Video. In *Image and Vision Computing*, 2006.
- [64] G. Littlewort, J Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. Computer Expression Recognition Toolbox. In *Proc. FG*, 2011.
- [65] G. Littlewort, J. Whitehill, Tingfan Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The Computer Expression Recognition Toolbox (CERT). In *Proc. FG*, 2011.

- [66] Gwen Littlewort, Jacob Whitehill, Ting-Fan Wu, Nicholas Butko, Paul Ruvolo, Javier Movellan, and Marian Bartlett. The Motion in Emotion c A CERT Based Approach to the FERA Emotion Challenge. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [67] C. Liu, J. Yuen, and A. Torralba. SIFT Flow: Dense Correspondence across Scenes and its Applications. *IEEE Trans. PAMI*, 2011.
- [68] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
- [69] D.G. Lowe. Object recognition from local scale-invariant features. In *IEEE Int. Conf. on Computer Vision.*, volume 2, pages 1150–1157, 1999.
- [70] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010.
- [71] S. Lucey, I. Matthews, Changbo Hu, Z. Ambadar, F. De La Torre, and J. Cohn. AAM derived face representations for robust facial action recognition. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2006.
- [72] Brais Martinez, Michel Valstar, Xavier Binefa, and Maja Pantic. Local evidence aggregation for regression based facial point detection. *IEEE Trans. PAMI*, 2013.
- [73] Iain Matthews and Simon Baker. Active Appearance Models Revisited. *IJCV*, 2003.
- [74] D. McDuff, R. Kaliouby, T. Senechal, D. Demirdjian, and R. Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 2014.
- [75] D. McDuff, R.E. Kaliouby, and R.W. Picard. Crowdsourcing facial responses to online videos. *IEEE Trans. on Affective Computing*, 2012.
- [76] Daniel McDuff, Rana El Kaliouby, Thibaud Senechal, May Amr, Jeffrey F. Cohn, and Rosalind W. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild. In *CVPR Workshops*, pages 881–888, 2013.
- [77] Hongying Meng, Bernardino Romera-Paredes, and Nadia Bianchi-Berthouze. Emotion Recognition by Two View SVM 2K Classifier on Dynamic Facial Expression Features. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [78] Hongying Meng, John Shawe-Taylor, Sandor Szedmak, and Jason Farquhar. Support Vector Machine to Synthesise Kernels. In Joab Winkler, Mahesan Niranjan, and Neil Lawrence, editors, *Deterministic and Statistical Methods in Machine Learning*. Springer Berlin / Heidelberg, 2005.
- [79] P. Negri, X. Clady, M. Milgram, and R. Poulencard. An oriented-contour point based voting algorithm for vehicle type classification. In *Proc. ICPR*, 2006.
- [80] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. PAMI*, 2002.
- [81] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In *Proc. ICISP*, 2008.

- [82] M. Pantic and I. Patras. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences. *IEEE Trans. SMC-B*, 2006.
- [83] M. Pantic and L.J.M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. PAMI*, 2000.
- [84] M. Pantic and L.J.M. Rothkrantz. Facial Action Recognition for Facial Expression Analysis from Static Face Images. *IEEE Trans. SMC-B*, 2004.
- [85] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based Database for Facial Expression Analysis. In *IEEE Int. Conf. on Multimedia and Expo.*, 2005.
- [86] Max Pashkevich, Sundar Dorai-Raj, Melanie Kellar, and Dan Zigmond. Empowering online advertisements by empowering viewers with the right to choose. *Journal of Advertising Research*, 2012.
- [87] G. Pearce and N. Pears. Automatic make and model recognition from frontal images of cars. In *IEEE AVSS*, pages 373–378, Aug 2011.
- [88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- [89] Yigang Peng, A. Ganesh, J. Wright, Wenli Xu, and Yi Ma. RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images. In *Proc. CVPR*, 2010.
- [90] V.S. Petrovic and T. F. Cootes. Analysis of features for rigid structure vehicle type recognition. In *Proc. BMVC*, 2004.
- [91] Tuan Q. Pham, Lucas J. van Vliet, and Klammer Schutte. Robust Fusion of Irregularly Sampled Data Using Adaptive Normalized Convolution. *EURASIP JASP*, 2006.
- [92] K Poels and Siegfried Dewitte. How to capture the heart? reviewing 20 years of emotion measurement in advertising. Technical report, Katholieke Universiteit Leuven, 2006.
- [93] James A. Russell and José Miguel Fernández-Dols. *The Psychology of Facial Expression*. Cambridge University Press, 1997.
- [94] J.M. Saragih, S. Lucey, and J.F. Cohn. Face Alignment through Subspace Constrained Mean-Shifts. In *Proc. ICCV*, 2009.
- [95] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. *IJCV*, 2002.
- [96] Caifeng Shan. Smile detection by boosting pixel differences. *IEEE Trans. on Image Processing*, 2012.
- [97] P. Siebert, T.N.M. Van Caenegem, and M. Wagner. Analysis and improvements of zapping times in IPTV systems. *IEEE Trans. on Broadcasting*, 2009.

- [98] Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim. Accumulated motion images for facial expression recognition in videos. In *Proceedings of Facial Expression Recognition and Analysis Challenge*, Santa Barbara, CA, Mar 2011.
- [99] Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim. Accumulated Motion Images for Facial Expression Recognition in Videos. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [100] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image Super-resolution Using Gradient Profile Prior. In *Proc. CVPR*, 2008.
- [101] Richard Szeliski. Image alignment and stitching: A tutorial. *Found. Trends. Comput. Graph. Vis.*, 2006.
- [102] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S. Huang, Xutao Lv, and Tony X. Han. Emotion Recognition from an Ensemble of Features. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [103] Thales Teixeira, Michel Wedel, and Rik Pieters. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research*, 2012.
- [104] N. Thakoor and Jean Gao. Automatic video object shape extraction and its classification with camera in motion. In *Proc. ICIP*, 2005.
- [105] N.S. Thakoor and B. Bhanu. Structural signatures for passenger vehicle classification in video. *IEEE Trans. Intelligent Transportation Systems*, 2013.
- [106] Carlos Eduardo Thomaz and Gilson Antonio Giralaldi. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 2010.
- [107] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. PAMI*, 2001.
- [108] Yan Tong, Jixu Chen, and Qiang Ji. A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding. *IEEE Trans. PAMI*, 2010.
- [109] Michihiro Uenohara and Takeo Kanade. Real-time vision based object registration for image overlay. In *1995 Conference on Computer Vision, Virtual Reality and Robotics in Medicine*, April 1995.
- [110] M. F. Valstar and M. Pantic. Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. In *Proc. CVPR workshop on Human Computer Interaction*, 2007.
- [111] M.F. Valstar, B. Jiang, M. Méhu, M. Pantic, and K Scherer. The First Facial Expression Recognition and Analysis Challenge. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [112] M.F. Valstar, M. Mehu, Bihan Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Trans. SMC-B*, 2012.
- [113] M.F. Valstar and M. Pantic. Fully Automatic Recognition of the Temporal Phases of Facial Actions. *IEEE Trans. SMC-B*, 2012.

- [114] M.F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, Workshop on Vision for Human-Computer Interaction*, page 76, Jun 2005.
- [115] M.F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2005.
- [116] Patrick Vandewalle, Sabine Süsstrunk, and Martin Vetterli. A Frequency Domain Approach to Registration of Aliased Images with Application to Super-Resolution. *EURASIP JASP*, 2006.
- [117] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/>, 2008.
- [118] Paul Viola and Michael Jones. Robust Real-time Face Detection. *IJCV*, 2004.
- [119] A. Wagner, J. Wright, A. Ganesh, Zihan Zhou, and Yi Ma. Towards a practical face recognition system: Robust registration and illumination by sparse representation. In *Proc. CVPR*, 2009.
- [120] J.Y.A. Wang and E.H. Adelson. Representing Moving Images with Layers. *IEEE Trans. IP*, 1994.
- [121] Mei Wang, Y. Iwai, and M. Yachida. Expression Recognition from Time-sequential Facial Images by Use of Expression Change Model. In *Proc. FG*, 1998.
- [122] J. Whitehill, Gwen Littlewort, Ian Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009.
- [123] Tingfan Wu, M.S. Bartlett, and J.R. Movellan. Facial Expression Recognition Using Gabor Motion Energy Filters. In *Proc. CVPRW*, 2010.
- [124] Tingfan Wu, N.J. Butko, P. Ruvolo, J. Whitehill, M.S. Bartlett, and J.R. Movellan. Action Unit Recognition Transfer across Datasets. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [125] Xuehan Xiong and F. de la Torre. Supervised descent method and its applications to face alignment. In *Proc. CVPR*, 2013.
- [126] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proc. CVPR*, 2013.
- [127] Y. Yacoob and L.S. Davis. Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow. *IEEE Trans. PAMI*, 1996.
- [128] Peng Yang, Qingshan Liu, and D.N. Metaxas. Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition. In *Proc. CVPR*, 2007.
- [129] S. Yang and B. Bhanu. Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image. *IEEE Trans. SMC-B*, 2012.
- [130] Songfan Yang, Le An, Bir Bhanu, and Ninad Thakoor. Improving action units recognition using dense flow-based face registration in video. In *Proc. FG*, 2013.

- [131] Songfan Yang and Bir Bhanu. Facial Expression Recognition Using Emotion Avatar Image. In *Proc. FG Workshop on FERA Challenge*, 2011.
- [132] Jiangang Yu and Bir Bhanu. Evolutionary Feature Synthesis for Facial Expression Recognition. *Pattern Recognition Letters*, 2006.
- [133] Iffat Zafar, Eran A. Edirisinghe, and B. Serpil Acar. Localised contourlet features in vehicle make and model recognition. In *Image Processing: Machine Vision Applications II, Proc. of SPIE-IS&T Electronic Imaging*, 2009.
- [134] Z. Zhang, X. Liang, A. Ganesh, and Y. Ma. TILT: Transform Invariant Low-rank Textures. In *Proc. ACCV*, 2010.
- [135] Guoying Zhao and M. Pietikäinen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Trans. PAMI*, 2007.
- [136] Guoying Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI*, 2007.
- [137] Feng Zhou, F. De la Torre, and J.F. Cohn. Unsupervised Discovery of Facial Events. In *Proc. CVPR*, 2010.
- [138] J. Zhu, M.R. Lyu, and T.S. Huang. A Fast 2D Shape Recovery Approach by Fusing Features and Appearance. *IEEE Trans. PAMI*, 2009.
- [139] Jianke Zhu, Luc Van Gool, and S.C.H. Hoi. Unsupervised Face Alignment by Robust Nonrigid Mapping. In *Proc. ICCV*, 2009.
- [140] Xiangxin Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, 2012.
- [141] Barbara Zitová and Jan Flusser. Image Registration Methods: A Survey. *Image and Vision Computing*, 2003.
- [142] A. Zomet, A. Rav-Acha, and S. Peleg. Robust Super-resolution. In *Proc. CVPR*, 2001.