UNIVERSITY OF CALIFORNIA
RIVERSIDE

Biologically Inspired Facial Emotion Recognition

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

Doctor of Philosophy

in

Electrical Engineering

by

Alberto C Cruz

December 2014

Dissertation Committee:
Dr. Bir Bhanu, Chairperson
Dr. Anastasios Mourikis
Dr. Aaron Seitz

The Dissertation of Alberto C. Cruz is approved:

____

____

____

Committee Chairperson

University of California, Riverside

Acknowledgements

I am grateful to my dissertation chair, Dr. Bir Bhanu, who motivated me to do great things. I would like to thank Dr. Anastasios Mourikis and Dr. Aaron Seitz for being my committee members during my thesis defense and my candidacy examination. I would also like to thank my other candidacy committee members: Dr. Gerardo Beni, Dr. and Dr. Jay Farrell for conferring my doctoral candidacy. I am grateful to Dr. Ninad S. Thakoor who inspired me and provided thought provoking discussion. I would like to thank Jhon Gonzalez for his assistance as VISLAB's administrative assistant. I am grateful for my colleagues who, at many times in my career here, provided writing assistance and thoughtful discussion: Anthony C. Bianchi, Benjamin X. Guan, Dr. Hoang T. Nguyen, and Songfan Yang. I would like to thank the IGERT Videobioinformatics faculty and participants for nominating me to compete in two national video poster competitions at the NSF: Dr. Katherine Borkovich, Dr. David Carter, Dr. Dimitrios Morikis, Dr. Andre Obenaus, Dr. Hyle Park, Dr. Victor Rodgers, Dr. Aaron Seitz, Dr. Prue Talbot and Dr. Victor Zordan. Finally, I am grateful to my colleagues whose data enabled my research: Geoffrey Harlow and Ilva E. Cabrera. Support for this work was provided for in part by NSF grants 0727129, 0905671 and NSF IGERT: Video Bioinformatics grant DGE 0903667. Portions of this dissertation appeared in "Facial emotion recognition with anisotropic inhibited Gabor energy histograms", "Vision and attention theory based sampling for continuous facial emotion recognition" and "Score-based facial emotion recognition" by A. Cruz et al. © 2013, 2014 the IEEE.

To my parents, Amethyst C. Cureg and Roberto V. Cruz, and to Laura E. Blough for all the support and never doubting my brilliance, and to the memory of Suresh Kumar Ramachandran Nair.

ABSTRACT OF THE DISSERTATION

Biologically Inspired Facial Emotion Recognition

by

Alberto C Cruz

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

When facial emotion recognition is performed in unconstrained settings, humans outperform state-of-the-art algorithms. The major technical problems of a state-of-the-art system are that: (1) frames in the training data are used build a model for prediction of emotion, including the frames that are redundant or not necessary to describe a person's emotions. (2) The Gabor filter, a frequently used feature descriptor in the field of computer vision, captures background texture as important edge information when it is noise. Additionally, the amount of computer memory required to describe faces using the Gabor filter is undesirably high. (3) Most of the current algorithms do not generalize to unconstrained data because each person expresses his/her emotions in different ways, and the persons in the testing data are not the same persons encountered in the training data. These technical challenges cause current approaches to perform inadequately. We address each of these problems by presenting novel algorithms that are based on the human visual system. The first system, called vision and attention theory, down samples the training and testing data temporally to reduce the memory cost in the same way that the human visual system pays attention to scenes. The second system, called background-suppressing Gabor filtering, represents the face in the same way the human visual system's non-classical receptive field represents a grating to overcome background texture. The third system, called score-based facial emotion recognition, scores a frontal face image's relationship to references of a face. It addresses the issue of a person not being present in the training data. We thoroughly test all systems on four different, publicly available datasets: the Japanese Female Facial Expression Database, Cohn-Kanade+, Man-Machine Interface and the Audio/Visual Emotion Challenge. We find that our systems perform better than other state-of-the-art systems. This work shows promise for the detection of facial emotion in unconstrained settings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS AND ABBREVIATIONS

F_1 F_1 score. See Equation 4.

Affective computing Systems designed to accept human non-verbal input or designed to react to human emotional states.

AIR Avatar image registration. See [10].

Apex labels (With emotions) The time point where an emotion is most visibly intense. Also known as apex.

Arousal (With emotions) a person's interest in the situation. Also known as activation.

AU Facial action unit. Facial action units are the minimal set of possible facial muscle movements. They are numbered, e.g. AU7 would refer to the seventh facial action unit defined in [3].

AVEC The Audio/Visual Emotion Challenge. A yearly grand challenge dataset for video-based facial emotion recognition in terms of the Fontaine emotional model [1].

Background texture Repeated patterns with the same orientation.

Bank (With Gabor filters) a set of filters.

Big-six The concept posited by Paul Ekman that there are six basic emotions that are universal across all cultures. See [2].

Bootstrapping A machine learning technique to subsample the training data.

Cells A portion of the V1 area that responds to a specific grating.

CK+ The Cohn-Kanade+ database. A database for video-based facial emotion recognition in terms of Ekman big-six emotions [2].

Correlation Pearson product-moment correlation coefficient. See Equation 3.

DBN Dynamic Bayesian network. See [7].

DC-offset (With Fourier transform) a phenomena that would cause the lowest frequency of the DFT to have the highest energy.

DCT Discrete cosine transform. See [5].

DFT Discrete Fourier transform.

Dominant frequency (With Vision and Attention Theory) The frequency corresponding to the frequency of highest energy of the DFT of the temporal feature.

Duchene smile A smile involving raised cheek muscles. Sometimes considered to be an authentic smile.

Dynamic (With temporal sampling) sampling at a rate which changes as opposed to uniform sampling.

Expectancy (With emotions) a person’s feelings of familiarity.

Expression energy An improved SIFT-Flow algorithm for measuring facial motion.

Expression Facial muscle movement. In this dissertation, expression specifically refers to a facial action unit. See AU.

FERA Facial emotion recognition and analysis 2011 grand challenge held at the IEEE AFGR workshops. A video-based facial emotion recognition database in terms of Ekman big-six emotions [2].

Fiducial feature points Landmark points on the face that were found to be important for facial emotion recognition in [4].

Fixations When an eye fixates on an object.

Fontaine emotion model The Fontaine emotional model. A system for describing emotion that is capable of describing complex emotions. See [1].

FPLBP Four-patch local binary patterns. An improved LBP that models relationships to other pixels as opposed to just encoding local texture.

Fusion (In classification) when a method fuses multiple sources of information.

Gabor energy filter A filter that approximates how gratings are perceived by the visual cortex at a low level.

Gesture Facial or bodily movements.

Grating A pattern of parallel lines with a specific width and orientation.

HMM Hidden Markov model.

JAFFE The Japanese Female Facial Expression database. A database for image-based facial emotion recognition in terms of Ekman big-six emotions [2], consisting of images of only Japanese females.

LBP Local binary patterns. A facial feature descriptor commonly used in facial emotion recognition.

LBP Local binary patterns. A facial feature descriptor that encodes local texture.

LBP-TOP Local binary patterns from three-orthogonal planes. A spatiotemporal version of LBP.

LDCRF Latent-dynamic conditional random field. An improvement of conditional random fields.

LFW Labelled faces in the wild. An image-based face verification database. See [65].

Local appearance features (With facial feature descriptors) when an algorithm describes the face by examining local pixel neighborhoods as opposed to the global representation of the face.

Local (With facial feature descriptors) considering the neighboring pixels.

LPQ Local phase quantization. A facial feature descriptor based on local patches described in the frequency domain.

Maximal response (With background suppressing Gabor filtering) the detected edge at a pixel with the highest magnitude.

MMI-DB The Man-Machine Interface database. A database for video-based facial emotion recognition in terms of Ekman big-six emotions [2].

NCRF Non-classical receptive field. The ability of the human visual system to remove background texture from a stimulus.

Nyquist rate A should be sampled at twice the highest frequency present to avoid introducing errors. Also known as Nyquist frequency.

Optical flow An algorithm that computes the motion between two frames.

OSE One-shot emotion. A method for person-independent classification based on OSS.

OSS One-shot scores. See [13].

Person-dependent Recognition when the individuals in testing are the same individuals in training.

Person-independent Recognition when the individuals in testing are not the same individuals in training.

Power (With emotions) a person’s feeling of control over oneself or the situation. Also known as potency or power-control.

Proposed The method proposed in the chapter.

RBF Radial basis function. A kernel function for SVM. See [9].

ROI Region of interest. (With face detection) the region where the face(s) is located in an image.

Saccades Rapid eye movements.

Segmented (With video data) when a video is given in a cut form that covers on the significant portions of the video.

SIFT Scale invariant feature transform. See [6].

SIFT-Flow An algorithm that computes the motion between two frames similarly to optical flow. See [11].

Sub-regions The process for extracting facial features where the face is divided into equally sized non-overlapping sub-regions to account for face morphology. Also known as cells, and gridding.

SVM Support vector machine. A machine learning algorithm.

SVR Support vector regression. A machine learning algorithm related to SVM.

Temporal feature (With Vision and Attention Theory) a feature describing the amount of change from frame to frame.

TPLBP Three-patch local binary patterns. An improved LBP that models relationships to other pixels as opposed to just encoding local texture.

UA Unweighted accuracy. Average true positive rate among all classes.

Valence (With emotions) a person's feelings toward the situation.

Viola-Jones The Viola and Jones face detector. See [8].

WA Weighted accuracy. Also known as classification rate.

XOR Exclusive OR.

CHAPTER 1 INTRODUCTION

Among the first researchers in the field of facial emotion and expression recognition was Charles Darwin who published, “Expression of the Emotions in Man and Animals,” in 1872 [7]. Darwin connected human expressions and bodily movements to emotional states. Communication between two humans is a complex process that involves more than just speech. We communicate non-verbally with gestures, pose and expressions. *Gestures* are a general term for motion of the body. For example, a person could give thumbs up while communicating and this is considered a gesture. *Pose* refers to the position and orientation of the body. For example, a person could pose by turning his/her face away from another person while communicating. *Expressions* are facial muscle movements, e.g. muscle movements corresponding the opening of a mouth. Expressions and emotions are not the same. *Emotions* are the underlying feelings of a person, which may be revealed his/her expressions, pose and gestures.

The understanding of human expressions and emotions is a biological process—particularly when framed in the context of their origins in mammals as Charles Darwin studied them. When two humans communicate, they use their hands to gesture, they use their facial muscles to form expressions, they focus their gaze, and they pose their face. Facial expressions are critically important in this non-verbal communication between humans.

Video-based facial emotion recognition is an important field of study where face video of a human is captured and computer algorithms must detect his/her facial expressions to infer his/her underlying emotional state. Expression/emotion recognition has applications in medicine (Asperger's Syndrome [8], Autism Spectral Conditions [9]), video games (Xbox Kinect [10]), human-computer interaction (embodied conversational agents [11]), deception detection [12] and *affective computing*. Affective computing is a field where computer interfaces are able to project expressions to facilitate non-verbal communication with a human. There has been an increased interest in facial emotion recognition and the field has seen great advances. However, a system that can detect facial emotions in unconstrained settings has yet to be seen.

Novel computer algorithms for facial emotion and expression recognition based on the human visual system are the focus of this dissertation. In Chapter 2 we discuss the related work for sampling methods, feature representations and classification schemes. In Chapter 3 we discuss the categories and systems for quantifying emotion, the datasets and the metrics to be used in the dissertation. In Chapter 4 we discuss a method for downsampling video data based on the human attention. In Chapter 5 we discuss a method for feature representation based on non-classical receptive field, the ability of the human visual system to remove background texture when perceiving gratings. In Chapter 6 we discuss a method for classification that can be used when a person has no representation in the training data. In Chapter 7 we give the conclusions of the dissertation.

CHAPTER 2 RELATED WORK

In the baseline visual system for the Facial Emotion Recognition and Analysis 2011 (FERA) [13] and the Audio/Visual Emotion Challenge (AVEC) datasets [14] [15], face region-of-interest (ROI) is extracted which is then aligned by eye corner points. Subsequently, Local Binary Patterns (LBP) [16] are extracted as histogram-based features, and the emotions are classified with a support vector machine (SVM). In [17], the top approach for discrete emotions on the FERA dataset, Yang and Bhanu introduced a novel registration procedure called avatar image registration. It was found that a better registration method greatly improved performance. In [4], Valstar et al. tracked 20 fiducial facial points and classified them using a probabilistic actively learned SVM.

In [18], Ramirez et al. quantified eye gaze, smile and head tilt with a commercial software (Omron OKAO Vision and Fraunhofer Sophisticated High-speed Object Recognition Engine) and used a Latent-Dynamic Conditional Random Field (LDCRF) [19] classifier. It was concluded that properly choosing classifier had a significant impact on performance. In [20], Glodek et al. modeled their system after the human perception's capability to separate form and motion. Gabor filters captured spatial information, and correlation features captured temporal information. The features were fed into multiple stages of filtering and non-linear pooling to further simulate human perception. It was found that the emotions expressed in AVEC were subtle and difficult to detect. In [21], Dahmane and Meunier proposed an approach for representation of the response to a bank of Gabor energy filters with histograms. An SVM with a radial basis function was used as a classifier. In that work, succeeded in compressing the feature vector length of the Gabor filter.

In [22], Nicolle et al. used 3-D model fitting, and global and local patch-based appearance features. These features were extended temporally with log-magnitude Fourier spectrum. A correlation based feature selector was proposed and a Nadaraya-Watson estimator was used as a classifier. During ground-truth labeling, the expert watches the video, and then notes changes in the label. There is a time delay between the actions in the video and when the expert notes the change. They found that accounting for this delay improved classification rate. In [23], Soladie et al. employed two active appearance models, one to quantify head pose, and one to quantify smile. A *Mamdani* type fuzzy inference system was used. The features included who the person was speaking with, duration of sentences, and how well engaged the person was in the conversation with the embodied agent. It was found that this situational context greatly improved performance. In [24], Maaten used the baseline features, the derivative of features, and L_2-regularized linear least-squares regression. In [25], Ozkan et al. proposed a concatenated hidden Markov model (co-HMM). The label intensity values were discretized into bins. An HMM was trained to detect a specific bin, e.g., if there were ten quantization levels, then there would be ten classifiers each detecting if that specific level was present. A final HMM fused these outputs at the decision level. However, it was found normalizing the ground-truth intensities per-person had a better impact on performance. In the video-based approach in [26], Savran et al. extended local appearance features to the temporal domain by taking the mean and standard deviation in sliding temporal windows. AdaBoost was used as a feature selector, and ϵ-support vector regression (SVR) was used to regress the labels.

A. Sampling Methods

In this section we discuss state-of-the-art methods with a particular interest to how samples are selected for model training and for testing. Some approaches have attempted to address the sampling issue. In [20], Glodek randomly sampled the video frames. In [21], a downsampling method was proposed that changed granularity of sampling based on whether or not a change was detected in the predicted label. A limitation of this system is that it assumes that the system can correctly predict the label. In [27], Zhu et al. reduced the number of frames in the dataset with a bootstrapping procedure. This method requires the apexes to be labeled. The apex is the time point where emotion is most intense. We propose a method that does not require peak frame labeling. In [26], Savran et al. down sampled the training data to frames that had an emotion label intensity greater than $\pm\sigma$ from the mean emotion intensity. No framework for downsampling test data was provided. In [28], Jiang et al. proposed a texture descriptor that extended Local Phase Quantization (LPQ) features to the temporal domain. It was called Local Phase Quantization from Three Orthogonal Planes. The paper also investigated three downsampling methods: randomly selecting frames, bootstrapping, and a heuristic approach that found two subsets of the data to describe static appearance descriptors and dynamic appearance descriptors. It was found that the heuristic method was the best performer. All of the methods in [28] focused on training data selection, and required apex labels.

B. Feature Representations

In this section we discuss state-of-the-art methods with a particular interest in the feature representation used by that method. A comprehensive survey of audio and visual emotion recognition methods and categorization of human emotions are given in [29]. Facial features in state-of-the-art methods can be grouped into two categories: geometric features, e.g. tracking of fiducial feature points [4], or local appearance features, e.g. texture and color features. Of particular interest to this work are local appearance features, the most commonly used of which are Local Binary Patterns (LBP) [16]. Though the features are often referred to as LBP features, they are actually histograms of an LBP coded image. LBP quantifies textures at a pixel level by encoding the microtexture of a pixel and its neighborhood with an 8-bit code. Current methods often divide the frontal face into sub-regions and compute the histogram of LBP codes for each sub-region. For example, in [15], the face was divided evenly into 10×10 sub-regions, or grids, and the outer regions were discarded because they corresponded to the regions of a face where there were no facial expressions. Uniform LBP features have been used as the baseline for recent facial emotion recognition grand challenges [15]. There have been many improvements to the original LBP feature. [30] proposed Three-Patch and Four-Patch Local Binary Patterns (TPLBP, FPLBP). Whereas LBP encodes a microtexture of a single pixel, TPLBP and FPLBP encode larger patterns and homogeneity of a region by comparing a pixel's microtexture to the microtextures of neighboring pixels. [31] proposed extending LBP to a spatiotemporal feature with the use of three orthogonal planes (LBP-TOP). [32] extended LBP to the spatiotemporal domain with monogenic signals analysis and phase-quadrant encoding and a local XOR operator in three orthogonal planes (STLMMBP).

Not all facial emotion recognition methods use LBP as a local appearance feature. The top approach for the Facial Emotion Recognition and Analysis challenge for discrete emotions used Local Phase Quantization (LPQ) [17]. In LPQ, the phase of a per-pixel discrete Fourier transform (DFT) quantifies the texture. It was found that the phase of DFT of a local neighborhood is invariant to centrally symmetric blur. Sub-region histograms give LPQ a compact representation. [5] used a difference image to quantify facial motion, and a discrete cosine transform (DCT) to compress the feature vector size. [6] proposed the scale-invariant feature transform (SIFT), which quantifies local features with the maxima and minima of a difference-of-Gaussians. Recently, it was used by [33], where the SIFT features were computed at 83 fiducial feature points.

C. Classification Schemes

In this section we discuss related work with a particular interest on cross-database testing and classification schemes. In [8], Ghanem used optical flow to track facial points, and focused on segmenting the video temporally. Results were reported on interdatabase experiments on MMI and CK+. In [34], Li et al. made a distinction between facial expressions, gestures caused by emotion, and facial feature points. A dynamic Bayesian network modeled the relationship between each of the three levels. In [35], Miao et al. proposed a novel supervised extension to kernel mean matching. A class to class matching was performed with a limited number of samples. Wolf et al. [30] proposed learning-with-side-information, and four-patch and three-patch local binary patterns (LBP). It is not a facial emotion recognition approach, but it is notable because it is the face processing pipeline that inspired the score-based approach presented later in the dissertation. In [20], Glodek et al. separately processed visual information along two separate pathways that represented appearance and dynamic information, resembling the processing pathways in the human visual system.

CHAPTER 3 EMOTIONS, DATASETS AND METRICS

A. Quantifying Emotion

Before emotions and expressions are detected, they must first be quantified. This is an ongoing field of research in psychology and neuroscience, and we will highlight three ways to quantify emotions and expressions. We use discuss three labeling systems: action units [3], emotions based on the Ekman big six [2] and the Fontaine emotional model [1], which is an extension of affective dimensions.

Expressions are facial muscle movements. Ekman and Friesen [3] defined the minimal set of facial muscle movements, or action units (AUs), that are used in expressions. This is the Facial Action Coding System. For example, a smile consists of AU 6 and AU 12. AU 6 indicates that a person's cheeks are raised. AU 12 indicates that the corners of a person's lips are being pulled outward. This often occurs when smiling. Emotion differ from expressions in that they are the underlying mental states that may illicit expressions. A common system for discrete emotional states is the Ekman big six: happiness, sadness, fear, surprise, anger and disgust. Ekman posits that these six emotions are basic emotions that span across different cultures [2].

A different system for emotion labels is the Fontaine emotional model [1] with four affect dimensions: valence, arousal, power and expectancy. An emotion occupies a point in this four-dimensional Euclidean space. Valence, also known as evaluation-pleasantness, describes positivity or negativity of the person's feelings or feelings of situation, e.g., happiness versus sadness. Arousal, also known as activation-arousal, describes a person's interest in the situation, e.g., eagerness versus anxiety. Power, also known as potency-control, describes a person's feeling of control or weakness within the situation, e.g., power versus submission. Expectancy, also known as unpredictability, describes the person's certainty of the situation, e.g., familiarity versus apprehension. For a more detailed explanation, the reader is referred to [1]. With this system, multiple emotions can be expressed at the same time. An Ekman big six emotion occupies a region in each of these four dimensions.

For example, happiness would be positive valued valence because the person would feel positive about the situation. It would have positive arousal, because the person would enjoy the situation. It would have positive power, because a person would likely need to feel in control of himself to feel happy. It may be any value for expectancy, because a person may or may not be both surprised and happy.

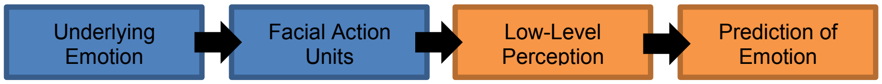


Figure 1: Overview of how emotions are projected and perceived by other humans. (Orange) Prediction of another humans' emotion is a two-step process of perceiving a person's face a low level and predicting what emotion is being expressed.

In the previous paragraphs, we discussed the process by which a human expresses their emotions with their face. An overview of how humans communicate their emotions non-verbally is given Figure 1. First, a human has an underlying emotion. That human will move their facial muscles, which are quantified by AUs. Groups of AUs form a *gesture*, such as a smile. These muscle movements are projected and perceived by another human. It is processed by the human visual system at a low-level, and a judgment is made by the other human as to what emotion the person is expressing. It is possible that the expressions projected by the human are not the underlying emotion, such as when a person is acting or when a person is being deceptive. In certain cultures outward displays of emotion are frowned upon. This is why detecting the emotions of another human can be a difficult task.

Table 1: Comparison of Publicly Available Data Sets

					Class Percentage (%)					
Name	Type	Apex	Labels	# Samples	Anger	Disgust	Fear	Happy	Sad	Sur.
AVEC [14]	Continuous	No	Affective Dimensions	270225	Regression problem, class percentage N/A.					
CK+ [36]	Segmented	Yes	Big-six	296	14.9	15.4	14.9	20.7	15.4	18.6
MMI [37]	Segmented	Yes	Big-six	118	14.5	18.9	8.1	22.6	9.1	26.7
JAFFE [38]	Images	No	Big-six	198	25.3	18.7	18.2	14.6	19.7	3.5

Table 2: Summary of experiment parameters of related work.

Method	# Video/Image		I/D	Validation	Classes
Ghanem [8]	100		D	3 fold	Joy, anger, sadness, disgust
Li et al. [34]	309		I	Leave-one-subject-out	Big-six
Miao et al. [35]	-		I	Leave-one-subject-out	Big-six
Poursaberi et al. [39]	96		D	Leave-one-out	Big-six
Yang and Bhanu [17]	316		I	Leave-one-subject-out	Big-six
This Dissertation	See Table 3		I	Leave-one-subject-out	Big-six

Table 3: Class percentage for positively expressed AU for CK.

AU1	AU2	AU4	AU5	AU6	AU7	AU9
29.2	19.6	31.7	16	22.7	22.1	10.2
AU10	AU12	AU15	AU20	AU24	AU25	AU27
2.5	23.1	15.1	14.1	8.6	60.1	15.5

B. Datasets

The field of facial emotion and expression recognition has advanced with the help of publicly available data sets. Among the first was the First Japanese Female Expression Data set [40]. Since then, there have been many data sets available: Cohn-Kanade+ (CK+), MMI Facial Expression Database (MMI-DB), the First Facial Expression Recognition and Analysis grand challenge (FERA), the Audio/Visual Emotion Challenge (AVEC), ordered by date. The field has moved toward more spontaneous, naturally collected data. A comparison of publicly available data sets is given in Table 1. A comparison of the experimental parameters of related work is given in
Table 2.

1. The Japanese Female Facial Expression Database

JAFFE is an image dataset that has images of varying emotion intensity. We follow the data methodology in Miao et al. [35], where the person 'NM' is not used. We use only images/video where there is a frontal face. We use a single multiclass classifier.

2. Cohn-Kanade+

The second dataset used is CK [36]. We use this database to test the quality of results of the proposed sampling method, when apex labels are provided. The length of segments range from 3 frames to over 100. We also use it to for cross-database tests when comparing results to other classification schemes. The percent of positively expressed action unit (AU) are given in
Table 3. We follow the testing methodology in Koelstra et al. [41]. An AU is selected if it

has more than 10 positive examples. We focus on the following actions units (AU): {1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 20, 24, 25, 27}. The reader is referred to Lucey et al. [36] for a more detailed explanation of the data. We use leave-one-person-out cross-validation. A binary classifier is used for each AU.

3. Man-Machine Interface Database

MMI-DB [37] is frontal face video data similar to CK. For most videos, the emotion peaks near the middle of the video. The percentage of class for each emotion is given in Table 4. We use leave-one-person-out cross validation. We use all sessions that have emotion labels, and we consider the classes with at least 10 positive examples. We use only frontal faces. A multi-class classifier is used.

Table 4: Class percentage for MMI-DB emotions.

Anger	Disgust	Fear	Happy	Sad	Surprise
21.1	13.9	13.0	19.7	14.4	17.9

4. The Audio/Visual Emotion Challenge

AVEC 2011 [14] and 2012 [15] are grand challenge datasets. In this chapter, they are used to compare the performance of a proposed method to other state-of-the-art methods. It is a non-trivial, unconstrained dataset: (1) the frame rate is too high to load all frames into memory. For example, if AVEC 2012 has 1351129 frames, if LBP features and baseline audio features [15] are used which have 7841 dimensions, and if double floating points are used for each feature, it would require 8.48 GB to load all frames into memory. This exceeds the memory of most computers (88.9% of computers have up to only 8 GB of computer memory according to a recent hardware survey [42]). (2) The subjects are free to change pose, and use hand gestures, and (3) the videos are not acted.

The videos are not pre-cut, and a person can express multiple emotions per video. In the AVEC datasets, a person is presented with an embodied agent who engages the person in conversation, and causes emotionally colored conversations by being biased to express a particular emotion, such as belligerence or sadness. Emotions expressed in this scenario are natural, continuous, and spontaneous. An example is available online in Figure 2. In this example, a person is interacting with a specific character named Spike. Spike is confrontational, and aggravates the person during conversation. Note that the person is smiling, but not from being pleased. The smile is caused by the person being polite and exercising restraint in response to hostility. A classifier is used for each affect dimension.

Figure 2: QR code leading to the web page at:
https://www.youtube.com/watch?v=6KZc6e_EuCg

The AVEC datasets are divided into three partitions: (a) 31 interviews of 8 different individuals form the training set. It is used as samples for a training model. (b) 32 interviews of 8 individuals, who are different from the training set form the development set. It is used as the testing fold in the training phase, and (c) 32 (AVEC 2012) or 11 (AVEC 2011) interviews of new individuals who are not in the development or training set form the testing set. The testing set is the official validation fold with which algorithms are compared to each other. The average length of all the videos in AVEC 2011 is 14.6×10^3 frames with a standard deviation of 5.2×10^3. All results are given in terms of the frame level subchallenge. The percentage of positively expressed affective dimension for the training and development datasets for AVEC 2011 dataset are given in Table 5. The percentages for the testing set are not available because the labels are withheld by the challenge organizers.

Table 5: Class percentage rate for AVEC 2011.

Sets	Arousal	Expectancy	Power	Valence
Training	47	46	51	55
Develop	56	40	59	64

C. Metrics to Quantify Performance

The AVEC datasets have two scoring systems. In AVEC 2011 [14] the metrics are weighted accuracy (WA) and unweighted accuracy (UA). Weighted accuracy is the classification rate, and is also known as percent correct, calculated as follows:

$$\mathrm{WA} = \frac{1}{n_c}\sum_{i=1}^{n_c} p(c_i)\frac{tp_i}{tp_i + fp_i} \qquad 1$$

where tp_i is the number of true positives of class i, fp_i is the number of false positives of class i, and $p(c_i)$ is the percentage of class. Unweighted accuracy is defined as:

$$\mathrm{UA} = \frac{1}{n_c}\sum_{i=1}^{n_c} \frac{tp_i}{tp_i + fp_i} \qquad 2$$

This metric is used because some classes in the data have disproportionate percentage. For example, positive valence has a percentage of class higher than 60% in the training fold.

In the AVEC 2012 scoring system, each emotion’s value is given a real number. The task is a regression problem. The algorithm detects the real valued emotion on a per-frame basis. While there are many metrics that could be used to quantify performance, the official metric for AVEC 2012 is Pearson product-moment correlation coefficient with the ground-truth. It is calculated as:

$$\rho = \frac{E[(c - \mu_c)(\hat{c} - \mu_{\hat{c}})]}{\sigma_c \sigma_{\hat{c}}} \qquad 3$$

where c is the vector of the ground-truth labels across all videos concatenated into a single vector; $\hat{c}$ is the vector of predicted labels; μ_c is the mean of c; and σ_c is the standard deviation of c. Results for CK+ are given in terms of true positive rate, false positive rate, false negative rate, true negative rate, precision, recall and F_1-score. The F_1-score is:

$$F_1 = 2\frac{(\text{precision} \ * \ \text{recall})}{(\text{precision} + \ \text{recall})} \qquad 4$$

CHAPTER 4 BIOLOGICALLY INSPIRED FRAME SELECTION

Current methods perform well on datasets acquired in controlled situations, e.g. the Japanese Female Facial Expression database [43], Cohn-Kanade (CK) [36], the MMI Facial Expression Database (MMI-DB) [37], and the Facial Emotion Recognition and Analysis (FERA) challenge dataset [13]. However, the Audio/Visual Emotion Challenge (AVEC) datasets [14] [15] present difficult challenges. With previous datasets, each dataset was small enough to be loaded into memory at once, even for cases of high feature dimensionality. Previous approaches could reduce the number of frames to be processed by taking advantage of apexes of emotions, such as in CK. The most intense and discriminative frames corresponding to the apexes were labeled so a method could choose to retain them only.

The AVEC datasets explore the problems of a continuous emotion dataset, where it is computationally undesirable to select all the frames for processing. There are approximately one and a half million frames of video. The expressions in the dataset are subtle, spontaneous, and difficult to detect. The people in the videos are expressing emotions in a natural setting. The videos are not segmented. The apex labels are not given and it may be difficult to detect them automatically. In this chapter, we propose a principled method for downsampling the frames for facial emotion and expression recognition. The method is inspired by the behavior of the human visual system. It can take advantage of apexes if they are provided, but they are not required.

We propose emulating the behavior of the human visual system to address the challenges in the AVEC datasets. The focus of work in this chapter is video-based temporal sampling. The contributions of the method discussed in this chapter are:

(1) We exploit vision and attention theory [44] [45] from perceptual psychology to determine an appropriate sampling rate. We assign a dynamic, temporal granularity that is inversely proportional to how frequent the visual information on a person's face is changing. The method improves average correlation with the ground-truth for all affect dimensions on the AVEC 2012 frame-level subchallenge testing set over the baseline approach by a factor of 2.7.

(2) We provide a framework for the method to integrate information from apex labels, if they are provided. The method improves average F_1 measure across 14 different classes by 7.6 over [17].

(3) We provide a framework for using match-score fusion temporally. The method improves average weighted accuracy on all classes on the AVEC 2011 frame-level subchallenge development set over the use of uniform sampling of 1 frame per segment and no fusion by 5.4%.

A. Motivation

In the AVEC datasets, videos are captured at a high frame rate and over a long period of time. This makes it difficult to train a model for classification using all the frames in the dataset. An easy solution is to temporally down sample the video at a uniform, low frame rate. Unfortunately, this procedure results in a loss of precision as it does not have the ability to precisely detect when the emotion changes. A dynamic sampling rate is desired that assigns a lower frame rate to parts of the video where the person is idle, and a higher frame rate to parts of the video where the person is animated. For example, in Figure 3, there are two different segments of the same video which merit different sampling rates. In Figure 3-(a), the person is changing his pose, opening his mouth, furrowing his brow, using his cheek muscles, and raising his eyebrows. Many frames are needed to describe this segment. In Figure 3-(b), the person holds his expression, so this segment would need only a few frames to be described. Therefore, we propose a method that applies a dynamic sampling rate which would allocate fewer frames for data analysis when the individual is idle, and more when the individual is active. The large volume of data poses the following problems to a downsampling procedure:

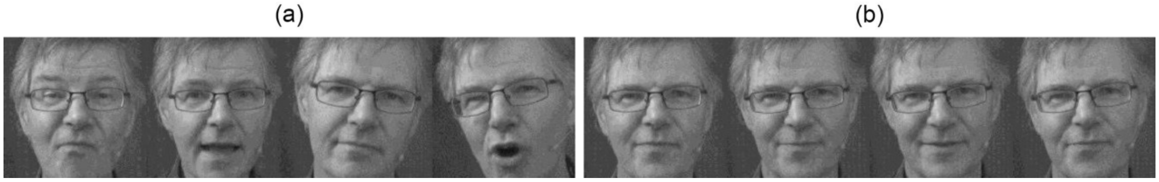


Figure 3: Two different segments of AVEC [14] development video 14. (a) Many frames are required to describe the person's pose change and facial expressions. (b) The person is less expressive and the segment needs few frames to be described.

1. Processing Cost

Though related work [27] propose dynamic downsampling, these methods prune samples *late* in the recognition pipeline, i.e. the decision to remove a sample from consideration occurs at the very end of the recognition pipeline. With the AVEC datasets, processing each frame would be too costly. To prevent unnecessary computation the downsampling should occur as early as possible in the video processing pipeline.

2. Must Not Use Apex Labels

Use of the apex label is popular in facial expression and emotion recognition, and results show that features from the apex region improve classification rates [46] [47] [48]. However, the apexes must be *manually* labeled by an expert. If an algorithm is used to detect the apexes, the labeling can have errors. Situations may arise in the AVEC datasets where expressions are so subtle that extracting apex information is a difficult task for both humans and computers. There is a need for annotation free facial emotion and expression recognition. Our method does not require apex labels.

B. Technical Approach

When viewing a natural scene, the human visual system exhibits a saccade-fixation-saccade pattern [49]. *Fixations* are moments where the eye fixes on an object, and visual information is processed. *Saccades* are rapid movements of eyes, where information is not being processed. First the eyes saccade, then fixate, and this procedure is repeated. The latency between two saccades decreases with the increasing frequency of temporal changes of visual information in the scene. We propose a method that emulates this process for emotion and expression recognition. Human perception of faces is different than recognition of scenes or other objects. However, the focus of work is the concept of *attention*, the length of focus on a scene, not recognition. The temporal frequency of visual information in the scene affects the amount of attention given to a part of the scene. Our algorithm is inspired by this physical process and emulates attention by downsampling a video.

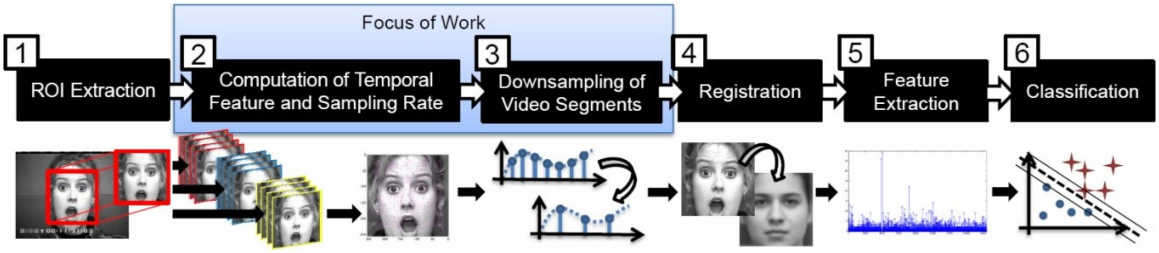

Focus of Work
1
ROI Extraction
2
Computation of Temporal Feature and Sampling Rate
3
Downsampling of Video Segments
4
Registration
5
Feature Extraction
6
Classification

Figure 4: System overview. (1) Extraction of ROI. (2) Partitioning of video into smaller segments, formation of temporal feature that quantifies motion, and computation of the dominant frequency of the temporal feature. (3) Downsampling of the video segment. (4) Registration of frames. (5) Appearance feature extraction. (6) Classification/regression.

The overview of this work is shown in Figure 4: (1) face ROI is detected with Viola-Jones [50]. (2) The video is partitioned into segments. Within each segment, the visual information is quantified with temporal features. We apply a discrete Fourier transform to the temporal feature to find the *dominant frequency*, the frequency of the temporal feature with the most energy. (3) The video is down sampled at the dominant frequency. (4) The selected frames after the downsampling are aligned with avatar image registration [17]. (5) Appearance features are generated in local regions, for each selected frame. (6) Initial *a posteriori* probabilities of emotion labels in each frame in the video segment are generated from SVM [51]. The results are temporally fused at the match-score level [52] to generate the final predicted labels. Section Chapter 4 B.1 discusses downsampling for continuous videos, Section Chapter 4 B.2 discusses downsampling when apex labels are given. The full emotion recognition pipeline is described in Section Chapter 4 B.3.

1. Downsampling Continuous Video

Downsampling of a continuous video without time annotations for apexes is done as data comes in. The videos are segmented into uniformly sized smaller segments. Each segment is down sampled *dynamically*, and each segment has its own appropriate downsampling factor. Conventionally, each segment would be processed with a *uniform* downsampling factor. Psuedocode for the downsampling method is given in Algorithm 1.

Algorithm 1: Computing the sampling rate for single segment/single apex

```
procedure downsampleSegment($I_\Phi$)
    for all frames $n \in \Phi$ do
        $\Delta I_n \leftarrow$ optical flow from $n-1$ to $n$
        $f(n) \leftarrow \sum_x \|\Delta I_n(\boldsymbol{x})\|_2$
    end for
    $\boldsymbol{f}_\Phi \leftarrow$ vector corresponding to all features $f$
    $\tilde{\boldsymbol{f}}_\Phi \leftarrow \boldsymbol{f}_\Phi$ – mean of $\boldsymbol{f}_\Phi$
    $\boldsymbol{F}_\Phi \leftarrow$ Discrete Fourier transform of $\tilde{\boldsymbol{f}}_\Phi$
    $\beta \leftarrow \arg\max_k \|\boldsymbol{F}_\Phi(k)\|$
    if $n_0$ is given then
        $\Phi^*_{Apex} \leftarrow$ range $n_0 - \beta/2 < n \leq n_0 + \beta/2$
        $\Phi^* \leftarrow \Phi^*_{Apex}$
    else
        $M \leftarrow N/\beta$ {Downsampling factor}
        $\Phi^* \leftarrow \Phi \downarrow M$ {Every $M$-th frame}
    end if
    return $I_{\Phi^*}$
end procedure
```

a) Time Partitioning Procedure

The video I is segmented into equally sized non-overlapping segments of N frames. The segment of video I_Φ contains the frames at indices Φ where: $\Phi = \{m_o, m_o+1, \ldots, m_o+N+1\}$. The down sampled video segment I_{Φ^*} contains the frames at indices Φ^*, where Φ^* is an ordered subset of Φ. Initially, the system delays for N

frames, and processes a video segment of N frames at a time. We start with $m_o = 0$, so the first N frames form one segment. Then $m_o = N$, so the frames from N to $2N - 1$ form another segment and so on, until the end of the video. If there is a remainder, it forms its own segment. We chose parameter N such that the duration of each segment is 1 s because 1 Hz is the maximum bound of the HVS according to vision and attention theory [49].

b) Computing the temporal feature

I_{Φ^*} is created by resampling I_Φ at a lower frequency. The first step is to quantify facial expressions into a signal that varies with time. The signal's frequency must respond to changes of facial expression. Because the frame rate is high, and the ROI is a frontal face, optical flow can be exploited to quantify the facial expressions [53]. ΔI_n is optical flow between the frames I_n and I_{n-1}. It outputs a motion vector. The magnitude is summed for all pixels in an image to form a 1-D signal:

$$f(n) = \sum_{\boldsymbol{x}} \|\Delta I_n(\boldsymbol{x})\|_2 \qquad 5$$

where $f(n)$ is the temporal feature for a single frame, $\boldsymbol{x}$ is a pixel, and $\|.\|_2$ is the magnitude. For the entire segment I_Φ, the temporal feature $\boldsymbol{f}_\Phi$ is indicated by: $\boldsymbol{f}_\Phi \equiv [f(m_0), f(m_0+1), \dots, f(m_0+N-1)]$. Figure 5 shows how the video is segmented, how the optical flow is computed, and how the temporal feature is generated. As registration is costly, to reduce the number of frames to be registered, we compute the optical flow before registration. We do not use optical flow as a feature for classification, nor for alignment.

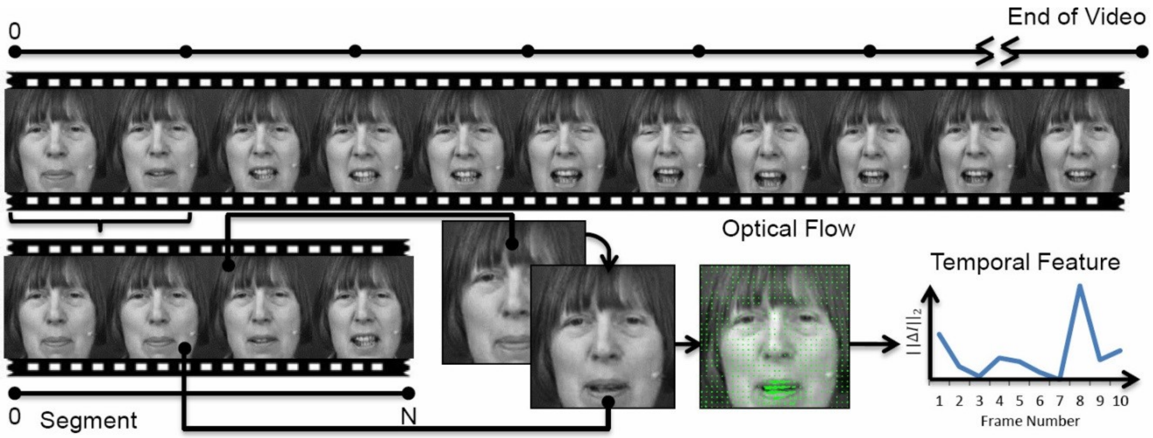


Figure 5: Overview of how the temporal feature is computed. The video is segmented into non-overlapping segments of length N. Optical flow is computed using a pair of adjacent frames. The result of the optical flow forms the temporal feature.

c) *Downsampling the video segment*

To compute the dominant frequency, first, the DC-offset is removed:

$$\tilde{\boldsymbol{f}}_{\Phi} = \boldsymbol{f}_{\Phi} - E(\boldsymbol{f}_{\Phi}) \tag{6}$$

where $E(.)$ is the expected value operator. It is important to remove the DC-offset for two reasons: (1) it normalizes the temporal feature and (2) for real data, the $\boldsymbol{F}_{\Phi}(0)$—corresponding to the coefficient at 0 Hz, the DC-offset—will be greater than other values of $\boldsymbol{F}_{\Phi}$, causing it to be selected as the dominant frequency. $\boldsymbol{F}_{\Phi}$ is the discrete Fourier transform of $\tilde{\boldsymbol{f}}_{\Phi}$: $\boldsymbol{F}_{\Phi} = DFT(\tilde{\boldsymbol{f}}_{\Phi})$, where $DFT(.)$ is the discrete Fourier transform, and k is the frequency index. The frequency index corresponding to the frequency with the most energy β and is computed as follows:

$$\beta = \arg\max_{k} \|\boldsymbol{F}_{\Phi}(k)\| \tag{7}$$

where $\|\boldsymbol{F}_{\Phi}(k)\|$ is the magnitude of $\boldsymbol{F}_{\Phi}(k)$. Note that the frequency in Equation 7 is not the Nyquist rate. The Nyquist rate applies to sampling a continuous signal in order to accurately reconstruct that signal. In this chapter we are downsampling a discrete signal by removing samples in the signal which have not changed much. For this reason, we sample at the dominant frequency itself.

The downsampling factor M is given by: (maximum frequency/dominant frequency). The frequency index β can be converted to the dominant frequency as: $2\pi\beta/N$. The maximum frequency index N corresponds to frequency 2π. It follows that: $M = N/\beta$. Let $\Phi^* \leftarrow \Phi \downarrow M$. That is, Φ^* is every M-th frame of Φ. When the temporal feature has a high frequency, $\beta \to N$, the downsampling factor is near 1, and all of the frames are preserved. When the temporal feature has a low frequency, the downsampling factor increases, and most of the frames are removed.

2. Downsampling with Apex Labels

When apex label information is given, instead of segmenting the video evenly, the system segments the video into durations centered at each apex. Instead of downsampling the segment evenly, the dominant frequency effects the duration of the segment. If the dominant frequency is high, then the method will select many frames at the apex; if low, only the frames nearest to the apex are selected. The human visual system has dynamic attention based on spatiotemporal changes of visual information. We realize attention as the number of selected frames. If there is not much change in the visual information, there is less attention given, and fewer frames are selected.

a) Time partitioning procedure

If apexes are provided, the video is partitioned into uniform segments of N frames, centered at the midpoint of the apex frames. There is a segment for each apex, and each segment is centered at that apex. Frames that are not near an apex will be removed. Let n_0 be the location of an apex. It now follows that:

$$\Phi_{Apex} = \left\{n: n_0 - \frac{N}{2} < n \ \leq n_0 + \frac{N}{2}\right\} \qquad 8$$

Ordinarily we down sample the segment evenly. However, when apex labels are given we reformulate the downsampling method to take advantage of these labels. At the apex, the expressions are strong and the emotion is more easily detected. For this reason, the frames in the duration centered at the apex should be retained, rather than downsampling uniformly, which may retain frames further away from the apex where emotions are more difficult to detect. An example comparing sampling at a uniform rate versus sampling at the apex is given in Figure 6. There is no change in the way β is computed.

b) *Downsampling the video segment*

In this formulation, Φ^*_{Apex} varies in duration according to β, and is defined as follows:

$$\Phi^*_{Apex} = \left\{n: n_0 - \frac{\beta}{2} < n \ \leq n_0 + \frac{\beta}{2}\right\} \qquad 9$$

If apex labels are given, Φ^*_{Apex} is taken to be Φ^* . When the temporal feature has a high frequency, N frames are preserved and Φ^* is equivalent to Φ^*_{Apex}. When the feature has a low frequency, the number of frames approaches $1$, and most of the frames are removed.

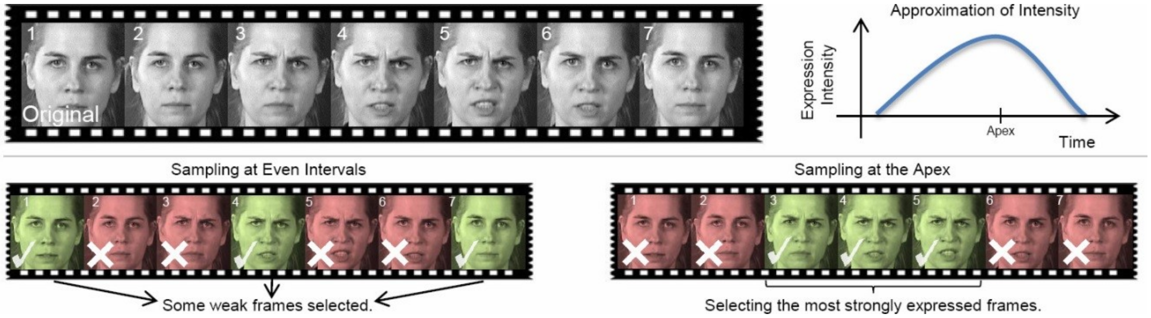

1
2
3
4
5
6
7
Original
Approximation of Intensity
Expression Intensity
Apex
Time
Sampling at Even Intervals
Sampling at the Apex
Some weak frames selected.
Selecting the most strongly expressed frames.

Figure 6: Comparison of sampling at even intervals versus sampling at the apex. A video is given, and its expression intensity is given. Sampling at even intervals retains frames that are further away from the apex. They are weakly expressed, and they are not a good representation of the emotion being expressed. Sampling at the apex retains the frames where the emotion is most strongly expressed.

3. Emotion Recognition Pipeline

a) Face ROI extraction, registration and features

Faces are detected with a boosted cascade of Haar-like features [50]. If a face is not detected in the frame, we assign the expected label to that frame. For classification, we assign the class label that has the highest percentage of class occurrence. For regression, we assign the average value of the emotion intensity from the training data. A better method for assigning the label in this situation would be a first-order Markov assumption, but this is not the focus of work (see [54]). If ROI is detected, faces are registered with avatar image registration. The reader is referred to [17] for a more in depth explanation. We use Local Binary Patterns (LBP) because they are the most popular features in the field for representing a face. The reader is referred to [31] for an in depth explanation. The features are computed for each frame in I_{Φ^*}.

b) Fusion

A method is needed to temporally fuse and smooth the estimated emotions. For each segment I_{Φ^*}, we propose fusing the *a posteriori* probabilities for each frame computed by the classifier. *A posteriori* probabilities are obtained with SVM [51]. The *a posteriori* probabilities are fused with combination-based match-score fusion [52], in which the scores, or *a posteriori* probabilities, from different matchers are weighted and combined to obtain a final, single score. Let $\boldsymbol{y}_j$ be the feature vector of LBP features of frame j in I_{Φ^*}. Let Ψ be the set of features of Φ^*: $\Psi \equiv \{\boldsymbol{y}_j : j \in \Phi^*\}$. c_i is the class label from one of the classes: $c_1, \dots, c_{n_c}$. n_c is the number of classes. The estimated label for all the frames in Φ^* is $\tilde{c}$. Note that this assigns labels to all frames Φ, including those that were not selected for processing. Temporal smoothing is introduced by assigning all the frames in I_{Φ^*} the same label. $p(c_i|\boldsymbol{y}_j)$ is the *a posteriori* probability of a class c_i. The first step of fusion is estimation of $p(c_i|\boldsymbol{y}_j)$ for each frame in I_{Φ^*} with the method in [55].

The second step aggregates the *a posteriori* probabilities from the selected frames into a single score. The classification rule for match-score fusion is:

$$\tilde{c} = \arg\max_{c_i} h\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) \qquad 10$$

where $h(.)$ is the rule for aggregation, and n_f is the number of frames in Φ^*. The *Sum rule* is as follows:

$$h_{Sum}\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) = \frac{1}{n_f} \sum_{j \in \Phi^*} p(c_i|\boldsymbol{y}_j) \qquad 11$$

The *Product rule* is as follows:

$$h_{Product}\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) = \prod_{j \in \Phi^*} p(c_i | \boldsymbol{y}_j) \tag{12}$$

The *Min and Max rules* are as follows:

$$h_{Min}\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) = \min_{j \in \Phi^*} p(c_i | \boldsymbol{y}_j) \tag{13}$$

$$h_{Min}\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) = \max_{j \in \Phi^*} p(c_i | \boldsymbol{y}_j) \tag{14}$$

The *Mode rule* h_{Mode}, differs from the above rules by assigning the most common label to each frame in the segment.

The approach can be applied to regression by taking the result of the aggregation rule to be the final decision value. This replaces Equation 10, where a second classifier is applied:

$$\tilde{c}_{Regression} = h\left(c_i, \Phi^*, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_f}\right) \tag{15}$$

Note that, for regression, we do not estimate the *a posteriori* probability. $p(.)$ in the above equations is replaced with the decision values from SVR [51].

C. Experiments

1. Experimental Parameters

After ROI extraction, all face images are resized to 200×200 with bicubic interpolation. For avatar image registration, we train the avatar reference image from the development data subsampled at 12 fps for detection. The parameters specific to avatar image registration are: $\alpha = 2, \frac{1}{\sigma^2} = .005$, and the number of iterations is 3. All three of these parameters are empirically selected from the previous work [17]. The parameters specific to LBP [16] are: the number of local regions is 8, patterns are computed for 8 neighbors at a radius of 1, and there are 10×10 sub-regions on the entire face image. All classifiers are SVM [51]. The parameters specific to the SVM are: an RBF kernel is used, the cost $c = 1$, and $\gamma = 2^{-8}$. The feature vectors are normalized to $[-1,1]$. For regression, an ϵ-SVR is used [51]. The parameters specific to the regression algorithm are: $\epsilon = 0.1$.

Selecting the initial number of frames N: There should be enough frames in Φ to describe the expression in progress. In the unconstrained case, an expression can be very quick. If that expression was a microexpression, it could be as fast as 1/25th of a second, requiring 25 fps [56]. MMI-DB videos were captured at 24 fps, so we recommend $N > 24$ for MMI-DB. We chose $N = 50$ frames. It is validated empirically.

AVEC 2012 is also used for selecting parameter N. A value is selected empirically by varying N in powers of 2 seconds: $\{2^{-3}, \ldots, 2^{8}\}$. The results are given in Figure 7. N=50 gives the best performance. The performance decreases as N is reduced below 50 frames. For decreasing values of N, the upper bound of β decreases, and more frames are to be selected. The worst performer is 6 frames per segment.

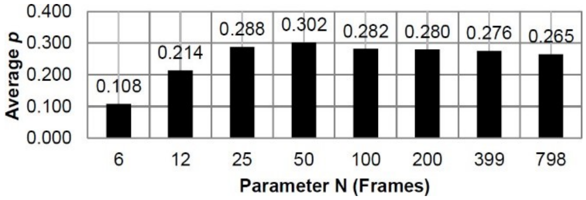


Figure 7: Average correlation of all affect dimensions on development set, AVEC 2012 frame-level subchallenge for varying values of N.

2. Experimental Results

Training results that select the best performing combination of registration method and fusion rule are given in Section Chapter 4 C.2.a). Results comparing temporal feature methods on MMI-DB are given in Section Chapter 4 C.2.b). Testing results on AVEC 2011 and AVEC 2012 are given in Section Chapter 4 C.2.c). Testing results on CK are given in Section Chapter 4 C.2.d). A discussion on memory cost and visual examples of the proposed downsampling method are given in Section Chapter 4 C.2.e).

Table 6: Weighted Accuracy Results for Various Sampling Methods, Registration Methods and Fusion Methods for AVEC 2011 Development set.

			WA Result				
Sampling	Reg	Rule	Arousal	Expectancy	Power	Valence	Average
Proposed	AIR	Sum	71.7	62.1	63.4	65.3	65.6
Proposed	AIR	Max	71	60.7	63.2	64.8	64.9
Proposed	RST	Min	70.1	61	62.1	65	64.5
Proposed	RST	Mode	71	61.9	61.8	62.6	64.3
Proposed	RST	Sum	70.7	60.2	63	63	64.2
Proposed	RST	Prod	69.6	61.9	61.2	62.8	63.9
Proposed	RST	Max	69	60.1	61.6	64.6	63.8
Proposed	AIR	HMM	68.5	62	59.8	64.9	63.8
Proposed	AIR	Prod	70.2	59.8	60.5	64.3	63.7
Proposed	AIR	Mode	71.6	59.5	60.9	62.6	63.6
Proposed	RST	No	69	59.6	62.1	63.6	63.6
Proposed	AIR	Min	70.1	59.2	60.8	62.6	63.2
Proposed	AIR	No	69.1	55.5	62.5	64.7	62.9
Uniform 3	AIR	Sum	69.3	57.7	61	63.7	62.9
Uniform 6	AIR	Sum	67.7	60	57.9	62.9	62.1
Uniform 9	AIR	Sum	67.6	57.2	60.2	61.4	61.6
Uniform 6	AIR	Mode	67.9	56.7	58.7	62.3	61.4
Uniform 3	AIR	Mode	65.9	61.6	59	58.5	61.2
Uniform 9	AIR	Mode	68.3	55.6	58.8	58.6	60.3
Uniform 1	AIR	No	65	56.3	57	62.4	60.2

Sampling: sampling rate. Uniform: uniform number of frames. Reg: registration method. AIR: avatar image registration. RST: similarity transform. Rule: fusion rule. HMM: hidden Markov model. WA: weighted accuracy.

a) Selection of registration method and fusion rule

The selection of the best performing combination of registration method and fusion rule is made with the development set on AVEC 2011. This experiment also tests the performance gain when using the proposed method versus a uniform sampling rate.

The results for different registration techniques, sampling methods, and rules are given in Table 6. The methods are ranked in descending order of average performance across all four classes. Under sampling method, Uniform indicates that a uniform number of frames were selected for each segment, Proposed indicates that the proposed method was used. RST indicates that a similarity transform was used with eye points as control points. Sum refers to the sum rule; Product, product rule; Min, min rule; Max, max rule; Mode, the mode rule; and no fusion, the labels are assigned without any fusion. HMM indicates hidden Markov model fusion detailed in [54].

The best performer (Proposed + AIR + Sum) improves classification rate by 5.4% versus Uniform 1 + AIR + No fusion. This is the combination that is used in the following experiments, except for AVEC 2011 testing results, which are the original, official entry results of the challenge that used the Max rule. The combinations can be grouped into three categories: (1) dynamic downsampling with avatar image registration, (2) dynamic downsampling with similarity transform based registration, and (3) uniform downsampling with avatar image registration. It is clear that methods with the proposed dynamic sampling rate (groups 1 and 2) are better than methods that sample uniformly (group 3). While the two best performers use AIR registration, the difference between avatar image registration (group 1) and similarity transform registration (group 2) is not as clear. Replacing avatar image registration with similarity registration does not cause a significant drop in performance. Proposed + AIR + Sum and Proposed + RST + Sum have a difference of 1.4\% on the average. For AVEC 2011, we conclude that intelligent selection of frames is a greater contributor to classification rate than a better registration algorithm.

Table 7: Confusion Matrices for MMI-DB.

(a)
Yang and Bhanu [17]

	An	Di	Fe	Ha	Sa	Su
An	71.7	2.2	2.2	6.5	4.4	13
Di	12.9	48.4	16.1	6.5	0	16.1
Fe	27.6	0	58.6	3.5	0	10.3
Ha	9.5	0	4.8	76.2	0	9.5
Sa	25	0	6.3	6.3	59.4	3.1
Su	18.4	2.6	7.9	0	5.6	65.8

(b)
Uniform Sampling of 1 Frame

	An	Di	Fe	Ha	Sa	Su
An	76.4	4.7	6.8	2.2	2.2	8.7
Di	9.7	64.5	9.7	3.2	3.2	9.7
Fe	24.1	0	55.2	0	6.9	13.8
Ha	11.9	0	2.4	76.2	2.4	7.1
Sa	28.1	0	6.3	3.1	53.1	9.4
Su	21.1	7.9	5.3	0	0	65.8

(c)
Proposed with Frame Differencing as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	78.3	6.5	0	4.4	4.4	6.5
Di	9.7	67.7	12.9	0	0	9.7
Fe	27.6	0	58.6	3.5	0	10.3
Ha	14.3	7.1	9.5	61.9	0	7.1
Sa	21.9	0	6.3	0	62.5	9.4
Su	15.8	2.6	2.6	0	2.6	76.3

(d)
Proposed with Dense-SIFT as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	76.1	6.5	0	0	4.4	13
Di	9.7	58.1	16.1	3.2	0	12.9
Fe	17.2	0	69	3.5	0	10.3
Ha	14.3	4.8	2.4	69.1	0	9.5
Sa	21.9	3.1	0	3.1	59.4	12.5
Su	18.4	0	2.6	0	0	79

(e)
Proposed with Optical Flow as Temporal Feature

	An	Di	Fe	Ha	Sa	Su
An	73.9	4.4	4.4	0	8.7	8.7
Di	6.5	74.2	6.5	0	0	12.9
Fe	17.2	3.5	69	0	0	10.3
Ha	9.5	4.8	2.4	76.2	0	7.1
Sa	21.9	0	0	3.1	71.9	3.1
Su	21.1	2.6	5.3	2.6	2.6	65.8

An: anger. Di: disgust. Fe: fear. Ha: happiness. Sa: sadness. Su: surprise.

b) *Evaluation of temporal feature*

We evaluate the use of optical flow as a temporal feature versus SIFT flow and frame differencing with MMI-DB empirically in Table 7. Weighted and unweighted accuracies are given in Table 8. When using a different temporal feature, ΔI_n is replaced by the new method (frame differencing or dense SIFT), the L_2-norm of the difference between frames n and $n-1$ is still used. For uniform sampling of 1 frame, the frame at the apex is the only frame used. In [17], Yang and Bhanu temporal smoothed the result by taking the mode predicted label and assigning it to the video. It is similar to the proposed approach, but it does not incorporate downsampling and uses Local Phase Quantization (LPQ) features. It has the worst performer because it uses all the frames, including the frames furthest away from the apex. Frame differencing is the fastest method for computing the temporal feature, but it has the worst performance among other temporal features.

Table 8: Weighted accuracy and unweighted accuracy on MMI-DB for varying temporal features.

Method	WA	UA
Yang and Bhanu [17]	63.4	64.8
Uniform Sampling of 1 Frame	65.2	66.6
Prop. + Frame Differencing Temporal Feature	67.6	68.4
Prop. + Dense-SIFT Temporal Features	68.4	69.4
Prop. + Optical Flow Temporal Feature	71.8	72

Prop.: Proposed. UA: unweighted accuracy. WA: weighted accuracy.

SIFT flow improves performance, but it is the slowest temporal feature optical flow has a better performance and speed. Retaining only 1 frame is worse than the proposed downsampling method. We conclude that, for MMI-DB, there are instances where retaining more than 1 frame can improve classification rate, if those frames are intelligently selected.

c) Results without apex labels

Results on the official AVEC 2011 testing and development sets are given in Table 9. The proposed method is compared to the two other entries that employed a dynamic sampling rate and it is always the best or second best performer for the development set. On the testing set, it improves weighted accuracy by 9.8%, and unweighted accuracy by 7.0% over the baseline approach. In [21], the method pays more attention when the predicted label changes, which assumes that the prediction is accurate, which is not always the case, especially for a difficult dataset such as AVEC 2011. We believe that the proposed method does well because it is the only downsampling method based on changes of visual information of the face.

Table 9: Comparison to Other Methods on AVEC 2011 Frame-level Subchallenge Testing Set.

(a) Development Set

	Arousal		Expectancy		Power		Valence		Average	
Method	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Proposed Method	71.7	67.8	62.1	59.8	63.4	61.8	65.3	60.7	65.6	62.6
Glodek et al. [20]	58.2	53.5	53.6	53.2	53.7	53.8	53.2	49.8	54.7	52.6
Dahmane and Meunier [21]	54.9	55	51.8	51.2	53.2	52.8	56.6	55.5	46.6	53.6
Baseline [14]	60.2	57.9	58.3	56.7	56	52.8	63.6	60.9	59.5	57.1

(b) Testing Set

	Arousal		Expectancy		Power		Valence		Average	
Method	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Proposed Method	56.5	56.9	59.7	55.1	48.5	49.4	59.2	56.7	56	54.5
Glodek et al. [20]	56.9	57.2	47.5	47.8	47.3	47.2	55.6	55.6	51.8	52
Dahmane and Meunier [21]	63.4	63.7	35.9	36.6	41.4	41.1	53.4	53.6	48.5	48.8
Baseline [14]	42.2	52.5	53.6	49.3	36.4	37	52.5	51.2	46.2	47.5

Table 10: Comparison to Other Methods on AVEC 2012 Video-based Frame-level Subchallenge Testing and Development Sets.

Video-only Development Set

Method	Arousal	Expectancy	Power	Valence	Average
Baseline [14]	0.151	0.122	0.031	0.207	0.128
Proposed Method	0.379	0.199	0.244	0.385	0.302
Nicolle et al. [22]*	0.354	0.538	0.365	0.432	0.422
Ozkan et al. [25]	0.117	0.076	0.062	0.2	0.114
Savran et al. [26]	0.306	0.215	0.242	0.37	0.283
Yang and Bhanu [17]		0.173	0.099	0.164	0.198

Video-only Testing Set

Method	Arousal	Expectancy	Power	Valence	Average
Baseline [14]	0.077	0.128	0.03	0.134	0.093
Proposed Method	0.302	0.244	0.199	0.279	0.252
Nicolle et al. [22]**	-	-	-	-	-
Ozkan et al. [25]**	-	-	-	-	-
Savran et al. [26]	0.251	0.153	0.099	0.21	0.178
Yang and Bhanu [17]	0.19	0.105	0.142	0.177	0.154

*Best performing video feature.

**Video-only testing set not reported.

Results on AVEC 2012 frame-level subchallenge are given in Table 10. For the development set, Nicolle et al. [22] has the best performance, but they did not provide video-only testing results. They noted that the ground-truth labelers had a time delay when recording the label, and they incorporated meta-data of who the user was speaking with, e.g. if the embodied agent speaking to them was belligerent. Though this improved performance, it is ad hoc in the sense that rater time delay may be specific to AVEC 2012, and meta-data about who the person is speaking to may not be available with other datasets.

Table 11: Apex label results compared to other methods for 14 AUSs on CK

	Facial Action Unit														
Method	1	2	4	5	6	7	9	10	12	15	20	24	25	27	Avg
Proposed	85.3	93	87.7	69.6	90.5	62.4	68.5	43.5	76.9	71	74	65.2	93.6	84.2	76.1
Koelstra et al. [41]	86.8	90	73.1	80	80	46.8	77.3	48.3	83.7	70.3	79.4	63.2	95.6	87.5	75.9
Valstar et al. [4]	87.6	94	87.4	78.3	88	76.9	76.4	50	92.1	30	60	12.3	95.3	89.3	72.7
Yang et al. [17]	82	92.1	82	58.6	84.9	52.5	68.4	34.8	68.2	66.7	65.7	51.1	85.6	67.2	68.6

Avg: Average of all AUs.

d) *Results with apex labels*

The efficacy of the proposed method with apex labels on CK is given in Table 11. A comparison is made with other methods according to F_1 measure. Yang and Bhanu [17] method does not take advantage of apex frame labeling. The proposed method takes advantage of apex labeling and it performs better. We performed best for 4 AUs. Valstar and Pantic [4] perform best for 6 AUs. However, the proposed method has a higher average F_1 measure among all the other works. The comparison to [17] demonstrates the importance of incorporating temporal information. Intuitively, assuming that each frame is equally discriminative, selecting as many frames as possible, such as in Yang and Bhanu [17], should increase the true positive rate by introducing more samples for the fusion. However, samples that are further away from the apex contain less relevant information of the expression being captured. Frames further away from the apex are close to neutral. They are not good examples of the expression being expressed, and they reduce accuracy. The proposed method sampled frames at the apex, and Koelstra et al. [41] modeled the temporal phases including the apex. This may explain the gap in performance, because [17] does not model the apex.

e) Memory cost savings and temporal feature results

In the following, we discuss the memory cost saving for each dataset, and show examples of the temporal feature. For AVEC 2011, the total number of frames for the development, training and testing (video sub-challenge) partitions are $\{449074, 501277, 140125\}$, respectively. The proposed method down sampled the number of frames by a factor of 16.6, retaining $\{27412, 30076, 8383\}$ frames. For CK, the proposed method sampled 3.4 ± 2.2 frames. For MMI-DB the proposed method sampled 3.4 ± 1.5. A comparison of the number of frames reduced by the proposed method is given in Table 12.

Table 12: Summary of Frames Used for Each Dataset

	AVEC 2011	AVEC 2012	CK	MMIDB
# of Videos	74	95	488	222
# of Frames	1090476	1351129	8795	23466
Proposed	65871	76960	1536	764
Dahmane and Meunier [21]	196051	239920	-	-
Savran et al. [26]	-	232600	-	-
Glodek et al. [20]	740	950	4930	2220

Because the method in [26] retains outliers based on the regression label, it can only be applied to continuous label intensities, such as in AVEC 2012. The method would process each testing frame uniformly. In [21], for continuous data, we categorized the labels into 10 bins. This method is not applicable to apex labeled data, where the videos are segmented and have a single class label. In [20], frames are sampled uniformly. The method's memory cost is proportional to the number of videos, so the method does not reduce memory cost well for datasets with many videos, such as CK and MMI-DB. Though the method has the least number of frames for AVEC 2011 and AVEC 2012, it may sample the long videos too sparsely to precisely detect when emotion changes. The proposed method can be used to reduce the number of frames on all four datasets, both on continuous and discrete data, and on segmented and unsegmented data. It is the best or second best method for reducing memory cost on all four datasets.

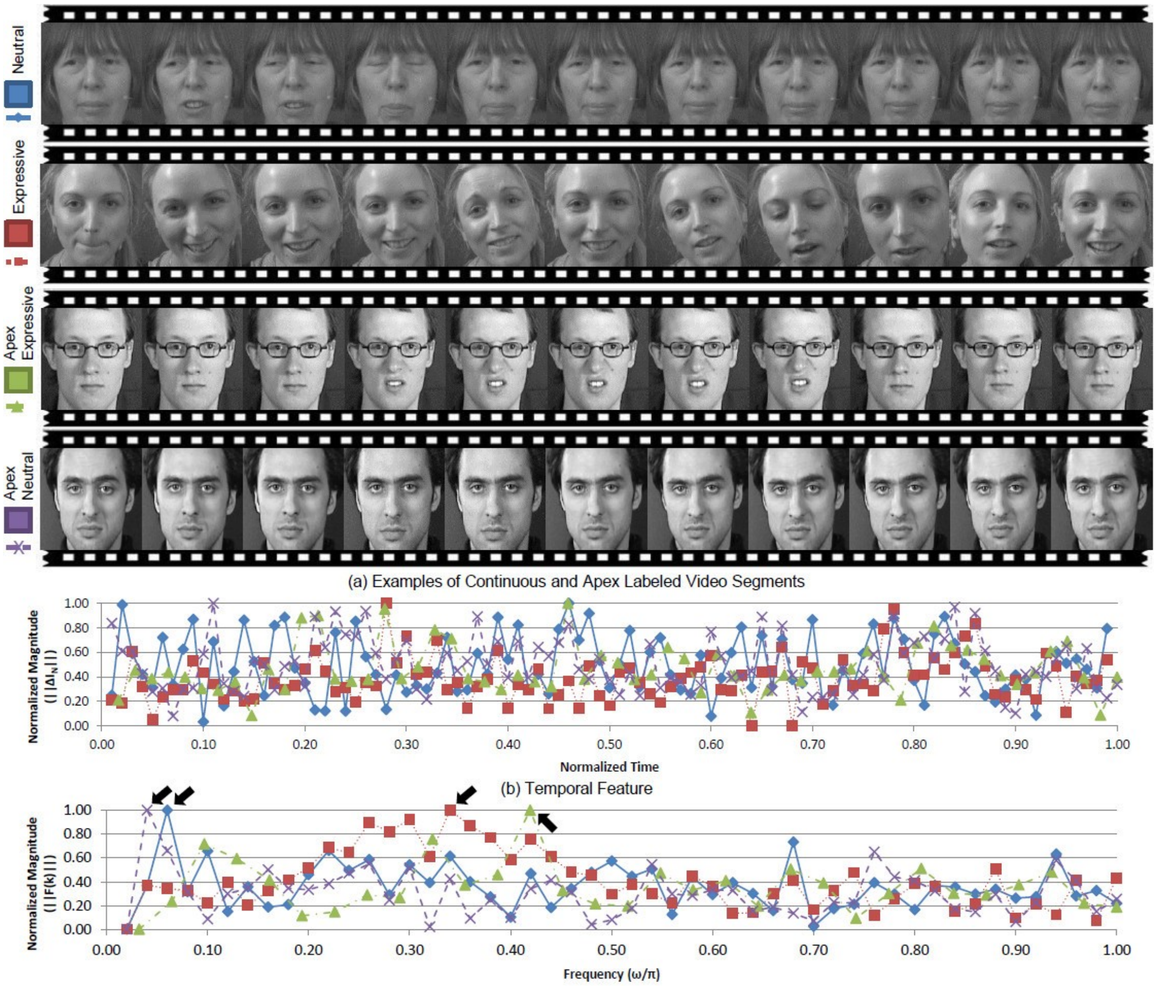
A detailed example of two continuous video segments from AVEC 2012, and two apex labeled segments from MMI-DB is given in Figure 8. The magnitude has been normalized to provide a better understanding of the results. The time range has been normalized because MMI-DB segments are of different lengths. For the discrete Fourier transform, the frequency is normalized to $[0,1]$. The first example in Figure Figure 8 (a) is of a person who does not use many expressions (Neutral). In this case the dominant frequency is at .06 cycles/frame, so only a few frames would be selected. The second row is of a person who is using many expressions and changing her pose (Expressive). Intuitively, many frames will be required to describe this segment, which is corroborated by the dominant frequency being at .34 cycles/frame. The third row is of a person who holds his expression for a long time at the apex (Apex Expressive). The dominant frequency is at .42 cycles/frame. In this example, there are 62 frames in the cycle, thus

$.42 \times 62 \approx 26$ frames would be selected. It can be observed from the example frames that his expression is held at the apex for roughly half of the frames, corroborating keeping 26 of the 62 frames. The fourth row is of a person who weakly expresses his emotion (Apex Neutral). In this case, the dominant frequency is .04 cycles/frame, so very few frames would be selected.

D. Conclusions

In this chapter, vision and attention theory was employed to temporally down sample the number of frames for video-based emotion and expression recognition. It was found that a uniform frame rate decreases performance and can unnecessarily increase memory cost for high frame rates. With the proposed method, AVEC 2011 is down sampled by a factor of 16.6 and weighted accuracy is improved over the baseline approach by 9.6% on the testing set. AVEC 2012 is down sampled by a factor of 17.6 and correlation is improved over the baseline by .159 on the testing set. CK is down sampled by a factor of 5.72 and the F_1 measure for AU detection is improved by 0.3. MMI-DB dataset is down sampled by a factor of 30.1 respectively and weighted accuracy for emotion recognition is increased over [17] by 8.4\% for all sessions. Unlike previous works, we reported results on all four datasets.

The conventional process of using a short duration of frames centered at the apex was corroborated with the proposed sampling method and extended to allow for an increase in duration when appropriate. It was found that top methods from previous challenges [17] did not generalize to continuous data sets. In that challenge, registration was found to be a significant contributor to performance, whereas, in the AVEC datasets, we have found that registration does not significantly contribute to performance. Previous datasets were segmented to the time points of most significance, and we posit that, for continuous datasets, a method must be critical in its selection of frames. A limitation of the current work is that the frames are processed in evenly sized segments, which may cause a boundary effect if an unlabeled apex is close to the segmentation boundary. However, this can be addressed by using overlapped boundary segments.



(a) Examples of Continuous and Apex Labeled Video Segments

(b) Temporal Feature

(c) Fourier Transform of Temporal Feature

Figure 8: (a) From top to bottom, a continuous video segment of a neutrally expressive person; a continuous video segment of an expressive person; an apex segment of a person who is expressive; an apex segment of a person who is less expressive. (b) The temporal feature of each of the examples, and (c) the discrete Fourier transform of the temporal feature. Note that both the continuous neutral and apex labeled less expressive examples have a low dominant frequency, whereas the other two expressive examples have a higher dominant frequency. Black arrow indicates dominant frequency.

CHAPTER 5 BIOLOGICALLY INSPIRED FEATURES

State-of-the-art methods use features with two properties: the first property is the generalization to factors encountered in unconstrained facial emotion recognition, and the second property is a compact feature representation. For example, Local Phase Quantization (LPQ) features are robust to blur, and use histograms to reduce feature vector size [57]. Local Binary Pattern (LBP) features can be rotation invariant, robust to monotonic grayscale transformation from shadows, and also use histograms to reduce feature vector size [16]. The original formulation [38] of the Gabor energy filter does not have either of these properties. We propose background suppressing Gabor energy filtering. The proposed method removes background texture with a generalization step, and reduces feature vector size with a computational efficiency step. We improve performance over other frontal face feature representations used for the Audio/Visual Emotion Challenge (AVEC) 2012 grand-challenge dataset [15]. We compare the performance of the generalization step and computational efficiency step on the Cohn-Kanade+ (CK+) dataset [36].

A. Motivation

In recent emotion recognition grand challenges [15] [13], Gabor filters were not the most commonly used local appearance feature. Out of the top six approaches for AVEC 2011, only one approach used a Gabor energy filter [20]. Approaches preferred LPQ, LBP or active appearance models. The Gabor filter is historically important and it is utilized in various computer vision applications. We assert that it can still be effectively applied to facial emotion if the following technical challenges are addressed:

(1) *Generalization:* Gabor energy filters do not generalize well in unconstrained settings, because a Gabor energy filter captures edge magnitudes at almost all orientations, including edges from noise due to background texture. Current local appearance features have additional steps in an effort to be more generalizable and robust. Wu et al. [58] addressed this by extending the Gabor filter to temporal domain with Gabor motion energy (GME) features. However, the feature vector size was increased by the number of temporal scales over the original Gabor energy filter, which already has a large feature vector size. For example, the feature vector was increased by a factor of 3.72 between Lyon and Akamatsu [38] and Wu et al. [58]. We address this technical challenge with background suppressing Gabor energy filtering, which removes the edges due to background noise but retains the significant edges that correspond to the edges of the objects in a scene. We also compute texture at a pixel, microtexture level, so the method is invariant to monotonic grayscale transformations.

(2) *Computational efficiency*: Gabor energy filters produce a response for each filter in its bank. The feature dimensionality of a Gabor feature vector is a product of the

size of the image by the number of scales and the number of orientations. For example, a Gabor energy filter bank at 6 orientations, 3 scales, and a square image of 150×150 results in a dimensionality of 405000. The dimensionality of LBP is 5900 in Yang and Bhanu [17] regardless of the image size. Dahmane and Meunier [21] addressed this with sub-region histograms, similar to LBP. However, their approach lacked a generalization step. We propose to combine maximal edge response and soft orientation histograms to create a compact representation for emotion recognition.

We contribute a novel method that improves the Gabor energy filter. It generalizes well because of its ability to suppress background texture. It has a low feature vector dimensionality because of soft orientation histograms. We demonstrate its efficacy on the non-trivial AVEC 2012 grand-challenge dataset [15]. We thoroughly examine the impact on performance of each part of the algorithm on the CK+ dataset. Additionally, we apply the method to bio-imaging data and examine the quality of edges extracted.

B. Technical Approach

The proposed system overview for extracting local appearance is described in Figure 9: In the generalization step, (1) the input image is filtered by a bank of Gabor filters, all fixed in scale at the pixel-level and varying for N different orientations. (2) Background texture of the input image is estimated on a per-pixel basis and removed from the result of each filtered image. In the computational efficiency step, (3) the bank of responses is condensed into a *maximal response*, a representation that retains the most intense edges and their orientations across all of the filters in the bank. (4) The image is divided into $M \times M$ subregions to account for face morphology, and *soft orientation histograms*, where bin counts are weighted by the magnitudes of their edges, are computed for each region. The histograms from each sub-region are concatenated to form the feature vector for the input image.

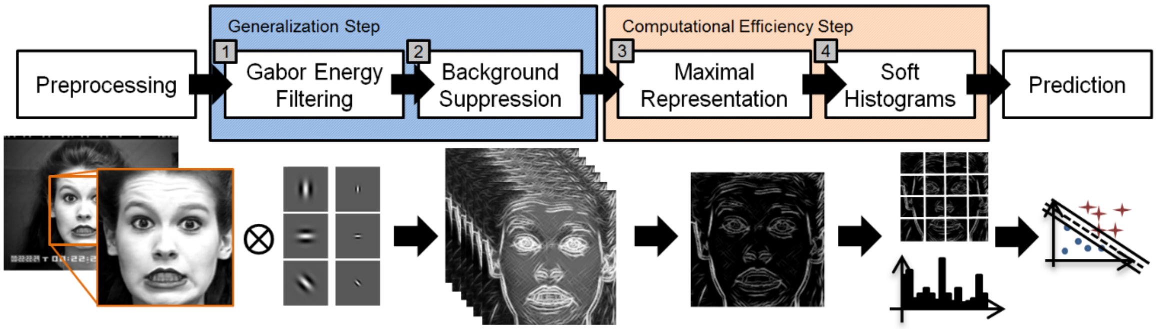


Figure 9: System overview of the proposed texture descriptor. ⊗ denotes convolution operation.

1. Gabor Energy Filter

A Gabor filter is a band-pass filter that can detect edges of a specific orientation and scale. Conventionally, an image is filtered by many Gabor filters with different parameters, and the collection of filters is called a *bank.* Each filter in the bank is tuned to a different orientation and scale. Under specific conditions, the Gabor filter can approximate the behavior of the human visual system [59]. The first component of the human visual cortex that processes visual information is the V1 area, located in the occipital lobe [59]. Parts of the V1 area form *cells*, and each cell responds to edges of a specific magnitude and orientation, called a *grating*. This is referred to as the Classical Receptive Field [60]. The Gabor energy filter emulates this process by creating a bank of filters where each filter responds to a specific grating. Let f be an input image. A Gabor energy filter for a specific magnitude and orientation is:

$$g(x,y;\gamma,\theta,\lambda,\sigma,\phi) = e^{\frac{\tilde{x}^2+\gamma^2\tilde{y}^2}{2\sigma^2}} \cos\left(2\pi\frac{\tilde{x}}{\lambda}+\phi\right) \qquad 16$$

where x and y are the pixel location. γ is the spatial aspect ratio that is a constant, taken to be 0.5. It effects the eccentricity of the filter. θ is the angle parameter that tunes the filter to specific orientations. λ is the wavelength parameter that tunes the filter to specific spatial frequencies, or magnitudes. In pattern recognition, this is also referred to as scale. σ^2 is the variance. It determines the size of the filter. ϕ is the phase offset taken to be 0 and π. $\tilde{x}$ and $\tilde{y}$ are defined as:

$$\tilde{x} = x\cos\theta + y\sin\theta \qquad 17$$

$$\tilde{y} = -x\sin\theta + y\cos\theta \qquad 18$$

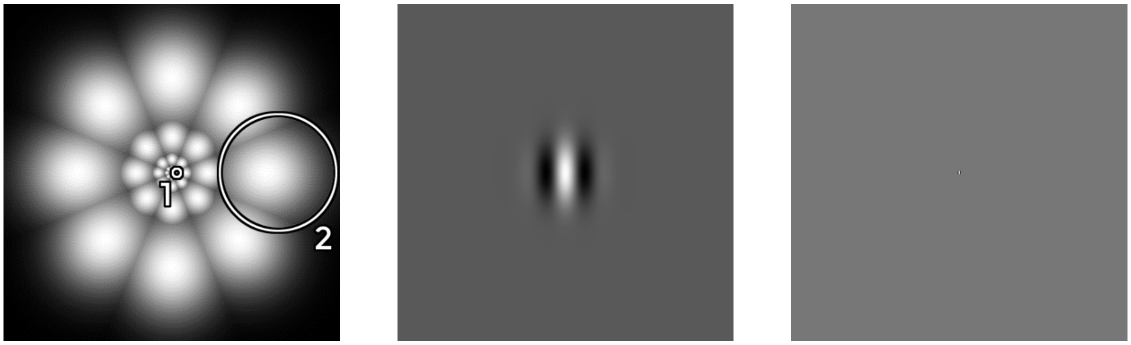
Conventionally, the scales and orientations in the bank are selected such that the half-magnitude of each filter overlaps with others [59]. An example of a Gabor filter bank is given in Figure 10. The Gabor filter can be used as local appearance filter by tuning the filter to a local neighborhood while still varying the orientation: $\sigma/\lambda = 0.56$, and varying θ. By considering pixel intensities in a local neighborhood in the same way that LBP computes a microtexture, the proposed method is invariant to monotonic grayscale transformations. For the rest of the paper, $g(x,y;\theta,\phi)$ is shorthand for the following: $g(x,y;0.5,\theta,7.14,3,\phi)$. $f(x,y)$ is filtered by $g(x,y;\theta,0)$, and by $g(x,y;\theta,\pi)$ and the magnitude of both is taken to be the result. This is called the Gabor energy filter:

$$E(x,y;\theta) = \sqrt{\big((f*g)(x,y;\theta,0)\big)^2 + \big((f*g)(x,y;\theta,\pi)\big)^2} \qquad 19$$

where $(f*g)(x,y;\theta,\phi)$ is the convolution of $f(x,y)$ and $g(x,y;\theta,\phi)$.

2. Generalization Step

Equation (19) captures the edge information. It responds to edges in the same way a simple cell in the human visual system responds to a grating. However, the human visual system is able to detect edges in the presence of background texture. This is called the *pop-out effect* [60], and an example is given in Figure 11. In Figure 11(a), a series of vertical edges are presented, but the edge which is perpendicular to the other edges appears to pop out of the image. In Figure 11(b), a triangle is presented over a background texture that is parallel to one of the sides of the triangle. The pop-out effect makes it difficult to detect the one side of the triangle. It appears as a capital ‘L’. The complex cells in the human visual cortex estimate background texture to focus on edges that are not consistent with the background texture. If the Gabor energy filter from equation (19) were applied to the images in Figure 11, it would detect a high energy in the direction of the background texture, and a low energy for the orientations associated with the perpendicular line, or the ’L’. The background texture is referred to as the *Non-Classical Receptive Field*. In some conditions, with the stimulus presented in Figure 11, the human visual system suppresses the Non-Classical Receptive Field to better represent the edge information. The proposed feature should emulate this effect.



(a) (b) (c)

Figure 10: (a) Visualization of a bank of Gabor filters in the frequency domain for 4 scales and 8 orientations, with parameters selected so that the half magnitudes overlap [59]. The visualization was created by taking the maximum value of the energy of each filter. (b) The grating corresponding to the spatial representation of the Gabor filter marked ‘1’ in (a). (c) The grating of ‘2’ in (a). Note that the grating in (c) is barely visible. This is because (c) is tuned to detect high frequency edges, which correspond to a pixel neighborhood in the spatial domain.

The Non-Classical Receptive Field t is estimated as a weighted Gabor filter:

$$t(x,y;\theta) = (E * w)(x,y) \qquad 20$$

where the weight function w is:

$$w(x,y) = \frac{1}{\left\| g\big(\mathrm{DoG}(x,y)\big) \right\|_1} g\big(\mathrm{DoG}(x,y)\big) \qquad 21$$

where $g(z) = H(z) * z$, where $H(z)$ is the Heaviside step function. $\mathrm{DoG}(x,y)$ is a Difference of Gaussians:

$$\mathrm{DoG}(x,y;K,\sigma) = \frac{1}{2\pi K^2\sigma^2} e^{-\frac{x^2+y^2}{2K^2\sigma^2}} - \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad 22$$

where K is a weight. σ^2 is the variance, the same as in equation 16. This ensures that the filter is bounded within the original Gabor filter. For visual examples the reader is referred to [62]. $w(x,y)$ resembles the ridges of a Mexican hat filter. When applied as the weight, equation (20) captures the edge information surrounding the current pixel. This allows background texture to be estimated on a per-pixel basis. The background suppressing Gabor energy filtered result $\tilde{b}$ is:

$$\tilde{b}(x,y;\theta) = g\big(E(x,y;\theta) - \alpha t(x,y)\big) \qquad 23$$

where α is a parameter that effects how much of the background texture is removed. When $\alpha = 0$, there is no background texture suppression, and the result is a Gabor energy filter.

3. Computational Efficiency Step

Equation (23) retrieves the significant gratings of f less background texture. It is computed for N different orientations. Conventionally, the responses from the N orientations would be concatenated and taken to be the feature vector. A method is needed to reduce the feature size. A representation of $\tilde{b}$ is created that retains edges with maximum magnitude, for each pixel:

$$b(x, y) = \max\{\tilde{b}(x, y; \theta) |\ \theta = \theta_1, \ldots, \theta_N\} \qquad 24$$

Equation (24) is called the maximal response. Separately, an orientation map $\Theta(x, y)$ is constructed that contains the orientation of the dominant edge in the maximal response, for each pixel:

$$\Theta(x, y) = \text{argmax}_\theta\{\tilde{b}(x, y; \theta) |\ \theta = \theta_1, \ldots, \theta_N\} \qquad 25$$

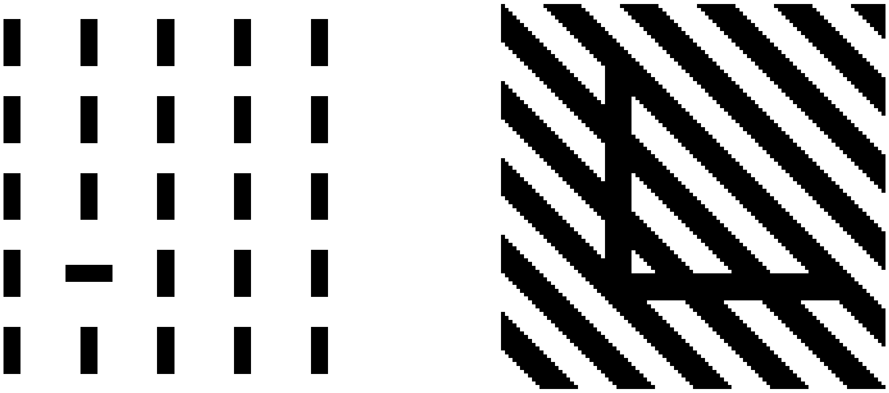

(a) (b)

Figure 11: Two examples of the pop-out effect. (a) In this image, the eye is drawn to the horizontal line because the repeated vertical lines form a background texture that is suppressed by the human visual system. (b) In this image, a triangle is presented along

with a diagonal pattern. The removal of background texture suppresses one side of a triangle to give the illusion of an ‘L’.

Equations (24) and (25) retain the information of the most dominant edge. b retains the value of the maximum edge intensity, across all orientations, and Θ stores the specific orientation of the maximal edge. The image f is divided into M^2, equally sized, non-overlapping sub-regions. LBP and LPQ features use a *hard histogram*. That is, a histogram is computed that counts the number of microtextures. We use a *soft orientation histogram* to represent each sub-region. Instead of equally counting the presense of each microtexture, the votes are weighted by their magnitude from the maximal representation:

$$h(\theta_i) = \sum\nolimits_{\forall(x,y)|\Theta(x,y)=\theta_i} b(x,y)$$

2

6

where $h(\theta)$ is an N bin histogram. A histogram is computed in each grid. The $M \times M$ grids are concatenated to form the feature vector for f.

4. *Emotion Recognition Pipeline*

Face regions-of-interest are detected with a cascade of Haar-like features [50]. The faces are registered with Avatar Image Registration, which is run for three iterations, based on tests in [17]. For facial emotion recognition, the following features are compared: (1) Gabor filter based features, (2) the contour map of a Gabor filter with Non-Classical Receptive Field inhibition [60] (NCRF), this is the proposed method without the computational efficiency step, (3) the proposed method, (4) Local Binary Patterns (LBP) [61], (5) Three-Patch (TPLBP) and Four-Patch Local Binary Patterns (FPLBP)

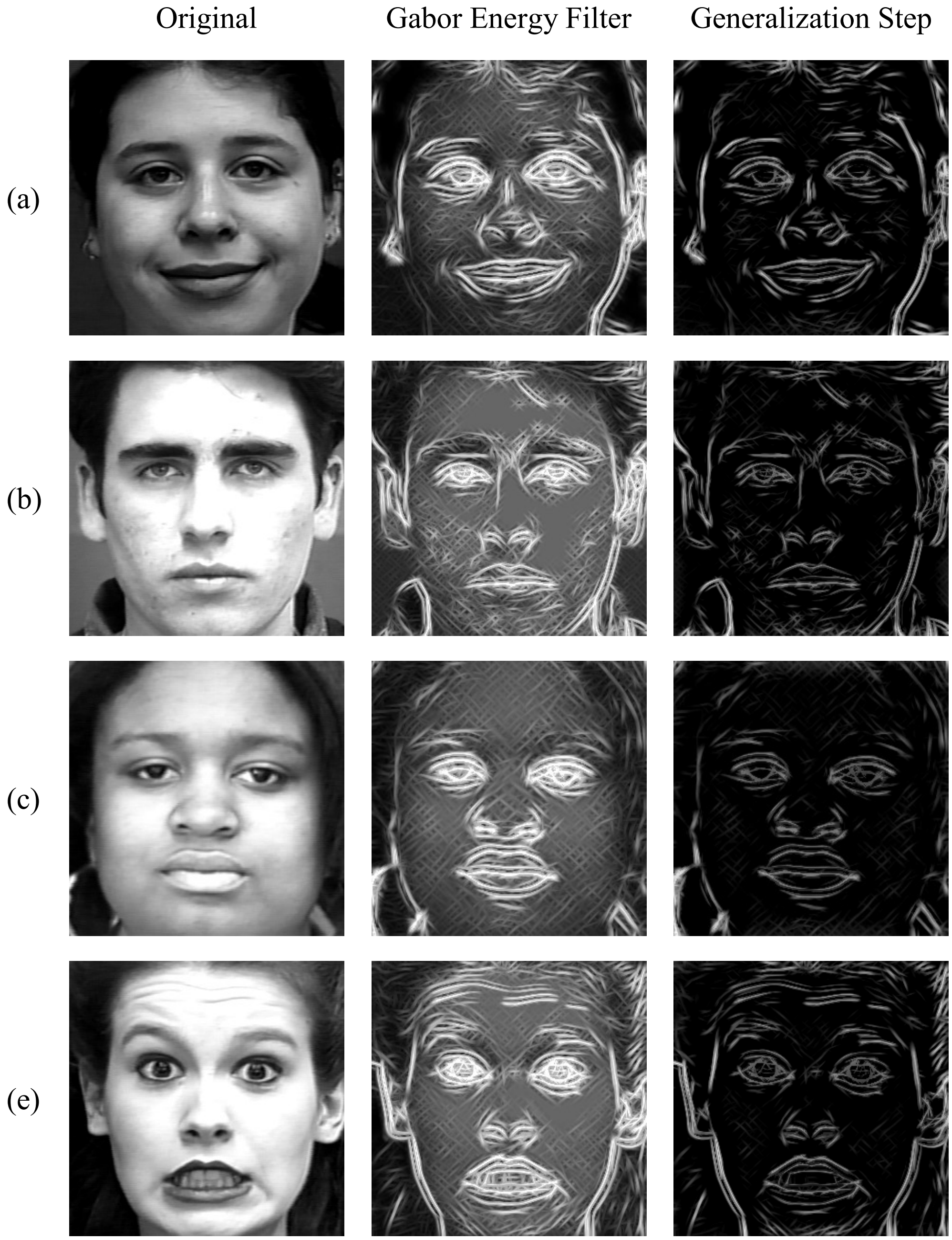


Figure 12: Maximal response of the generalization step applied to faces from the Cohn-Kanade+ dataset. (a-d) Note that the Gabor energy filter detects strong edges on the cheek, despite there being no visible edges in the image. These edges are detected from noise. Note that they are removed in the generalization step. (e) Teeth form a pattern that is detected as background noise and removed.

[30], (6) Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [31], (7) Discrete Cosine Transform (DCT) [5], (8) dense SIFT features [6], and (9) sub-region histograms of a Gabor energy filter [21]. For regression, an ϵ-SVR detects the emotion label intensity and, for classification, a linear SVM detects classes [51].

C. Experiments

1. Experimental Parameters

For the background suppressing Gabor energy filtering: $\sigma/\lambda = 0.56$, values of θ were selected such that $\theta_{N+1} = \pi$, $\sigma = 4$, $\alpha = 1$. For equation (22), $K = 4$ and is chosen from previous work [60]. N in the computational efficiency step is taken to be 64. All local histograms are calculated in neighborhoods of 8×8. For LBP, patterns are 8-neighbors with a radius of 1. For TPLBP and FPLBP, parameters are the same parameters in Wolf et al. [30]. Images are resized to 128×128 before processing. For the Gabor energy filter, there are 4 scales and 8 orientations, with all of the responses concatenated to make the feature vector. For LBP-TOP, the radii parameter for x and y is 1.

2. Experimental Results

a) Results on the Audio/Visual Emotion Challenge

Results on AVEC 2012 are given in Table 13. Results are given on the development set, and they are generated with a 3-fold cross validation. We use the same folds from previous work [62]. In Table 13, average indicates the average correlation among the four labels of arousal, expectancy, valance and power. Size indicates the feature vector size. There is a clear dichotomy in the performance. There are three categories of performance. The proposed method, FPLBP, and LBP, LBP-TOP and Gabor histograms are the best performers. DCT and LPQ features perform badly. TPLBP, SIFT, NCRF and the Gabor filter are the worst performers. The proposed method does better than the other methods in the categories of expectancy, valence and power. FPLBP performs better for arousal, but its variance is higher. Note that the pairing of LPQ and Avatar Image Registration was the best performer in the Facial Emotion Recognition and Analysis, discrete emotional states sub-challenge [17]. There is a relationship between the size of the feature vector and the performance.

LBP and FPLBP are comparable in performance to the proposed method. However, LBP and FPLBP rely on the existence of coded microtextures. An LBP image of 8 neighbors and a radius of 1 is challenging to understand with the human eye. The proposed method produces a visually understandable contour map to humans. An example of background texture suppression is given in Figure 12. Note that for all examples the dominant contours from the eyes and mouth are extracted, but the Gabor filter detects many false alarms. In Figure 12(e), the teeth are detected as a background texture and removed in the generalization step. This is desirable because we want to detect the facial expressions from facial muscle movements. The teeth pattern in Figure 12(e) would be detected as edges by the Gabor energy filter, falsely detecting edges that may indicate facial expressions in the center of the mouth.

b) Impact of Generalization and Computational Efficiency with CK+

In this section, we explore the impact on performance from the generalization step and computational efficiency step. The three methods are compared: (1) a background suppressing Gabor energy feature bank is used as a feature. The response of each filter is concatenated to form the feature vector. This represents the generalization step without the computational efficiency step. (2) The second method is the computational efficiency step applied to a Gabor energy filter, without the background suppression. This method represents the computational efficiency step without the generalization step. (3) The third method is the proposed method.

Table 13: Results on AVEC 2012 development set frame-level sub-challenge. For correlation, higher is better. Bold: best performer. Underline: second best performer. Factor: the downsampling factor applied to the frame rate to fit all of the feature vectors into memory; smaller is better and 1.0 indicates that all the feature vectors fit into memory without requiring downsampling. Size: the size of the feature vector, smaller is better.

	Pearson Product-moment Correlation Coefficient					Feature	
Feature	Arousal	Expectancy	Valence	Power	Average	Factor	Size
DCT [5]	0.034 ± 0.015	0.078 ± 0.024	0.076 ± 0.024	0.063 ± 0.035	0.063	1.1	8192
FPLBP [63]	0.425 ± 0.037	0.108 ± 0.050	0.291 ± 0.066	0.093 ± 0.033	0.229	1.0	200
Gabor [38]	0.059 ± 0.047	0.019 ± 0.008	0.063 ± 0.043	0.012 ± 0.009	0.036	70.3	5.2 x105
Gabor Hist. [21]	0.171 ± 0.058	0.080 ± 0.092	0.082 ± 0.073	0.067 ± 0.053	0.100	1.0	2048
LBP [16]	0.434 ± 0.039	0.072 ± 0.030	0.257 ± 0.064	0.088 ± 0.032	0.213	1.0	5900
LBP-TOP [31]	0.389 ± 0.016	0.092 ± 0.084	0.177 ± 0.013	0.084 ± 0.069	0.186	1.0	177
LPQ [57]	0.032 ± 0.029	0.085 ± 0.014	0.072 ± 0.047	0.076 ± 0.026	0.066	3.5	25600
SIFT [6]	0.037 ± 0.036	0.038 ± 0.024	0.073 ± 0.043	0.048 ± 0.028	0.049	3.5	25600
TPLBP [63]	0.024 ± 0.034	0.047 ± 0.025	0.086 ± 0.030	0.039 ± 0.028	0.049	283.7	2.1x106
Proposed	0.417 ± 0.035	0.143 ± 0.051	0.347 ± 0.062	0.124 ± 0.033	0.258	1.0	6400

The true positive rate, false positive rate, false negative rate, true negative rate, precision, recall and F_1-score are given in Table 14. The negative samples greatly outnumber the positive samples, so the true negative rate is very high for all the AUs, except for AU17, which has a positive rate of 0.61. For this reason, more attention should be paid to the true positive, false negative and F_1-score. A summary comparing the average F_1-score values is given in Table 15. The generalization step by itself without the computational efficiency step is the worst performer of the three in all metrics. This is due to the large feature dimensionality. Because there are 64 filters in the bank, the feature vector size is 1048576. The computational efficiency step by itself and the proposed method has a feature vector size of 6400. Also, because each pixel is taken to be a feature, there is an extreme sensitivity to alignment. Histograms in local regions allow for some tolerance of registration errors, which is why histograms were adopted for use in LBP and LPQ features. The pairing of generalization and computational efficiency is always the best performer.

Table 14: Breakdown of performance of the different parts of the proposed method for different Facial Action Units on CK+.

(a) Generalization Step Only

AU	True Positive	False Positive	False Negative	True Negative	Precision	Recall	F_1-score
1	0.69	0.31	0.04	0.96	0.69	0.94	0.79
2	0.67	0.33	0.03	0.97	0.67	0.96	0.79
4	0.51	0.49	0.02	0.98	0.51	0.96	0.67
5	0.57	0.43	0.06	0.94	0.57	0.91	0.70
6	0.76	0.24	0.05	0.95	0.76	0.94	0.84
7	0.80	0.20	0.08	0.92	0.80	0.91	0.85
12	0.23	0.77	0.03	0.97	0.23	0.90	0.37
17	0.90	0.10	0.21	0.79	0.90	0.82	0.86
25	0.76	0.24	0.69	0.31	0.76	0.53	0.62

(b) Computational Efficiency Step Only

AU	True Positive	False Positive	False Negative	True Negative	Precision	Recall	F_1-score
1	0.73	0.27	0.05	0.95	0.73	0.94	0.82
2	0.74	0.26	0.03	0.97	0.74	0.96	0.83
4	0.60	0.40	0.03	0.97	0.60	0.96	0.74
5	0.64	0.36	0.06	0.94	0.64	0.91	0.75
6	0.82	0.18	0.05	0.95	0.82	0.94	0.88
7	0.84	0.16	0.09	0.91	0.84	0.91	0.87
12	0.33	0.67	0.03	0.97	0.33	0.91	0.49
17	0.93	0.07	0.22	0.78	0.93	0.81	0.86
25	0.86	0.14	0.02	0.98	0.86	0.97	0.91

(c) Proposed Method

AU	True Positive	False Positive	False Negative	True Negative	Precision	Recall	F_1-score
1	0.77	0.23	0.06	0.94	0.77	0.93	0.84
2	0.79	0.21	0.04	0.96	0.79	0.95	0.86
4	0.67	0.33	0.03	0.97	0.67	0.95	0.78
5	0.69	0.31	0.07	0.93	0.69	0.91	0.78
6	0.76	0.24	0.05	0.95	0.76	0.94	0.84
7	0.80	0.20	0.08	0.92	0.80	0.91	0.85
12	0.41	0.59	0.04	0.96	0.41	0.91	0.56
17	0.95	0.05	0.24	0.76	0.95	0.80	0.87
25	0.92	0.08	0.03	0.97	0.92	0.97	0.94

Table 15: Summary of results from Table 3 in terms of average F_1-score across all AUs. Higher is better.

Method	Average F_1-score
Generalization Step Only	0.72
Computational Efficiency Step Only	0.80
Proposed method	0.81

D. Conclusions

In this paper, we proposed a novel procedure that extended the Gabor filter to be robust against background noise and reduced the feature vector size by a factor of 126.56, when comparing the proposed method to a Gabor energy filter [59]. The proposed texture descriptor was found to have competitive performance on the AVEC 2012 dataset. It was demonstrated that the generalization step and the computational efficiency step improved classification accuracy, and that even more performance is improved by combining the two parts of the proposed algorithm. It was also shown that the edges detected by the proposed method are more meaningful than a Gabor filter on bio-imaging data.

CHAPTER 6 SCORE-BASED FACIAL EMOTION RECOGNITION

A major problem in video-based facial emotion recognition is that two people can express their emotions in entirely different ways. When a recognition model is trained, there is no guarantee that the expressions encountered in the testing data will be properly described a model that is developed using the training data. This is corroborated by a concept called idiosyncratic gestures [64]. A person's expressions are so unique to that person that they can be used for identification purposes. In the related field of face verification, *learning with side information* [30] has found success in overcoming the generalization problems encountered with the Labeled Faces in the Wild (LFW) challenge dataset [65]. In a state-of-the-art method, features, such as geometric or appearance features, are extracted and used as the feature vector for classification. In learning with side information, a score is computed that compares a queried frontal face image to background data, a reference frontal face images that are not the training frontal face images. Since the score does not directly compute similarity to training data samples, it alleviates problems from unique expressions and a lack of sufficient training examples from a specific person.

We propose a method to overcome the generalization problem by comparing the frontal face image to background information and previous frames to capture temporal face dynamics. We compute three scores: neutral score, temporal score, and one-shot emotion score. The method is inspired by One Shot Similarities (OSS) [30] which are motivated by a similar issue in face verification, but the method is significantly different. Efficacy of the method is tested on the Audio/Visual Emotion Challenge (AVEC) 2012 [15], and inter- and intradatset experiments on Cohn-Kanade+ (CK) [36], the Man-Machine Interface Facial Expression Database (MMI) [37], and the Japanese Female Facial Expressions dataset (JAFFE) [38].

A. Motivation

Despite the numerous advances to the state-of-the-art, a video-based approach has yet to be seen that performs well on AVEC, or in interdatabase experiments. There are two challenges to state-of-the-art methods:

Uniqueness of expressions. A model is trained from the training data that best describes the emotions of the persons encountered in the training data. In testing, a person could be encountered that was never encountered in training data. A person's expressions and gestures are so unique to a person that they can be used for identification [64]. An example of different variations of a smile are given in Figure 13. This is encountered in the AVEC datasets [14, 15] and in interdataset experiments, where the persons in the training fold data are not the same persons in the testing data. In the presence of this technical challenge, state-of-the-art methods that perform well in training do not generalize to these experiments. In some cases, uniqueness of expressions resulted in classification rates below the class percentage rate for the AVEC datasets. This is similar to the face recognition challenge Labeled Faces in the Wild (LFW) [65], where most individuals have only a single sample in the data.

Insufficient examples. Unconstrained facial emotion recognition is *person-independent*. The system must predict emotions of new individuals that are not in the training data. Because of uniqueness of expressions, in testing, there are insufficient examples to properly describe the emotion of a new individual. The top approach for discrete emotions on the FERA challenge [17] achieved 96% with *person-dependent* experiments–shared persons between folds, but the performance dropped 21% when conducting person-independent experiments.

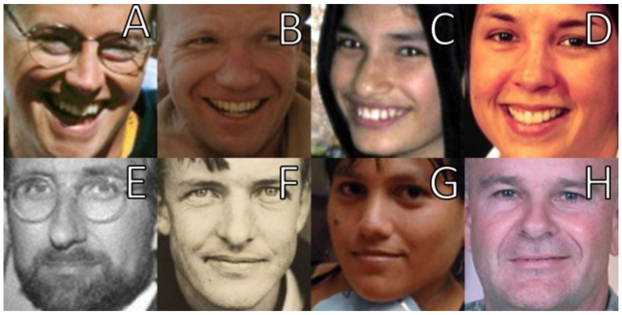


Figure 13: Examples of people smiling in different ways. (A) A strong, open mouthed smile with eyes squinted. (B) A strong smile. (C) A strong smile and pose change. (D) A strong smile. Though the top row (A-D) appears related, the faces in (A-C) are Duchene smiles, and (D) is not. Duchene smiles are characterized by use of mouth muscles and eye muscles. Note the formation of crows feet around the outer eye corners (B). (D) Does not use eye muscles, and is called a Pan-American smile. Duchene smiles are thought to be more genuine than Pan-American smiles, but this is not always the case. (E) A slight smile. (F) A crooked smile. (G) A subtle smile. (H) A crooked, slight smile. Though (E-H) appear related, only (E), (G) and are Pan-American smiles. Note that (F) has slight crows feet forming in below his right outer eye corner. This is a weak Duchene smile. Similarly, (H) has crows feet forming on his left outer eye corner.

We propose a method for comparing two face images for facial emotion recognition. It is compared to other face similarity metrics [66, 30] on MMI and found to better quantify the intensity of facial expressions. We propose three scores that can be used as features to address the generalization problem: one that measures the spatial change from a neutral face (neutral score), one that measures temporal changes between frames (temporal score), and one that compares the face to references of emotion (one-shot emotion score). The method is demonstrated to generalize better than related work [67, 34, 35, 22, 25, 39, 26, 17], on AVEC, and on intra- and interdataset experiments with CK, MMI and JAFFE. Performance is improved over a state-of-the-art method that combines many features by 23.6% on interdataset experiments when training with MMI and testing on CK+.

B. Technical Approach

The system for facial emotion recognition follows (see Figure 14): (1) In training, reference faces of the big-six emotions, and a neutral reference are estimated from all the images in the training data. (2) The frontal face image is extracted with a boosted cascade of Haar-like features [50]. (3) Neutral score captures spatial motion of the frontal face with a comparison to a neutral reference face. (4) Temporal score captures temporal motion of the frontal face with a comparison to previous frames. (5) One-shot emotion (OSE) scores compare the face to background data of big-six emotions. (6) Neutral, temporal and OSE scores are measured with a novel SIFT-flow objective function. We supplement a state-of-the-art approach with side information to improve performance. The state-of-the-art approach in this work follows: (7) faces are registered with a SIFT-based warping process [17]. (8) Local appearance features are extracted. We use LBP, four patch LBP (FPLBP) and three patch LBP (TPLBP) [30, 31] and other features. (9) The features are fused at the feature level [52]. (7) A support vector machine predicts the labels by fusing the scores and an initial prediction by the state-of-the-art method at the decision-level [52].

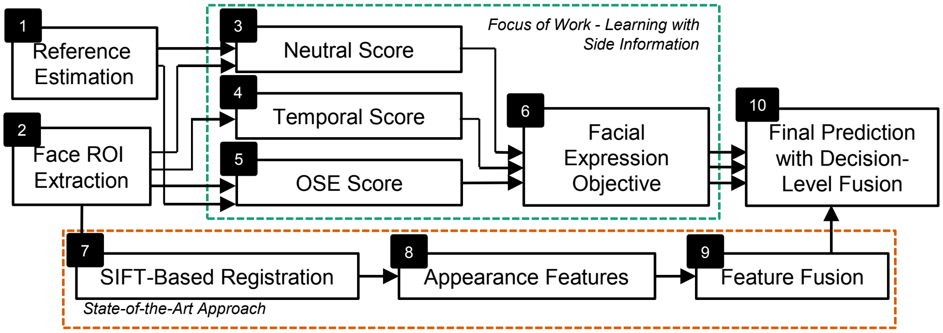


Figure 14: The proposed system integrates learning with side information (neutral, temporal, and one-shot emotion scores) with a state-of-the-art approach. (1) In training, references of characteristic emotion, and a neutrally expressive face are estimated. (2) In testing, a frontal face image is extracted. Learning with side information extracts the score features: (3) neutral score measures spatial motion from a reference face, (4) temporal score measures temporal motion between frames, and (5) one-shot emotion scores measure the distance to the reference of each emotion. (6) All three are measured with expression energy. The state-of-the-art approach extracts appearance information: (7) face images are registered, (8) appearance features are extracted, and (9) fused at the feature-level. (10) Final fusion of motion and appearance is done at the decision- level.

1. Quantifying the Score of Two Images

A function is needed to determine if a query frontal face image $Q(\boldsymbol{x})$ is similar to the target frontal face image $T(\boldsymbol{x})$. Q and T are of the same size, and $\boldsymbol{x}$ is a pixel location. The function $d(Q,T)$ computes a score that gauges whether or not Q and T are similar images. We compare two face images with an improved version of SIFT-Flow [66]. SIFT-flow warps the dense SIFT features between two images. Its original purpose was to align two images of similar content, and has been applied to face registration [17]. It resembles optical flow performed on dense SIFT descriptors. It is described as follows: (1) 128 dimensional SIFT descriptors of Q and T are computed densely, for each pixel. They are s_Q and s_T respectively. (2) The SIFT descriptors are matched using max-product loop belief propagation [68]. An example is given in Figure 15. SIFT-Flow quantifies the changes between Figure 15-(A) and -(B). For a more detailed description, the reader is referred to [66].

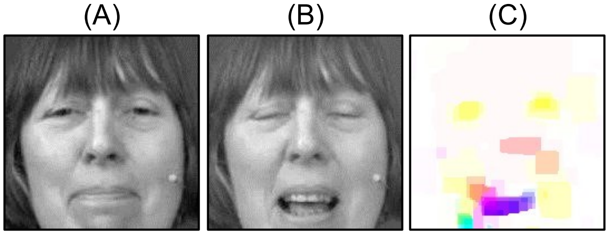


Figure 15: An example of how expression energy can measure the difference between two frontal face images. (A) The query image. (B) The target image. (C) Visualized flow. Color indicates a change detected between (A) and (B). Note that the changes in eyes and mouth are detected. There is also a change detected her right nasolabial furrow–the crease in the cheek–because her jaw extends to her right between (A) and (B).

For a frontal face image score, the objective function, of the SIFT-flow warping process can be used to describe the similarity between two images. If the objective function has a low score, the images are similar; if high, the images are different. SIFT-flow finds an optimal flow field $\boldsymbol{w}$ warping the SIFT features of Q and T. The flow field has vertical and horizontal components: $\boldsymbol{w}(\boldsymbol{x}) = |u(\boldsymbol{x}), v\ (\boldsymbol{x})|$. We propose modifying the SIFT-flow objective function to better detect changes in appearance of the face due to facial expressions. It is called *expression energy*, and is defined as follows:

$$\begin{aligned} \mathrm{E}\big(\boldsymbol{w}(\boldsymbol{x})\big) = &\sum\nolimits_{\boldsymbol{x}} \big\| s_Q(\boldsymbol{x}) - s_T\big(\boldsymbol{x} + \boldsymbol{w}(\boldsymbol{x})\big)\big\|_1 \\ &+ \gamma \sum\nolimits_{\boldsymbol{x}} \left((u(\boldsymbol{x}) - \mu_u)^2 + (u(\boldsymbol{x}) - \mu_v)^2\right) \end{aligned} \qquad 27$$

where $\mathrm{E}\big(\boldsymbol{w}(\boldsymbol{x})\big)$ is the objective function that is minimized; $\|.\|_1$ is the L^1 norm; γ is a parameter; $s_Q\big(\boldsymbol{x} + \boldsymbol{w}(\boldsymbol{x})\big)$ are the dense SIFT features of Q at $\boldsymbol{x}$ offset by $\boldsymbol{w}(\boldsymbol{x})$; μ_u and

μ_v are the mean motion in the x and y directions respectively. The proposed score function d is:

$$d(Q,T) = E\big(\boldsymbol{w}(\boldsymbol{x})\big) \qquad 28$$

Assuming the images are of similar content, there is some SIFT feature in s_Q that matches with the feature in $s_{T.}$ With this assumption, the optimal $\boldsymbol{w}(\boldsymbol{x})$ causes $s_Q(\boldsymbol{x}) - s_T\big(\boldsymbol{x} + \boldsymbol{w}(\boldsymbol{x})\big) = 0$. The first term iterates for each pixel and finds the SIFT feature that matches according to an L^1 norm difference. The second term constrains the motion to punish large changes. It sums the magnitude of all the motion vectors in $\boldsymbol{w}(\boldsymbol{x})$. In a video sequence, assuming that the frame rate is sufficiently high, the optimal match should be spatially near to $\boldsymbol{x}$.

Is Equation 2 a metric? Equation 1 follows most properties of a metric. The non-negativity principle requires that $E(\boldsymbol{w}) \geq 0$. Note that the terms of Equation 1 are an L^1 norm and a squared magnitude. Because all terms of $\boldsymbol{w}$ are absolute value or squared, all values of $\boldsymbol{w}$ cause $E(\boldsymbol{w}) \geq 0$. The identity of indiscernibles property requires that $E(\boldsymbol{w}) = 0$ when computing the score between two identical images. If the images are the same, $\boldsymbol{w} = \boldsymbol{0}$, it follows that $E(\boldsymbol{w}) = 0$, so this property is satisfied. The symmetry property requires that, $E(\boldsymbol{w})$ is the same when warping $Q \rightarrow T$ and vise versa. $Q \rightarrow T$ will have the same magnitude as $T \rightarrow Q$, but an inverted angle. Note that the terms of Equation 25d o not incorporate the angle of $\boldsymbol{w}$. Only the magnitude is considered, so this property is satisfied.

However, it is not clear if the triangle inequality is satisfied. It would require that $d(Q,T) \leq d(Q,R) + d(R,T)$, given that R is another image. Q, R and T can be any image, so there may be a condition that does not satisfy this property. For this reason, we do not refer to Equation 25 as a metric. It is a score. Despite this, the original formulation of SIFT-Flow has been used as a metric for scene and face recognition [66].

2. Discussion of Score Fusions

The score function of the original SIFT-Flow method $d_{\text{SIFT−Flow}}(Q,T)$ uses the following objective function:

$$\begin{aligned} \mathrm{E}\big(\boldsymbol{w}(\boldsymbol{x})\big) = \sum\nolimits_{\boldsymbol{x}} \big\| s_Q(\boldsymbol{x}) - s_T\big(\boldsymbol{x} + \boldsymbol{w}(\boldsymbol{x})\big)\big\|_1 \\ + \gamma \sum\nolimits_{\boldsymbol{x}} \Big(\big(u(\boldsymbol{x})\big)^2 + \big(u(\boldsymbol{x})\big)^2\Big) \\ + \alpha \sum\nolimits_{\boldsymbol{x}} \sum\nolimits_{\forall k : k \in N_k} \big(\min|u(\boldsymbol{x}) - u(\boldsymbol{k})| + \min|v(\boldsymbol{x}) - v(\boldsymbol{k})|\big) \end{aligned} \qquad 29$$

where α is a parameter; and N_k is the neighborhood of pixels at $\boldsymbol{k}$. The third term is the homogeneity term. It iterates for each pixel $\boldsymbol{x}$, and then iterates for each pixel in its neighborhood N_k, where $\boldsymbol{k}$ is the iterator. It enforces $\boldsymbol{w}(\boldsymbol{x})$ to have a similar value as its neighbors. The proposed objective function (Equation 27) differs from the original objective function of SIFT-Flow (Equation 29) in two ways. The first difference is that, in Equation 27, the motion vectors are offset by the mean translation of the flow field. In Equation 29, if there is a translation error in frontal face image detection, the translation would affect each $\boldsymbol{w}(\boldsymbol{x})$, and increase the value of the objective function. This is not desirable because we want to detect motion due to motion of parts of the face, not a translation error from a coarsely registered frontal face image. A translation error would detect motion for all $\boldsymbol{x}$. The improvements made in Equation 27 allow the score to be computed on images extracted with Viola and Jones [50] without fine registration, where there is not much out-of-plane pose change. The ROI detector suffers from slight translation errors from frame to frame, which are handled by Equation 27. The second difference is that Equation 1 lacks the third term from Equation 3. It was removed because SIFT-flow was designed for scene alignment. In scene alignment, there are multiple objects in the frame that need their local structure preserved. This is achieved by ensuring that $\boldsymbol{w}(\boldsymbol{x})$ warps similarly to its neighbors. Since we are only looking at a single object, the face, it is not needed. Because the third term was removed, we have reduced the number of weight parameters and reduced computation time.

Wolf et al. [30] employed the Minkowski metric for comparing two images:

$$d_{Minkowski}(Q,T) = \left(\sum_{\boldsymbol{x}} |Q(\boldsymbol{x}) - T(\boldsymbol{x})|^p\right)^{\frac{1}{p}} \qquad 30$$

where p is the degree. It is also referred to as the L^p norm, or p-norm. For $p = 1$ this is the Manhattan distance, for $p = 2$ this is the Euclidean distance.

a) Neutral Score

The key motivation for learning with side information is that the feature space is not discriminative. Side information is used to create a more discriminative feature space. Neutral score addresses the question, “how intense is the facial expression?” The score quantifies spatial motion in Q with a comparison to a reference of a neutrally expressive reference frontal face image. Because the score measures the degree of spatial motion, it is proportional to the intensity of the emotion being expressed, which is useful when the system has to predict the intensity value of emotion in a regression problem. It is also useful because certain emotions cause significant distortion of the face, such as surprise.

The neutral reference is a representation of a person's face without expressions, which is estimated from all the images in training data D. There exists an neutral reference $T_{Neutral}$, which is a representation of a face without expressions, that which can be used to gauge how expressive a face image is. A pixel of a frame from a video $R(\boldsymbol{x})$, in the data D, is an independent observation of $T_{Neutral}$, subject to noise from pose, gaze, expression, lighting conditions, etc. Assuming that the noise is additive and normally distributed, the observed pixel $R(\boldsymbol{x})$ is distributed according to:

$$R(\boldsymbol{x}) \sim N(T_{Neutral}(\boldsymbol{x}), \sigma) \qquad 31$$

where $N(T_{Neutral}(\boldsymbol{x}), \sigma)$ is a Normal distribution of mean $T_{Neutral}(\boldsymbol{x})$ and standard deviation of σ. Each pixel $T_{Neutral}(\boldsymbol{x})$ must be estimated. The minimum variance, unbiased estimator solution to $T_{Neutral}(\boldsymbol{x})$ is:

$$\tilde{T}_{Neutral}(\boldsymbol{x}) = \frac{1}{n_D} \sum\nolimits_{\forall R \in D} R(\boldsymbol{x}) \qquad 32$$

where $\tilde{T}_{Neutral}(\boldsymbol{x})$ is the estimated value of $T_{Neutral}(\boldsymbol{x})$ and n_D is the number of training images in the sequence.

$T_{Neutral}(\boldsymbol{x})$ is sensitive to registration errors. Without registration, $T_{Neutral}(\boldsymbol{x})$ will be a blurry image because the facial features have not been aligned. To address this, we exploit avatar reference image [17]. It generates the reference of an image by computing an initial reference image, then warps the training images to the initial reference to create a better reference. Specifically, first, $\tilde{T}_{Neutral}(\boldsymbol{x})$ is estimated from the images in D. Next, all the images in D are warped to $\tilde{T}_{Neutral}(\boldsymbol{x})$. Let the warped images be D_0. This results in D_0 being coarsely registered, so a better version of

$\tilde{T}_{Neutral}(\boldsymbol{x})$ is recomputed from D_0. The process iterates. Let D_1 be the result of warping the images in D_0 to the better version of $\tilde{T}_{Neutral}(\boldsymbol{x})$. The number of iterations is a parameter that is selected empirically. Let S_{NS} be the neutral score:

$$S_{NS} = d\left(Q, \tilde{T}_{Neutral}\right) \tag{33}$$

Examples of S_{NS} for frontal face images from CK+ and the AVEC datasets are given in Figure 16. Estimated $\tilde{T}_{Neutral}$ is given in Figure 17 for three iterations.

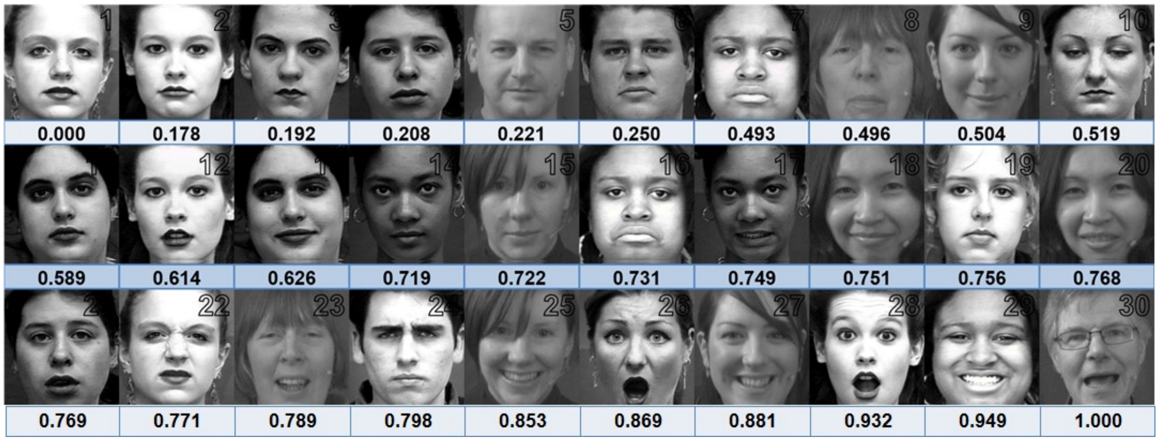


Figure 16: Frames from CK+ and AVEC with normalized neutral score S_{NS} below each frame, in ascending order. Neutral score should be proportional to distance from a neutrally expressive face. In general, faces 1-15 are characterized by having no expressions, or a single slight expression; faces 16-23 are characterized by having a strong expression; and faces 24-30 are characterized by having multiple strong expressions.

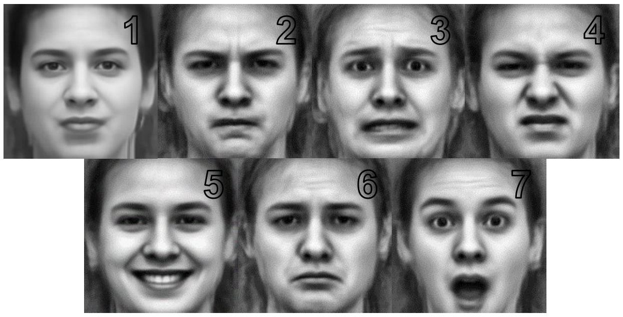


Figure 17: Estimated target reference images from CK+ and MMI. (1) Neutral reference $\tilde{T}_{Neutral}$. (2-6) $T_{a_{j^j}}$ for disgust, fear, anger, happiness, sadness, and surprise.

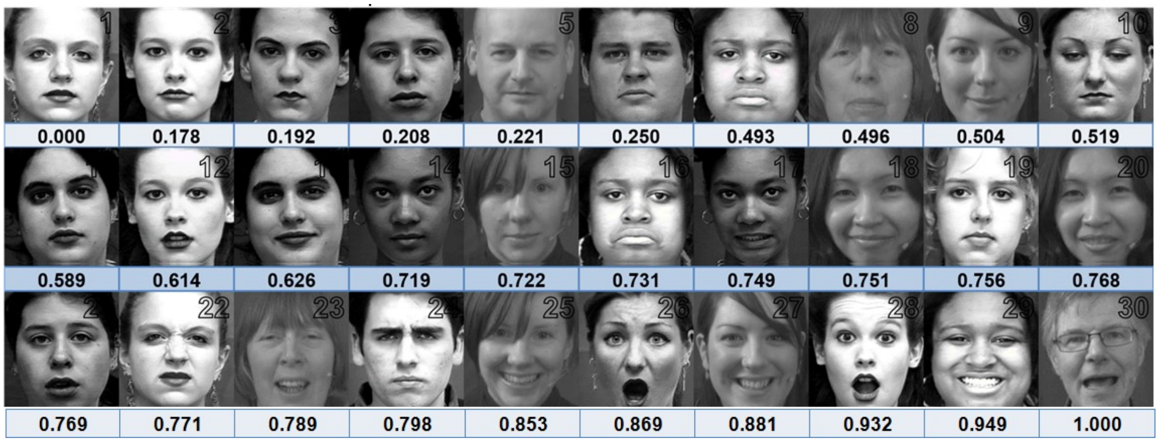


Figure 18: Frames from CK+ and AVEC with normalized temporal score S_{TS} below each frame. For CK the apex frame is compared to the first frame. For AVEC $\delta = 5s$. Temporal score should be proportional to the amount of temporal change between the two faces. For faces 1-4 a single part of the face has a weak expression. In faces 5-9, there are one or more regions of the face that are subject to strong facial expressions. In faces 10-15, the face has significantly changed from the previous time point.

b) Temporal Score

Temporal score addresses the question, “what has changed from the previous frame?” It measures the temporal changes between two frames of the same video. Temporal score can be discriminative for emotions that cause sudden changes in expression. For example, a sudden change in temporal score can indicate surprise, because the expressions of the face suddenly change. The method should report a low score when the person has little or no change between frames, e.g. a smirk, and should report a higher score in situations when there is a large changes between frames, e.g. a full, open-mouth smile. For this score, Q_{t_0} indicates the query image from a video at the current frame t_0, and $Q_{t_0-\delta}$ a frame from the same video from δ frames before t_0. Let S_{TS} be the temporal score:

$$S_{TS} = d\left(Q_{t_0}, Q_{t_0 - \delta}\right) \tag{34}$$

δ is a parameter. Examples of S_{TS} score are given in Figure 18.

c) One-Shot Emotion Scores

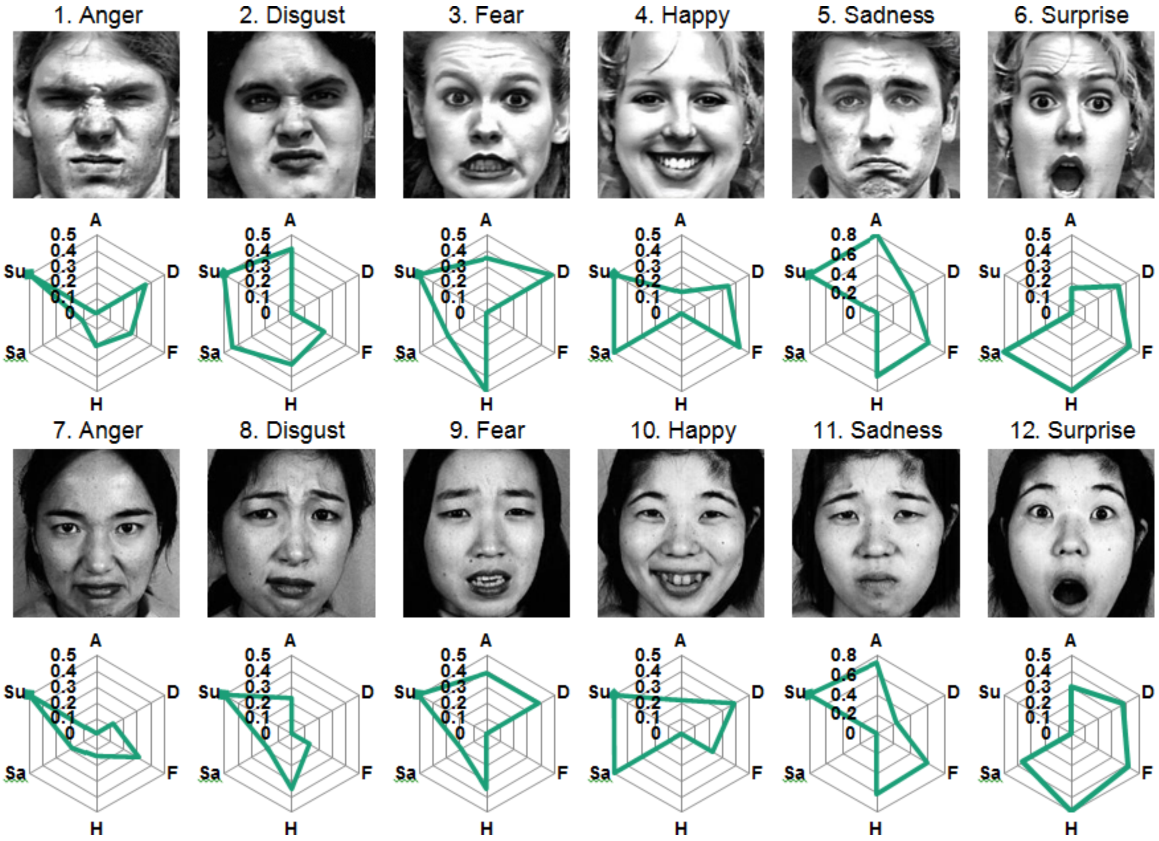
One-shot emotion scores address the question, "is this emotion similar to other emotions?" It answers this in a way that does not directly compare Q to specific examples in D. We do not want to generate a training model from positive examples from the persons in D, because each person expresses their emotion in a unique way. We overcome this by comparing Q to a set of reference faces that describe the big-six emotions. The set of faces is called *background data*.

Let Φ be the set of big-six emotions: $\Phi = \{a_1, \ldots, a_{n_e}\}$, where n_e is the number of emotions. Let T_{a_j} be the reference of emotion a_j. Let the set of background emotions A be: $\Phi - \{a_j\}$. The one-shot emotion scores are:

$$S_{Q\to A} = \arg\min_{a_i \in A} d\left(Q, T_{a_i}\right) \tag{35}$$

$$S_{Q\to T_{a_j}} = d\left(Q, T_{a_j}\right) \tag{36}$$

A visual example is given in Figure 8. If a face is not similar to T_{a_j}, $S_{Q\to A}$ should have a lower score than $S_{Q\to T_{a_j}}$. If a face is similar to T_{a_j}, $S_{Q\to T_{a_j}}$ should have a lower score than $S_{Q\to A}$. T_{a_j} is estimated for each emotion in Φ with the same method that computes $\tilde{T}_{Neutral}$, but, instead of D, a subset of D is used where only the faces that positively express a_j are used. Estimated T_{a_j} for each emotion is given in Figure 17. The OSES scores are the pair $S_{Q\to T_{a_j}}$and $S_{Q\to A}$. Equations 35 and 36 form the feature vector and an initial classification is done by SVM.



1. Anger 2. Disgust 3. Fear 4. Happy 5. Sadness 6. Surprise

7. Anger 8. Disgust 9. Fear 10. Happy 11. Sadness 12. Surprise

Figure 19: A: anger, D: disgust, F: fear, H: happiness, Sa: sadness and Su: surprise. Radar graphs of showing $S_{Q\rightarrow T_{a_j}}$ for twelve face images. Each dimension on the radar graph indicates $S_{Q\rightarrow T_{a_j}}$ of the given frontal face image computed from the references in Figure 17. A low score indicates that the face is similar to the reference for that emotion. The values are normalized to [0,1]. The top row is from CK+ and the bottom row is from JAFFE. Note that JAFFE is sometimes less expressive than CK+, causing different score values for the same emotion. However, surprise was the maximum score for both datasets, and was clipped at .5 (.8 for sadness) to better visualize the results. This is because it causes the most distortion of the face. Anger and fear are scored similar to sadness because the brow is lowered. The examples of happiness are Duchene smiles, so the cheeks are raised causing the appearance of squinted eyes, which is similar to the reference

d) Comparison to Learning with Side Information

Examples of side information include one-shot similarity (OSS) and two-shot similarity [30]. In [30], Q and T are two face images, and the system must determine whether they are the same. Q is the query image, and T is the target image. In OSS, instead of utilizing the features for classification, Q is compared to a set of unlabeled background data called junk faces, faces that are not of class Q or T. The distance to the junk faces, in the feature space, is used as the feature for verification. These similarities cannot be directly applied for facial emotion recognition. We are querying one face, not a pair of faces. In OSE scores, we compare Q to A and Q to T_{a_j}, whereas OSS compares Q to A and T to A.

3. Emotion Recognition Pipeline

Faces are extracted with Viola and Jones [50]. In training, if there is no frontal face detected, the frame/image is removed from the training data. In testing, the frame/image is not classified, and the missing values are interpolated with the nearest prediction from a frame/image with a successfully detected frontal face. The state-of-the-art approach in the paper is described as follows: faces are registered with avatar image registration [17]; LBP features [16], TPLBP and FPLBP features [30], and frame number are used as features; and features are fused at the feature level [52].

The state-of-the-art approach and the proposed approach are fused. An initial prediction of the label is generated using a support vector machine (SVM) for classification, or -support vector regression (SVR) for regression [51]. There are n_e + 3 initial predictions, one for S_{NS}, one for S_{TS}, one from the state-of-the-art approach, and one initial prediction for each pair of OSE scores (there are n_e emotions, so there are n_e initial predictions from OSE scores). The final prediction is made by fusing the $n_e + 3$ initial predictions the decision level [52] with another SVM/SVR.

Additionally, for continuous datasets, temporal smoothing is introduced by aggregating all the initial predictions in a duration of $2l$: if the current frame is t_0, the $n_e + 3$ initial predications for the frames in the range of $t_0 - l < t < t_0 + l$ are concatenated to form the features for the SVM/SVR. When the range contains values before or after the start of the video, the missing values are interpolated with nearest neighbor.

C. Experiments

Section 1 gives the parameters used in the paper. Section a) gives unconstrained, continuous data results on AVEC 2012. Section b) gives generalization results with interdataset experiments on CK+, MMI and JAFFE.

1. Parameters

For the parameters specific to AVEC 2012: videos are subsampled by a factor of 5, the delay for temporal score $\delta = 5$, and the temporal smoothing parameter $l = 2$. $\tilde{T}_{Neutral}$ and $\tilde{T}_{a_j}$ are estimated the apex frames in CK+ and MMI. For CK+ and MMI: for S_{TS}, t_0 is the apex frame and δ is the neutral frame. There is no temporal smoothing. $\tilde{T}_{Neutral}$ and $\tilde{T}_{a_j}$ are estimated from the apex frames of which ever dataset is used for training, e.g. in C2M, D are the apex frames of CK+. For JAFFE: S_{TS} is not used. There is no temporal smoothing. The class label of an image is taken to be the emotion label with the highest intensity. For SIFT-flow: α = 510 and γ = 1.257. For LBP: the radius is 1 and there are 8 neighbors. We divide the image into 10 × 10 sub-regions, and discard the outer 36 sub-regions, because these correspond to the outer regions of the frontal face where there are usually no facial expressions. For FPLBP and TPLBP: we use the same parameters as Wolf et al. [30]. For computing the avatar reference image, $\tilde{T}_{Neutral}$ and $\tilde{T}_{a_j}$: the algorithm is run for three iterations, and is trained on CK+ and MMI. For the SVM [51]: an RBF kernel is used, the cost c = 1, and $\gamma = 2^8$. The feature vectors are normalized to [-1,1]. For the SVR: ϵ-SVR is used, where $\epsilon = 0.1$.

2. *Experimental Results*

a) *Continuous Dataset Results on the Audio/Visual Emotion Challenge*

The proposed method is compared to three other top performers that reported a video-only result for the frame-level subchallenge on AVEC 2012. We also compare the results from different initial predictions. For OSE scores, the n_e initial predictions for all Φ are taken to be the feature vector. The results are given in terms of the average across all classes, and in terms of correlation the ground-truth in Table 4.

For the development set, Nicolle et al. [22] is the best performer for average correlation. This is because they took advantage of metainformation, such as who the user was speaking with. In this specific case, the person in video is speaking with a character who expresses only one emotion, and there may be a typical response based on who they were speaking with. The metainformation may not be available for other datasets. The proposed method fusing all the features and scores is the second best performer for average correlation.

For the testing set, Nicolle et al. [22] and Ozkan et al. [25] do not report results using video-only features. The final prediction method has the best average performance, except for arousal. Both the neutral and temporal only, and OSE score only methods are better than the state-of-the-art only method. We conclude that the use of scores is a better contributor to correlation than the use of appearance features.

Table 16: Comparison to other methods on AVEC 2012 video-based frame-level subchallenge testing and development sets.

Video-only Development Set					
Method	Arousal	Expectancy	Power	Valence	Avg
Baseline [15]	0.151	0.122	0.031	0.207	0.128
Nicolle et al. [22]*	**0.354**	**0.538**	**0.365**	**0.432**	**0.422**
Ozkan et al. [25]	0.117	0.076	0.062	0.200	0.114
Savran et al. [26]	0.306	0.215	0.242	0.370	0.283
State-of-the-art	0.140	0.160	0.073	0.178	0.138
N/T score only	0.160	0.280	0.258	0.253	0.238
OSE score only	0.283	0.279	0.224	0.340	0.282
Proposed fusion	0.332	0.372	0.278	0.349	0.333
Video-only Testing Set					
Method	Arousal	Expectancy	Power	Valence	Avg
Baseline [15]	0.077	0.128	0.030	0.134	0.092
Nicolle et al. [22] *,**	-	-	-	-	-
Ozkan et al. [25]**	-	-	-	-	-
Savran et al. [26]	**0.251**	0.153	0.099	0.210	0.178
State-of-the-art	0.117	0.133	0.082	0.122	0.114
N/T score only	0.258	0.274	0.187	0.181	0.225
OSE score only	0.218	0.247	0.209	0.309	0.246
Proposed fusion	0.240	**0.346**	**0.215**	**0.334**	**0.284**

N/T: Neutral and temporal score. Italic indicates results from this paper. Bold indicates the best performer, underline indicates the second best performer.

**Video-only testing set not reported.

*Best performing video feature.

b) *Interdataset Experiments*

We test the person-independent generalization capability of the algorithm with interdataset experiments. There is no official testing methodology for interdataset experiments, though some works use CERT [69], which is a combination of many datasets including CK+ [36] and MMI [37]. However, CERT uses some non-public datasets, so we cannot use this methodology. While CK+ and MMI datasets have been thoroughly addressed with intradataset experiments, with high classification rates [70, 71], intradataset experiments with CK and MMI can have a classification rate as low as 44.1% for a state-of-the-art system (Table 17). For these reasons, we conduct interdataset experiments with CK+, MMI and JAFFE. We conduct a 2-run testing validation: one dataset is used as the testing fold and one or more other datasets are used as the training fold. We compare our results to related work and also give results when using the initial predictions from different parts of the method. We also conduct intradataset experiments on CK+ and MMI.

Results for inter- and intradatasets experiments on CK+, MMI and JAFFE are given in Table 17. In leave-one-subject-out, folds are created for each individual, to ensure person-independent generalization. The state-of-the-art only method does well on intradataset CK+, but not on interdataset MMI or any interdataset experiment. Both of the score only methods outperform the state-of-the-art only method for all interdataset experiments. A score-based feature space describes the differences between classes better than the state-of-the-art feature space. An example showing 5 faces for each of the 6 big-six emotions from CK+ and JAFFE are given in Figure 20. From Figure 20, graphing the samples based on their OSE scores form more distinct clusters than graphing the samples based on their feature vector from the state-of-the-art only approach.

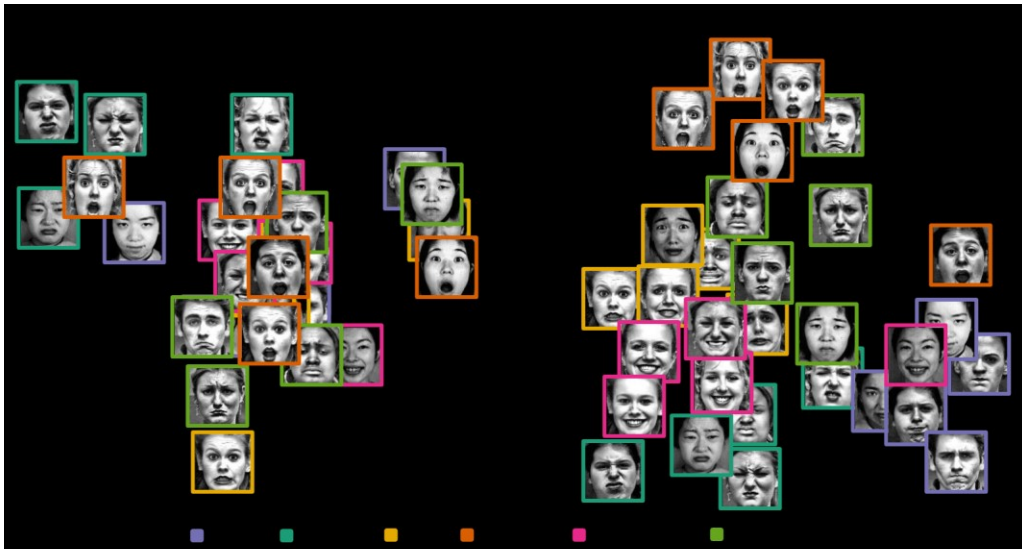


Figure 20: A comparison of the feature representations from the state-of-the-art approach versus the OSE scores for 30 samples from CK+ and JAFFE, using and MATLABs mdscale function that creates a 2-D representation of the samples based on Euclidean distance. (Left) The classes in the state-of-the-art only representation form clusters, but, with the exception of disgust, the clusters intersect and are congested in the center of the graph. (Right) The classes in the OSE score representation also form clusters. Though there are a few errors, the clusters are well separated.

Table 17: Interdatabase testing for score-based facial emotion recognition on CK+, MMI-DB and JAFFE databases.

Method	CK+	MMI	C2M	C2J	MC2J	M2C
Ghanem [67]	-	83.1	53.1	-	-	85.1
Li et al. [34]	87.4	-	-	-	-	-
Miao et al. [35]	-	-	55.7	**58.5**	58.3	-
Poursaberi et al. [39]	-	87.7	-	-	-	-
Yang and Bhanu [17]	82.6	-	-	-	-	-
State-of-the-art only	89.8	62.5	43.4	44.1	45.5	64.9
N/T score only	89.0	78.0	56.8	51.0*	52.0*	72.0
OSE score only	90.9	86.4	57.6	56.1	57.1	85.8
Proposed fusion	**92.4**	**90.5**	**61.9**	58.1	**60.1**	**88.5**

Acronym indicates which dataset was used for training and which was used for testing. N/T: Neutral and temporal score. C: CK+. M: MMI. J: JAFFE. For example, C2M indicates CK+ was used for training and MMI was used for testing. *Temporal score not used because dataset is images.

D. CONCLUSION

In conclusion, we describe a method for learning-with-side-information that computes scores from background information (references faces and a previous frame).

We demonstrate incorporating the scores with a state-of-the-art approach improves performance. Correlation with the ground-truth increased by a factor of 2.41 on AVEC2012. Classification rate increased by 5.1% when training with CK+ and testing on MMI; it increased 23.9% training on CK+ and testing on MMI. The scores are quantified with an improved version of SIFT-flow. We demonstrate that the improvements represent the intensity of an emotion better versus related work [66, 30]. The work shows promise for unconstrained person-independent emotion recognition.

CHAPTER 7 CONCLUSIONS AND FUTURE WORK

One of the major themes of computer science is that algorithms are designed as either: (1) heuristics or (2) designed after biological and human systems. There are some cases where human-like systems do not perform well when compared to their heuristic counterparts. An example of this is Deep Blue and other chess playing algorithms, which use alpha-beta pruning. These heuristic algorithms perform better than human-like logic (IF-THEN) for chess. However, we found that facial emotion recognition systems designed to emulate the human visual system perform better than heuristics. Human facial expressions are intended to be understood by other humans, so it is possible that this is the reason why non-heuristics perform better than heuristics for facial emotion recognition.

It was found that more samples is not better for building a training model for prediction of human emotions with the Audio/Visual Emotion Challenge. In Chapter 4 we found that the majority of samples in a continuous video set are redundant and reduce performance a facial emotion recognition system. This suggests that HD quality video is not necessary for facial emotion recognition. High quality image resolution is not required for facial recognition, which is a similar field [72]. Microexpressions occur at a rate which requires between 15-25 fps, so it is necessary to capture a base frame rate of at least 25 fps. However, there are time segments during speech where a person is not expressive. These redundant, non-discriminative samples outnumber the samples where a person is expressive and it is possible that this caused a decrease in performance. We do not recommend using HD spatial resolution and we recommend compressing the video temporally, leaving expressive parts of the video intact. It is possible that HD video is not necessary to recognition facial emotion.

We presented a method that was able to detect facial emotion of individuals who had no representation in the training data. The algorithm in Chapter 6 shows promise for real world applications. In real world applications, you cannot expect an individual in testing to be represented in the training data.

An anecdote is often said that computer vision is about the features. Consider the performance of the system that focused strictly on the best feature with AVEC 2012 in Chapter 5 : the average correlation across all classes was 0.258. Compare this to the system that focused on the classifier in Chapter 6 : 0.284 with similar parameters. The system that focused on sampling in Chapter 4 achieved an average correlation across all classes of 0.252. This seems to contradict the anecdote. The classifier was found to be the most important part of the system. Furthermore, using a state-of-the-art system and downsampling the videos temporally, in an intelligent way, has a similar impact on performance. This suggests that choosing the most representative samples for your problem is as important as the feature you select, and that choosing the right classifier is also as important.

A. Future Work

With respect to Vision and Attention Theory downsampling, future work should investigate a better method for segmenting the videos. Currently, a video is segmented into uniformly sized non-overlapping subsegements, and the frames are subsampled within that subsegment. However, there may be an instance where the boundary of the subsegment falls at a time point where there is an important expression. Future work should dynamically segment the regions. Furthermore, future work should investigate the importance of specific subfeatures of the face to determine if subfeatures would require specific sampling rates from other subfeatures. With respect to background suppressing Gabor filtering, future work should investigate applications of the non-classical receptive field to improve the acutance of blurry images. With respect to score-based facial emotion recognition, future work should investigate the importance of what feature is used when computing the similarity between the face and the generalized representation of an emotion.

REFERENCES

[1] J. Fontaine, K. Scherer, E. Roesch and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science,* vol. 18, no. 12, pp. 2050-1057, 2007.

[2] P. Ekman, "Basic Emotions," in *The Handbook of Cognition and Emotion*, New York, John Wiley & Sons, 1999, pp. 45-60.

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

[4] M. 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. CVPR*, 2005.

[5] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. Systems, Man, and Cybernetics B,* vol. 34, no. 3, pp. 1588-1595, 2004.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l. J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[7] C. Darwin, The Expression of the Emotions in Man and Animals, John Murray, 1872.

[8] R. e. Kaliouby and P. Robinson, "The Emotional Hearing Aid: An Assistive tool for Children with Asperger Syndrome," *Universal Access in the Information Society,* vol. 4, no. 2, pp. 121-134, 2005.

[9] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilley, P. Robinson, I. Davies, O. Golan, S. Friedenson, S. Friedenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri and S. Piana, "ASC-Inclusion: Interactive emotion games for social inclusion of children with Autism Spectrum Conditions," in *Intelligent Digital Games for Empowerment and Inclusion*, 2013.

[10] J. Shotton, A. Fitzgibbon, M. Cook, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Proc. IEEE Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011.

[11] G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schröder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent," *IEEE Trans. on Affective Computing,* vol. 3, no. 1, pp. 5-17, 2012.

[12] A. C. Elkins, Y. Sun, S. Zafeiriou and M. Pantic, "The Face of an Imposter: Computer Vision for Deception Detection," in *Proc. Hawaii Int'l. Conf. on System Sciences*, Grand Wailea, HI, 2013.

[13] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic and K. Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge," *IEEE SMC B,* vol. 42, no. 4, pp. 966-979, 2012.

[14] B. Schuller, M. Valstar, F. Eyben, R. Cowie and M. Pantic, "AVEC 2011 – the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*, 2011.

[15] B. Schuller, M. Valstar, F. Eyben, R. Cowie and M. Pantic, "AVEC 2012 – the continuous audio/visual emotion challenge," in *ACM ICMI*, Santa Monica, CA, 2012.

[16] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI,* vol. 24, no. 7, pp. 971-987, 2002.

[17] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Trans. SMC Part B,* vol. 42, no. 4, pp. 980-992, 2012.

[18] G. A. Ramirez, T. Baltrusaitis and L. Morency, "Modeling latent discriminitive dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction Workshops*, 2011.

[19] L. P. Morency, A. Quanttoni and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. CVPR*, 2007.

[20] M. Glodek, S. Tschenchne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kachele, M. Schmidt, H. Neumann, G. Palm and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*, 2011.

[21] M. Dahmane and J. Meunier, "Continuous emotion recognition using Gabor energy filters," in *Affective Computing and Intelligent Interaction Workshops*, Memphis, TN, 2011.

[22] J. Nicolle, V. Rapp, K. Bailly, L. Prevost and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[23] C. Soladie, H. Salam, C. Pelachaud, N. Stoiber and R. Seguier, "A multimodal fuzzy inference system using a continuous facial," in *ACM Int'l. Conf. Multimodal Interaction Workshops*, 2012.

[24] L. Maaten, "Audio-visual emotion challenge 2012: a simple approach," in *ACM Int'l. Conf. Multimodal Interaction Workshops*, 2012.

[25] D. Ozkan, S. Scherer and L. Morency, "Step-wise emotion ecognition using concatenated-HMM," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[26] A. Savran, H. Cao, M. Shah, A. Nenkova and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in

ACM Int'l. Conf. Multimodal Interaction, 2012.

[27] Y. Zhu, F. D. l. Torre, J. F. Cohn and Y. Zhang, "Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior," *IEEE Trans. Affective Computing,* vol. 2, no. 2, pp. 79-91, 2011.

[28] B. Jiang, M. Valstar, B. Martinez and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. SMC B,* 2013.

[29] Z. Zeng, M. Pantic, G. Roisman and T. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. PAMI,* vol. 31, no. 1, pp. 39-58, 2009.

[30] L. Wolf, T. Hassner and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. PAMI,* vol. 33, no. 10, pp. 1978-1990, 2011.

[31] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI,* vol. 29, no. 6, pp. 915-928, 2007.

[32] X. Huang, G. Zho, W. Zheng and M. Pietikainen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *IEEE Signal Processing Letters,* vol. 19, no. 5, pp. 243-246, 2012.

[33] W. Zheng, H. Tang, Z. Lin and T. Huang, "Emotion recognition from arbitrary view facial images," in *European Conf. Computer Vision*, 2010.

[34] Y. Li, S. Wang, Y. Zhao and Q. Ji, "Simulataneous facial feature tracking and facial expression recognition," *IEEE Trans. IP,* p. 2559–2573, 2013.

[35] Y. Miao, R. Araujo and M. S. Kamel, "Cross-domain facial expression recognition using supervised kernel mean matching," *IEEE Int'l Conf. Machine Learning Applications,* 2012.

[36] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih and Z. Ambadar, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit," in *IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

[37] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the MMI facial expression database," in *Proc. Int'l. Language Resources and Evaluation Conference*, Malta, 2010.

[38] M. Lyons and S. Akamatsu, "Coding facial expressions with Gabor wavelets," in *IEEE Conf. Automatic Face and Gesture Recognition*, 1998.

[39] A. Poursaberi, H. A. Noubari, M. Gavrilova and S. N. Yanushkevich, "Gauss-laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP J. Image and Video Processing,* 2013.

[40] M. Lyons and S. Akamatsu, "Coding facial expressions with Gabor wavelets," in *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, Nara, JP, 1998.

[41] S. Koelstra, M. Pantic and I. Patras, "A dynamic texture-based approach to

recognition of facial actions and their temporal models," *IEEE Trans. PAMI,* vol. 32, no. 11, pp. 1940-4954, 2010.

[42] "Steam hardware & software survey: July 2013," 1 July 2013. [Online]. Available: http://store.steampowered.com/hwsurvey. [Accessed 2 August 2013].

[43] J. Yu and B. Bhanu, "Evolutionary feature synthesis for facial expression recognition,," *Pattern Recognition Letters,* vol. 27, no. 11, pp. 1289-1298, 2006.

[44] J. Findlay and I. Gilchrist, The Psychology of Looking and Seeing, Oxford: Oxford University Press, 2003.

[45] N. Ghosh and B. Bhanu, "A psychological adaptive model for video analysis," in *Int'l. Conf. Pattern Recognition*, 2006.

[46] W. Tingfan, M. S. Bartlett and J. R. Movellan, "Facial expression recognition using Gabor motion energy filters," in *IEEE CVPR*, San Francisco, CA, 2010.

[47] P. Lucey, S. Lucey and a. J. F. Cohn, "Registration invariant representations for expression detection," in *IEEE Conf. Digital Image Computing: Techniques and Applications*, 2010.

[48] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers," *IEEE Int'l. Conf. Systems, Man and Cybernetics,* 2005.

[49] R. Haber and M. Hershenson, The Psychology of Visual Perception, New York: Rinehart and Winston Inc., 1973.

[50] J. Viola and P. Jones, "Robust real-time object detection," *Int'l. J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2001.

[51] C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 1-27, 2011.

[52] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition,* vol. 38, no. 12, pp. 2270-2285, 2005.

[53] T. Gautama and M. A. V. Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. Neural Nets,* vol. 13, no. 5, pp. 1127-1136, 2002.

[54] A. Cruz, B. Bhanu and N. Thakoor, "Facial emotion recognition in continuous video," in *Int'l. Conf. Pattern Recognition*, 2012.

[55] H. T. Lin, C. J. Lin and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning,* vol. 5, pp. 975-1005, 2004.

[56] P. Ekman, "What are micro expressions?," Paul Ekman Group LLC., [Online]. Available: http://www.paulekman.com/me-historymore/.

[57] J. Heikkila and V. Ojansivu, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, New York, NY, 2008.

[58] T. Wu, M. S. Bartlett and J. R. Movellan, "Facial expression recognition using Gabor motion energy filters," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2010.

[59] J. R. Movellan, "Tutorial on Gabor filters," MPLab, 2008.

[60] C. Grigorescue, N. Petkov and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE Trans. Image Processing,* vol. 12, no. 7, pp. 729-739, 2003.

[61] M. Pietikainen and G. Zhao, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 915-928, 2007.

[62] A. C. Cruz, B. Bhanu and N. S. Thakoor, "Facial emotion recognition with anisotropic inhibited Gabor energy histograms," in *IEEE Int'l. Conf. Image Processing*, 2013.

[63] L. Wolf, T. Hassner and Y. Taigman, "Descriptor based methods in the wild," in *European Conf. Computer Vision Workshops*, 2008.

[64] A. J. O'Toole, F. Jiang, D. Roark and H. Abdi, "Predicting human performance for face recognition," in *Face Processing: Advanced Models and methods*, Academic Press, 2006.

[65] G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Amherst, University of Massachusetts, 2007.

[66] C. Liu, J. Yuen and A. Torralba, "SIFT flow: dense correspondence across scenes and its applications," *IEEE Trans. PAMI,* 2011.

[67] K. Ghanem, "Hidden Markov models for modeling occurance order of facial temporal dynamics," in *Conf. Adv. Concepts Int'l. Vis Systems*, 2013.

[68] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int'l. J. Computer Vision,* pp. 41-54, 2006.

[69] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett and J. R. Movellan, "Action unit recognition transfer across datasets," in *IEEE Conf. AFGR*, 2013.

[70] Y. Guo, G. Zhao and M. Pietikainen, "Dynamic facial expression recognition using logitudinal facial expression atlases," in *IEEE Conf. ECCV*, 2012.

[71] A. R. Rivera, J. R. Castillo and O. Chae, "Local directional number pattern for face anaylsis: face and expression recognition," *IEEE TIP,* 2013.

[72] S. Biswas, G. Aggarwal and P. Flynn, "Face recognition in low-resolution videos using learning-based likelihood measurement model," in *IEEE IJCB*, 2011.