From Facial Expression to Level of Interest: A Spatio-Temporal Approach

M. Yeasin\textsuperscript{1} B. Bullot\textsuperscript{2} and R. Sharma\textsuperscript{2}

\textsuperscript{1}Computer Science, State University New York Institute of Technology
yeasinm@cs.sunyit.edu

\textsuperscript{2}Computer Science and Engineering, Pennsylvania State University
bbullot@amadeus.net rsharma@cse.psu.edu

ABSTRACT: This paper presents a novel approach to recognize the six universal facial expressions from visual data and use them to derive the level of interest using psychological evidences. The proposed approach relies on a two-step classification built on the top of refined optical flow computed from sequence of images. First, a bank of linear classifier was applied at frame level and the output of this stage was coalesced to produce a temporal signature for each observation. Second, temporal signatures thus computed from the training data set were used to train discrete Hidden Markov Models (HMMs) to learn the underlying models for each universal facial expressions. The average recognition rate of the proposed facial expression classifier is 90.9\% without classifier fusion and 91.2\% with fusion using a five fold cross validation scheme on a database of 488 video sequences that include 97 subjects. Recognized facial expressions were combined with the intensity of activity (motion) around the apex frame to measure the level of interest. To further illustrate the efficacy of the proposed approach two set of experiments, namely, Television (TV) broadcast data (108 sequences of facial expression containing severe lighting conditions, diverse subjects and expressions) analysis and emotion elicitation on 21 subjects were conducted.

1 Introduction

Research in social psychology [1] has shown that facial expressions form the major modality in human communication. In particular, facial expressions provide a very strong cue about the level of interest of the user. Automated analysis of shown emotions could bring affective dimension into man-machine interaction [2, 3]. It is expected that introducing affective dimension can significantly enhance the users’ interaction experience also can make the interaction tighter and more efficient. Such a system could also make classification of facial expressions widely accessible as a tool for research in behavioral science and medicine [4].

Recognizing facial expression with a very high accuracy is a relatively hard problem even for human experts [4-7]. Automatic recognition of facial expression and its usage in affective computing presents a number of difficult challenges. One of the major problems in designing a robust facial expression classifier is the effective use of non-rigid complex motion pattern that evolve during the production of facial expression. Several approaches have been reported in the literature to automate the recognition of facial expression. The early methods use mug shot of each expression that captures characteristic image at the apex [8-10]. However, according to psychologists [5, 11], analysis of sequence of images produces more accurate and robust recognition of facial expressions. Based on this philosophy several researchers reported their works with varying degrees of success. For example, some researchers [11-14] built complex parameterized models in two or three dimensional space to track the changes between images.

To avoid the computational complexity and image quality required by these methods, many approaches only consider facial features [15-17] or facial points [18-20] and the relationships between them to classify expressions. Raw optical flow have been used past to design classifier for recognizing facial expression [21-23]. In contrast, this paper presents a two stage approach based on the following assumption:

- Using only visual sensors, we can only hope to recognize the apparent emotion, which may or may not be the person’s true emotion.
- Theories of emotion often claim that there is a small set of basic expressions [4, 5], even if psychologists sometimes disagree on the list [6]. A recent cross-cultural study [7] confirms that some emotions have a universal facial expression across cultures and that the original set proposed by Ekman [1] is a very good choice. Hence, the six basic expressions, namely, surprise, happiness, sadness, fear, anger, and disgust were considered.

Proposed two-stage approach judiciously exploits the temporal information from video sequences. It is easy to see that the production mechanism of facial expression evolves over time which produces complex motion pattern between images. First, optical flow is computed and projected (using the PCA and NMF) onto a lower dimensional space to extract the intrinsic dimensionality of the motion information that can capture the evolution of complex motion. A set of naïve classifier (k-Nearest Neighbours) were used on the refined optical flow to derive characteristic temporal signature for every video sequences. Second, the temporal signature thus computed was used to train discrete hidden Markov models (HMMs) to learn the underlying model of facial expressions. Psychological evidence has been used to evaluate the level of interest by combining the observed facial expressions with the intensity of activity (motion) around the apex frame. Additionally, emotion elicitation experiment was conducted using data captured by showing carefully chosen clips of videos to the subjects that are designed to arouse more spontaneous reactions to measure the level of interest. The proposed approach is computationally efficient compared to existing techniques [11-23] and is suitable for real-time applications.

The rest of the paper is organized as follows. Section 2 presents the state of the art in the facial expressions. Following this, the proposed two-stage approach for facial expression recognition and its usage in deriving the level of interest is described in section 3. Section 4 discusses the results obtained from three experiments conducted on different datasets, namely, the Cohn-Kanade database, emotion elicitation experiment and TV broadcasting analysis. Finally, section 5 concludes the paper with few remarks on the future directions.
2 Literature Review

Psychological studies have suggested that the facial motion is fundamental to the recognition of the facial expression. Experiments conducted by Bassili [11] demonstrate that the humans do better job in recognizing expressions from dynamic images as opposed to mug shot. The Facial Action Coding System (FACS) developed by Ekman and Friesen [24] is one of the popular methods in characterizing facial expressions. In the past, the FACS representation has been adopted to recognize facial expressions. Based on this, Essa et al. [12, 25] developed an automated system using optical flow coupled with a physical model of face muscles to describe the facial motions. Darrel et al. [13] augmented the approach by view-based representation of the facial expression using a correlation network. Tian et al. [29] developed a system to recognize upper and lower facial action units (AUs) based on detailed parametric models of the facial features.

In [26], Avent et al. developed a low-level system to detect the edge clusters corresponding to the eyes, eyebrows, and lips and use them to perform classification using Neural Network (NN). Lisetti and Rumelhart [27] propose another NN-based approach to recognize the facial expressions using different facial regions (manually cropped). In [16], Lin et al. proposed PCA and hierarchical Radial Basis Function (RBF) network for recognition of facial expression. Kumar and Poggio [15] use Support Vector Machines (SVMs) to track facial features for similar purpose. Moses et al. [17] use valley contour for tracking facial features and use it to recognize the facial expression. Yoshitomi et al. [28] presented a method based on Neural Networks (NNs) that uses differential images generated by the discrete cosine transformation (DCT). In [18,19] Yacoob et.al. use the local motions in the rectangles enclosing the facial features, to derive an intermediate description of the facial motions and use some heuristics to classify the overall expression. Learning approaches based on Hidden Markov Models (HMMs) to capture the temporal dynamics of the facial expression has gained significant popularity in recent years. For example, Oliver et al. [23] track the mouth motion based as a two-dimensional blob features and use HMMs to recognize facial expression. Cohn et al. [21] proposed a multilevel HMM that uses the state sequence of independent HMMs to automatically segment and recognize facial expressions from video sequences. Hoey [30] proposed another multi-level method using an unsupervised classification of sequences with different levels of dynamics at the image level, at the sequence level, and at a higher level corresponding to juxtaposition of the sequences with transitions. It was shown that the recognition accuracy ranges between 81%-91% depending of the way transitions occur in the higher level. Lien et al. [22] compared various methods that use optical flow. The key difference between the proposed approach and existing approaches are in the usage of information to form “temporal signature” for achieving robust recognition accuracy with computational cost graceful enough for real-time implementation. Next section describes the proposed approach in great details.

3 Proposed Approach

The key components of the proposed facial expression classifier design and determining the level of interest are described below.

3.1 Pre-Processing and Feature Extraction

Feature extraction stage is designed to obtain a meaningful representation of observations (raw sequence of images of all facial expression) and also to reduce the dimension of the feature vector. It is assumed a classifier that usage a smaller dimensional feature vector would run faster and use less memory, which is very desirable for any real-time system. In the proposed method, following techniques were implemented to construct feature vector from the observation.

The pre-processing stage segments the faces from the background, removes the noise, and normalizes the face patterns. Face detector developed in [35] was applied on raw images to localize and segment the face region from the background. Following this illumination gradient correction was performed to compensate or to reduce the effect of lighting variations within the window where face was detected. Also histogram equalization was performed to reduce the non-uniformity in the pixel distributions that may occur due to various imaging situations. Additionally, normalization on “training set” was performed to align all facial features (based on manual labeling) of the face with respect to a canonical template, so that they form a good cluster in the high dimensional feature space. A modified sum of squared distance based technique was used to compute the optical flow from consecutive image pairs. Principle component analysis (PCA) was used to reduce the dimension of the optical flow. Selected components of the PCA were used to form the feature vector.

Experiments were conducted to identify the component of optical flow that can faithfully capture the motion pattern of facial expression. First, the horizontal and vertical components of the optical flow were considered. Second, the phase and amplitude of the flow along with the horizontal and vertical flow was considered. It was found that first set of feature yields a recognition accuracy of 90.9% compared to 88.7% for the second case. The results for a given combination assume that the other parameters of the system are optimized. From the result it is seems that the vertical and horizontal components of the optical flow capture the motion pattern better than the second feature for the facial expression recognition. Hence, the horizontal and vertical components of the flow were used.

3.2 Classifier Design

To better model facial expressions a two-step classification strategy was proposed. First, a set of naïve classifiers were used on consecutive frames for entire sequences that produce characteristic temporal signature. To uncover the hidden patterns associated with each expression discrete HMMs were used to model the facial expressions. The choice of using HMMs stems from the fact that HMMs have been successfully
used in speech recognition, and more recently for visual recognition purposes such as in gesture recognition, activity recognition etc. An average estimate of the error rate of the five fold cross validation was used to report the results in percentage. A set of experiment was performed with using the Cohn-Kanade database to fine tune the parameters of the classifier. At the frame level, the free parameter of the k-NN rule, namely, the number of neighbours of the individual motion pattern was determined empirically. At the sequence level, discrete HMMs were trained to recognize the temporal signatures associated with each of the basic expressions using the signature obtained in the first stage.

3.3 Level of Interest

Instead of producing a result from the set of basic expressions, the level of interest was derived based on psychological evidences using the recognized facial expression as its input. To derive this value, facial expression recognized by the classifier discussed in the previous subsection was associated with the intensity of the expression. Psychologists [33, 34] posit that facial expression have a systematic, coherent, and meaningful structure that can be mapped to affective dimensions. Some mappings have been proposed in the context of psychological evaluation, learning and marketing.

Figure 1: Affect Space [33].

Three-dimensional affect space shown in Figure 1 discussed in [33] was adopted in this work. Each affective state is characterized by three tags associated with intensities that scale the contribution to the overall affective state. The arousal tag denotes the activity of the autonomous nervous system. Positive values correspond to a high arousal stimulus, associated with a positive interest level. The valence tag specifies how pleasant the percept is to the emotional system. The stance tag corresponds to advance, whereas negative values correspond to retreat, respectively, considered positively and negatively in terms of interest level. A weight \( W \), between -5 and 5 was assigned to individual expression. The weight assigned to a particular expression is computed by summing the contributions from the different dimensions. Expressions with a negative value thus correspond to a negative reaction in terms of arousal, valence, and stance, while positive values correspond to a positive reaction. To compute the intensity \( I \) of an emotion, relative number of images has been considered which concentrates most of the motion in the sequence. If the number is low, we associate a high coefficient for the intensity otherwise a low coefficient was used. The value of coefficients ranges between [0-1]. The intensity coefficient takes into account the range of expressions going from slightly expressed reactions with hardly visible motion to exaggerated expressions where motion concentrates on a few frames. The level of interest \( L \) is produced as \( L = W \times I \). The proposed level of interest provides tangible interpretation of the human emotional reaction and can be used in affective computing.

4 Results

A set of experiment was performed with the proposed classification strategy (discussed in the previous section) and was tested on a number of data set with different lighting environments, subjects, and expressions. First, data captured in the laboratory environment [32] were used to show the performance of the proposed approach. To report the performance a five fold cross validation approach was adopted. To further illustrate the efficacy of the proposed approach two sets of experiments, namely, TV broadcast analysis and emotion elicitation were conducted. The main aim of the experiments was to understand the effects of a number of factors that are detrimental to the recognition of facial expression recognition using visual data. Emotion elicitation experiment was conducted on 21 subjects by showing the subjects six different clips of movies carefully chosen in a manner to arouse spontaneous emotional reactions that would produce natural facial expression. TV broadcast analysis was conducted on 108 meticulously collected facial expressions from TV broadcast and labeled by human coder for subsequent analysis.

<table>
<thead>
<tr>
<th></th>
<th>Surprise</th>
<th>Happy</th>
<th>Sad</th>
<th>Fear</th>
<th>Anger</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surprise</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>96.6</td>
<td>0</td>
<td>3.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>96.2</td>
<td>3.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>11.8</td>
<td>0</td>
<td>76.4</td>
<td>0</td>
<td>11.8</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>37.5</td>
<td>0</td>
<td>0</td>
<td>62.5</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix: Illustrating the distribution of failure for six universal facial expressions considered for experimentation.

Table 1 summarizes the results obtained using a five fold cross validation scheme using Cohn-Kanade database obtained from CMU [32]. The model achieves 90.9% overall recognition of facial expression. The confusion matrix confirms that some expressions are harder to differentiate than others. Expressions labeled as surprise, happiness, sadness, and anger are recognized with very high accuracy (96.2 % - 100%), whereas fear and disgust present significant failure. The reaction associated with fear presents comparable horizontal motion in the mouth area as
with happiness and the upward motion of the corner may not by sufficient to discriminate between them. Similarly, the disgust motion pattern in the mouth area is very close to the upward motion of the mid mouth area, while the motions occurring in other parts of the face may not concentrate enough energy to be discriminative.

To validate the hypothesis that a two-step classification performs better than a single stage classification, the two stage classifier, namely, k-NN followed by discrete HMMs was replaced by continuous HMMs. The inputs to both classifiers are the same feature vectors obtained from the feature extraction stage discussed in the previous section (section 3.1). Figure 2 summarizes the results obtained for the different number of coefficients. From Figure 2, it is easy to note that proposed two-step classification improves the recognition rate by about 4% compared with a direct approach using continuous HMMs. To validate the use of a temporal classifier after the k-NN classification, an experiment was conducted using k-NN followed by a voting scheme, which classifies the sequence according to the most common expression in the time series. The performance achieved using this method was very poor (75.3%), which validates the use of a temporal classifier at the sequence level.

To improve the performance of the classifiers several classifier fusion strategies were implemented. Table 2 summarizes the results obtained for the different combination strategies tested for different model. The poor improvement may be due to the similarity of the model and insufficient balanced training data to create diverse models for each expression. It is expected that combination of classifiers performs better when the models are diverse; that can be potentially obtained by using a diverse set of features or sensorial diversity.

4.1 Level of interest

The level of interest was derived by associating the recognized facial expression with the level of activity based on the proposed approach discussed in section 3.3. Figure 3 presents some representative examples of the measured interest level from video sequences to subjectively assess the interest level derived by the proposed approach.

4.2 Emotion Elicitation Experiment

Most facial expression recognition systems have focused on deliberately expressed emotion posed in front of a camera, and not on those that arise in natural situations. Many laboratory tests in psychology was done by showing films to people and analyzing their facial behaviors, recorded by video cameras [14, 26, 28]. Showing emotionally loaded films is regarded as the least problematic emotion induction technique. Most reports in psychology consist of global ratings for one or more films that last for several minutes.

The study was carried out on 21 subjects using an HCI setting. Subjects sat on a chair in front of a large screen display and a video camera. The subjects were presented six carefully chosen movies clips that are supposed to arouse spontaneous and natural patterns of emotions. To get unaffected results, the test subjects were not told the real purpose of the experiment. To have the subjects’ concentration on the scene, she was rather told that her eye movements were recorded. Three persons unaware of the movie contents were asked to classify the sequences into the 6 categories of Ekman to create ground truth.

Other works using the same database have achieved comparable results. For example, the approach in [22] achieves 85% recognition and the approach by Tian et al. [29] achieves 95%.
for subjective evaluation. It was found that only 82% of the time three persons agreed with the type of expressions.

Fully automated system developed based on the approach discussed in section 3 was used to test the sequences obtained from test subjects. Table 2 summarizes the results obtained from the analysis of the sequences. The overall recognition rate is 82.1% which is marginally better than the human coders. The performance decrease can partly be attributed to the inherent difficulty of the classification into a few categories of expressions even for human beings. Analysis of recognition accuracy reveals the similar trends with the results obtained using Cohn-Kanade database. It was observed that the surprise, happiness, sadness, and anger result in less confusion than fear and disgust. Once again, some sequences showing fear and happiness are misclassified, and disgust is often recognized as sadness.

Human subjects express the facial reactions without acting in a natural context. It results in more natural expressions, but also produce more complex deformations of facial muscle, which may include blends of emotions or facial deformations without emotional content. Furthermore, the expressions exhibited by the subjects during emotion elicitation experiment are different (the range of expressions varies from hardly expressed reactions to exaggerated expressions) from those in the training set. The intensity of expression range from no expression to almost exaggerated expression for different subjects. Also it is easy to imagine that the observed emotional reaction does not always correspond to the emotion targeted by the excerpt.

4.3 Analysis of TV broadcast data:

The goal of the second series of experiments was to validate the classification system by testing it with real world data instead of laboratory data. To do that 108 sequences of facial expressions were collected from the TV broadcast. These expressions contain extreme lighting conditions, diverse subjects and expressions. The Table 3 summarizes the results obtained from the analysis based on the level of interest classifier. The overall agreement among three labelers was 89%, in contrast with the 72% overall recognition rate by the fully automatic system.

<table>
<thead>
<tr>
<th></th>
<th>Surprise</th>
<th>Happy</th>
<th>Sad</th>
<th>Fear</th>
<th>Anger</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surprise</td>
<td>66.4</td>
<td>0</td>
<td>11.2</td>
<td>22.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>75.9</td>
<td>9.6</td>
<td>9.6</td>
<td>0</td>
<td>4.9</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>10.0</td>
</tr>
<tr>
<td>Fear</td>
<td>7.7</td>
<td>0</td>
<td>19.1</td>
<td>60.7</td>
<td>0</td>
<td>12.5</td>
</tr>
<tr>
<td>Anger</td>
<td>4.9</td>
<td>0</td>
<td>9.7</td>
<td>14.6</td>
<td>61.1</td>
<td>9.7</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>5</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 3: Similar results shown in Table-1 using TV broadcast data.

The confusion matrix illustrates the most common misclassifications. In general, surprise, fear, anger, and disgust have the lowest recognition rates, as they do not occur naturally alone but associated with other emotions. For example, it appears that surprise is often confused with fear. It is easy to note that even though sadness is the best recognized expression, some confusion occur with disgust as it happen within the Cohn-Kanade database.

This experiment exhibits some limitations of the proposed method with real world data, more challenging in terms of environment conditions. In a natural situation, problems related to out of plane head motion, rotation, widely varying lighting conditions and change of scale may arise. Even though face regions are preprocessed the classifier experiences problems dealing with above mentioned variabilities. One of the interesting case where the approach produces wrong results is when people express an emotion while speaking, which occurs relatively frequently in real life. In that case, the classifier may detect the deformations related to the speech as the prominent motion and misclassify the expression. The facial deformation related to the verbal communication counterbalances the upper motion of the eyebrows. With similar motions in the upper part of the face for sadness and surprise, the motion present in the mouth area can help to discriminate between both. However, in this case it is closely comparable with a sadness expression, which causes misclassification. Finally, even though the failure may come from different sources but the interpretation of the confusion remains subjective and qualitative.

5 Conclusions

This paper, implements and experiments with a two-stage classification approach, that recognize six universal facial expressions proposed by Ekman [1], from previously unseen observations of facial expressions and derives the “level of interest” based on psychological evidences [34]. It was found that the temporal signature derived from the observations by concatenating the output of naïve classifiers at frame level is robust compared to the raw representation of the optical flow using continuous HMMs. Experiments on laboratory data (Cohen-Kanade) shows 90.9% recognition accuracy on 488 video sequences including 97 subjects. A number of experiments, namely, emotion elicitation and analyses of TV broadcast, have been conducted on additional data sets containing variability in terms of lighting conditions, subjects (different age group, gender and ethnicity), and expressions (showing expression while talking). The emotion elicitation experiment revealed the limitations of the classifier in handling spontaneous reactions and also was useful in evaluating the interest level as the ground truth information was gathered while collecting the data. Sequences collected from TV broadcast, expose the model to very different test data with respect to the recording conditions and diversity of subject presented to the classifier. Finally, the estimation of the interest level based on psychological evidence may prove valuable once incorporated in affective computing. While the proposed system performs reliably in many real-world situations, it may be useful to combine with prosodic information form speech to correlate the emotional state with facial expressions to achieve more robustness.
Acknowledgements: Authors gratefully acknowledges the Cohn-Kanade facial expression database provided by CMU.

References: