A METHOD TO INFER EMOTIONS FROM FACIAL ACTION UNITS

Sudha Velusamy, Hariprasad Kannan, Balasubramanian Anand, Anshul Sharma, Bilva Navathe

Frontier Research Group,
Samsung India Software Operations, India.
{sudha.v, h.kannan, balu.anand, anshul, bilva.n}@samsung.com

ABSTRACT

We present a robust method to map detected facial Action Units (AUs) to six basic emotions. Automatic AU recognition is prone to errors due to illumination, tracking failures and occlusions. Hence, traditional rule based methods to map AUs to emotions are very sensitive to false positives and misses among the AUs. In our method, a set of chosen AUs are mapped to the six basic emotions using a learned statistical relationship and a suitable matching technique. Relationships between the AUs and emotions are captured as template strings comprising the most discriminative AUs for each emotion. The template strings are computed using a concept called discriminative power. The Longest Common Subsequence (LCS) distance, an approach for approximate string matching, is applied to calculate the closeness of a test string of AUs with the template strings, and hence infer the underlying emotions. LCS is found to be efficient in handling practical issues like erroneous AU detection and helps to reduce false predictions. The proposed method is tested with various databases like CK+, ISL, FACS, JAFFE, MindReading and many real-world video frames. We compare our performance with rule based techniques, and show clear improvement on both benchmark databases and real-world datasets.

Index Terms— Affect Recognition, Facial Action Coding System, Facial Action Units, String Edit Distance.

1. INTRODUCTION

Machine analysis of human affective behavior has potentially wide variety of applications such as human-computer interaction, health-care, computer assisted learning, anomalous event detection, and interactive computer games. Among various cues that express human emotion, nonverbal information like facial cues play an important role in analyzing human behavior. Facial Action Coding System (FACS) [1] provides a method for objective measurement of facial expression.

Emotion recognition from facial cues based on FACS rules can be classified as: a) single-phase, where emotions are recognized directly; and b) two-phase, where the facial action units (AUs), which are considered as building blocks of facial expressions, are detected first and then the output emotion is inferred from the detected AUs. Latter approach is found to be more practical than the former, as most of the facial expressions can be described using a sub-set of 44 AUs [1]. Detecting AUs prior to emotion makes a recognition system more suited to a culture independent interpretation. Besides, it reduces the amount of independent training data required to model each emotion as there are around 7000 emotions in practical.

Though there are several works present in the area of AU detection [2, 3, 4], robust methods mapping AUs to emotions is still largely unexplored. There are few deterministic rule based techniques [5, 6] which maps computed facial AUs to emotion categories. For example, Valstar et al [7] have formulated mapping rules based on EMFACS (Emotional FACS). They also presented an Artificial Neural Network (ANN) based method which is similar to Piat et al’s [8] work for examining the underlying emotions from facial expressions. Chang et al [9] used partially-observed hidden conditional random fields for facial expression recognition. Based on the various combinations of 15 most frequently occurring AUs, they prepared an extensive set of 100 hidden labels within their graphical model.

The primary challenge of deterministic rules based techniques is their sensitivity to noises due to the uncertainties in tracking and localization of facial features in AU detection. Likewise, fuzzy rule based techniques like ANN require extensive training to accommodate various possible combinations of AUs for predicting the emotions. For example, given a happy face with ground truth AUs, \{AU6, AU7, AU12\}, an automatic AU detector might result in outputs like, i) \{AU12\}, where many AUs are missed; ii) \{AU1, AU5, AU6, AU7, AU12\}, where there are wrongly inserted AUs like AU1 (inner brow raise), AU5 (eye-lid raise); and iii) \{AU6, AU7, AU20\}, where AU12 (lip corner pull) is wrongly replaced by AU20 (lip stretch). Considering such practical issues, we present a method that approaches AU to emotion mapping as a problem of approximate string matching while using learned statistics that quantify the match. The main contribution of this work is a novel method of mapping AUs to emotions using: i) the concept of discriminative power for deriving the template strings of AUs; and ii) approximate string matching approach to the problem of mapping AUs.
to emotions, as it accommodates errors (false positives and misses) in detected AUs. In the following section, we present an emotion recognition system developed for the proof of concept. Section 3 explains our proposed method in detail. Experimental results and analysis are present section 4.

2. THE PROPOSED METHOD

We now present a method to map a set of AUs to one or more final emotions using a learned statistical relationship and a suitable distance measure. For the proof of concept, we developed an emotion recognition system which comprises of: a) an AU detection module based on [2]; and b) the proposed AU to emotion mapping module. Fig. 2 shows the integrated emotion recognition system.

Fig. 1. The System Flow Chart.

For training the AU detector, we use 580 images from the CK+ database [10]. For each training image, the system localizes and aligns the face region according to the pair of detected pupils. The aligned faces are rescaled to the size of 96x96 pixels. Gabor filter bank of size 56 (7 scales, 8 orientations) is applied on each cropped face template to extract the feature vectors of size 96x96x56. AdaBoost as a feature selection method is applied to reduce the feature vector dimension. We use Support Vector Machines (SVM) [11] to model the AUs. Based on our detailed study of FACS, we chose 15 AUs (See Table 1) that are found to be sufficient in representing the six basic emotions (Anger, Fear, Happy, Sad, Surprise, Disgust). 15 SVM classifiers are trained to model each of the 15 selected AUs.

Given a test input image, its feature vector is obtained by the procedure similar to training, and it is input to each of the 15 trained SVMs to predict the AUs. Each SVM classifier outputs the presence or absence of its corresponding AU according to the intensity of the AUs present in the face (test image). The detected facial AUs are then processed by the AU to emotion mapping module to predict the final emotion state. Detailed description on inferring emotions from AUs is given in the following sections.

3. INFERRING EMOTIONS FROM ACTION UNITS

Many of the existing heuristic rule based methods suffer for noisy AU inputs which are unavoidable due to the errors in tracking and localization of the faces. To generalize the solution to robust scenario, we present a principled approach to infer emotions from AUs. The method consist of two parts: (1) Deriving AU-Emotion relationship; and (2) Predicting the emotions.

3.1. Deriving AU-Emotion relationship

The relationship between the AUs and the six basic emotions is obtained by a statistical analysis of a benchmark database (CK+ [10]) labeled for both emotions and AUs. The relations are obtained in the form of a relation matrix which is derived using a concept called discriminative power [4]. The discriminative power is defined as,

\[ H = P(Y_j|X_i) - P(Y_j|\bar{X}_i) \]  \hspace{1cm} (1)

where \( P(Y_j|X_i) \) is the probability of action \( Y_j \), given that the emotion \( X_i \) has occurred, and \( P(Y_j|\bar{X}_i) \) is the probability of action \( Y_j \), given that the emotion \( X_i \) has not occurred. The magnitude of \( H \) quantifies the discriminative power of an AU for an emotion, and the sign depicts whether an AU increases or decreases the probability of mapping to an emotion. The relation matrix is derived by normalizing \( H \) across all the AUs for each of the emotions. Hence, non-linear association weights are obtained for each of the AUs based on their relevance to the emotions calculated using Equ. (1).

Figure 2 shows the relation matrix calculated for the six basic emotions. Here, the positive intensity (white) indicates the high probability or weight associated for an AU belonging to an emotion and negative intensity (black) indicates the high probability for an AU not belonging to an emotion. For example, happy emotion has \( AU_6, AU_7, AU_{12}, AU_{26} \) as positive associations and \( AU_1, AU_2, AU_5, AU_9 \) as negative associations. The experimental results present in section 4 proves that derived relation matrix is efficient in identifying highly relevant facial actions and their association weights for various emotions, and hence infer emotions correctly.

Table 1. Visual Examples of Facial Action Units (AU)
For each emotion, we select the top \( N \) entries of highly discriminative AUs and store them as template strings of AUs for future matching. We use template string length \( N \) of 5 for our experiments. Example templates for the six basic emotions are given in Fig. 3.

![Relation Matrix (Action Unit Vs Emotion)](image)

For each emotion, we select the top \( N \) entries of highly discriminative AUs and store them as template strings of AUs for future matching. We use template string length \( N \) of 5 for our experiments. Example templates for the six basic emotions are given in Fig. 3.

### 3.2. Predicting the emotions

Given a test string of AUs, it is matched against the template strings to find the emotion. We use Longest Common Subsequence (LCS) \([12]\) to measure the similarity between the strings. Given two or more sequences, LCS is a method to find the longest subsequence that is common to all sequences. Here, a subsequence is defined as a sequence in which units appear in the same relative order, but not necessarily contiguous. For example, in a string \( ACTTGCG \), ‘ACT’, ‘ATT’ and ‘ACTTG’ are all subsequences. LCS allows only ‘insertion’ and ‘deletion’, but not (or conditional) ‘replacement’ of units between the strings. This characteristic of LCS is found to be very suitable for our problem of predicting emotions from strings of AUs.

For example, given a test image shown in Fig. 4(a), automatic AU detectors might produce combinations of AUs as given in Fig. 4(b). Here, ‘deletion’ property becomes important to map various AU strings like \{AU12\} or \{AU6, AU12\} or \{AU6, AU12, AU26\} to the emotion happy, as all of them indicates happy though there are certain AUs missing in some of the strings. Similarly, ‘insertion’ plays a role in accommodating erroneously detected AUs like \{AU1, AU4\}

![Example AU detections](image)

while still mapping the emotion in Fig. 4(a) to happy. Also, conditionally allowing ‘replacement’ helps to correct the detection errors involving visually close AUs like AU12, AU20, AU23 (see, Fig. 4(b)), and predicts the emotion with high confidence value. It also avoids wrongly mapping to many emotions when the test string has considerably high errors.

For example, in unconditional replacements, a test string like \{AU4, AU6, AU12, AU17, AU23\} can get mapped to \{AU1, AU4, AU10, AU15, AU17\} (sad) and \{AU4, AU6, AU7, AU9, AU17\} (disgust) with the equal replacement cost of ‘3’. On the other hand, under conditional replacement of visually close AUs like AU12 and AU15, the output emotion will only be (sad). In the following section, we present experimental results proving the efficacy of our approach.

### 4. EXPERIMENTS AND RESULTS

#### 4.1. Databases

For our experiments, we collected a large set of images and videos from standard databases like CK+ \([10]\), ISL \([13]\), FACS \([1]\), JAFFE \([14]\), MindReading \([15]\), and videos captured in our lab. Images/frames which have frontal faces with head rotation not more than \(\pm 10^\circ\) are selected from these databases. We use a portion (580 images) of CK+ database labeled for AU and emotion to learn the statistical relationship between the AUs and emotions. For validating the performance, we use the remaining databases CK+ (images which were not used for learning), ISL, FACS, JAFFE, MindReading, and real-world videos.

#### 4.2. Results

For comparison, we implemented EMFACS based technique present in \([2]\) and Moon et al’s method \([5]\) which are based on deterministic rules for mapping AUs to emotions. The correct detection rate with respect to the ground truth emotion labels is used as a metric for performance measurement.

Table 2 shows the comparison results of AU to emotion mapping techniques. The superior performance of the proposed method can be seen from the detection rate. The proposed mapping technique performs very well for the images from CK+ database which have minimal errors in AU detection (since a similar set of images were used for training
the AU detector). The method also shows good performance for FACS database as the expressions present in the images are performed by FACS trained people and are peak in intensity. Also, we tested the mapping techniques with various challenging databases like ISL with highly contrast images, JAFFE with people of different ethnicity and MindReading with close to tolerance head movements, and report comparably better performance than the rule based techniques. The primary advantage of the proposed method as compared to the rigid rule based techniques is that, the method does not require a large look-up tables to accommodate combinations of AUs that may occur due to errors in AU detection. Also, the proposed method of non-rigid template AUs with weight association helps to infer non-peak emotions and as well handles errors in AU detection.

4.3. Conclusion

The way to infer emotions from facial cues is to first detect AUs and then map them to emotions. In this paper, we presented a novel method to map detected AUs to Emotions. We used a learned statistical relationship between AUs and emotions to build template strings of AUs for six basic emotions. An input test string of detected AUs, which is prone to have false positives and misses, is matched to these strings using Longest Common Subsequence (LCS). The suitability of the method was proven with tests on various facial expression databases like CK+, ISL, JAFFE, MindReading and also real world data. Benchmark comparison with existing techniques, showed comparably better performance. In the immediate future, we hope to extend this work to inferring emotions in videos by considering temporal dynamics.

5. REFERENCES


