A BIOLOGICALLY INSPIRED APPROACH FOR FUSING FACIAL EXPRESSION AND APPEARANCE FOR EMOTION RECOGNITION

Albert Cruz and Bir Bhanu

Center for Research in Intelligent Systems, University of California, Riverside Riverside, CA 92521-0425, USA

ABSTRACT

Facial emotion recognition from video is an exemplar case where both humans and computers underperform. In recent emotion recognition competitions, top approaches were using either geometric relationships that best captured facial dynamics or an accurate registration technique to develop appearance features. These two methods capture two different types of facial information similarly to how the human visual system divides information when perceiving faces. In this paper, we propose a biologically-inspired fusion approach that emulates this process. The efficacy of the approach is tested with the Audio/Visual Emotion Challenge 2011 data set, a non-trivial data set where state-of-the-art approaches perform under chance. The proposed approach increases classification rates by 18.5% on publicly available data.

Index Terms— Representation, Supplemental Information Hypothesis, Cognitive Science, Emotion Recognition, Match-Score Fusion

1. INTRODUCTION

Facial expression recognition has applications in human-computer interaction, consumer electronics, and intelligent tutoring systems. While there has been advances in the past decade, an approach has yet to be seen which performs well on challenge data sets. These data sets, such as the Audio/Visual Emotion Challenge 2011 [1] and FERA 2011 [2] contain continuous footage of spontaneously expressed, natural emotions, which are difficult to classify.

In previous challenges [1, 2], state-of-the-art approaches fall into two groups: The first group uses facial point data, e.g. active shape models (ASM), for alignment or features. These methods are well suited to detecting strong facial expressions. However, ASM is difficult to initialize. The second group uses SIFT Flow [3]. In this method, images are well aligned, and allow discriminative and relevant features to be generated even in cases of extreme pose, where the first method would fail. However, dynamics are lost in the process and the second method does not perform as well as the first for strong facial expressions.

We posit that the difference in performance of the first and second methods are due to the methods capturing different types of facial information. The differences between the two information available to the two methods resembles the division of information in the human visual system when it processes faces. O’Toole et al. [4] groups facial feature information processed by the brain as being either: (1) static information, also known as facial appearance, referring to invariant facial features of the face, such as eyebrows, iris color, etc. or (2) dynamic information, referring to variant facial expression information, as well as gesture and pose, e.g. the facial muscle dynamics associated with smile. In that work, a crossover was theorized where static information and dynamic information are combined under non-optimal conditions with a phenomenon called the Supplemental Information Hypothesis. It suggests that humans represent idiosyncratic gestures, called, dynamic facial signatures, to a specific person. Specifically, the Haxby et al. distributed neural system for faces [5] was amended to include a crossover from the motion-computing oriented middle temporal visual area to the facial appearance oriented fusiform face area. Our approach emulates this process by discriminating static and dynamic facial information, and fusing them to improve emotion recognition rates.

In this paper, we are the first to acknowledge the importance of discriminating static and dynamic information for automatic facial emotion recognition. We propose a novel approach for estimating static information from video of a dynamic, expressive individual. The proposed approach is motivated from a high-level cognitive neuroscience background and from experimental results from previous challenges showing that the two methods performed well under different, exclusive conditions [1, 2].

2. TECHNICAL APPROACH

The proposed system overview is shown in Fig. 1: (1) Face ROI is detected with a boosted cascade of Haar-like features. Dynamic and static information are computed in separate pathways. Dynamic information is quantified with (2) active shape models (ASM); (3) facial points detected with ASM are used for registration, and appearance features are developed from the registered images. Static information is obtained by (4) estimating a static representation of the face and (5) warping each face to minimize dynamics. Appearance features are generated from this representation. (6) The two approaches are fused at the match-score level and (7) emotion labels are
Facial Appearance (Static Information) 
Facial Expression (Dynamic Information) 
Static Reference

Face Estimation ...
dynamic information from Sec. 2.1 and the static information
from Sec. 2.2 are fused to create a more robust classification

\[ \mathbf{μ} \]
mixture distribution: \( \mathbf{μ} \)

physical attributes, and lighting conditions. We assert that this
variables that alter the intensity value such as facial motion,

Distributions in Eq. 1 are assumed to be normally distributed,
\( \mathbf{A} \)
of \( \mathbf{A} \) distribution in Eq. 1 with the highest weight be the distribution

sis. We hypothesize that, for each person in video, there exists
static information can be estimated on a frame-by-frame ba-
are not available. A reference image must be generated so that

Static information features are more difficult to obtain, as ob-
observed frames are expressive and static references of subjects
are not available. A reference image must be generated so that
static information can be estimated on a frame-by-frame basis. We hypothesize that, for each person in video, there exists
some image which is the static reference face of that person, \( \mathbf{A}(\mathbf{x}) \), where \( \mathbf{x} = \{x_1, ..., x_m\} \). However, in the frame \( f(x_i) \), a given pixel \( x_i \) is observed in the presence of other random variables that alter the intensity value such as facial motion, physical attributes, and lighting conditions. We assert that this relationship is additive and model the intensity of \( f(x_i) \) as a mixture distribution:

\[ f(x_i) = \sum_{j=1}^{k} w_j p_j(x_i) \tag{1} \]

where \( f(x_i) \) is the observed pixel intensity of the image at
pixel \( x_i \), \( p_j(x_i) \) is the distribution function of the \( j \)-th term, \( w_j \) is the weight of \( j \) and \( k \) is the number of terms. Let the
distribution in Eq. 1 with the highest weight be the distribution
of \( A(x_i) \):

\[ w_{A} = \max \{ w_0, w_1, ..., w_k \} \tag{2} \]

Distributions in Eq. 1 are assumed to be normally distributed, so \( A(x_i) \) is distributed according to:

\[ p_A(x_i) \sim N(\mu_{A_i}, \sigma_{A_i}) \tag{3} \]

where \( \mu_{A_i} \) and \( \sigma_{A_i} \) are the mean and variance of \( p_A(x_i) \) respectively. For the static reference face, \( E(A(x_i)) \) is used as
the pixel intensity, i.e. take \( A(x_i) \) to be the mean of the the
distribution in Eq. 1 with the highest weight.

After the static reference face is estimated, static information is developed per frame by warping each frame to the
static reference face in such a way as to minimize dynamic information. This is done with SIFT Flow [3], where a given
frame is warped to a the static reference face similarly to optical flow. A given frame is registered to the static reference
face by warping \( f(x) \) to match \( A(x) \) by minimizing the

\[ E(w) = \sum_{x_i \in \mathbf{x}} \min \| s_A(x_i) - s_f(x_i + w_{x_i}) \|_1 + \frac{1}{\sigma^2} \sum_{x_i \in \mathbf{x}} u_{x_i}^2 + v_{x_i}^2 + \sum_{x_i \in \mathbf{x}} \sum_{x_j \in N_{x_i}} \min (\alpha|u_{x_j} - u_{x_i}|) + \min (\alpha|v_{x_j} - v_{x_i}|) \]

where \( x_i \) is a pixel in the image, \( w_{x_i} \) is the motion vector at
pixel \( x_i \) between the query and target where \( w_{x_i} = [u_{x_i}, v_{x_i}] \),
\( s_f \) and \( s_A \) are the dense SIFT descriptors of the given frame
and the static reference face, \( \sigma^2 \) is a normalization constant, \( \alpha \) is a parameter that controls the homogeneity of the motion
and \( N_{x_i} \) is the 4-member neighborhood about \( x_i \). The a query
image which has been warped to match the target image has diminished dynamic information. After the warping process,
ULBP features are developed from the frame and used as the
feature vector for static information.

Results of the proposed static information estimation approach are given in Fig. 2. In pairs 1, 2 and 5, pose has been
corrected so that the face plane is parallel to the image plane.
In pairs 3 and 4, jaw drop and mouth stretching has been

\[ E(w) = \sum_{x_i \in \mathbf{x}} \min \| s_A(x_i) - s_f(x_i + w_{x_i}) \|_1 + \frac{1}{\sigma^2} \sum_{x_i \in \mathbf{x}} u_{x_i}^2 + v_{x_i}^2 + \sum_{x_i \in \mathbf{x}} \sum_{x_j \in N_{x_i}} \min (\alpha|u_{x_j} - u_{x_i}|) + \min (\alpha|v_{x_j} - v_{x_i}|) \]

2.1. Dynamic Information Features

To quantify facial dynamics, ASM features are developed for
each frame. ASM feature points are registered to inner eye
points using a similarity transform. Inner eye points are fur-
ther used to align each frame, and uniform local binary pattern
(ULBP) features are generated. The dynamic information
feature vector includes ULBP feature points, ASM features,
and head tilt in terms of pitch, yaw and roll, calculated from
ASM.

2.2. Static Information Features

Static information features are more difficult to obtain, as ob-
served frames are expressive and static references of subjects
are not available. A reference image must be generated so that
static information can be estimated on a frame-by-frame ba-

Fig. 1. Proposed system overview.
scheme. Additionally, because an emotional state is consistent with its temporal neighbors, in that a subject is not likely to make high frequency emotional state changes, there should be temporal averaging. The posterior probabilities of both dynamic and static information, over a time interval, are fused with combination-based match-score fusion, where the posterior probabilities from different matchers combined to obtain a final, single score as the a posteriori probability. Let $X_t$ be the feature vector of modality $i$ at time $t$. Let $\tilde{y}$ be the assigned label from one of the classes $\{y_1, ..., y_o\}$. Let $p(y_j|X_{it})$ be the output of the matcher of modality $i$ at time $t$. The classification rule for match-score fusion is:

$$\tilde{y} = \text{argmax}_i K(p(y_j|X_{11}), ..., p(y_j|X_{1T}))$$

(5)

where $K(\cdot)$ is the rule for aggregation. We use the sum rule, which is defined as follows:

$$K(p(y_j|X_{11}), ..., p(y_j|X_{1T})) = \frac{1}{Z} \sum_{i=1}^{l} \sum_{\tau=t-T}^{t} p(y_j|X_{\tau})$$

(6)

where $Z$ is a normalization term to constrain the probability to $[0, 1]$ and $T$ is an experimental parameter.

3. RESULTS AND DISCUSSION

Features extracted are ULBP features of 8 neighbors and a radius of 1, with the face partitioned to $10 \times 10$ regions. Feature dimensionality for static information is 5900. Feature dimensionality for dynamic information includes ULBP features, 113 facial points, pitch, yaw and roll resulting in a feature dimensionality of 6129. In Eq. 1, $k$ is selected s.t. the Bayesian information criterion is minimized, to a maximum of $k = 5$. In Eq. 6, $T = 24$, averaging the matchers over a period of 1s. This was selected experimentally.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat.+Dyn.</td>
<td>65.4</td>
<td>76.7</td>
<td>69.3</td>
<td>69.7</td>
<td>70.3</td>
</tr>
<tr>
<td>Dynamic</td>
<td>62.6</td>
<td>69.2</td>
<td>67.7</td>
<td>66.9</td>
<td>66.7</td>
</tr>
<tr>
<td>Static</td>
<td>53.7</td>
<td>63.7</td>
<td>68.9</td>
<td>58.2</td>
<td>61.1</td>
</tr>
<tr>
<td>Schuller [1]</td>
<td>60.2</td>
<td>58.3</td>
<td>56.0</td>
<td>63.6</td>
<td>59.3</td>
</tr>
<tr>
<td>Dahmane [6]</td>
<td>54.9</td>
<td>54.8</td>
<td>53.2</td>
<td>56.6</td>
<td>54.1</td>
</tr>
<tr>
<td>Glodek [7]</td>
<td>56.9</td>
<td>47.5</td>
<td>47.3</td>
<td>55.6</td>
<td>51.8</td>
</tr>
</tbody>
</table>

Table 1. Testing results on the AVEC2011 development data set for activation-arousal (Act.), unpredictability (Unpred.), potency-control (Pot.) and evaluation-pleasantness (Eval.).

We evaluate results of the proposed method with the AVEC2011 [1] development data set which consists of 32 interviews of 8 different individuals, resulting in roughly half a million frames. Unlike previous data sets, the data is naturally expressed. Subjects are being engaged by an interviewer, so expressions are spontaneous, continuous, and natural. An example video is given on YouTube [8]. (2) The subjects are free to change pose, and use hand gestures, and (3) emotion is quantized in terms of: evaluation-pleasantness, activation-arousal, potency-control and unpredictability. An emotion is binary valued along these four dimensions. Evaluation-pleasantness, describes positivity or negativity of the subjects feelings or situation, e.g. happiness versus sadness. Potency-control describes a subjects feeling of control of the situation, e.g. power versus submission. Activation-arousal describes a subjects interest in the situation, e.g. eagerness versus anxiety. Unpredictability describes the subject’s certainty of the situation, e.g. familiarity versus apprehension. A subject can express multiple emotions at once, requiring four binary classifiers.

Results are generated using a six-fold cross validation with a 33/66 testing/training split. Classification rates are given in Tab. 1. “Dynamic” refers to a fusion scheme using matchers with dynamic information; “Static”, refers to a fusion scheme using matchers with static information; “Stat.+Dyn.”, refers to a fusion scheme using matchers of both static and dynamic information types. The ROC curves of the proposed approach with the sum rule are given in Fig. 3.

Suitability of pairing static and dynamic information is compared with $Q$-statistics. This metric reflects how well two different classifiers would perform if fused. It is a correlation measure between the two classifiers. Two classifiers that would benefit from being fused should complement by classifying samples that the other did not. Given two classifiers $i$ and $j$, the $Q$-statistic for a fusion scheme is:

$$Q_{ij} = \frac{n_{00}n_{11} - n_{10}n_{01}}{n_{10}n_{11} + n_{01}n_{00}}$$

(7)

where, in this paper, $i$ is static and $j$ is dynamic. $n_{00}$ is the rate of errors in both $i$ and $j$; $n_{11}$, correct classification in both $i$ and $j$; $n_{10}$ and $n_{01}$ rates for when classification occurs
Fig. 3. ROC curves for the AVEC2011 development data set on: (I) activation-arousal, (II) unpredictability, (III) power-control and (IV) evaluation-pleasantness.

### Table 2. Q-statistics of the Proposed Approach

<table>
<thead>
<tr>
<th></th>
<th>n_{10}</th>
<th>n_{11}</th>
<th>n_{10}</th>
<th>n_{11}</th>
<th>Q_{ij}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act.</td>
<td>.087</td>
<td>.233</td>
<td>.151</td>
<td>.529</td>
<td>.129</td>
</tr>
<tr>
<td>Unpred.</td>
<td>.051</td>
<td>.200</td>
<td>.152</td>
<td>.597</td>
<td>.005</td>
</tr>
<tr>
<td>Pot.</td>
<td>.053</td>
<td>.158</td>
<td>.145</td>
<td>.643</td>
<td>.202</td>
</tr>
<tr>
<td>Eval.</td>
<td>.059</td>
<td>.181</td>
<td>.143</td>
<td>.617</td>
<td>.173</td>
</tr>
</tbody>
</table>

with either static or dynamic matchers only. If two classifiers perform well, they minimize Eq. 7. Q-statistics are given in Tab. 2. The pairing of dynamic and static information is most minimal for unpredictability, foreshadowing that the most performance gain is to be had in this class, which is confirmed by the ROC plot of unpredictability in Fig. 3.

### 4. CONCLUSION

In this paper we proposed an approach that fused facial expression and facial appearance—which was estimated using static reference face and SIFT Flow. It was motivated experimentally from related work, and from cognitive neuroscience’s Supplemental Information Hypothesis. Efficacy of the approach was demonstrated on the non-trivial AVEC2011 data set, where the proposed approach improved classification results by 18.5% on the development set.

**Acknowledgements.** Support for this work was provided for in part by NSF grants 0727129 and NSF IGERT: Video Bioinformatics Grant DGE 0903667. The contents and information do not reflect the position or policy of the U.S. Government.

### 5. REFERENCES


