Automatic 3D facial expression recognition based on a Bayesian Belief Net and a Statistical Facial Feature Model

Xi Zhao, Di Huang, Emmanuel Dellandrea and Liming Chen
Université de Lyon, CNRS, Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France
\{xi.zhao, di.huang, emmanuel.dellandrea, liming.chen\}@ec-lyon.fr

Abstract—Automatic facial expression recognition on 3D face data is still a challenging problem. In this paper we propose a novel approach to perform expression recognition automatically and flexibly by combining a Bayesian Belief Net (BBN) and Statistical facial feature models (SFAM). A novel BBN is designed for the specific problem with our proposed parameter computing method. By learning global variations in face landmark configuration (morphology) and local ones in terms of texture and shape around landmarks, morphable Statistic Facial feAture Model (SFAM) allows not only to perform an automatic landmarking but also to compute the belief to feed the BBN. Tested on the public 3D face expression database BU-3DFE, our automatic approach allows to recognize expressions successfully, reaching an average recognition rate over 82%.

Keywords—3D facial expression recognition, Bayesian Belief Net, Statistical Face Model

I. INTRODUCTION

Facial Expression Recognition (FER) is an important part of affective computing aiming at integrating emotion into artificial intelligence systems, and has many possible applications, particularly concerning intelligent human-machine interactions.

The number of works dealing with 3D FER has recently significantly increased since 3D faces carry richer information allowing to overcome 2D FER difficulties related to pose and illumination variations. As an emergent domain, many challenges remain to be addressed. Indeed, there is no consensus on how to exploit the third dimension, and most of FER approaches require human intervention (manual landmarking for example) preventing any fully automatic recognition which is however necessary for realistic applications.

FER approaches can globally be divided into two categories: feature-based and model-based FER. Approaches based on features rely on the extraction of facial features which are further fed into classifiers. Among the features characterizing different expressions, one can mention distances between landmarks [1], face curvature [2], texture [3], LBP [4], HOG [4] and SIFT [4]. Once these features have been extracted, expression recognition is performed by using a classifier, such as SVM [1], Adaboost [5], K-NN [5], LDA [2], [5], [4], Modified PCA [3] or Neural Networks [6]. On the other hand, model-based FER approaches [7], [8] rely on the deformation of a 3D morphable face model to fit the test face. Fitted model parameters are further used to perform the FER.

In most of these studies, the six universal expressions are considered: anger, sadness, fear, disgust, surprise, happiness. Global recognition rates vary for the different studies from 80% to 95% [3][7][1][5][2][6], with different constraints. For example, some methods require manual landmarking [1][5][2][6], while others rely on automatic landmarking but requiring a dense correspondence which leads to a high computation cost [7].

Our goal is to propose an efficient approach that is flexible and applicable to real-use cases. Thus, manual landmarking should not be necessary and the FER system should be easily extended to handle more expression classes as well as more features to enrich the face knowledge for better performance. To do this, we propose an automatic approach making use of a BBN combined with the SFAM. Indeed, thanks to the proposed structure of this graphical model, adding new features and new classes for the expression recognition problem is naturally handled. Moreover, automatic landmarking is achieved by SFAM. It is a morphable partial face model which learns the global variations in landmark configuration (morphology) and local ones in terms of texture and shape around landmarks. This statistical model is further used to compute the belief to feed the BBN. Such an approach can be seen as a hybrid approach combining both two types of FER approaches mentioned previously.

The main contribution of this paper consists of 3 aspects: we first apply the Bayesian Belief Network for 3D facial expression recognition in 3D environment, which has a novel structure that allows to add new feature and new expression flexibly; we proposed to a novel method to compute parameters in BBN for inference belief of expression states based on the SFAM; combined BBN with our SFAM, this approach is capable of recognizing facial expression in 3D automatically and efficiently.

The rest of the paper is organized as follows. The BBN for FER is presented in section II. The models for landmarking and feature extraction are described in section III. Experiments are given in section IV followed by the conclusion in section V.
II. BAYESIAN BELIEF NET

A. Background of BBN

A BBN [9] is a probabilistic graphical model with the topology of a directed acyclic graph. Nodes in the net represent a set of random variables and directed edges represent their conditional independencies.

In a given BBN, the 'belief' of variables \( X = (x_1, x_2, ..., x_n) \) on a node \( X \) describes the probability distribution of the states of \( X \) conditioned by the known evidences \( e \) (observations) on the neighboring nodes of \( X \). We divide these nodes into parents (nodes connecting directly to \( X \) via an edge) and children (nodes connecting directly from \( X \) via an edge) and compute the belief as in Eq. 1, where \( e^p \) is the evidence on all parents and \( e^c \) is the evidence on all children.

\[
P(X | e) \propto P(e^c | X) P(X | e^p)
\]

B. BBN for Expression Recognition

The BBN we propose is structured as shown in fig. 1. The node \( X \) represents the facial expression variable and can have 6 states, corresponding to the 6 universal expressions: anger, disgust, fear, happiness, sadness, surprise. The node \( S \), is \( X \)'s parent, representing human subjects that we explore. It has as many states as the number of the subjects. \( X \)'s children \( F_1, F_2, ... , F_{NF} \) represent the different facial features that are considered.

Since there is only one parent for the node \( X \), the parent factor in Eq. 1 can be expressed as \( P(X | e^p) = \sum_{i=1}^{N_s} P(X | p_{S1}^i) P(p_{S1}^i | e^p) \), where \( N_s \) is the number of subjects. \( P(p_{S1}^i | e^p) \) is the probability of the \( i \)th subject given the evidence, and \( P(X | p_{S1}^i) \) is the probability distribution of expressions given the subject. Since all tested subjects perform the same number of expressions, \( P(X | e^p) \) follows a uniform distribution. Moreover, the factor related to children can be rewritten as \( P(e^c | X) = \prod_{i=1}^{N_c} P(e^c_i | X) \), where \( e^c_i \) is the evidence or observation of the \( i \)th child node, \( N_c \) is the number of children, \( P(e^c_i | X) \) is the conditional probability of evidence knowing the state \( X \). Therefore, Eq. 1 can be rewritten as follows:

\[
p(X | e^K) \propto \prod_{l=1}^{N_c} P(e^c_l | X)
\]

\( e^K \) refers to observations from a given face \( \kappa \). Thus, the belief for each expression state is computed from \( e^K \) and the state having the highest belief is considered as the most probable expression of the face \( \kappa \). The computation of \( P(e^c_i | X) \) will be detailed in section III-D.

III. STATISTICAL FACIAL FEATURE MODEL

A. Model building

In order to efficiently learn variations on global morphology, local texture and local shape among training faces, a preprocessing stage is first performed to exclude other variations introduced by global factors like head pose or face scale. Local grids are then used to remesh local regions centered at 19 landmarks placed around eyebrows, eyes, nose and mouth. Intensity and range data are finally extracted from these grids. This process ensures that the same number of points are sampled from all training faces and that they are point-to-point corresponded. SFAM is then learnt by applying Principle Component Analysis (PCA) respectively on the three types of features from training faces, preserving 95% of major components for each one. The resulting model is given in Eq.3

\[
s = \bar{s} + P_s b_s, \quad g = \bar{g} + P_g b_g, \quad z = \bar{z} + P_z b_z
\]

where \( \bar{s}, \bar{g}, \bar{z} \) are respectively the mean morphology, mean intensity and mean range value while \( P_s, P_g, P_z \) are their learnt variation components respectively obtained from PCA. \( b_s, b_g, b_z \) are the corresponding sets of controlling parameters.

Partial face instances, corresponding to local face regions with texture and shape configured by its morphology, can be estimated and synthesized by a linear combination of these components for a given face. The model building is similar as the one we proposed in [10].

B. Automatic landmarking

Automatic landmarking on new 3D faces can be considered as the fitting process of the SFAM. Two response meshes are first computed, the first one describing the similarity between local texture and a template generated from SFAM, and the second one describing the similarity between local shape and its corresponding template. Then, optimal landmark locations, depending on the morphology model parameters, are obtained by optimizing the sum of similarities on both meshes. Fig. 2 shows the flowchart for the automatic landmarking method using SFAM. The detailed fitting algorithm for neutral faces is given in [10].
The feature instances \( F_{i\kappa} \) are extracted by concatenating \( l\kappa \) on local grids of SFAM. We choose shape index because it has been proven to be an efficient feature to describe local curvature information and is independent of the coordinate system [12].

Moreover, we have considered Multi-Scale LBP [13] which is an improved representation of facial texture as compared to the standard LBP [14]. It describes local texture structure more comprehensive and meanwhile has computational simplicity as well as tolerance to monotonic lighting and skin color variations. In our case, we use five different LBP operators \( (\text{LBP}_{(16,1)}, \text{LBP}_{(16,2)}, \ldots, \text{LBP}_{(16,5)}) \) to achieve multi-scale face description.

To summarize, we have computed 9 features, \( D, R, T, SI, \text{LBP}_{(16,1)}, \text{LBP}_{(16,2)}, \ldots, \text{LBP}_{(16,5)} \) as children nodes in our BBN.

D. Belief Computation for BBN

To compute the belief for each expression, we need to compute \( P(e_{\kappa} | X) \) for each feature. Therefore, given a training set for each of the features \( F_t \), we divide it into \( Nc \) (number of expression) subsets. For each subset \( I_{\kappa} \), PCA is applied similar to SFAM: \( F_{i\kappa} = F_t + \frac{1}{\sigma} \).

The feature instances \( F_{i\kappa} \) can be generated from the above equation using feature \( F_{i\kappa} \) to estimate the best parameter \( b_{i\kappa} \) [15]. We set a boundary \((\pm 0.5\sigma)\) for the parameter to increase the separability among instances.

The probability \( P(e_{\kappa} | X) \) can be considered as the probability of matching the feature \( F_{i\kappa} \) to its instances \( F_{i\kappa} \), knowing the expression state \( X \), which follows a Gibbs distribution \( e^{A_t Q_t} \). \( Q_t \) is the match quality, computed as the correlation response between evidence \( F_{i\kappa} \) and its instance \( F_{i\kappa} \). \( A_t \) is a normalizing constant. Inserting the Gibbs distribution into Eq. 2 and taking logarithm gives:

\[
\log p(X|e_{\kappa}) = \log \left( \prod_{t=1}^{Nc} P(e_{\kappa} | X) \right) + c = \sum_{t=1}^{Nc} A_t Q_t + c \quad (4)
\]

Finally, the belief for each state of the node \( X \) is computed as in Eq.4. As the purpose is to discover the highest belief among all expression states, the constant \( c \) can be omitted.

IV. EXPERIMENTAL RESULTS

We tested our approach on the BU-3DFE database [16] using both manual landmarks and automatic landmarks. For each test, we used the data of 60 subjects with scans of two high-intensity expressions for each of the six universal facial expressions to make our results comparable with other works in Table II. In order to train SFAM for automatic landmarking, 143 face scans from another 11 subjects are used. These subjects do not belong to any of the training or test sets for FER. Both FER tests followed a 10-fold person-independent cross-validation process. In each round, 90% subjects (54) are used for training and 10% subjects (6) are used for testing. The tests are repeated for 10 times so that every subject has been tests once. During the experiment, subjects in the training set always differ from those in the testing set.

<p>| Table I: Confusion Matrix of the Person-Independent Expression Recognition |
|------------------|----------------|----------------|----------------|----------------|----------------|----------------|</p>
<table>
<thead>
<tr>
<th>Input ( \backslash ) Output</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>79.2 %</td>
<td>2.5 %</td>
<td>3.3 %</td>
<td>0.0 %</td>
<td>15.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>Disgust</td>
<td>71.7 %</td>
<td>5.0 %</td>
<td>3.3 %</td>
<td>0.0 %</td>
<td>20.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>Fear</td>
<td>3.3 %</td>
<td>87.6 %</td>
<td>5.8 %</td>
<td>0.8 %</td>
<td>2.5 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>Happiness</td>
<td>5.0 %</td>
<td>85.0 %</td>
<td>8.3 %</td>
<td>1.7 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>Sadness</td>
<td>1.7 %</td>
<td>85.5 %</td>
<td>79.2 %</td>
<td>5.8 %</td>
<td>8.7 %</td>
<td>0.8 %</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.0 %</td>
<td>0.0 %</td>
<td>6.7 %</td>
<td>93.3 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>4.2 %</td>
<td>0.8 %</td>
<td>4.2</td>
<td>0.0 %</td>
<td>90.8 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>13.3 %</td>
<td>0.0 %</td>
<td>5.0 %</td>
<td>0.0 %</td>
<td>81.7 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>0.0 %</td>
<td>1.7 %</td>
<td>5.0 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
<td>93.3 %</td>
<td></td>
</tr>
<tr>
<td>0.8 %</td>
<td>1.7 %</td>
<td>3.7 %</td>
<td>0.0 %</td>
<td>0.0 %</td>
<td>93.8 %</td>
<td></td>
</tr>
</tbody>
</table>
less than 3mm in average for locating all 19 landmarks. These results not only show the efficiency of the proposed landmarking, but also that our recognition framework is robust to small displacements of landmarks. Most of the expressions are indeed identified with high accuracy in both tests, while anger and fear have comparatively lower recognition rates. Anger is classified more likely into sadness because their confusion, even for humans, is much larger than for other expressions. The ease of fear is different. Indeed, the motions of this expression are moderate compared to happiness or surprise for example, and thus much more difficult to discriminate.

From Table II, which presents a comparison with typical results of the literature, it is found that the proposed automatic approach leads to a comparable accuracy for 3D facial expression recognition with the approaches requiring a number of manual landmarks. Moreover, as compared to automatic approaches, such as the one of [7] which relies on dense point correspondence, our approach with SFAM is likely less time consuming since a simple local remeshing process around few landmarks (19) is performed.

V. CONCLUSION

We have presented in this paper a new automatic 3D facial expression recognition approach based on a BBN combined with a morphable SFAM. This statistical model which learns the global variations in landmark configuration (morphology) and local ones in terms of texture and shape around landmarks, allows not only an automatic landmarking but also to compute the belief to feed the BBN. Our experiments have brought to the fore the efficiency of our approach for recognizing expression since recognition rates of 87.2% and 82.3% have been reached respectively with a manual landmarking and with the automatic landmarking by SFAM. Moreover, the structure we proposed for the BBN allows an interesting flexibility since knowledge carried by new features can be easily integrated (by adding new children nodes of the X node), as well as new expressions (by adding new states in the X node) to be recognized. We will take benefit of this in our future work since we envisage to explore other type of features to recognize more realistic (non-prototypical) emotions.

ACKNOWLEDGMENT

This work has been partially supported by the French ANR Omnia project under the grant ANR-07-MDCO-009-02.

REFERENCES


