HUMAN VISION INSPIRED FRAMEWORK FOR FACIAL EXPRESSIONS RECOGNITION

Rizwan Ahmed Khan[1,2], *Alexandre Meyer*[1,2], *Hubert Konik*[1,3], *Saida Bouakaz*[1,2]

[1] Université de Lyon, CNRS
[2] Université Lyon 1, LIRIS, UMR5205, F-69622, France
[3] Université Jean Monnet, Laboratoire Hubert Curien, UMR5516, 42000 Saint-Etienne, France
{Rizwan-Ahmed.Khan, Alexandre.Meyer, Saida.Bouakaz}@liris.cnrs.fr, Hubert.Konik@univ-st-etienne.fr

ABSTRACT

We present a novel human vision inspired framework that can recognize facial expressions very efficiently and accurately. We propose to computationally process small, salient region of the face to extract features as it happens in human vision. To determine which facial region(s) is perceptually salient for a particular expression, we conducted a psycho-visual experimental study with an eye-tracker. A novel feature space conducive for recognition task is proposed, which is created by extracting Pyramid Histogram of Orientation Gradients features only from the salient facial regions. By processing only salient regions, proposed framework achieved two goals: (a) reduction in computational time for feature extraction (b) reduction in feature vector dimensionality. The proposed framework achieved automatic expression recognition accuracy of 95.3% on extended Cohn-Kanade (CK+) facial expression database for six universal facial expressions.

Index Terms— facial expression recognition, human vision, eye-tracker, pyramid histogram of oriented gradients

1. INTRODUCTION

Humans are blessed with the amazing ability to decode facial expressions across different cultures, in diverse conditions and in a very short time. Human visual system (HVS) has limited neural resources but still it can analyze complex scenes in real-time. As an explanation for such performance, it has been proposed that only some visual inputs are selected by considering "salient regions" [1], where "salient" means most noticeable or most important.

In this paper, we propose very efficient and simple framework for automatic facial expression recognition (FER) based on HVS. We propose a new feature space which is created by computationally processing salient facial regions with Pyramid Histogram of Orientation Gradients (PHOG) [2] operator. To determine which facial region(s) is the most important or salient according to HVS, we conducted a psycho-visual experiment using an eye-tracker. We have considered six universal facial expressions for psycho-visual experimental study as these expressions are proved to be consistent across cultures [3]. These six expressions are anger, disgust, fear, happiness, sadness and surprise.

There are two main approaches to extract facial features: appearance-based methods or geometric feature-based methods. One of the widely studied method to extract appearance information is based on Gabor wavelets [4, 5]. Another promising approach to extract appearance information is by using Haar-like features [6]. Recently texture descriptor "Local Phase Quantization" [7] is also studied to extract appearance-based facial features. For geometric feature-based methods [8, 9], shapes and locations of facial components are extracted. Research has been done with success in recent times to combine features extracted using appearance-based and geometric feature-based methods [10].

We have found one shortcoming in the reviewed methods for facial expression recognition that none of them try to mimic HVS in recognizing them. Rather all of the methods, spend computational time on whole face image or divides the facial image based on some mathematical or geometrical heuristic for features extraction. We argue that the task of expression analysis and recognition could be done in more conducive manner, if only some regions are selected for further processing (i.e. salient regions) as it happens in human visual system. Thus, our contribution in this study is twofold:

a. Through psycho-visual experiment we determined which facial region(s) is salient for a particular expression.

b. We show that very high facial expression recognition accuracy is achievable by using proposed framework.

The next section provides the details related to psycho-visual experiment. Results obtained from the psycho-visual experiment are presented in the Section 3. Section 4 presents the proposed framework for FER. Experimental results of the proposed approach for the expression recognition on the classical databases are presented in the Section 5. This is followed by the conclusion.

2. PSYCHO-VISUAL EXPERIMENT

The aim of our experiment was to record the eye movement data of human observers in free viewing conditions. The data

were analyzed in order to find which components of face are salient for specific displayed expression.

2.1. Participants, apparatus and stimuli

Eye movements of fifteen human observers were recorded using video based eye-tracker (EyelinkII system, SR Research), as the subjects watched the collection of 54 videos selected from the extended Cohn-Kanade (CK+) database [11], showing one of the six universal facial expressions [3]. Observers include both male and female aging from 20 to 45 years with normal or corrected to normal vision. All the observers were naïve to the purpose of an experiment. CK+ database contains 593 sequences across 123 subjects. Each video showed a neutral face at the beginning and then gradually developed into one of the six facial expression.

2.2. Eye movement recording

Eye position was tracked at 500 Hz with an average noise less than 0.01°. Head mounted eye-tracker allows flexibility to perform the experiment in free viewing conditions as the system is designed to compensate for small head movements.

3. PSYCHO-VISUAL EXPERIMENT RESULTS

In order to statistically quantify which region is perceptually more attractive for specific expression, we have calculated the average percentage of trial time observers have fixated their gazes at specific region(s) in a particular time period. As the stimuli used for the experiment is dynamic i.e. video sequences, it would have been incorrect to average all the fixations recorded during trial time (run length of the video) for the data analysis as this could lead to biased analysis of the data. To meaningfully observe and analyze the gaze trend across one video sequence we have divided each video sequence in three mutually exclusive time periods. The first time period correspond to initial frames of the video sequence i.e. neutral face. The last time period encapsulates the frames where the expression is shown with full intensity (apex frames). The second time period is a encapsulation of the frames which has a transition of facial expression i.e. transition from neutral face to the beginning of the desired expression (i.e neutral to the onset of the expression). Then the fixations recorded for a particular time period are averaged across fifteen observers. For drawing the conclusions we considered second and third time periods as they have the most significant information in terms of specific displayed expression. Conclusions drawn are summarized in Table 1. Refer [12] for the detailed explanation of the psycho-visual experimental study.

Expression	Salient facial region(s)
Happiness	Mouth region.
Surprise	Mouth region.
Sadness	Mouth and eye regions. Biased towards mouth region.
Disgust	Nose, mouth and eye regions. Wrinkles on the nose region gets little more attention than the other two regions.
Fear	Mouth and eye regions.
Anger	Mouth, eye and nose regions.

Table 1. Summary of the facial regions that emerged as salient for six universal expressions

4. EXPRESSION RECOGNITION FRAMEWORK

Feature selection along with the region(s) from where these features are going to be extracted is one of the most important step to successfully recognize expressions. As the proposed framework draws its inspiration from the human visual system, it processes only perceptual salient facial region(s) for the feature extraction. The proposed framework creates a novel feature space by extracting Pyramid Histogram of Orientation Gradients (PHOG) [2] features from the perceptually salient facial regions. PHOG features are selected as they have proven to be highly discriminative for FER task [13, 14]. Schematic overview of the proposed framework is illustrated in Figure 1. Steps of the proposed framework are as follows:

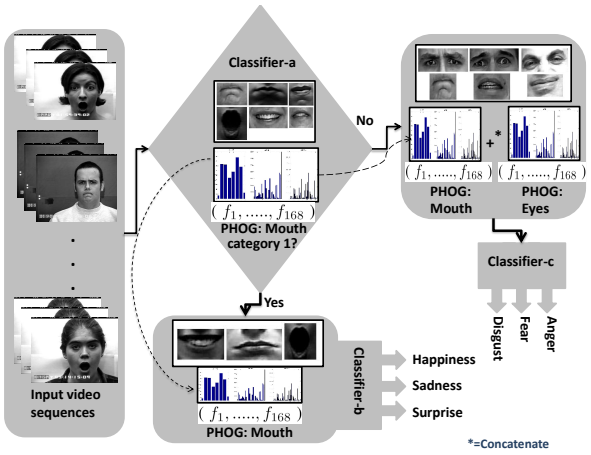

Fig. 1. Schematic overview of the proposed framework

Step 1: The framework initializes with the localization of the mouth region from the input sequence. Then, the PHOG features are extracted from the localized mouth region. The classification ("Classifier-a" in the Figure 1) is carried out on the basis of extracted features in order to make two groups of facial expressions. First group comprises of those expressions that has one perceptual salient region i.e. happiness, sadness and surprise while the second group is composed of those ex-

pressions that have two or more perceptual salient regions i.e. anger, fear and disgust. Purpose of making two groups of expressions is to reduce feature extraction computational time.

Step 2: If the sequence is classified in the first group, then it is classified either as happiness, sadness or surprise by the "Classifier-b". Classification is carried out on the already extracted PHOG features from the salient mouth region.

Step 3: If the input sequence is classified in the second group, then the framework extracts PHOG features from the eyes region and concatenates them with the already extracted PHOG features from the mouth region. Then, the concatenated feature vector is fed to the classifier ("Classifier-c") for the final classification of the sequence.

4.1. Feature extraction using PHOG

PHOG [2] is a spatial shape descriptor. It first extracts Edge contours of the given stimuli using the Canny edge detector. Then, the image is divided into finer spatial grids by iteratively doubling the number of divisions in each dimension. The grid at level l has 2^l cells along each dimension. Afterwards, a histogram of orientation gradients (HOG) are calculated using 3 x 3 Sobel mask and the contribution of each edge is weighted according to its magnitude. Within each cell, histogram is quantized into N bins. Each bin represents the accumulation of number of edge orientations within a certain angular range. To obtain the final PHOG descriptor, histograms of gradients (HOG) at the same levels are concatenated. The final PHOG descriptor is a concatenation of HOG at different pyramid levels. Generally, the dimensionality of the PHOG descriptor can be calculated by: $N \sum_l 4^l$. In our experiment we obtained 168 dimensional feature vector ($f_1, ..., f_{168}$) from one facial region, as we created two pyramid levels with 8 bins with the range of [0-360].

5. EXPRESSION RECOGNITION EXPERIMENT

To test the effectiveness of the proposed framework we conducted the expression recognition experiment on the CK+ database [11]. The performance of the framework was evaluated using four classifiers i.e. "Support vector machine (SVM)" with χ^2 kernel and γ=1, "C4.5 Decision Tree" with reduced-error pruning, "Random Forest" of 10 trees and "2 Nearest Neighbor (2NN)" based on Euclidean distance. The parameters of the classifiers were determined empirically.

For the experiment we used all the 309 sequences from the CK+ database which have FACS coded expression label [15]. The experiment was carried out on the frames which covers the status of onset to apex of the expression, as done by Yang et al. [6]. Region of interest was obtained automatically by using Viola-Jones object detection algorithm [16] and processed to obtain PHOG feature vector. The proposed framework achieved average recognition rate of 95.3%, 95.1%, 96.5% and 96.7% for SVM, C4.5 decision tree, random forest and 2NN respectively. These values were calculated using *10*-fold cross validation.

	Sa	Ha	Su	Fe	An	Di
Sa	**95.5**	0	0.5	0	4.0	0
Ha	0	**95.1**	0	4.1	0	0.8
Su	3.4	0	**96.6**	0	0	0
Fe	0	3.2	0	**94.6**	2.2	0
An	4.8	0	0	0	**95.2**	0
Di	0.8	0.9	0	0	3.4	**94.9**

Table 2. Confusion Matrix: SVM

For comparison and reporting results, we have used the classification results obtained by the SVM as it is the most cited method for classification in the literature. Table 2 shows the confusion matrix for SVM. In the presented table expression of Happiness is referred by "Ha", Sadness by "Sa", Surprise by "Su", Fear by "Fe", Anger by "An" and Disgust by "Di". Diagonal and off-diagonal entries of confusion matrix shows the percentages of correctly classified and misclassified samples respectively.

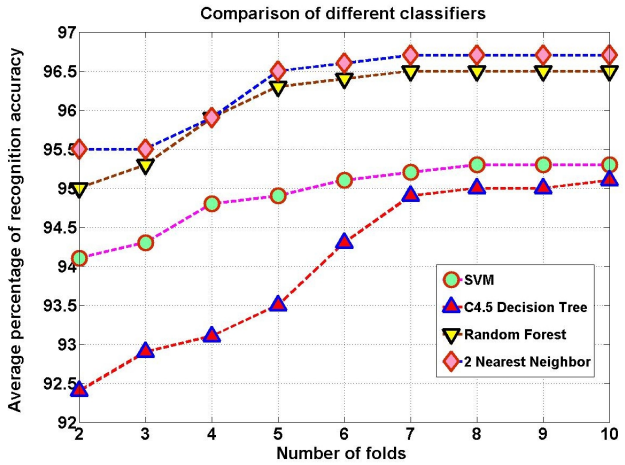

Fig. 2. Evolution of the achieved average recognition accuracy for the six universal facial expressions with the increasing number of folds for the *k*-fold cross validation technique.

Figure 2 shows the influence of the size of the training set on the performance of the four classifiers used in the experiment. For all the classifiers we have computed the average recognition accuracy using different number of folds (k's) for the k-fold cross validation technique and plotted them in the Figure 2. It can be observed that C4.5 decision tree classifier was influenced the most with less training data while 2NN classifier achieved highest recognition rate among the four classifiers with relatively small training set (i.e. 2-folds). This indicates how well our novel feature space was clustered.

Table 3 shows the comparison of the achieved average recognition rate of the proposed framework with the state-

	Sequence Num	Class Num	Performance Measure	Recog. Rate (%)
[4]	313	7	leave-one-out	93.3
[17]	374	6	ten-fold	96.26
[10]	374	6	five-fold	94.5
[5]	375	6	-	93.8
[6]a	352	6	66% split	92.3
[6]b	352	6	66% split	80
Ours	**309**	**6**	**ten-fold**	**95.3**

Table 3. Comparison with the state-of-the-art methods

of-the-art methods[4, 17, 10, 5, 6] using the same database (i.e Cohn-Kanade database). Results from [6] are presented for the two configurations. "[6]a" shows the reported result when the method was evaluated for the last three frames (apex frames) from the sequence while "[6]b" presents the reported result for the frames which encompasses the status from onset to apex of the expression. It can be observed from the Table 3 that the proposed framework is comparable to any other state-of-the-art method in terms of expression recognition accuracy. The method discussed in "[6]b" is directly comparable to our method, as we also employed the same approach. In this configuration, our framework is better in terms of average recognition accuracy.

6. CONCLUSION

In this paper we presented a novel framework for automatic and reliable facial expression recognition. Framework is based on a initial study of human vision. With the proposed framework high recognition accuracy, reduction in feature vector dimensionality and reduction in computational time for feature extraction is achieved by processing only perceptually salient region of face. Our proposed framework can be used for real-time applications since its unoptimized Matlab implementation run at 4 frames / second which is enough as facial expression does not change abruptly.

7. ACKNOWLEDGMENT

This project is supported by the Région Rhône-Alpes, France.

8. REFERENCES

[1] L. Zhaoping, "Theoretical understanding of the early visual processes by data compression and data selection," *Network: computation in neural systems*, vol. 17, pp. 301–334, 2006.

[2] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *6th ACM International Conference on Image and Video Retrieval*, 2007, pp. 401–408.

[3] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motivation*. 1971, pp. 207–283, Lincoln University of Nebraska Press.

[4] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, pp. 615–625, 2006.

[5] Y. Tian, "Evaluation of face resolution for expression analysis," in *Computer Vision and Pattern Recognition Workshop*, 2004.

[6] P. Yang, Q. Liu, and D. N. Metaxas, "Exploring facial expressions with compositional features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on Image and Signal Processing*, 2008.

[8] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, pp. 433–449, 2006.

[9] M.F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2005, pp. 76–84.

[10] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition*, vol. 41, pp. 833–851, 2008.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression.," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010.

[12] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Exploring human visual system: study to aid the development of automatic facial expression recognition framework," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2012.

[13] Y. Bai, L. Guo, L. Jin, and Q. Huang, "A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition," in *International Conference on Image Processing*, 2009.

[14] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *IEEE Automatic Face and Gesture Recognition Conference FG2011, Workshop on Facial Expression Recognition and Analysis Challenge FERA*, 2011.

[15] P. Ekman and W. Friesen, "The facial action coding system: A technique for the measurement of facial movements," *Consulting Psychologist*, 1978.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[17] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915–928, 2007.