Emotion Recognition Using PHOG and LPQ features

Abhinav Dhall1 Akshay Asthana2 Roland Goecke1,3 Tom Gedeon1

1School of Computer Science and 2School of Engineering, CECS, Australian National University, Australia
3Vision & Sensing Group, Faculty of Information Sciences and Engineering, University of Canberra, Australia

abhinnav.dhall@anu.edu.au, akshay.asthana@gmail.com, roland.goecke@ieee.org, tom.gedeon@anu.edu.au

Abstract—We propose a method for automatic emotion recognition as part of the FERA 2011 competition [1] . The system extracts pyramid of histogram of gradients (PHOG) and local phase quantisation (LPQ) features for encoding the shape and appearance information. For selecting the key frames, kmeans clustering is applied to the normalised shape vectors derived from constraint local model (CLM) based face tracking on the image sequences. Shape vectors closest to the cluster centers are then used to extract the shape and appearance features. We demonstrate the results on the SSPNET GEMEP-FERA dataset. It comprises of both person specific and person independent partitions. For emotion classification we use support vector machine (SVM) and largest margin nearest neighbor (LMNN) and compare our results to the pre-computed FERA 2011 emotion challenge baseline [1].

I. INTRODUCTION

We propose a method for automatic emotion recognition competition FERA 2011 [1] based on pyramid of histogram of gradients (PHOG) [2] and local phase quantisation (LPQ) [3] features. The experiment data set GEMEP-FERA [4] comprises of image sequences with actors speaking some dialogue exhibiting emotions, which in turn contains multiple units of individual facial expressions. Therefore, we quantize the number of frames in an image sequence using kmeans clustering algorithm. Further we compare the emotion recognition results of our method with the baseline results provided with the GEMEP-FERA dataset. For classification we experiment with support vector machines (SVM) [5] and largest margin nearest neighbor (LMNN) [6].

Automatic human emotion analysis finds a lot of application in HCI (Human Computer Interaction) Gaining insight into the state of the user’s mind via facial emotion analysis can provide valuable information for affective sensing systems. Many psychological studies [7], [8] have discussed the importance of using multimodal system for accurate emotion analysis. Here in our work we concentrate on automatic emotion analysis via facial expression recognition. Automatic emotion analysis finds applications in affective computing, intelligent environments, lie detection, psychiatry, emotion and paralinguistic communication and multimodal human computer interface (HCI).

Facial expression analysis has been an active field of research over the last two decades. It is mainly divided into two categories: image based and video based. In real world, human facial expressions are dynamic in nature. They constitute an onset, one or more apex (peaks) and an offset. Studies [9], [10], have proven the effectiveness of video based facial expression analysis over the static analysis. In [9], Bassili suggests that motions cues from a face image sequence are enough to recognise an expression even with minimal spatial information.

In one of the early works [11], Yacoob et al., facial parts are tracked and optical flow is calculated at high gradient values of the image sequence. Here the head were static.

The direction of the flow is quantised to eight levels in order to have a mid-level representation for high level facial expression classification. In [12], Black et al., use parametric models to extract parameters from facial features and use nearest neighbor classifier for FER. Different parametric models are used to differentiate between facial features relative to the head. In [13], comparison is performed for various facial expression analysis techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), optical flow and local filters such as Gabor wavelet representation via quantifying the Facial Action Coding System (FACS) [14]. In [15], Active Appearance Models are used for extracting facial features post fitting and machine learning techniques to classify emotions into FACS AU units [14]. Recently, [16] evaluate various state-of-the-art machine learning and image representation techniques on a new practical environment dataset for robust smile detection.

In [17], Pantic et al., propose an automatic AU detection method for profile face image sequences. Face tracking is dealt with as a segmentation problem where the profile face is the foreground. It finds largest connected component in HSV color space and then use the watershed segmentation algorithm to finally segment the face. Using contour based method, 20 points are extracted which are then used to identify AU using a rule based method. In [18], Pantic et al., propose two methods, first for automatic recognition of AU in video sequences and second for classifying AU coded expressions into learned emotion categories. The method is suitable for analysing temporal sequence pattern. Post face registration, temporal template called Motion History Image (MHI) [19] are constructed from the image sequence. Then temporal rules are used to identify AUs in the Cohn-Kanade [20] and the MMI [21] databases. The method achieves 90% recognition rate when detecting 27 AUs. Further the work is extended in [22], a wavelet based gentleboost template is used to track 20 facial fiducial points, which further are used to construct spatio-temporal features. A subset of features is selected using AdaBoost and SVM is used for checking presence of AUs.
In [23], authors propose volume LBP (VLBP) and LBP-TOP for dynamic facial expression analysis using LBP descriptors. In [24], authors propose use of LPQ for static facial expression analysis and extend LPQ to LPQ-TOP with the concept similar to LBP-TOP for extracting temporal information. In [25], a system based on PHOG and book of words BOW [26] features for robust static facial expression analysis. We use PHOG features in our temporal emotion analysis system. Also in the prior art the databases are fairly straightforward as compared to GEMEP-FERA dataset, as in it there are no pre-defined onset, apex and offsets and the actors are speaking sentences in which they exhibit same expression multiple times. The dataset has varied pose, dynamic head movement and occlusion. Our method gives good result on this dataset.

The rest of the paper is organised as follows: Section II presents the proposed technique in detail. Section III discusses the experimental results for emotion recognition of our method on GEMET-FERA dataset. Finally, Section IV provides the conclusions and future work.

II. SYSTEM

Our emotion recognition system starts with face tracking using CLM (Constraint Local Models) [27]. The shape vectors of the face in an image sequence are then normalised. Kmeans clustering algorithm is applied to the normalised shape vectors and images having face shape vectors closest to the cluster centers are chosen for further processing. Further, Viola Jones [28] face detector is applied to the chosen images. Then we compute the PHOG and LPQ features on the cropped faces. For classification we use SVM and LMNN. Figure 1 describes the test process of our method.

The system is explained in detail in the following sub sections.

1) Face tracking using CLM: In recent times, learnt model-based techniques have been extensively used in non-rigid deformable object fitting. We use the constrained local models (CLM) described in [27] for face tracking in the image sequences. It is based on fitting a parameterised shape model to the location landmark points of the face. It predicts locations of the model’s landmarks by utilising an ensemble of local feature detectors, which are then combined by enforcing a prior over their joint motion. The distribution of the landmark locations is represented non-parametrically and optimised via subspace constrained meanshifts. It fits well to various poses. We used a person independent model which was trained on the Multi-PIE database. Using the shape parameters from the fitted model a rough region of interest around the face is extracted. Figure 2, shows a snapshot of the face tracking using CLM. The fitting process gives a row vector $P$ containing location of each of the $n$ landmark points.

$$P = [x_1; y_1 \ldots x_n; y_n]$$

$P$ is then normalised by taking the horizontal Euclidean distance between the outer eye corners on the left and right side. The vertical distance is the Euclidean distance between the tip of the nose and the midpoint between the eyebrows. We denote the normalised shape vector as $P^\text{norm}$.

2) Clustering based sequence quantisation: $P^\text{norm}$ are calculated for all the images in an sequence. As the amount of motion in two consecutive frames is very sparse, we wish to remove the redundant frames so that the features are extracted on the frames which efficiently describe the
Fig. 2. The figure is the snapshot for CLM based face tracking. The image in the figure is from FEEDTUM[30] database.

temporal dynamics of the expression. Therefore for selecting key frames, we compute kmeans clustering algorithm on the normalised shape vectors. In [29], authors also perform kmeans clustering for computing their similarity features. But it is different from our as they compute it on the apex images, in order to divide them into various clusters. Our aim is different here, we want to remove the redundant frames. And also image sequences in the GEMET-FERA database, have multiple apex in an image sequence and exhibit an expressions multiple times. Post calculating the cluster centers we search for the nearest neighbor of the cluster centers. The algorithm is summarised in 1.

Post searching for the nearest neighbor of the cluster centers the Viola Jones face detector [28] detector is applied to the images corresponding to those nearest neighbors. Further we extract the features on the cropped faces computed by the face detector.

A. Shape feature extraction using PHOG

For extracting shape information we use PHOG [2] features. PHOG is a spatial pyramid extension of the histogram of gradients (HOG) [31] descriptors. The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image and has been used extensively in computer vision methods. PHOG features being an extension of HOG have shown good performance in object recognition [2] and static facial expression analysis [25], [32]. In the works, [25] and [32], PHOG descriptors have been used for static facial expression analysis. In the start canny edge detector is applied to the cropped face. Then the face is divided into spatial grids at all pyramid levels. After this a 3x3 Sobel mask is applied to the edge contours for calculating the orientation gradients. Then the gradients of each grid are joined together at each pyramid level. There is an option for two orientation ranges, [0-180] and [0-360]. In [25], [0-360] orientation range perform better then [0-180]. In our experiments, we use number of pyramids $L = 3$, the bin size $N = 8$ and the orientation range is [0-360].

B. Appearance feature extraction using LPQ

Local binary patterns (LBP) family of descriptors (LBP [33], LBP-TOP [23], LPQ [3] and LPQ-TOP [24]) have been extensively used for texture analysis, static and temporal facial expression analysis and face recognition. We use LPQ (Local Phase Quantization) appearance descriptor. Though LPQ-TOP [24] has been proposed for temporal data analysis, but as we do not have labeling of an onset, apex and offset in the database in our experiments, we use LPQ only. LPQ is based on computing short-term fourier transform (STFT) on local image window. At each pixel the local fourier coefficients are computed for four frequency points. Then the signs of the real and the imaginary part of the each coefficient is quantized using a binary scalar quantiser, for calculating the phase information. The resultant eight bit binary coefficients are then represented as integers using binary coding. This is step is similar to the histogram construction step in LBP. In the end we get a 256 dimensional feature vector. In our experiments we divided the cropped face of size 60x60 into four blocks. This gave us a vector dimension of 1024 for an image and 6144 for an image sequence where the number of cluster centers $m = 6$.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Anger</th>
<th>Fear</th>
<th>Joy</th>
<th>Relief</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Fear</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Joy</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Relief</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

TABLE I
CONFUSION MATRIX FOR PERSON INDEPENDENT TEST FOR OUR METHOD 1 (PHOG+SVM).

Algorithm 1: Clustering based sequence quantisation algorithm

Data: Shape vector $P_N$ from all the images of a sequence.
1. Normalise $P_N$ with horizontal and vertical euclidian distances.
2. Compute kmeans on $P_N$ for $m$ cluster centers and $i$ iterations.
3. Search one nearest neighbor $P_N$ for each cluster $m$.
4. Sort the nearest neighbors.
TABLE II
ACCURACY COMPARISON AMONG BASELINE (PI-BL, PS-BL, PO-BL), METHOD 1 (PI-M1, PS-M2, PO-M3), METHOD 2 (PI-M2, PS-M2, PO-M2), METHOD 3 (PI-M3, PS-M3, PO-M3). HERE, PI - PERSON INDEPENDENT PARTITION, PS - PERSON SPECIFIC PARTITION AND PO - PERSON OVERALL.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>PI-BL</th>
<th>PS-BL</th>
<th>PO-BL</th>
<th>PI-M1</th>
<th>PS-M2</th>
<th>PO-M2</th>
<th>PI-M3</th>
<th>PS-M3</th>
<th>PO-M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.86</td>
<td>0.92</td>
<td>0.89</td>
<td>0.714</td>
<td>0.518</td>
<td>0.857</td>
<td>0.846</td>
<td>0.851</td>
<td>0.928</td>
</tr>
<tr>
<td>Fear</td>
<td>0.07</td>
<td>0.40</td>
<td>0.20</td>
<td>0.666</td>
<td>1.0</td>
<td>0.80</td>
<td>0.333</td>
<td>0.80</td>
<td>0.520</td>
</tr>
<tr>
<td>Joy</td>
<td>0.70</td>
<td>0.73</td>
<td>0.71</td>
<td>0.60</td>
<td>0.454</td>
<td>0.548</td>
<td>0.70</td>
<td>0.545</td>
<td>0.645</td>
</tr>
<tr>
<td>Relief</td>
<td>0.31</td>
<td>0.70</td>
<td>0.46</td>
<td>0.687</td>
<td>1.0</td>
<td>0.807</td>
<td>0.687</td>
<td>1.0</td>
<td>0.807</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.27</td>
<td>0.90</td>
<td>0.52</td>
<td>0.667</td>
<td>0.70</td>
<td>0.68</td>
<td>0.666</td>
<td>1.0</td>
<td>0.807</td>
</tr>
<tr>
<td>Average</td>
<td>0.44</td>
<td>0.73</td>
<td>0.36</td>
<td>0.667</td>
<td>0.69</td>
<td>0.67</td>
<td>0.648</td>
<td>0.838</td>
<td>0.724</td>
</tr>
</tbody>
</table>

TABLE III
CONFUSION MATRIX FOR PERSON SPECIFIC TEST FOR OUR METHOD 1 (PHOG+SVM).

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Anger</th>
<th>Fear</th>
<th>Joy</th>
<th>Relief</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>14</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Fear</td>
<td>9</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Joy</td>
<td>3</td>
<td>2</td>
<td>17</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Relief</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

TABLE IV
OVERALL CONFUSION MATRIX FOR OUR METHOD 1 (PHOG+SVM).

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Anger</th>
<th>Fear</th>
<th>Joy</th>
<th>Relief</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Joy</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Relief</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

TABLE V
CONFUSION MATRIX FOR PERSON INDEPENDENT TEST FOR OUR METHOD 2 (PHOG+LPQ+SVM).

A. GEMEP-FERA dataset

The GEMEP-FERA dataset consists of recordings of 10 actors displaying a range of expressions, while uttering a meaningless phrase, or the word Aaah. There are 7 subjects in the training data, and 6 subjects in the test set, 3 of which are not present in the training set. The training set contains 155 image sequences and the testing contains 134 image sequences. The test contains six actors, three of them are also present in the training data. This enables testing of the method on both person independent and person specific settings. There are in total five emotion categories in the dataset: anger, fear, joy, relief and sadness. The baseline method was provided pre computed. The method uses LBP for feature extraction. It computes LBP feature for all the frames and then classifies each frame individually. Then maximum voting mechanism is used for deciding the final emotion category of the video sequence. The average person-specific and person-independent classification accuracy of the method are 0.73 and 0.44 and the overall baseline system accuracy is 0.56.

We experimented with three methods: a) PHOG features with SVM classification, b) PHOG+LPQ features with SVM classification and c) PHOG+LPQ with LMNN classification. First we track the face using CLM and then apply kmeans clustering to the shape parameters. In our experiments the number of clusters centers $m = 6$, this was chosen empirically. Then we apply kNN and find the nearest neighbour to the computed cluster centers. Post this we apply the Viola Jones face detector to the closest images. The cropped face's size was set to 60x60. Then for method 1, we compute the PHOG descriptors, here number of pyramids $L = 3$, the orientation range is [0-360]. PHOG descriptors for the images (6 images in this case) are then concatenated. In total we have 6800 dimension vector for each sequence. We then applied PCA to the data and 98% of the variance is maintained.

For classification we trained a SVM RBF model [5]. For parameter selection we used 10 cross-validation. The accuracy for method 1 comes out to be: 0.66 for person independent partition, 0.69 for person dependent partition and 0.67 as overall accuracy. In this method we are using shape features only. Tables I, III and IV are the confusion matrix for method 1 for person independent, person dependent and overall partitions. Next, in method 2, we added appearance features in the form of LPQ. In literature LPQ have shown to perform better [3] then LBP and are invariant to blur and illumination up to some extent. The LPQ were calculated on 60x60 cropped face. For LPQ, 6144 features were generated per sequence. These are then concatenated with the PHOG features. We again apply PCA and retain
We presented a novel method for automatic emotion recognition. We use unsupervised clustering on normalised shape vectors for choosing key frames. For capturing shape information we use, PHOG features and for appearance we use the recently proposed LPQ features. We test our method for person independent and person dependent classification performance on the GEMEP-FERA data set. The data set is a challenging data base with actors posing various emotions while uttering sentences. The proposed method performs better then the baseline results. For future work we want to explore methods for feature selection such as via boosting and mutual information. Also we will experiment on other databases such as the MMI database [21], Cohn-Kanade database [20] so as to check the generic performance of the proposed method.

### IV. CONCLUSIONS AND FUTURE WORK

We presented a novel method for automatic emotion recognition. We use unsupervised clustering on normalised shape vectors for choosing key frames. For capturing shape information we use, PHOG features and for appearance we use the recently proposed LPQ features. We test our method for person independent and person dependent classification performance on the GEMEP-FERA data set. The data set is a challenging data base with actors posing various emotions while uttering sentences. The proposed method performs better then the baseline results. For future work we want to explore methods for feature selection such as via boosting and mutual information. Also we will experiment on other databases such as the MMI database [21], Cohn-Kanade database [20] so as to check the generic performance of the proposed method.

### REFERENCES


