Facial Expression Recognition Using Emotion Avatar Image

Songfan Yang and Bir Bhanu
Center for Research in Intelligent Systems, University of California, Riverside n
Riverside, CA 92521, US
syang@ee.ucr.edu, bhanu@cris.ucr.edu

***Abstract*—Existing facial expression recognition techniques analyze the spatial and temporal information for every single frame in a human emotion video. On the contrary, we create the Emotion Avatar Image (EAI) as a single good representation for each video or image sequence for emotion recognition. In this paper, we adopt the recently introduced SIFT flow algorithm to register every frame with respect to an Avatar reference face model. Then, an iterative algorithm is used not only to super-resolve the EAI representation for each video and the Avatar reference, but also to improve the recognition performance. Subsequently, we extract the features from EAIs using both Local Binary Pattern (LBP) and Local Phase Quantization (LPQ). Then the results from both texture descriptors are tested on the Facial Expression Recognition and Analysis Challenge (FERA2011) data, GEMEP-FERA dataset. To evaluate this simple yet powerful idea, we train our algorithm only using the given 155 videos of training data from GEMEP-FERA dataset. The result shows that our algorithm eliminates the person-specific information for emotion and performs well on unseen data.**

Keywords-face registration; level of Emotion Avatar Image; Person-independent emotion recognition; SIFT flow

I. INTRODUCTION

Automatic recognition of emotion from human facial expression images has been an interesting and challenging problem for over 30 years. Aiming towards the application of safety, surveillance and human-computer interaction, this topic has drawn even more attention recently.

A literature review shows that early stage research on facial expression recognition focused on static images [1]. Both feature-based and template-based approaches have been investigated. Recently, researchers tend to work with image sequence or video data for developing the automated expression recognition systems. As demonstrated in the field of computer vision [2, 3, 13, 15] and psychology experiments [16, 17], various types of dynamic information is crucial for recognition of human expressions, such as dynamical appearance-based and dynamical geometric-based information.

However, extracting the facial dynamics from an expression sequence is not a trivial problem. It requires near perfect alignment for the head pose and facial features. The inherent challenge for facial expression recognition is the dilemma between rigid motion of the head pose and non-rigid motion of facial muscles. Though vast amount of head pose estimation algorithms have good performance, it is still not easy to recover the morphing of facial texture. On the other hand, the alignment based upon the facial features such as eyes or eye corners, nose, etc. encounters issues such as:

(1) Non-rigid morphing (although claimed to be "anchor features", meaning relatively stable compared to other facial features such as eyebrow or mouth).

(2) Person-specific appearance (the location of the tip of nose or the distance between eyes are absolutely not constant for different people).

Therefore, the aforementioned questions encourage us to develop techniques which not only correctly register the face image but also eliminate the person-specific effects. To pinpoint the key emotion of an image sequence while circumventing the complex and noisy dynamics, we also seek to summarize the emotion video containing hundreds of frames. If we can find a single good image representation based upon which we make judgments, we will be able to capture the emotion that a whole video is trying to express in a computationally efficient manner.

In this paper, we adopt the recently introduced SIFT flow algorithm [8] to register the facial images. By matching the dense SIFT descriptors across image pairs, this method is able to generate a favorable alignment result. Although SIFT flow is originally designed for image alignment at the scene level, it is reasonable to apply it here to facial expression recognition. It is capable of globally aligning the head/face region while maintaining the shape and motion of facial features. In order to solely extract the facial motion information irrespective of person-specific effects, we iteratively build a single Avatar reference face model, onto which we align all the face images. Later, by averaging, we update the Avatar reference face model, and also the single good representation, EAI, for each video. The reason we name the model Avatar is because the subjects are morphed towards homogeneity while the emotions are successfully maintained. Subsequently, the EAIs are passed through both LBP and LPQ texture descriptor for feature extraction. Finally, Support Vector Machine (SVM) with linear kernel is used for classification. As we demonstrate later by our experimental results, using a single representation for a facial expression session is a simple but powerful idea to recognize facial emotions. It is able to transform the whole expression recognition problem from image sequence back to static image that is amenable for real-time application.

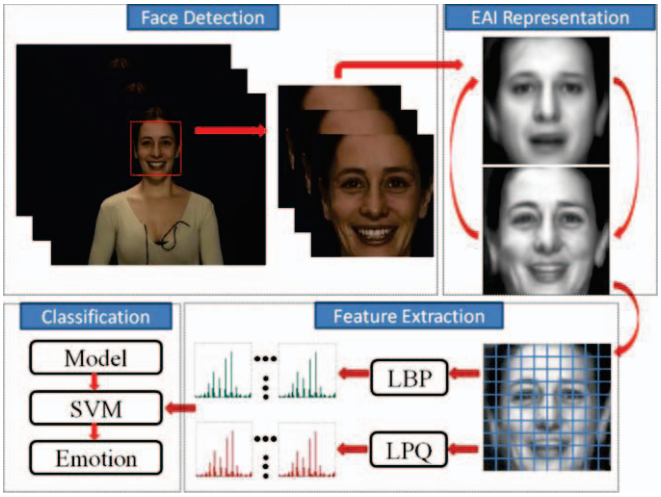

Figure 1. The overview of our approach

In what follows, we first discuss the related work (Section II) and subsequently introduce our iterative algorithm to build Avatar reference and EAIs in Section III. Several combinations of methods are tested and the classification results are demonstrated in Section IV. Finally, Section V provides the conclusions of this paper.

II. Related Work and Our Contribution

A large amount of effort has been focused on describing an emotion using two major ideas, i.e., geometric-based approaches and appearance-based approaches.

Geometric-based approaches track the facial geometry information over time and classify expressions based on the deformation. Chang et al. [6] defined a set of points as the facial contour feature, and an Active Shape Model (ASM) is learned in a low dimensional space. Lucey et al. [5] employed Active Appearance Model (AAM)-derived representation while Valtar, Patras, and Pantic [4] tracked 20 fiducial facial points on raw video using a particle filter.

On the other hand, appearance-based approaches emphasize on describing the appearance of facial features and their dynamics. Zhao and Pietikaninen [2] employed the dynamic LBP which is able to extract information along the time axis. Bartlett et al. [7] used a bank of Gabor wavelet filter to decompose the facial texture. More recently, Wu et al. [3] utilized Gabor Motion Energy Filters which is also able to capture the spatial-temporal information. But its feature dimension exceeds 2 million and, therefore, its applicability in real-time environment is unclear.

Existing work intensely emphasizes on analyzing the sequential change of the facial feature. But the onset and offset is hard to detect. If a near-apex frame is able to be picked up to represent an entire expression session, we can avoid extracting subtle sequential facial feature deformations from noisy environment, and describe emotions in a reliable manner.

The contributions of this work are:

(1) The idea of condensing a video sequence into single image representation, EAI, in expression recognition.

(2) Align each face image with respect to a reference face model, Avatar reference, which is able to eliminate the person-specific information.

To our best knowledge, until now, little work has been done to condense a video sequence into a tractable image representation for emotion recognition. As the results show later, if the expression is not extremely subtle such that even human visual system is unable to capture, our algorithm can distinguish the difference among expressions.

III. Technical Approach

Fig. 1 outlines our approach in four major steps. After extracting the face from a raw video, we create the EAI for each video as a single representation to eliminate the person-specific effect while maintaining the shape and texture information of facial features. Both LBP and LPQ texture descriptors are applied to generate features, and then, the linear SVM classifier is used for classification.

A. SIFT flow alignment

SIFT flow is recently introduced by Liu et al. [8]. It is originally designed to align an image to its plausible nearest neighbor for large image variations. The SIFT flow algorithm robustly matches dense SIFT features between two images, while maintaining spatial discontinuities.

In [8], the local gradient descriptor, SIFT [9], is used to extract pixel-wise feature component. For every pixel in an image, the neighborhood (e.g. 16×16) is divided into a 4×4 cell array. The orientation of each cell is quantized into 8 bins, generating a $4 \times 4 \times 8 = 128$-dimension vector as the SIFT representation for a pixel, or the so called SIFT image. The SIFT image has a high spatial resolution and can characterize the edge information.

After obtaining the per-pixel SIFT descriptors for two images, a dense correspondence is built to match the two images. The objective energy function similar to optical flow is designed as:

$$E(\boldsymbol{w}) = \sum_p \min\left(\left\|s_1(\boldsymbol{p}) - s_2(\boldsymbol{p}+\boldsymbol{w}(\boldsymbol{p}))\right\|_1, t\right) + \quad (1)$$

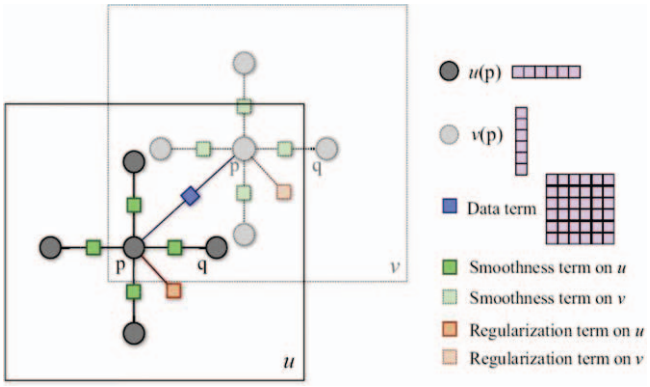

Figure 2. Dual-layer Belief Propagation. The designed objective function of SIFT flow is decoupled for horizontal (u) and vertical (v) components. [8]

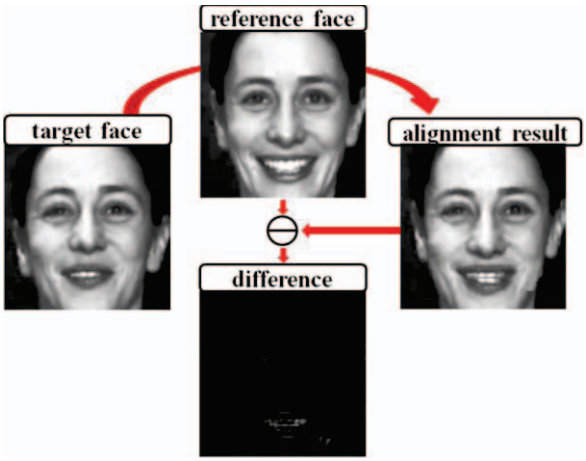

Figure 3. Face alignment using SIFT flow. Facial motion can be captured as shown by the difference between the reference face and the alignment result.

$$\sum_p \eta(|u(\boldsymbol{p})| + |v(\boldsymbol{p})|) + \tag{2}$$

$$\sum_{(\boldsymbol{p},\boldsymbol{q})\in\varepsilon} \min(\alpha|u(\boldsymbol{p}) - u(\boldsymbol{q})|, d) +$$

$$\min(\alpha|v(\boldsymbol{p}) - v(\boldsymbol{q})|, d) \tag{3}$$

where $\boldsymbol{p} = (x, y)$ is the grid coordinate of images, and $\boldsymbol{w}(\boldsymbol{p}) = \big(u(\boldsymbol{p}), v(\boldsymbol{p})\big)$ is the flow vector at $\boldsymbol{p}$. s_1 and s_2 are two SIFT images to be matched. ε contains all the spatial neighborhoods (a four-neighbor system is used). The *data term* in (1) is a SIFT descriptor match constraint that enforces the match along the flow vector $\boldsymbol{w}(\boldsymbol{p})$. The *small displacement constraint* in (2) allows the flow vector to be as small as possible when no other information is available. The *smoothness constraint* (3) takes care of the similarity of flow vectors for adjacent pixels. In this objective function, truncated L1 norm is used in both the data term and the smoothness term with t and d as the threshold of matching outliers and flow discontinuities, respectively.

The dual-layer loopy belief propagation is used as the base algorithm to optimize the objective function. Fig. 2 shows the factor graph of the model. Then, a coarse-to-fine SIFT flow matching scheme is adopted to improve the speed and the matching result.

Fig. 3 contains two frames (frame 1 and frame 30) in sequence 29 from GEMEP-FERA training dataset [19]. Assuming frame 1 is the reference and we align the target frame 30. As shown in Fig. 3, this alignment method is able to eliminate the small pose difference while maintaining the motion on mouth, eyes, and lower cheek.

B. Emotion Avatar Image Representation

SIFT flow has the potential to align images with large spatial variation. This is favorable for us to align the face image given the possibility of large head pose movement or occlusion. The question is how to eliminate the person-specific information. We seek to build a reference face with respect to which we can align each face image.

Algorithm 1: Avatar Reference and EAI

Given: face image $I^{(m,n)}$ from sequence m, frame n. The total number of image sequence is M, and the total number of frames in sequence m is N_m. Q is a user defined number of levels. Define A_i^{ref} as the Avatar reference at level-i; EAI_i^m as the single representation for sequence m based on the level-i Avatar reference A_i^{ref}; $I_{align}^{(m,n)}$ as the alignment result for a face image $I^{(m,n)}$ using SIFT flow.

Initialization: $A_0^{ref} = \frac{1}{\sum_{m=1}^{M} N_m} \sum_{m=1}^{M} \sum_{n=1}^{N_m} I^{(m,n)}$

for $i = 1$ to Q **do**

  for $m = 1$ to M **do**

    for $n = 1$ to N_m **do**

      $I_{align}^{(m,n)} \coloneqq SIFTflow\big(I^{(m,n)}, A_{i-1}^{ref}\big)$

    end for

    $EAI_i^m \coloneqq \frac{1}{N_m} \sum_{n=1}^{N_m} I_{align}^{(m,n)}$

  end for

  $A_i^{ref} \leftarrow \frac{1}{\sum_{m=1}^{M} N_m} \sum_{m=1}^{M} EAI_i^m$

end for

return EAI_Q^m

In Algorithm 1, we design an iterative average method to generate an Avatar reference face model. To put simply, we initialize our algorithm by averaging all possible face images in the training dataset. Using this average face as the reference, we align each face image in the video using SIFT flow. Then, we update the Avatar reference using the aligned face images.

The number of iterations defines the level of the Avatar reference (level 0 means the average of all the face images). The Avatar reference models for the first 3 levels are shown in row 1 of Fig. 4. One observation we made is that the Avatar reference does not necessarily have to be in neutral face. It simply captures the most likely facial appearance throughout the whole dataset, and therefore, it has less total variation in registration. For example, in this work, we train our algorithm solely using the training data. The mouth is open for the level-1 and level-2 Avatar reference face result (as shown in Fig. 4, row 1). This is because most of the subjects in the training data are uttering meaningless phrases [14], therefore, carry quite a lot of mouth movement.

In Algorithm 1, once the Avatar reference face model is obtained, we establish the single representation EAI for sequence of face images at the current level. Intuitively, we describe an image sequence as the average of all frames within this sequence. The formal description is defined by EAI_i^m in Algorithm 1. Moreover, EAIs of current level are obtained based on the Avatar reference from previous level.

In this work, we attempt to test the performance of EAIs of different level. As shown in Fig 4 (row 2 and row 3), the

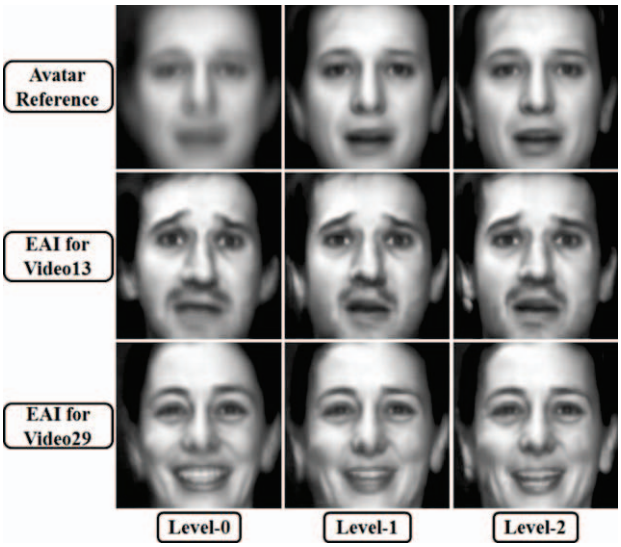

Figure 4. Avatar reference face model and EAI representations for the first three levels. For comparison, level-0 EAIs are the average of every face image from their corresponding videos.

quality of the EAIs improves as the level of Avatar reference becomes higher. A high level Avatar reference model tends to super-resolve the EAIs by enhancing the facial details, and attenuates the person-specific information. However, the emotion information is still retained. It is easy for human visual system to identify the emotion category based on the EAI representations.

C. Feature Extraction

We transform the facial expression recognition problem from video to static image. To describe the facial texture from a single image, we use the well-known texture descriptor LBP and recently proposed blur insensitive LPQ descriptor.

1) LBP

LBP is a powerful and well known texture descriptor. In this work, we used the extended version of basic LBP by Ojala et al. [10] where the LBP descriptor is gray-scale and rotation invariant. To briefly go over this extended work, the operator, denoted as $LBP^{riu2}_{P,R}$, is applied to an circularly symmetric neighborhood with P number of pixel on the circle of radius R. Superscript "$riu2$" informs the rotation invariant property. Rotation invariant LBP is favorable since it reduces the feature dimension. For example, the $LBP^{u2}_{8,1}$ adopted in this work will generate 59 basic patterns while the $LBP_{8,1}$ has 256 possibilities.

After thresholding each pixel in the neighborhood with respect to the center value, the histogram is used to accumulate the occurrence of the various patterns over a region. In our experiment, we restore the face images to 200×200, and each image is divided into size 20×20 blocks to capture the local texture pattern. Therefore, the LBP feature vector in use is of dimension $59 \times 10 \times 10 = 5900$.

2) LPQ

The blur insensitive LPQ descriptor is originally proposed by Ojansivu et al. in [11]. The spatial blurring is represented as multiplication of original image and a point spread function (PSF) in frequency domain. The LPQ method is based upon the invariant property of the phase of the original image when the PSF is centrally symmetric.

LPQ method examines a local M-by-M neighborhood $\mathcal{N}_x$ at each pixel position $\boldsymbol{x}$ of the image $f(\boldsymbol{x})$, and extracts the phase information using the short-term Fourier transform defined by

$$F(\boldsymbol{u},\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathcal{N}_x} f(\boldsymbol{x}-\boldsymbol{y}) e^{j2\pi \boldsymbol{u}^T \boldsymbol{y}} = \boldsymbol{w}_{\boldsymbol{u}}^T \boldsymbol{f}_{\boldsymbol{x}} \qquad (4)$$

where $\boldsymbol{w}_{\boldsymbol{u}}$ is the basis vector of the 2-D DFT at frequency $\boldsymbol{u}$, and $\boldsymbol{f}_{\boldsymbol{x}}$ another vector containing all M^2 image samples from $\mathcal{N}_{\boldsymbol{x}}$.

The local Fourier coefficients are at four frequency points: $\boldsymbol{u}_1 = [a, 0]^T$, $\boldsymbol{u}_2 = [0, a]^T$, $\boldsymbol{u}_3 = [a, a]^T$, and $\boldsymbol{u}_4 = [a, -a]^T$, where a is a sufficiently small scalar. We use $a = 1/7$ in our experiment. The vector for each pixel is obtained as

$$\boldsymbol{F}_{\boldsymbol{x}} = [F(\boldsymbol{u}_1,\boldsymbol{x}), F(\boldsymbol{u}_2,\boldsymbol{x}), F(\boldsymbol{u}_3,\boldsymbol{x}), F(\boldsymbol{u}_4,\boldsymbol{x})] \qquad (5)$$

The phase information is recovered by a scalar quantizer

$$q_j(\boldsymbol{x}) = \begin{cases} 1, & if\ g_j(\boldsymbol{x}) \geq 0 \\ 0, & otherwise \end{cases} \qquad (6)$$

where $g_j(\boldsymbol{x})$ is the j th component of the vector $G_x = [Re\{\boldsymbol{F}_{\boldsymbol{x}}\}, Im\{\boldsymbol{F}_{\boldsymbol{x}}\}]$.

The result eight binary coefficients $q_j(\boldsymbol{x})$ are represented as integer values between 0-255 using binary coding

$$f_{LPQ}(\boldsymbol{x}) = \sum_{j=1}^{8} q_j(\boldsymbol{x}) 2^{j-1} \qquad (5)$$

Also, the de-correlation process is added to the original LPQ implementation to eliminate the dependency of the neighboring pixels. Similar to LBP, we divided the 200×200 face images into size 20×20 regions. Therefore, the LPQ feature vector is of dimension $256 \times 10 \times 10 = 25600$.

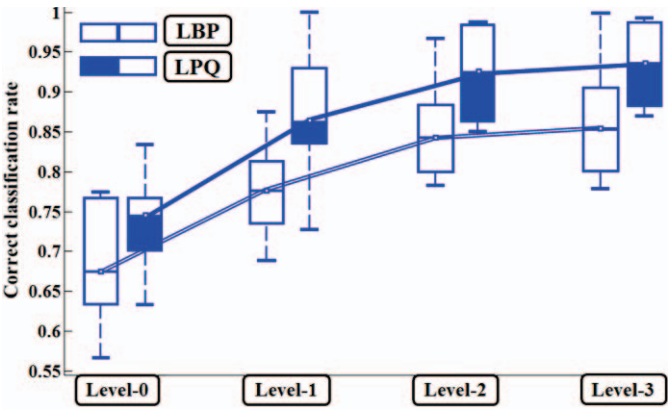

Figure 5. Box plot of 10-fold cross-validation results on 155 training videos using different level of EAIs. The average classification rate is connected for each texture descriptor to show the improvement on each level. This is to demonstrate that we adopt level-2 EAIs because of its potential of good performance and relative computational efficiency.

IV. EXPERIMENTAL RESULTS

A. Data

Our method is tested on the Facial Expression Recognition and Analysis Challenge (FERA2011) data, GEMEP-FERA dataset [19]. We submitted the results to the organization panel of FERA 2011. They provided the evaluation result in three different categories: person-independent (subjects in test data are not in the training), person-specific (subjects in test data are in the training), and overall.

GEMEP-FERA dataset consists of 10 actors displaying expressions including anger, fear, joy, relief, and sadness, while uttering some meaningless phrase. We train our model using the training data which contains 7 subjects with 155 videos. Unlike other datasets, where expressions start with neutral and end with apex or offset, GEMEP-FERA dataset does not enforce this requirement. Moreover, arbitrary head movement and facial occlusion are also uncontrolled. We compare our method to the baseline approach described in [14].

B. Results and Discussions

We first extract the face from the raw data using the Viola and Jones face detector [18] which achieved near perfect result on GEMEP-FERA dataset. The face images are then aligned to level-2 Avatar reference face model based on the aforementioned iterative algorithm, and single representation EAIs are generated. Subsequently, using both LBP and LPQ operators, we extract the feature from all the EAIs. Specifically, the $LBP_{8,1}^{u2}$ is used in our experiment. The parameters for the LPQ operator are $M = 9, a = \frac{1}{7}, \rho = 0.9$. Lastly, the classifier we used is linear kernel SVM [12] classifier with default parameter setting.

The reason why we use level-2 Avatar reference face model is statistically demonstrated in Fig. 5. We carry out a series of 10-fold cross-validation experiments on only the training set using first four levels of the Avatar reference face model and test on both LBP and LPQ texture descriptor. Fig. 5 shows that the performances for both descriptors improve as the level of the Avatar reference increases. This keeps consistency with our discussion on Avatar reference level in Section III part B. The reason we stop at level-2 is because the improvement is less noticeable, yet requires more running time as we use level-3 of Avatar reference.

Notice that the recognition rate in Fig. 5 is higher than the overall rate in test result. This is because the cross-validation on training data does not enforce the exclusiveness of test and training subjects, and therefore, this test result is subject to person-specific category.

The confusion matrices for EAI using LPQ operator are shown in Table I-III, with test results on person-independent, person-specific, and overall, respectively. Similarly, the confusion matrices for EAI using LBP operator are presented

TABLE I. CONFUSION MATRIX FOR EAI + LPQ (PERSON-INDEPENDENT)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**12**	3	0	0	1
	Fear	0	**7**	0	0	0
	Joy	0	5	**19**	4	0
	Relief	1	0	1	**11**	2
	Sadness	1	0	0	1	**12**
Total rate		0.8571	0.4667	0.95	0.6875	0.8
Average rate		**0.7523**				

TABLE II. CONFUSION MATRIX FOR EAI + LPQ (PERSON-SPECIFIC)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**13**	0	0	0	0
	Fear	0	**10**	0	0	0
	Joy	0	0	**10**	0	0
	Relief	0	0	1	**10**	0
	Sadness	0	0	0	0	**9**
Total rate		0. 9231	1	1	0.9091	1
Average rate		**0.9618**				

TABLE III. CONFUSION MATRIX FOR EAI + LPQ (OVERALL)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**25**	3	0	0	2
	Fear	0	**17**	0	0	0
	Joy	0	5	**29**	4	0
	Relief	1	0	2	**21**	2
	Sadness	1	0	0	1	**21**
Total rate		0. 9259	0.68	0.9355	0.8077	0.84
Average rate		**0.8378**				

TABLE IV. CONFUSION MATRIX FOR EAI + LBP (PERSON-INDEPENDENT)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**12**	4	4	0	5
	Fear	0	**8**	0	0	0
	Joy	1	3	**16**	0	0
	Relief	1	0	0	**14**	1
	Sadness	0	0	0	2	**7**
Total rate		0. 8571	0.5333	0.8	0.875	0.4667
Average rate		**0.7064**				

TABLE V. CONFUSION MATRIX FOR EAI + LBP (PERSON-SPECIFIC)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**11**	0	1	0	0
	Fear	0	**9**	1	0	0
	Joy	2	1	**8**	0	0
	Relief	0	0	1	**10**	1
	Sadness	0	0	0	0	**9**
Total rate		0.8462	0.9	0.7273	1	0.9
Average rate		**0.8747**				

TABLE VI. CONFUSION MATRIX FOR EAI + LBP (OVERALL)

		Truth				
		Anger	Fear	Joy	Relief	Sadness
Prediction	Anger	**23**	4	5	0	6
	Fear	0	**17**	1	0	1
	Joy	3	4	**24**	0	0
	Relief	1	0	1	**24**	2
	Sadness	0	0	0	2	**16**
Total rate		0.8519	0.68	0.7742	0.9231	0.64
Average rate		**0.7738**				

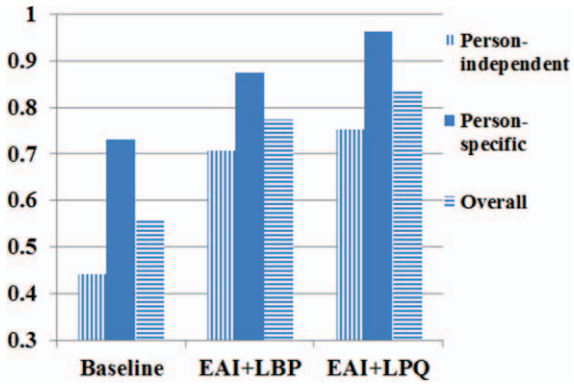

Figure 6. Comperision of test results. The value for each class is the "Avergae rate" from its corresponding confusion matrix. The EAI representation possesses the largest imporvement (31% in for LPQ, 26% for LBP) in predictiing the unseen person-independent data .

in Table IV-VI. In both LBP and LPQ cases, our EAI representation achieves the most improvement on baseline result in the person-independent test (from 0.44 to 0.75 in LPQ; from 0.44 to 0.71 in LBP). This improvement can also be visualized in Fig. 6. This is a positive evidence that our approach eliminates the person-specific effect and captures the facial expression information. Also, this demonstrates the desired ability of EAI in predicting the unseen data in real applications.

Moreover, the inherent characteristic of our approach is to condense the training information while maintaining the key emotion factor. Unlike the baseline approach which treats each frame as a single training instance (total of 8995 frames from 155 videos if all the images in the training set are used), our model only considers it as 155 EAIs. Given more videos of training data, chances are high that our approach will perform better since 155 videos of five emotions (approximately 30 video/emotion on average) may not be sufficiently large to represent a single emotion across a large population in EAIs.

V. CONCLUSIONS

In this paper, we explore the new idea of condensing a video sequence to be a single EAI representation. We adopt SIFT flow for aligning the face images which is able to compensate for large global motion and maintain facial feature motion detail. Then, an iterative algorithm is used to generate an Avatar reference face model onto which we align each face image. The 10-fold cross-validation result on the training data shows that the higher resolution Avatar reference has the potential to generate better quality of EAI, and subsequently, higher classification rate. Both LBP and LPQ texture descriptor are tested on our EAI representation for feature extraction. Based on the result of the challenging facial expression recognition dataset, GEMEP-FERA dataset [19], the performance is dramatically improved by the idea of single EAI representation compared with the baseline approach. In the future, we will give the theoretical proof for the characteristic of Emotion Avatar Image and also, expand our experiments on larger datasets.

ACKNOWLEDGMENT

This work was supported in part by NSF grant 0727129. The authors would like to thank the organizers of FERA 2011 Grand Challenge for providing the training data, test data, and evaluating the results. All this effort has been highly valuable in advancing the field of emotion recognition.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424-1445, Dec. 1996.

[2] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition using Local Binary Patterns with an Application to Facial Expressions," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 915-928, 2007.

[3] T. F. Wu, M. S. Bartlett, J. R. Movellan, "Facial Expression Recognition using Gabor Motion Energy Filters," IEEE Int'l Conf. Computer Vision and Pattern Recognition, workshop for Human Communicative Behavior Analysis, 42-47, 2010.

[4] M.F. Valstar, I. Patras, and M. Pantic, "Facial Action Unit Detection Using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data," IEEE Int'l Conf. Computer Vision and Pattern Recognition, workshop for Human-Computer Interaction, 2005.

[5] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, J. Cohn, "AAM Derived Face Representations for Robust Facial Action Recognition," International Conference on Automatic Face and Gesture Recognition, 2006, pp. 155–160.

[6] C. Hu, Y. Chang, R. Feris, and M. Turk, " Manifold based analysis of facial expression," IEEE Int'l Conf. Computer Vision and Pattern Recognition,Workshop on Face Processing in Video, 2004.

[7] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," International Conference on Automatic Face and Gesture Recognition, 2006, pp. 223–230.

[8] C. Liu, J. Yuen, A. Torralba, "SIFT Flow: Dense Correspondence across Scenes and its Applications," IEEE Trans. Pattern Analysis and Machine Intelligence, Aug. 2010.

[9] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," IEEE International Conference on Computer Vision (ICCV), pages 1150–1157, Kerkyra, Greece, 1999.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Grey-scale and Rotation Invariant Texture Classification with Local Binary Patterns," IEEE Trans. Pattern Analysis and Machine Intelligence, 24(7):971– 987, 2002.

[11] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," In Proc. Int. Conf. on Image and Signal Processing, volume 5099, pages 236–243, 2008.

[12] C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines.

[13] Y. Tong, J. Chen, and Q. Ji. "A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding," IEEE Trans. Pattern Analysis and Machine Intelligence, 99(1), 2010.

[14] Michel F. Valstar, Bihan Jiang, Marc Méhu, Maja Pantic, and Klaus Scherer, "The First Facial Expression Recognition and Analysis Challenge", in Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2011, in print

[15] P. Yang, Q. Liu, and D. Metaxas, "Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition," IEEE Computer Vision and Pattern Recognition, pages 1–6, 2007.

[16] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the Enigmatic Face: the Importance of Facial Dynamics in Interpreting Subtle Facial Expressions," Psychological Science, 16:403–410, 2005.

[17] J. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," Journal of personality and social psychology, 37(11):2049–2058, 1979.

[18] P. Viola and M. Jones, "Robust Real-time Face Eetection," International Journal of Computer Vision, 57(2):137–154, 2004.

[19] FERA2011 Challenge data: http://sspnet.eu/fera2011/fera2011data/