Emotion Recognition from an Ensemble of Features

Usman Tariq, *Student Member, IEEE,* Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, *Student Member, IEEE,* Thomas S. Huang, *Life Fellow, IEEE,* Xutao Lv and Tony X. Han

***Abstract*—This work details the authors' efforts to push the baseline of expression recognition performance on a realistic database. Both subject-dependent and subject-independent emotion recognition scenarios are addressed in this work. These two happen frequently in real life settings. The approach towards solving this problem involves face detection, followed by key point identification, then feature generation and then finally classification. An ensemble of features comprising of Hierarchial Gaussianization (HG), Scale Invariant Feature Transform (SIFT) and Optic Flow have been incorporated. In the classification stage we used SVMs. The classification task has been divided into person specific and person independent emotion recognition. Both manual labels and automatic algorithms for person verification have been attempted. They both give similar performance.**

I. INTRODUCTION

Automated expression recognition shall very soon have its sizeable impact in areas ranging from psychology to HCI (human computer interaction) to HRI (human-robot interaction). In psychology, for instance, the applications include autism early intervention techniques, etc. While in HRI and HCI there is an ever increasing demand to make the computers and robots behave more human-like. For instance in automated learning, the computer should ideally be able to identify the cognitive state of the student. Say, if the student is gloomy, it might tell a joke, etc.

The multi-modal computer-aided learning system at the Beckman Institute in University of Illinois at Urbana-Champaign, USA is one landmark example of computer aided learning (http://itr.beckman.uiuc.edu). The computer avatar offers an appropriate tutoring strategy based upon the users facial expressions, task state, eye-movement and keywords [1].

Psychologists and linguistics have various opinions about the importance of different cues in human affect judgment [1]. But there are some studies (e.g. [2]) which indicate that facial expression in the visual channel is the most affective and important cue and correlates well with the body and voice. In this work, we also use features extracted from the facial region.

II. BACKGROUND WORK

Emotion recognition using visual cues has been receiving a great deal of attention in the past decade. Most of the existing approaches do recognition on six universal basic emotions because of their stability over culture, age and other identity related factors. The choices of features employed for emotion recognition are classified by Zeng et al. [1] into two main categories: geometric features and appearance features. In this section, we closely follow that taxonomy to review some of the notable works on the topic.

The geometric features are extracted from the shape or salient point locations of important facial components such as mouth and eyes. In the work of Changbo et al. [3], 58 landmark points are used to construct an active shape model (ASM). These are then tracked and give facial expressions recognition in a cooperative manner. Pantic and Barllet [4] introduced a set of more refined features. They utilized facial characteristic points around the mouth, eyes, eyebrows, nose, and chin as geometric features for emotion recognition.

The appearance features representing the facial characteristics such as texture and other facial miniatures are also employed in many works. Among them, another work of Bartlett and colleagues [5] highlights Gabor wavelets extracted after warping the image in 3D into canonical views. Also, the work by Anderson and McOwan [6] introduces a holistic spatial ratio face template. In this work, the movement of identified regions of the face are extracted out from rigid head movement through tracking and used as feature for SVM classification. Usage of temporal templates by Valstar et al., in [7], is another example. They used multilevel motion history images to study the subtle changes in facial behavior in terms of action units.

Beside geometric and appearance based, the hybrid features are also used and have shown impressive recognition results. In [8], Tian et al. combined shapes and the transient features to recognize fine-grained changes in facial expression. In an intuitive way of analyzing facial expressions, several other works, such as [9] and [10], follow the traditional approach of using 3D face models to estimate the movements of the facial feature points. These features are related to the AUs and their movements control the emotional states of the subject.

In this work we used an ensemble of features extracted from the facial region. These include both appearance and motion features. The feature set comprises of SIFT features at the key points, Hierarchical Gaussianization (HG) features and motion features. Classification is carried out using SVMs. The final video emotion is computed based upon majority voting of detected emotion in the frames from the respective video. Our approach has proven its significance over the baseline methodology [12].

U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang are with Department of Electrical and Computer Engineering, Coordinated Science Laboratory and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 N. Mathews Ave., Urbana, IL 61801, USA. Email: {utariq2, klin21, zhenli3, xizhou2, wang308, vuongle2, t-huang1}@illinois.edu

X. Lv and T. X. Han are with Department of Electrical and Computer Engineering, University of Missouri, Engineering Building West, Columbia, MO 65201, USA. Email: xlyp2@mail.mizzou.edu, hantx@missouri.edu

III. DATABASE

The database used in this work is the GEMEP-FERA database [11] [12]. It consists of the video-recordings of 10 actors. They are displaying a range of expressions, while uttering the word 'Aaah', or a meaningless phrase. There are 7 subjects in the training data (3 males and 4 females). While the test data-set has 6 subjects. 3 out of those 6 are not present in the training set. The total number of videos in the training partition is 155 while that in the testing partition is 134.

There are five discrete, mutually-exclusive emotion categories that are staged in the database[12]. These categories are: Anger, Fear, Joy, Relief, and Sadness. Emotions are labeled on a per video basis. In the training partition each emotion appears 30-32 times.

IV. PRE-PROCESSING

A number of pre-processing steps were carried out before the feature extraction phase. We observed interlacing [13] in the training videos. Thus de-interlacing [13] was performed for each video to improve image quality for feature extraction in the later stage. Specifically, we extracted two horizontal fields (one is comprised of odd lines and the other is of even lines) from each frame, and then resized them to one half in the horizontal direction in order to keep the original aspect ratio. In this way, we obtained double the temporal resolution at the cost of losing one half of the spatial resolution.

After de-interlacing, the face area and location of the nose and two eyes on each frame are located by the Pittpatt face detection and tracking library [14],[15]. This area is then in-plane rotated so that the face will have straightened up pose.

On a located and straightened face, 83 facial feature points (on the face contour, eyes, nose and lips) are detected using an adaptation of active shape model (ASM). Face detection, in plane rotation and key point identification is shown for a frame from one of the test videos in fig. 1. A selected subset of these points were later used for face alignment and local SIFT feature extraction steps. The extracted faces were aligned using five key points in a least square sense. These points include, two eye-corners, one nose-tip and two mouth corners. For extraction of motion flow features, all detected key points were used.

Fig. 1. (a) is an input frame, (b) shows the result for face detection followed by in-plane rotation, (c) shows key points identification.

V. FEATURES

We extracted an ensemble of features from the faces detected in the videos. These include SIFT at selected key points, HG features and motion (Optic Flow) features. Following subsections give a brief description of these features.

A. Scale Invariant Feature Transform(SIFT)

Since its inception, SIFT [16] has proven its significance in a vast majority of applications in computer vision. We extracted SIFT descriptors on 7 selected key points. This subset of points consisted of the following: one point at the center of each eye, one point on the nose base and 4 points around the lips (two points at the lip corners and two points at the centers of upper and lower lips). The SIFT descriptor extracted on these points was concatenated into one long vector (resulting into a $128 \times 7 = 896$ dimensional feature vector). These points were selected based upon their better performance on the training data.

B. Hierarchical Gaussianization

The novel Hierarchical Gaussianization (HG) [17] representation is a locality sensitive patch-based approach, where the patch locality information is characterized jointly with the appearance information. The proposed representation has the distinct advantage of being insensitive to scale, pose and appearance. The robust representation is generic, which supports wide range of imagery processing. It is worthwhile to mention here that HG features have been a key component in our research group's top performance in a number of recent competitions. These include the following:

1) Large Scale Visual Recognition Challenge (First position), 2010
2) PASCAL Visual Object Classes Challenge (First position), 2009
3) StarChallenge Multimedia Retrieval Competition (Bronze Medal), 2008

In the process of extraction of these features, first we adopt a hierarchical GMM for feature vectors at difference levels: the whole corpus, each image and individual patches. We learn the image-specific GMM in a Bayesian framework to allow information sharing across different images and to bridge the universal and individual information retrievals. Given an image-specific GMM, each patch of that image is assigned to a Gaussian component with respect to a posterior probability. All these probabilities constitute a set of so-called *Gaussian maps* over the entire patch grid. After obtaining a GMM and Gaussian maps for each image which we term as a Hierarchical Gaussianization (HG) process, we extract the appearance information from the GMM parameters, and the spatial information from global and local summary statistics over Gaussian maps. Finally, all parameters of the GMM and statistics of the Gaussian maps are concatenated as a super-vector, followed by a supervised dimension reduction to further enhance the discriminating power of the representation. An illustration of this new representation is shown in figure 2.

The three major components of the representation: 1) Gaussian Mixture Model for appearance modeling; 2) Gaussian maps for spatial representation; and 3) Discriminant attribute projection, will be described respectively here.

1) GMMs for appearance representation: Let z denotes a p-dimensional feature vector from the I-th image. We model

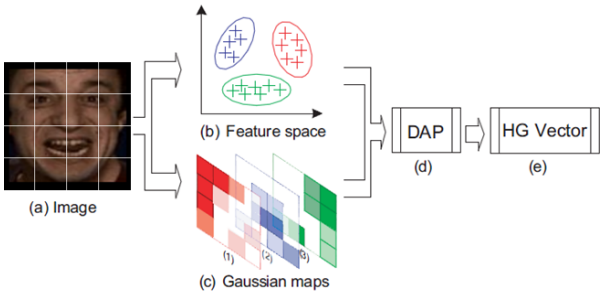

Fig. 2. (a) is an input image. (b) shows the patch features in the feature space. Each "+" denotes a feature vector, whose distribution is approximated by a GMM. (c) shows a set of Gaussian maps, each of which corresponds to one Gaussian component in (b). A supervised dimension reduction algorithm, DAP, is performed in (d) to form the final image representation, hierarchical Gaussianization vector.

z by a GMM, namely,

$$p(z|\Theta) = \sum_{k=1}^{K} w_k^I \mathcal{N}(z; \mu_k^I, \Sigma_k^I), \tag{1}$$

where K denotes the total number of Gaussian components, and $(w_k^I, \mu_k^I, \Sigma_k^I)$ are the image-specific weight, mean and covariance matrix of the kth Gaussian component, respectively. For computational efficiency, we restrict the covariance matrices Σ_k^I to be a diagonal matrix Σ_k shared by all images.

We estimate the prior mean vector μ_k, prior weights w_k and covariance matrix Σ_k by fitting a global GMM based on the whole corpus, and the remaining parameters by solving the following *Maximum A Posteriori* (MAP) loss,

$$\max_{\Theta} \Big[\ln p(z|\Theta) + \ln p(\Theta) \Big].$$

The MAP estimates can be obtained via an EM algorithm: in the E-step, we compute

$$Pr(k|z_i) = \frac{w_k^I \mathcal{N}(z_i; \mu_k^I, \Sigma_k)}{\sum_{j=1}^{K} w_j^I \mathcal{N}(z_i; \mu_j^I, \Sigma_j)}, \tag{2}$$

$$n_k = \sum_{i=1}^{N} Pr(k|z_i), \tag{3}$$

and in the M-step, we update

$$\hat{w}_k^I = \gamma_k n_k/N + (1-\gamma_k) w_k, \tag{4}$$

$$\hat{\mu}_k^I = \alpha_k m_k + (1-\alpha_k)\mu_k, \tag{5}$$

where

$$m_k = \frac{1}{n_k} \sum_{i=1}^{N} Pr(k|z_i) z_i,$$

$$\alpha_k = n_k/(n_k+r), \ \gamma_k = N/(N+T).$$

If a Gaussian component has a high probabilistic count, n_k, then α_k approaches 1 and the adapted parameters emphasize the new sufficient statistics m_k; otherwise, the adapted parameters are determined by the global model μ_k.

After Gaussinization, we can calculate the similarity between a pair of images via the similarity between two GMMs. In our experiments, we follow the suggestion in [21] and choose the appearance vector for an image, x^I, to be

$$m(x^I) = [\sqrt{w_1^I}\Sigma_1^{-\frac{1}{2}}\mu_1^I; \cdots; \sqrt{w_K^I}\Sigma_K^{-\frac{1}{2}}\mu_K^I]. \tag{6}$$

2) Gaussian maps for spatial representation: According to equation (2), the feature vector at each patch is again modeled by a mixture of Gaussians with a mixture probability $Pr(k|z_i)$. For a fixed k, all such probabilities $Pr(k|z_i)$ form a map over the patch locations, which is referred as a *Gaussian map*. For a GMM with K components, we have K Gaussian maps, and we can learn the spatial information of an image by analyzing each of these Gaussian maps.

We follow the suggestion in [17], hierarchically split a Gaussian map and extract summary statistics over local regions. Specifically, each of the K Gaussian maps is divided into subregions based on a sequence of increasingly coarser grids; assume there are M subregions in total, then we calculate some summary statistic ν over each of the M regions. As a parallel form to (6), we define $v(x^I)$, a vector expressing spatial information of image x^I as follows,

$$v(x^I) = [\nu_{11}^I; \cdots; \nu_{M1}^I; \nu_{12}^I; \cdots; \nu_{M2}^I; \cdots; \nu_{MK}^I] \tag{7}$$

3) Discriminant attribute projection: We concatenate the appearance vector $m(x^I)$ and the spatial vector $v(x^I)$ as a super-vector

$$\phi(x^I) = [m(x^I); v(x^a)],$$

To enhance the discriminating power of our representation, we project $\phi(x^I)$ to a subspace that depresses the directions with high inter-category variabilities. Let V denote the projection matrix toward the subspace with high inter-category variabilities, that is, $(I-V)\phi(x^I)$ is the discriminant projection we are looking for. We solve V via the following objective function

$$V = \arg\max_{V^T V = I} \sum_{i \neq j} ||V^T \phi(x^i) - V^T \phi(x^j)||^2 W_{ij}, \tag{8}$$

where W_{ij}=1 when x^i and x^j belong to the same category, otherwise $W_{ij} = 0$. Let $\Phi = [\phi(x^1), \phi(x^2), \cdots, \phi(x^N)]$, a matrix with N columns where N is the total number of training images. It can be shown that the optimal solution for V consists of the top eigenvectors corresponding to the largest eigenvalues of matrix $\Phi(D-W)\Phi^T$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}, \forall i$.

Suppose we use the dot product as a similarity measure between super-vectors. After applying discriminant attribute

projection (DAP), the similarity between two images, x^a and x^b, is equal to

$$D(x^a, x^b) = \phi(x^a)^T (\mathrm{I} - VV^T)\phi(x^b). \quad (9)$$

That is, the projection toward V, which is irrelevant to the classification, is discarded in the similarity calculation. Thus, the HG feature vector, used for classification in later stages, is essentially $\phi(x)$ with its those dimension suppressed which are irrelevant for classification.

C. Motion feature

From our experiments on the training videos, we found motion features useful in increasing the classification accuracy. We used local statistics of the optical flow of the regions of interest as the motion feature in each frame. Here optical flow estimation is used to compute an approximation to the motion field from the intensity difference of two consecutive frames. The main concern of using optical flow feature is that it usually requires heavy computational loading. For this reason we use the algorithm implemented on GPU [18] which decreases computation time by orders of magnitude.

Following steps are taken for extraction of motion feature:

1) compute motion vector of each pixel using optical flow computation algorithm,
2) rotate the optical flow field to aligned the optical flow with the key points,
3) crop out the seven regions of interest including two eyebrows, two eyes, nose, mouth and the residual part of the face by taking the convex hull of the respective key points (or in other words, by connecting the key points around each region),
4) compute means and variances of horizontal and vertical components of the optical flow of each region of interest and then concatenating them as a feature vector.

The final motion feature for a frame has 28 dimensions.

VI. CLASSIFICATION

The feature vectors for the frames from the training set, obtained by the concatenation of the features outlined in the precious section, were fed into SVM classifiers for training. The frames in which no face was detected or where the motion feature was not available were left out in both the training and testing stages (for instance, there cannot be motion flow feature for the first frame in each video, for obvious reasons). An image specific approach was primarily adopted. The final decision was done on the basis of majority voting.

A hierarchial approach was followed for expression classification. Thus person specific and person independent classifiers were trained. To better conceptualize our system, please refer to the flowchart given in fig. 3. Given a test video, in the first step, features were extracted, and in parallel it was determined whether the subject appearing in the video, appeared in the training set or not. If it did, then it was found which one it was. Based upon the decision, person specific or person independent classifier was used. Both the manual and automated person ID and verification were experimented with (since manual person ID was allowed). It turns out that both give similar performance for expression recognition.

The parameters for SVMs for expression recognition were all tuned on the training videos, by following a leave-one-video-out for videos of each subject for subject-dependant classification and leave-one-subject-out for subject-independent classification.

A. Person ID and Verification

Since a hierarchial approach is adopted, so classifiers were needed for automated person identification and verification. SVMs ([19] and [20]) with holistic features were used for this task. More specifically, resized images to 32×32, were used as features in this stage.

To find out whether the subject in a video was in the training set or not; we trained probability models using leave-one-video-out (classifier 'A') and leave-one-subject-out (classifier 'B') on the original training set using linear SVMs [19] for person identification. The frames from the video, which was left during training, were fed into the two classifiers (classifiers 'A' and 'B') for testing. The probability outputs from both of these classifiers were sorted. Since there were 7 classes (subjects) for classifier 'A', the sorted probability outputs from this classifier were truncated to six largest probability values. The sorted probability outputs from classifier 'B' (6 outputs as there were 6 classes) served to represent the case when the subject was not in training. Also the remaining sorted probability values from 'A' gave examples of probability values when the subject was indeed present in training. This was repeated for all the videos in training set.

The hypothesis for such an approach was that if the subject does appear in the training set then the probability values for the actual class (actual subject) would be very high and the rest will be quite small. On the contrary, the probability values would not be too high for one particular class if the subject does not appear in training.

After obtaining the probability values for each frame (where face was detected) in each video in the training set, as outlined above, an SVM classifier was trained on probability outputs. This was a binary classifier (classifier 'C'), that would decide if a subject appeared in the training set or not. Since the decision was to be made at the video level, a majority voting decision criterion was adopted.

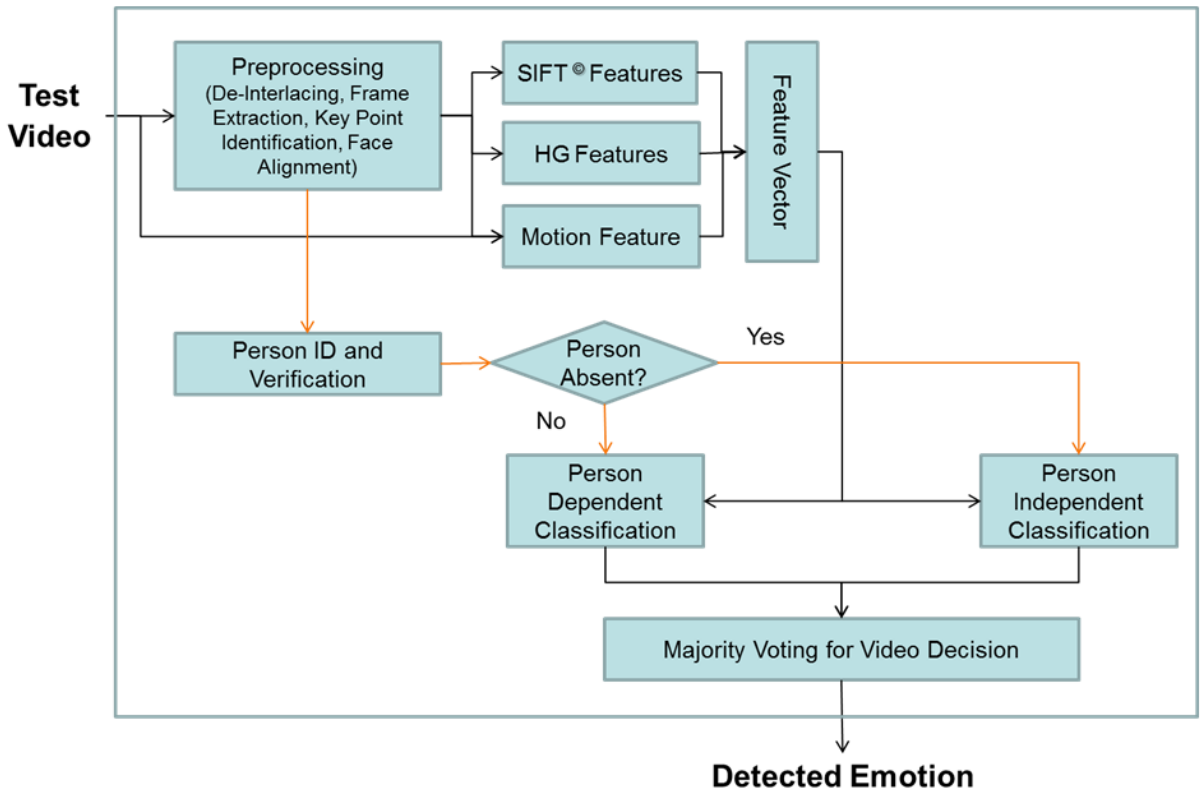

Fig. 3. Flowchart of our system

Once if it was established if a person in a video did appear in the training set; another classifier was used to establish its person ID (classifier 'D'). All the frames from the training set were used for training SVMs [20]. The parameters were tuned by adopting a leave-one-video out approach on the training set.

The above two stage classification procedure verifies the origin of the subjects and finds correctly the person ID of 129 out of 134 videos in test data-set. The five error cases all stem from the classifier 'C'. Four of the videos containing subjects who did not appear in the training set were labeled otherwise. While one video which contained a subject who appeared in the training was labeled otherwise.

Since manual person ID labeling was permitted, for the manual ID case, the videos were labeled manually to find out which videos contain subjects which appear in the training and what is their person ID.

B. Person Specific Results

This section details the results of the person specific classification. This approach was adopted for the videos of the subjects who were present in the training set. The 'present-or-absent' decision was done using manual or automated process. Then a classifier trained on the videos of a particular subject in the training data-set was selected, based upon the person ID. There were seven subject-wise multi-class SVM classifiers (corresponding to seven subjects in the training set). Also there were 54 such videos in the test set, where the subjects appeared in the training set as well.

The confusion matrix for the results with manual person ID is given in table I; while for the results with automated person ID and verification is given in table II. A comparison with the F1 scores in the baseline results is given in table III.

C. Person Independent Results

If the subject in a test video is not found to be present in the training stage (by manual or automated person ID), then we resort to the person independent results. There were 80 such videos in the test set, where the subjects did not appear in the training set. The classifier here is again a multi-class SVM. It is trained on all the feature vectors extracted from the training data-set. The parameters are tuned using a leave-one-subject out training procedure on the training data-set.

TABLE I
CLASS CONFUSION MATRIX FOR PERSON SPECIFIC EXPRESSION RECOGNITION CLASSIFICATION (WITH MANUAL PERSON ID)

Person Specific classifiers with Manual ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	13	0	0	0	0
	Fear	0	10	0	0	0
	Joy	0	0	11	0	0
	Relief	0	0	0	10	0
	Sadness	0	0	0	0	10

TABLE II
CLASS CONFUSION MATRIX FOR PERSON SPECIFIC EXPRESSION RECOGNITION CLASSIFICATION (WITH AUTOMATED PERSON ID AND VERIFICATION)

Person Specific classifiers with Auto. ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	13	0	0	0	0
	Fear	0	10	0	0	0
	Joy	0	0	11	0	0
	Relief	0	0	0	10	0
	Sadness	0	0	0	0	10

TABLE III
COMPARISON IN TERMS OF F1 SCORES WITH THE BASELINE RESULTS FOR PERSON SPECIFIC RESULTS

Emotion	Baseline	Manual P.ID	Automated P.ID
Anger	0.92	1.0	1.0
Fear	0.4	1.0	1.0
Joy	0.73	1.0	1.0
Relief	0.7	1.0	1.0
Sadness	0.9	1.0	1.0
Average	0.73	1.0	1.0

TABLE IV
CLASS CONFUSION MATRIX FOR PERSON INDEPENDENT EXPRESSION RECOGNITION CLASSIFICATION (WITH MANUAL PERSON ID)

Person Indep. classifier with Manual ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	9	2	0	0	4
	Fear	0	4	0	1	0
	Joy	3	7	19	1	0
	Relief	0	1	1	12	1
	Sadness	2	1	0	2	10

The confusion matrix for the results with manual person ID is given in table IV; while for the results with automated person ID and verification is given in table V. A comparison with the F1 scores in the baseline results is given in table VI.

D. Overall Results

This section lists the combination of the results obtained from the person specific and person independent classification. The class confusion matrix for the results with manual person ID is given in table VII while for the results with automated person ID is given in table VIII. A comparison with the F1 score of the baseline results is also given in table IX.

The overall, person specific and person independent *classification rate* is given in table X for manual person ID, while in table XI for automated person ID and verification.

VII. DISCUSSION

The thing which stands out from the comparison outlined in tables III, VI and IX, is the substantial improvement over the baseline performance. For instance, the average F1 score is 1.00 for person specific classification as in table III compared to the average baseline score of 0.73 for person specific performance. It highlights one important aspect that emotion recognition becomes much easier, if one has the training examples of the same person. May be because, every person exhibits facial expressions in a slightly different fashion.

The person independent results are also much better than the baseline. For instance the average baseline F1 score for person independent results is 0.44 (table VI). Whereas, our performance is 0.64 (table VI). The same trend translates to the overall results. Our average F1 score for the overall results is 0.80, while the baseline average F1 score overall is 0.56.

Another thing worth mentioning is that the automated person identification and verification does not distort the results by a significant amount, mainly because the person ID is fairly accurate. It reduces the average overall classification rate from 0.798 for manual person ID to 0.775 for automated person ID and verification (tables X and XI). Since the emphasis of this work is on emotion recognition and not on person verification, more novel approaches shall be adopted in the future to improve the automated person verification algorithm. Also, please note that the automated person identification and verification does not affect the person specific recognition performance (tables X and XI).

By looking at the class confusion matrices in tables IV, V, VII and VIII, one can notice that the worst performer is

TABLE V
CLASS CONFUSION MATRIX FOR PERSON INDEPENDENT EXPRESSION RECOGNITION CLASSIFICATION (WITH AUTOMATED PERSON ID AND VERIFICATION)

Person Indep. classifier with Auto. ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	9	4	1	0	4
	Fear	0	2	0	1	0
	Joy	3	7	18	1	0
	Relief	0	1	1	12	1
	Sadness	2	1	0	2	10

TABLE VI
COMPARISON IN TERMS OF F1 SCORES WITH THE BASELINE RESULTS FOR PERSON INDEPENDANT RESULTS

Emotion	Baseline	Manual P.ID	Automated P.ID
Anger	0.86	0.62	0.56
Fear	0.07	0.4	0.22
Joy	0.7	0.76	0.73
Relief	0.31	0.77	0.77
Sadness	0.27	0.67	0.67
Average	0.44	0.64	0.59

the fear emotion. It is confused more with joy emotion than anger. On the other hand, in terms of classification rate, the best performer is the joy emotion, as can be noted in tables X and XI. However, in terms of F1 scores, the best performer is the relief emotion. This can be noted in table IX. The reason for joy and relief performing better than others, may stem from the hypothesis that there is lesser variance in expressing joy and relief.

TABLE VII
CLASS CONFUSION MATRIX FOR OVER-ALL CLASSIFICATION (WITH MANUAL PERSON ID)

Overall classification with Manual ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	22	2	0	0	4
	Fear	0	14	0	1	0
	Joy	3	7	30	1	0
	Relief	0	1	1	22	1
	Sadness	2	1	0	2	20

TABLE VIII
CLASS CONFUSION MATRIX FOR OVER-ALL CLASSIFICATION (WITH AUTOMATED PERSON ID AND VERIFICATION)

Overall classification with Auto. ID		Ground Truth				
		Anger	Fear	Joy	Relief	Sadness
Predicted	Anger	22	4	1	0	4
	Fear	0	12	0	1	0
	Joy	3	7	29	1	0
	Relief	0	1	1	22	1
	Sadness	2	1	0	2	20

TABLE IX
COMPARISON IN TERMS OF F1 SCORES WITH THE BASELINE RESULTS FOR OVER-ALL RESULTS

Emotion	Baseline	Manual P.ID	Automated P.ID
Anger	0.89	0.80	0.76
Fear	0.20	0.70	0.63
Joy	0.71	0.83	0.82
Relief	0.46	0.86	0.86
Sadness	0.52	0.80	0.80
Average	0.56	0.80	0.77

TABLE X
CLASSIFICATION RATE FOR EMOTION DETECTION WITH MANUAL PERSON ID

Emotion	Person independent	Person specific	Overall
Anger	0.643	1.000	0.815
Fear	0.267	1.000	0.560
Joy	0.950	1.000	0.968
Relief	0.750	1.000	0.846
Sadness	0.667	1.000	0.800
Average	0.655	1.000	0.798

TABLE XI
CLASSIFICATION RATE FOR EMOTION DETECTION WITH AUTOMATED PERSON ID AND VERIFICATION

Emotion	Person independent	Person specific	Overall
Anger	0.643	1.000	0.815
Fear	0.133	1.000	0.480
Joy	0.900	1.000	0.935
Relief	0.750	1.000	0.846
Sadness	0.667	1.000	0.800
Average	0.619	1.000	0.775

VIII. CONCLUDING REMARKS

In essence, this paper highlights the strength of our features and classification methodology over the baseline method. The dense-patch based feature, HG; the key-point based feature, SIFT; and motion feature, optical flow; are complementary to each other. The three sets of features, when evaluated on the training data separately, yielded worse performance. Their combination did indeed improve the results on the training data. Also, as expected, the person dependent emotion recognition shows better performance than person independent. By adopting our person ID based strategy, our system can automatically switch between person dependent and person independent classifiers, and therefore their combination achieves better performance.

ACKNOWLEDGMENT

We thank all of our lab-mates including Dennis J. Lin, Liangliang Cao, Shen-Fu Tsai and others for their useful suggestions.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, January 2009.

[2] P. Ekman and W. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[3] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image Vision Comput.*, vol. 24, pp. 605–614, June 2006.

[4] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," Vienna, Austria, pp. 377–416, July 2007.

[5] M. S. Bartlett, G. Littlewort, B. Braathen, T. J. Sejnowski, and J. R. Movellan, "A prototype for automatic recognition of spontaneous facial actions," in *Advances in Neural Information Processing systems*, 2003, pp. 1271–1278, MIT Press.

[6] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions." *IEEE Trans Syst Man Cybern B Cybern*, vol. 36, no. 1, pp. 96–105, 2006.

[7] M. F. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *SMC (1)*, 2004, pp. 635–640.

[8] Y.-L. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 97 – 115, February 2001.

[9] H. Tang and T. Huang, "3d facial expression recognition based on automatically selected features," in *3DFace08*, 2008, pp. 1–8.

[10] H.Tang and T.S.Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *FG'08*, 2008, pp. 1–6.

[11] T. Banziger and K. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus," in *Blueprint for affective computing: A sourcebook* (in press), K. R. Scherer, and T. Banziger, and E. Roesch, Ed.: Oxford University Press, Oxford, England.

[12] M. Valstar, B. Jiang, M. Mhu, M. Pantic, and K. Scherer, "The First Facial Expression Recognition and Analysis Challenge," in *Proc. IEEE Intl Conf. Automatic Face and Gesture Recognition* (in print), 2011.

[13] A. C. Bovik, *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. Academic Press, Inc. Orlando, FL, USA, 2005.

[14] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection." in *CVPR (2)'04*, 2004, pp. 29–36.

[15] H. Schneiderman, "Learning a restricted bayesian network for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2004.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[17] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang, "Hierarchical gaussianization for image classification," in *2009 IEEE 12th International Conference on Computer Vision*, 2010, pp. 1971–1977.

[18] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L 1 optical flow," in *Proceedings of the 29th DAGM conference on Pattern recognition*. Springer-Verlag, 2007, pp. 214–223.

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[20] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[21] Zhou, X., S. Yan, S. Chang, M. Hasegawa-Johnson, and T. Huang, 2008: SIFT-bag kernel for video event analysis. ACM New York, NY, USA.