Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image

Songfan Yang, *Student Member, IEEE*, and Bir Bhanu, *Fellow, IEEE*

***Abstract*—Existing video-based facial expression recognition techniques analyze the geometry-based and appearance-based information in every frame as well as explore the temporal relation among frames. On the contrary, we present a new image-based representation and an associated reference image called the emotion avatar image (EAI), and the avatar reference, respectively. This representation leverages the out-of-plane head rotation. It is not only robust to outliers but also provides a method to aggregate dynamic information from expressions with various lengths. The approach to facial expression analysis consists of the following steps: 1) face detection; 2) face registration of video frames with the avatar reference to form the EAI representation; 3) computation of features from EAIs using both local binary patterns and local phase quantization; and 4) the classification of the feature as one of the emotion type by using a linear support vector machine classifier. Our system is tested on the Facial Expression Recognition and Analysis Challenge (FERA2011) data, i.e., the Geneva Multimodal Emotion Portrayal-Facial Expression Recognition and Analysis Challenge (GEMEP-FERA) data set. The experimental results demonstrate that the information captured in an EAI for a facial expression is a very strong cue for emotion inference. Moreover, our method suppresses the person-specific information for emotion and performs well on unseen data.**

***Index Terms*—Avatar reference, emotion avatar image (EAI), face registration, person-independent emotion recognition, Scale-invariant feature transform (SIFT) flow.**

I. Introduction

FACIAL expression plays a significant role in human communication. It is considered the single most important cue in the psychology of emotion [1]. Automatic recognition of emotion from images of human facial expression has been an interesting and challenging problem for the past 30 years [2]. Aiming toward the applications of human behavior analysis, human–human interaction, and human–computer interaction, this topic has recently drawn even more attention.

A literature review shows that early-stage research on facial expression recognition focused on static images [2], [3]. Both feature-based and template-based approaches were investigated. Recently, researchers have been using image sequences or video data in order to develop automated expression recognition systems. As demonstrated in the fields of computer vision [4]–[7] and psychology [8], [9], various types of dynamic information, such as dynamic appearance and dynamic geometry, are crucial for the recognition of human expressions.

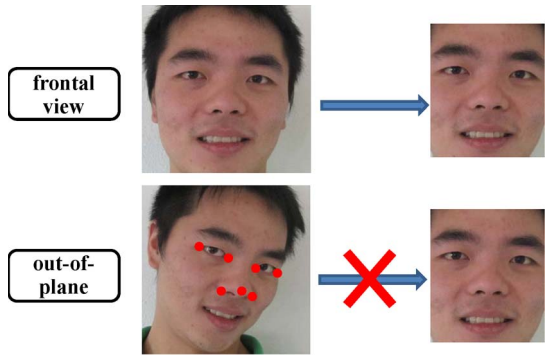

Fig. 1. Existing face registration techniques cannot handle out-of-plane head rotation.

However, extracting the facial dynamics from an expression sequence is not a trivial problem. There are two critical questions: First is how to aggregate the dynamic information from expressions of varying lengths and to create features with fixed length, and second is how to perform alignment since capturing the dynamics requires near perfect alignment for the head pose and facial features. The inherent challenge for facial expression recognition is the dilemma between compensating the rigid motion of the head pose and extracting the nonrigid motion of facial muscles. Most existing algorithms and real-time computer programs [10], [11] are only capable of analyzing a frontal face with a near upright angle. This is not due to the failure to detect a face but due to the failure to register the detected face reasonably in a video.

As shown in Fig. 1, when a subject's face is in frontal view, near frontal view, or has in-plane rotation, the alignment can be done easily by in-plane image transformation. We can detect both eye locations, scale the distance of both eyes to a constant value for every subject, and then rotate the image to guarantee that both eyes are horizontally aligned. Finally, we can translate the entire image such that the eyes are located at some predefined locations. This registration technique is suitable for some early-stage research experiments where facial expression data are acquired under controlled conditions. One restriction is that in the collected data, not much head movement should be involved. To accomplish this, data are collected by cameras mounted on the subject's head to eliminate the head motion [12].

Three types of data are used in the facial expression recognition community, namely posed data, acted data, and spontaneous data. For data sets that are collected from a stationary

Manuscript received June 1, 2011; revised November 14, 2011 and February 17, 2012; accepted March 6, 2012. Date of publication May 7, 2012; date of current version July 13, 2012. This work was supported in part by the National Science Foundation under Grant 0727129 and Grant 0903667. This paper was recommended by Associate Editor M. Pantic.

The authors are with the Center for Research in Intelligent Systems, University of California, Riverside, CA 92521 USA (e-mail: syang@ee.ucr.edu; bhanu@cris.ucr.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2012.2192269

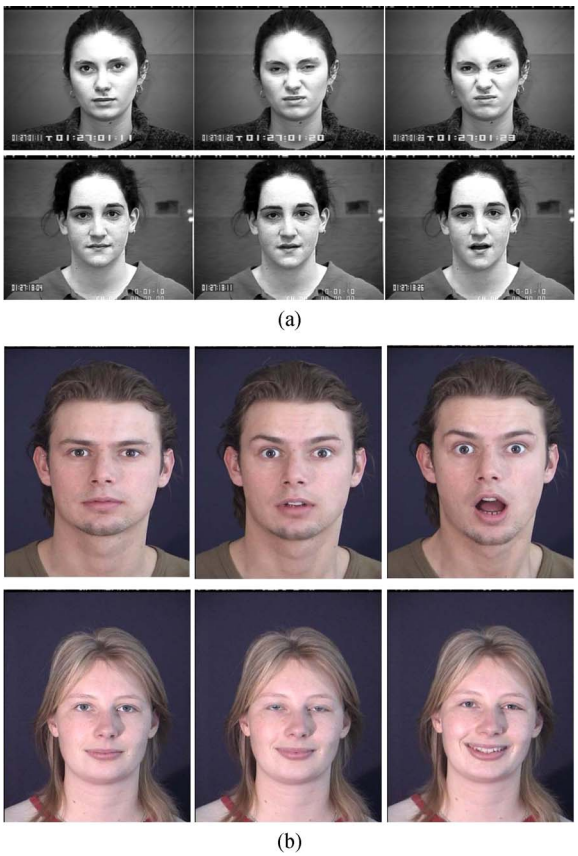

Fig. 2. Sample sequence for posed data. Very little head motion is involved. (a) CK data set [14]. (b) MMI data set [13].

camera such as the web-based database for facial expression analysis (MMI) data set [13] and the Cohn–Kanade (CK) data set [14] (see Fig. 2), the subjects show facial expressions with minimum head movement and therefore help researchers to focus on the nonrigid facial muscle movement. Thus, these data sets fall into the category of posed facial expressions, meaning that the subjects are given "instructions" before showing expressions. Subjects are conscious about controlling their facial muscle movement. All the expressions start from a neutral face, which provides a good reference for computing the nonrigid facial motion. However, experiments demonstrate that, in human–human interaction such as conversation, people tend to adapt their head movements and facial expressions in response to the stimulus [15]. This is a strong evidence of the fact that facial expression is correlated with head motion. This fact is also true in a more realistic data set such as the GEMEP-FERA challenge data set [16]. Therefore, registration techniques should take care of out-of-plane head rotation for realistic data.

One technique that state-of-the-art algorithms use is 2-D affine transformation. A number of facial "anchor points" are defined whose motion is relatively stable during facial expressions. Such anchor points include eye locations, inner and outer eye corners, and the tip of the nose. We could also define a corresponding target location for each anchor point. Once the anchor points are detected, the affine transformation matrix could be computed by minimizing the sum of the least square error of detected location and target location of the anchor points. This affine transform is subsequently applied to the entire face image to complete the registration step.

The affine-based registration performs quite well when in-plane or minor out-of-plane head motion is present. However, the anchor points are not entirely stable during a facial expression. The eye corner could be unstable if the subject is blinking or the tip of the nose could also be moving and so forth. The typical number of anchor point is around six. If not all points are detected correctly, a large transformation error will be generated, and the affine transformation of the original image will be unacceptable.

Moreover, affine-based registration is not temporally smooth. If a minor change occurred to an anchor point for two consecutive face images, the affine transform matrix will be off by a small amount. After applying this affine transform to the entire face image, every single pixel is affected due to this minor change. This will result in a fake motion for the stationary face regions. Therefore, the entire dynamic analysis based on this registration method will be imprecise.

Another registration technique is through the active appearance model (AAM) [17], [18]. The automatic AAM fitting process may perform poorly for person-independent cases. Thus, it may require manual labeling of a subset of the frames for each data [19] to improve the fitting result, which is undesirable in an automated system. Recently, a person-independent AAM approach has been developed [20]; however, this technique can be inaccurate due to false feature localization.

A significant issue in addition to face registration is the person-independent property (subjects in the test data are not used for training) of the algorithm. Computer algorithms cannot be trained with data for all human beings. The generalization ability must allow the system to predict for the unseen people. Thus, the computer algorithms are expected to extract person-invariant features. This property enables the system to carry out facial expression recognition from a person-dependent (or person-specific) environment to a person-independent environment.

The person-specific information, including facial geometry and facial appearance, can be eliminated at two steps in a system: face registration and feature extraction. In-plane image-transformation-based registration techniques do not change the geometry or appearance of facial features; therefore, the person-specific information is retained. An AAM-based approach can warp the facial appearance and align the landmark points. The affine-transformation-based registration algorithms are able to change the geometry and the appearance of a person to a limited extent. When a face is in a near frontal view (where the affine-based registration accomplishes the most plausible result) and only a simple transformation is needed, the face geometry is mostly unaltered. When faces are not in the frontal view (contain out-of-plane head rotation), the affine-based algorithm is able to change the face geometry by a large amount, but unfortunately, that is when this approach performs poorly and when most of the registration results are unacceptable.

The person-specific information can be also eliminated through feature extraction. Features that are extracted could be categorized into geometry based and appearance based. Geometry-based approaches track the geometry of landmark

points over time and use their geometric relations or dynamics as the feature. If the locations of the facial landmark points are normalized and only the amount of location change is considered to be the feature, it falls into the category of a person-independent feature. For instance, emotion "joy" is typically accompanied with a smile, which results in the mouth expanding and the lip corner being pulled up [21]. However, the geometry-based inference is sensitive to out-of-plane head rotation. This rigid head motion may disguise the true motion of the landmark points; therefore, it generates a large error in the extracted feature. On the other hand, the appearance-based approaches, such as local binary patterns (LBPs) [22], Gabor wavelets [23], and local phase quantization (LPQ) [24], concentrate on the dense response of filters to the intensity values of a face. These methods are inherently person dependent unless person-dependent information is eliminated during the face registration process.

The challenges aforementioned encourage us to develop a system that accurately registers face images even with out-of-plane rotation and, at the same time, eliminates the person-specific information. 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 a sequence of frames. If we can find a single good image representation based upon which we make judgements, we would be able to infer the emotion expressed through a sequence of facial expressions in a computationally efficient manner.

In this paper, we have adopted the recently introduced SIFT flow algorithm [25] to register the facial images. By matching the dense SIFT descriptors across image pairs, this method is able to generate satisfactory alignment results for facial features. Although the SIFT flow is originally designed for image alignment at the scene level, it is reasonable to apply it here to facial expression recognition since a human face can be considered as a scene in this case. It is capable of globally aligning the head/face region while maintaining the shape and motion of facial features for consecutive frames. In order to solely extract the facial motion information irrespective of person-specific information, we iteratively build a single "avatar reference" face model, onto which we align all the face images. Later, we update the avatar reference face model and also the single good representation, i.e., the emotion avatar image (EAI), for each video consisting of frames for an emotion. We name the model avatar because the subjects are morphed toward homogeneity, whereas the emotions are successfully retained. Subsequently, the EAIs are individually passed through LBP and LPQ texture descriptors for feature extraction. Finally, support vector machines (SVMs) with a linear kernel are used for classification. Our approach transforms the expression recognition problem from an image sequence back to a single image.

In what follows, we first discuss the related work, motivation, and the contribution of this paper (see Section II). Subsequently, we introduce the effectiveness of the data for facial expression and our iterative algorithm to build the avatar reference and EAIs in Section III. Two combinations of methods are tested, and the classification results of different techniques are compared in Section IV. The conclusions of this paper is provided in Section V.

II. Related Work, Motivation, and Our Contributions

A. Related Work

A large amount of effort has been focused on describing facial expression features. Based on the feature in use, as introduced earlier, we can broadly divide the methods into three categories, i.e., geometry-based approaches, appearance-based approaches, and the combination of the two. Geometry-based approaches track the facial geometry information based on a set of facial landmark points over time and classify expressions based on their deformation. On the other hand, appearance-based approaches use information from the facial texture described by various types of texture descriptors, such as LBP, Gabor wavelets, and LPQ. The dynamics of the texture deformation can also be included for feature extraction. In Table I, a thorough comparison of methods from the literature based on the usage of registration techniques, feature types, dynamic features, classifiers, and the data set is provided.

In this paper, the methods that are compared with the proposed method are listed in Table II. In this table, we also analyze their registration techniques, features, and classifiers similar to the comparison shown in Table I. In addition, we include the features and classifiers that are adopted in these papers. Later, in Section IV, we provide a comparison of the methods on the same data, which is the GEMEP-FERA challenge data set [16].

B. Motivation

Based on how the data are acquired, it can be categorized into three classes: posed data, acted data, and spontaneous data. When posed data are collected, subjects are given a series of "instructions" such as emphasize on the facial muscle movement and try not to move the head. Posed data played an important role in the early-stage research, because it provided researchers with more insights about the relation of expression to the muscle movement. The CK database [14] and the MMI database [13] fall into this category.

The ultimate goal of the research community is to recognize spontaneous facial expressions. However, spontaneous data are very hard to acquire. Facial expressions can be called spontaneous when subjects are not aware that they are being recorded, and naturally express emotions. Since it is very difficult to design a fully unaware environment when collecting data, no spontaneous data set coded with explicit emotions is publicly available.

The intermediate stage between the previous two, namely, the acted data, has less control than the posed data, but subjects are fully aware when data are being recorded. The GEMEP-FERA challenge data set [16] that this paper used belongs to this class and is shown in Fig. 3. In the process of data collection, subjects are not asked to control themselves but just to convey a certain emotion. These experiments have no control about the body pose, the head pose, the gesture, or occlusion and are therefore very challenging for expression recognition by an automated system.

TABLE I
COMPARISON OF SELECTED FACIAL EXPRESSION RECOGNITION TECHNIQUES

Authors	Registration	Feature	Dynamic feature	Classifier	Dataset
Yacoob and Davis [26]	not mentioned	**geometry**: region-based optical flow	yes	rule-based classifier	posed data: 32 subjects
Essa and Pentland [27]	3D model fitting	**geometry**: 3D motion and muscle models	yes	Euclidean norm	posed data: number of subjects not specified
Wang et al. [28]	not mentioned	**geometry**: B-spline curve	yes	Euclidean norm	posed data: 8 subjects
Hu et al. [29]	not mentioned	**geometry**: variation of active shape model	yes	probabilistic model	posed data: subject number not mentioned
Valstar et al. [30]	affine transform	**geometry**: dynamics of 20 facial points	yes	probabilistic actively learned Support Vector Machine (SVM)	MMI [13] and CK [14]
Pantic et al. [31]	affine transform	**geometry**: dynamics of 15 facial profile points	yes	rule based	MMI [13]
Donato et al. [21]	in-plane image transform	**appearance**: Gabor wavelets	no	nearest neighbor	posed data: 24 subjects
Zhao and Pietikäinen [4]	in-plane image transform	**appearance**: LBP on three orthogonal planes (LBP-TOP)	yes	SVM	CK [14]
Bartlett et al. [32]	in-plane image transform	**appearance**: Gabor wavelets	no	SVM and Adaboost	CK [14] + posed data: 24 subjects
Wu et al. [5]	in-plane image transform	**appearance**: Gabor motion energy	yes	SVM	CK [14]
Tian et al. [33]	in-plane image transform	**hybrid**: geometric + transient facial features	no	Neural network	CK [14] + posed data
Lucey et al. [19]	Active Appearance Model(AAM)	**hybrid**: 2D shape + 2D appearance + 3D shape	no	nearest neighbor or SVM	acted data: Rutgers University2 (RU), 20 subjects
Zhou et al. [34]	AAM	**hybrid**: geometry + SIFT [35]	yes	multidimensional assignment algorithm	CK [14] + RU (20 subjects)

TABLE II
COMPARISON OF METHODS PROPOSED BY DIFFERENT TEAMS WHOSE PAPERS WERE ACCEPTED IN THE FERA CHALLENGE [36]. THE RANKED RESULTS ON THE SAME GEMEP-FERA CHALLENGE DATA SET ARE GIVEN IN Fig. 11

Paper ID	Registration	Feature	Dynamic feature	Classifier
UCR: EAI [37]	SIFT flow [25]	**appearance**: LQP [22] or LBP [24]	no	SVM
UIUC-UMC [38]	affine transformation	**appearance**: SIFT + hierarchical Gaussianization + motion	no	SVM
KIT [39]	in-plane image transformation	**appearance**: discrete cosine transform (DCT)	no	SVM
UCSD-CERT [40]	affine transformation	**appearance**: Gabor wavelets	no	SVM
ANU [41]	constrained local model	**appearance**: pyramid of histogram of gradients + LPQ	no	SVM
UCL [42]	in-plane image transformation	**appearance**: edge orientation histogram of motion history histogram + motion change frequency of LBP	yes	SVM_2K [43]
UMont. [44]	in-plane image transformation	**appearance**: histograms of oriented gradients	yes	SVM
NUS [45]	in-plane image transformation	**hybrid**: accumulated motion image	no	SVM
QUT [46]	constrained local model	**appearance**: pixel based feature + LBP	no	SVM
Baseline [47]	in-plane image transformation	**appearance**: LBP	no	SVM
MIT-Cambridge [48]	not mentioned	**hybrid**: geometry + Gabor wavelets	no	Hidden Markov Model + Dynamic Bayesian Network

To motivate our method, we analyze the specifications of the GEMEP-FERA data set as follows:

- Ten subjects (five males and five females) are involved with their upper body visible.
- Subject's age is approximately between 25 and 60 years, as judged by observation.
- Video resolution is 720 $\times$ 576, and face resolution is around 200 $\times$ 200 pixels.
- Average video length is about 2 s with a frame rate of 30 fps.
- Each video contains one subject displaying expressions corresponding to a certain emotion.
- Five emotions are involved: Anger, Fear, Joy, Relief, and Sadness. This is different from the typical six basic emotions data sets.
- There are three to five videos for each subject with the same emotion.
- Most subjects are uttering meaningless phrases while displaying an expression [47].

Fig. 3. Uncontrolled acted data from the GEMEP-FERA data set. These data are used for testing.

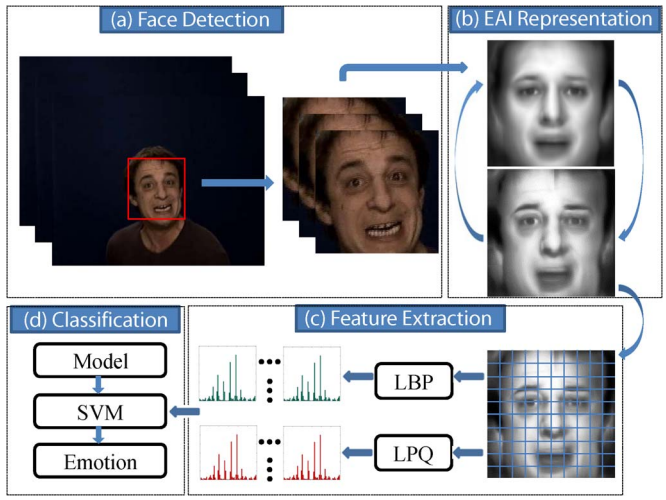

Fig. 4. Overview of our approach.

- Videos do not start with the neutral face or end at the apex or the offset. This is unlike the CK and MMI data sets.
- Multiple apexes are involved in some videos.
- The neutral face is not always available.

The given observations provide us the following key facts that inspire our system:

1) Good registration is demanding, and previous registration techniques (in-plane image transformation and affine-based transformation) are not suitable for this data set.
2) Dynamical changes are hard to recover because the neutral reference face is not always available.
3) Constant lip motion limits the geometry-based approaches.

C. Our Contributions

Existing work intensely emphasizes on analyzing the sequential change of the facial feature. Nevertheless, since the onset and the offset for realistic data are 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 and describe emotions in a reliable manner.

The contributions of this paper are the following. First, iteratively build a reference face model called the avatar reference. This homogenous reference face model can capture the nature of the entire data set. Second, condense a video sequence into a single image representation, i.e., an EAI, for facial expression recognition. The EAI representation registers the facial features at meaningful locations and maintains the nonrigid facial muscle movement. Third, the EAI representation is capable of aggregating dynamic facial expression information with various lengths into fixed length features. Fourth, being able to suppress the person-specific information, the EAI representation also allows the expression recognition tasks to be carried out in a person-independent manner.

To the best of our knowledge, until now, little work has been done to condense a video sequence into a tractable image representation for emotion recognition. As the results in Section IV show, our algorithm can distinguish most of the differences between expressions, as long as the expressions are not so subtle that even the human visual system is unable to detect them.

III. Technical Approach

In Fig. 4, our approach is outlined in four major steps. After automatically extracting faces from a raw video, we provide some insights about our EAI representation that suppresses the person-specific information while maintaining the shape and texture information on the facial features. Both LBP and LPQ texture descriptors are applied to generate the features; then, the linear SVM classifiers are used for classification. The model used for testing is trained with a 1-versus-1 SVM.

A. Face Detection

We first extract the face from the video using the Viola and Jones face detector [49] implemented in OpenCV. This algorithm achieves high-quality performance and is suitable for real-time processing. The detection rate is near perfect on the GEMEP-FERA [16] data set. Since the face resolution is around 200×200 pixels, we resize the detected face image exactly to this resolution using bicubic interpolation. This process removes the noise and smoothes the raw images.

B. EAI Representation

1) SIFT Flow Alignment: SIFT flow has been recently introduced in [25]. It is originally designed to align an image to its plausible nearest neighbor, which can have large variations. The SIFT flow algorithm robustly matches dense SIFT features between two images while maintaining spatial discontinuities.

In [25], the local gradient descriptor SIFT [35] is used to extract a pixelwise 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 eight 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. Similar to optical flow, the objective energy function that we attempt to minimize is designed as

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

$$+ \sum_{\mathbf{p}} \eta\left(|u(\mathbf{p})| + |v(\mathbf{p})|\right) \quad (2)$$

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

$$+ \min\left(\alpha\,|v(\mathbf{p}) - v(\mathbf{q})|, d\right) \quad (3)$$

where $\mathbf{p} = (x, y)$ is the grid coordinates of the images and $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at $\mathbf{p}$. $u(\mathbf{p}), v(\mathbf{p})$ is the flow vector for the x-direction and the y-direction, respectively. s_1 and s_2 are two SIFT images to be matched. ε contains all the spatial neighbors (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 $\mathbf{w}(\mathbf{p})$. 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 in (3) takes care of the similarity of flow vectors for adjacent pixels. In this objective function, the 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. η and α are scale factors for the small displacement and the smoothness constraint, respectively.

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

Two frames with a minor pose difference are shown in Fig. 5(a). We align the target frame with respect to a reference frame. For comparison purpose, we separately take the absolute difference between images before alignment and after alignment with respect to the reference. Comparing the two difference images in Fig. 5(a), the rigid head motion from the minor pose change is eliminated. Nevertheless, the difference image also shows that the SIFT flow alignment process is noisy.

Consider a case with a major pose change in Fig. 5(b), the head pose motion is out of plane, and the facial appearance significantly changes. The registration result is in the upright pose, and nonrigid motion in the mouth and eye areas can still be captured. Differences at the periphery are due to the lack of correspondences for SIFT flow vectors. However, this information is still useful as it captures the pose change, which is also an important cue in facial expression recognition [15]. Differences at the periphery show that the pose change and the true facial feature motion are separated. Similar to the minor pose change case, the noise and discontinuity are issues in the aligned result.

2) Avatar Reference and the EAI: SIFT flow has the potential to align images with large spatial variation. This is useful in aligning the face image given the possibility of a large head pose change or occlusion. However, the person-specific information still has to be eliminated. We seek to build a reference face with respect to which each face image can be aligned.

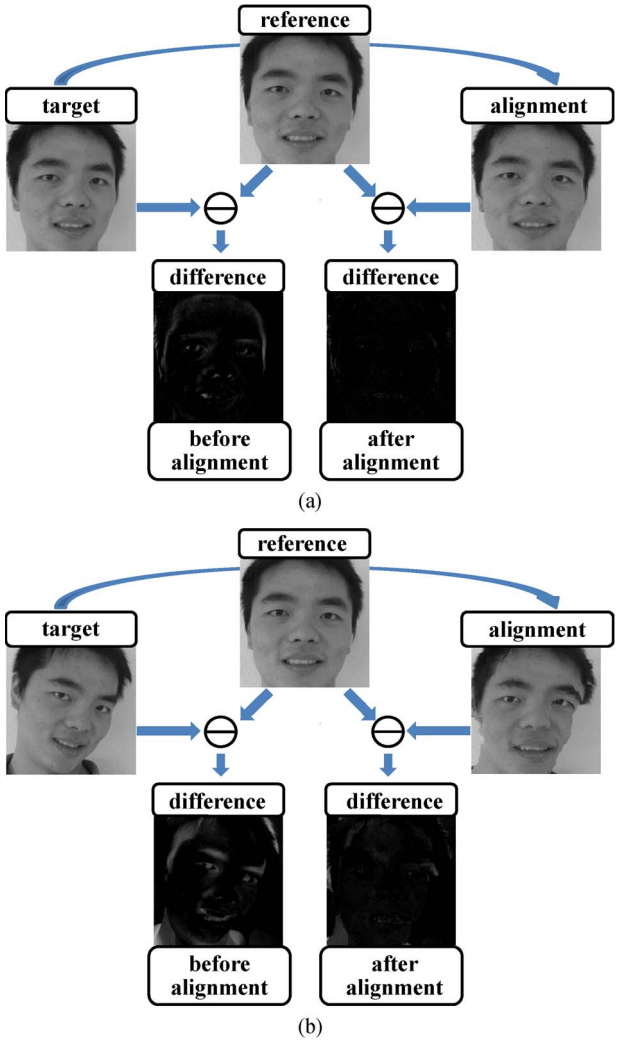

Fig. 5. SIFT flow face registration performs well when the pose change is small or large. It captures the facial muscle motion in both cases but the results are very noisy. (a) Minor difference. Only true facial motions are captured as shown by the corresponding difference image of before alignment and after alignment. (b) Major difference. (bottom right) Difference image of the reference and the alignment result shows the true facial motions are captured in the inner eye corner areas.

Algorithm 1 Avatar Reference and EAI

Given:

$I^{(m,n)}$: face image from sequence m, frame n

M: total number of image sequences

N_m: total number of frames in sequence m

Q: user-defined number of levels

A_i^{ref}: Avatar Reference at level-i

EAI_i^m: EAI representation for sequence m based on the level-i Avatar Reference A_i^{ref}

$I_{\text{align}}^{(m,n)}$: the alignment result for a face image $I^{(m,n)}$ using SIFT flow

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

for $i = 1 \rightarrow Q$ **do**

for $m = 1 \rightarrow M$ **do**

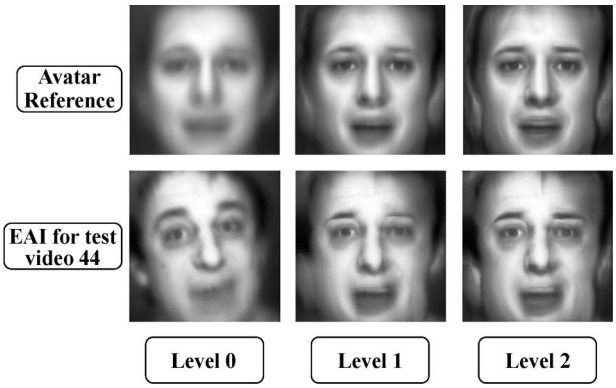

Fig. 6. Avatar reference face model and EAI representations for the first three levels. For comparison, level-0 EAIs are the average of every face image from their corresponding videos without alignment. Higher levels of EAI have more facial feature details and a homogenous face model.

for $n = 1 \rightarrow N_m$ **do**
$I_{\text{align}}^{(m,n)} \leftarrow SIFTflow(I^{(m,n)}, A_{i-1}^{\text{ref}})$
end for
$EAI_i^m \leftarrow 1/N_m \sum_{n=1}^{N_m} I_{\text{align}}^{(m,n)}$
end for
$A_i^{\text{ref}} \leftarrow 1/\sum_{m=1}^{M} \sum_{m=1}^{M} EAI_i^m$
end for

In Algorithm 1, we design an iterative averaging method to generate an avatar reference face model. To put it simply, we initialize our algorithm by averaging all possible face images in the training data set. Initially using this average face as the reference, we align each face image in the video using SIFT flow. After alignment, the user can update the avatar reference using all the aligned faces. The iteration number defines the level of the avatar reference (level 0 means the average of all the unaligned face images). The avatar reference models for the first three levels are shown in row 1 in Fig. 6. From our observation, the avatar reference is not always a neutral face. It captures the most likely facial appearance throughout the whole data set; therefore, it has less total variation in registration. The mouth is open for the level-1 and level-2 avatar reference face results (as shown in row 1 in Fig. 6). This is because most of the subjects in the training data are uttering meaningless phrases [16] and therefore have a lot of mouth movement.

In Algorithm 1, once the avatar reference face model is obtained, we establish the single-representation EAI for the sequence of face images at the current level. As demonstrated earlier, a single-aligned face image possesses errors and discontinuities. Therefore, we describe an image sequence as the average of all frames within this sequence. The statistical justification of the EAI representation is similar to [50]. Assume that the distribution of every aligned face frame is subject to an addition of a true face and additive noise. The noise is further assumed to be Gaussian. During the averaging process, the noise variance is reduced by a factor of N, where N is the number of face images. Thus, the alignment noise can be removed from our EAI representation.

3) Characteristics of EAI: In this paper, we attempt to test the performance of EAIs at different levels. As shown in Fig. 6 (row 2), the quality of the EAIs improves as the level of avatar reference becomes higher. A high-level avatar reference model enhances the facial details, corrects the rigid head pose change, and attenuates the person-specific information. Meanwhile, EAI representation retains the expression information that is recognizable by the human visual system. The EAI representations for five subjects with different emotions are shown in Fig. 7 (due to publication permission issue, we only show sample EAI representations for a subset of emotions in the CK+ data set. Please refer to [37] for the similar figure for the Facial Expression Recognition and Analysis Challenge (FERA) challenge data set). Since all the faces are aligned with respect to the same avatar reference, the EAI representation can be seen to align facial features, such as the nose, eyes, and the mouth reasonably. This lays the foundation for extracting meaningful facial feature motion. In addition, aligning every face image with the avatar reference allows the elimination of the person-specific information to a great extent.

The EAIs in Fig. 7 can also be observed to capture the nonrigid facial feature motion and the corresponding facial expression information. This is due to the small constraint intensity parameter η in (2). Larger values of η will penalize the large flow vectors more, which will result in less morphing for the alignment result. Ideally, if two face images are perfectly aligned, all the facial features should be at exactly the same locations. The facial feature motion will be eliminated in this case. In practice, the real facial feature motions during an expression are larger than the SIFT flow compensation and, subsequently, can be maintained in the noisy alignment results. The accumulation process will smooth the alignment results while capturing the real motion caused by a facial expression.

The reasons we decide to use EAI are given below. First, it is a morphed version or incarnation of the original person. Its identity is altered through the change of facial geometry. Facial features for every person are warped to a common reference. Second, the representation maintains the original emotion conveyed through facial expression. Thus, an emotion avatar is a subset of an avatar. Third, it is an image representation and not a 3-D model. The avatar reference and EAI are related as described in Algorithm 1.

C. Feature Extraction

The EAI representation allows us to represent the recognition problem with a single image rather than a video. To test the effectiveness of our single-image-representation EAI, we describe the facial texture from EAI using the well-known texture descriptor LBP and the recently proposed blur-insensitive LPQ descriptor. We expect to receive similar improvements for both methods.

1) LBP: The LBP is a powerful and well-known texture descriptor. In this paper, we used the extended version of the basic LBP in [22], where the LBP descriptor is uniform and grayscale invariant. To briefly go over this extended work, the operator, which is denoted as $\text{LBP}_{P,R}^{u2}$, is applied to a circularly symmetric neighborhood with P number of pixels on the circle of radius R. Superscript "$u2$" denotes the uniform property. A uniform LBP is favorable since it reduces the feature dimension. For example, the $\text{LBP}_{8,1}^{u2}$ adopted in this

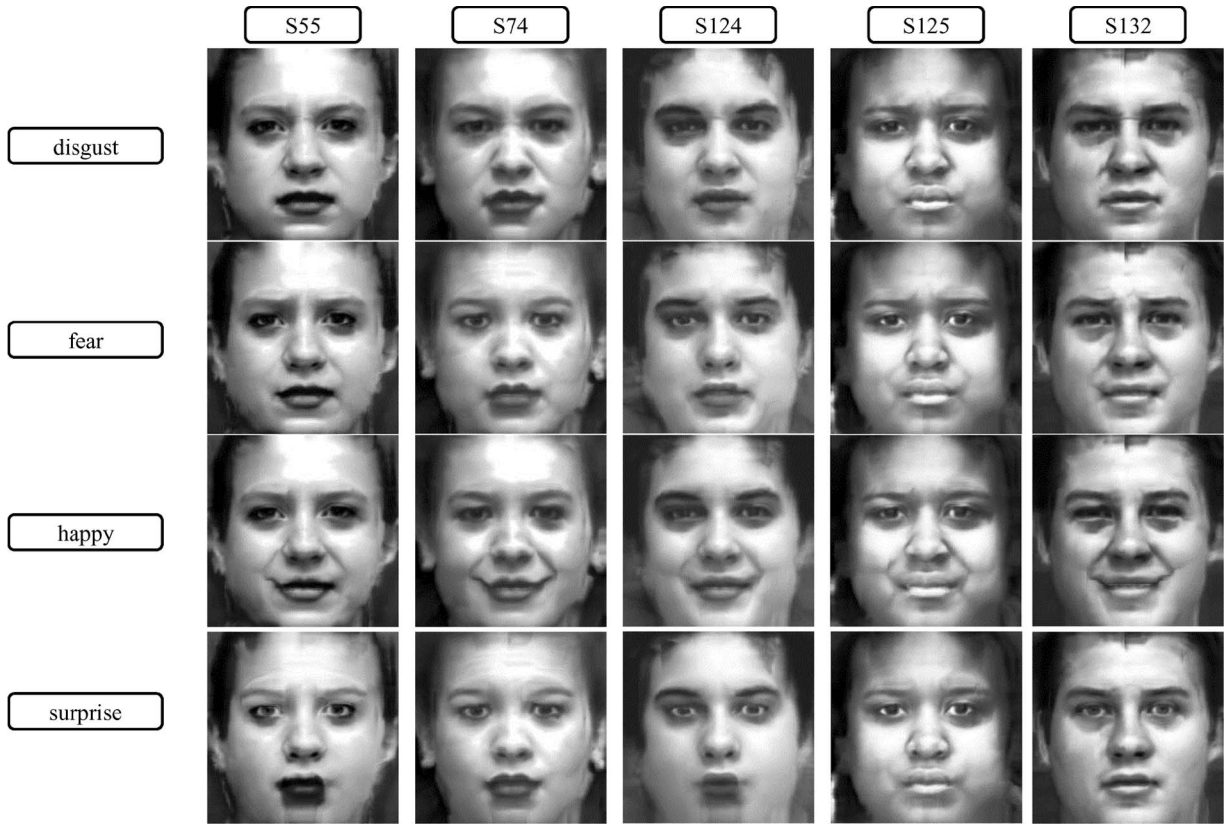

Fig. 7. Level-2 EAI representation for subjects in the CK+ data set. The facial features are reasonably aligned, and person-specific information is attenuated.

paper will generate 59 basic patterns, whereas the $\text{LBP}_{8,1}$ has 256 possibilities. Since these parameter settings are used in the baseline method [47], we adopt the same settings for better comparison.

After thresholding each pixel in its 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 resize the face images to 200×200, and each image is divided into blocks of size 20×20 blocks to capture the local texture pattern. Therefore, the LBP feature vector in use is of dimension $59 \times 10 \times 10 = 5900$. As mentioned earlier, the face resolution is close to 200×, hence we resize all face images to this uniform value to minimize the information loss.

2) LPQ: The blur insensitive LPQ descriptor is originally proposed in [24]. The spatial blurring is represented as multiplication of the original image and a point spread function (PSF) in the frequency domain. The LPQ method is based upon the invariant property of the phase of the original image when the PSF is centrally symmetric.

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

$$F(\mathbf{u}, \mathbf{x}) = \sum_{y \in N_{\mathbf{x}}} f(\mathbf{x} - \mathbf{y}) e^{-j2\pi \mathbf{u}^T y} = w_{\mathbf{u}}^T f_{\mathbf{x}} \qquad (4)$$

where $\omega_{\boldsymbol{u}}$ is the basis vector of the 2-D Discrete Fourier transform at frequency $\boldsymbol{u}$, and $f_{\boldsymbol{x}}$ is another vector containing all M^2 image samples from $N_{\boldsymbol{x}}$.

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

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

The phase information is recovered by a scalar quantizer, i.e.,

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

where $g_j(\boldsymbol{x})$ is the jth component of the vector $G_{\boldsymbol{x}} = [Re\{\boldsymbol{F}_{\boldsymbol{x}}\}, Im\{\boldsymbol{F}_{\boldsymbol{x}}\}]$. The resulting eight binary coefficients $q_j(\boldsymbol{x})$ are represented as integer values between 0–255 using binary coding as follows:

$$f_{\text{LPQ}}(\mathbf{x}) = \sum_{j=1}^{8} q_j(\mathbf{x}) 2^{j-1}. \qquad (6)$$

In addition, the decorrelation process is added to the original LPQ implementation to eliminate the dependence among the neighboring pixels. Similar to LBP, we divided the 200×200 face image into size 20×20 regions. Therefore, the LPQ feature vector is of dimension $256 \times 10 \times 10 = 25600$.

D. Classification

We train our multiclass linear SVM classifier [51] in the 1-versus-1 manner. The cost parameter C is chosen to be 1 for our system for the reason that, as demonstrated in Fig. 8, the

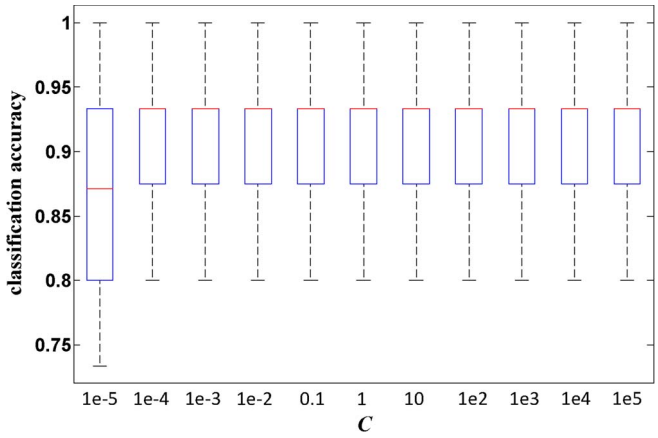

Fig. 8. Box plot of the tenfold cross-validation result on 155 GEMEP-FERA training data with respect to different values of the SVM parameter C.

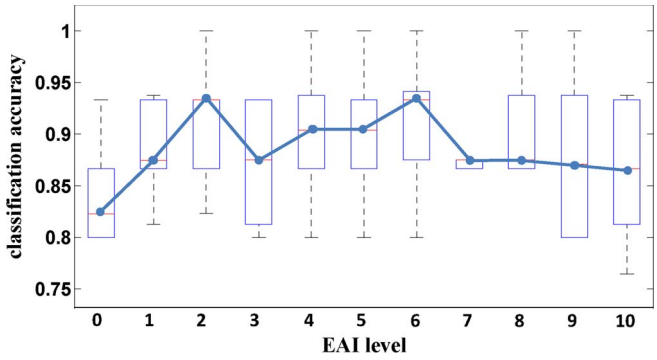

Fig. 9. Box plot of tenfold cross-validation results on 155 training videos using different levels of EAIs. The average classification rate is connected for the LPQ texture descriptor to show the improvement at each level. This is to demonstrate that we adopt level-2 EAIs because of its potential to good performance and relative computational efficiency.

tenfold cross-validation accuracy will not be degraded if C is not extremely small.

The iterative EAI algorithm is only executed during training. In the test phase, we register the detected faces using SIFT flow with respect to the level-1 avatar reference obtained from the training phase. Summing up all the registered faces from one sequence generates the corresponding level-2 EAI. We then extract LBP and LPQ texture features from every EAI representation for classification.

The reason why we use level-2 EAI face model is statistically demonstrated in Fig. 9. We carry out a series of tenfold cross-validation experiments on only the training of the GEMEP-FERA data set using the first 11 levels of the EAIs and the test on the LPQ texture descriptor. The cross-validation procedure results in person-specific category because we do not exclude the test subjects from the training. In Fig. 9, it is shown that the performance improves as the level of the EAI increases for the first three levels. This is consistent with our discussion on the avatar reference level in Section III-B. The performance peaks at both levels 2 and 6. After analyzing the avatar reference and the corresponding EAI representation, the overfitting issue occurs to the avatar reference as the level increases, as shown in Fig. 10. Artifact facial details are excessively displayed through

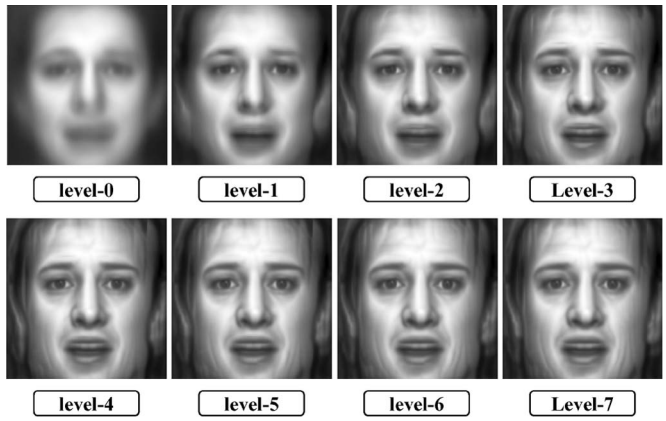

Fig. 10. Avatar reference from levels 0 to 7. A higher level of the avatar reference will have excessive facial details due to overfitting. Level 1 is used in our system.

the higher number of iteration in Algorithm 1. The system with level-6 EAI may not have a good generalization to unseen data.

IV. Experimental Results

A. System Implementation

Similar to our previous work [37], after extracting the faces from the raw data using the face detector in [49], the face images are then aligned to the level-1 avatar reference face model based on Algorithm 1, and the single-representation EAIs are generated. Subsequently, using both LBP and LPQ operators, we separately extract the feature from all the EAIs. Specifically, $\text{LBP}^{u2}_{8,1}$ is used in our experiment. The parameters for the LPQ operator are $M = 9$, $a = 1/7$, and $\rho = 0.9$. Lastly, as demonstrated in Section III-D, the classifier we used is the 1-versus-1 linear SVM [51] classifier with $C = 1$.

B. Challenge Evaluation Protocol

Our method and ten other methods (including the baseline [47]) are compared using the FERA2011 data, i.e., the GEMEP-FERA data set [16]. As part of the FERA2011 challenge, 155 training videos were given out a month before the deadline. Then, the 134 test videos were released one week before the deadline. There are seven subjects in the training data and six subjects in the test set, three of which are not present in the training set [16]. We ran the test videos using our system that takes each video session as the input and outputs the emotion label. All predicted labels were then submitted to the organization panel of FERA2011. After evaluation, the results were provided in three different categories: person independent, person specific, and overall.

C. Challenge Results

The confusion matrices for the EAI using the LPQ operator are shown in Tables III–V, with test results on person independent, person specific, and overall, respectively. Similarly, the confusion matrices for EAI using the LBP operator are presented in Tables VI–VIII.

TABLE III
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		0.86	0.47	0.95	0.69	0.8
average rate		**0.75**				

TABLE IV
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		0.92	1	1	0.91	1
average rate		**0.96**				

TABLE V
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		0.93	0.68	0.94	0.81	0.84
average rate		**0.84**				

TABLE VI
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		0.86	0.53	0.8	0.88	0.47
average rate		**0.71**				

TABLE VII
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		0.85	0.9	0.73	1	0.9
average rate		**0.87**				

TABLE VIII
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		0.85	0.68	0.77	0.92	0.64
average rate		**0.77**				

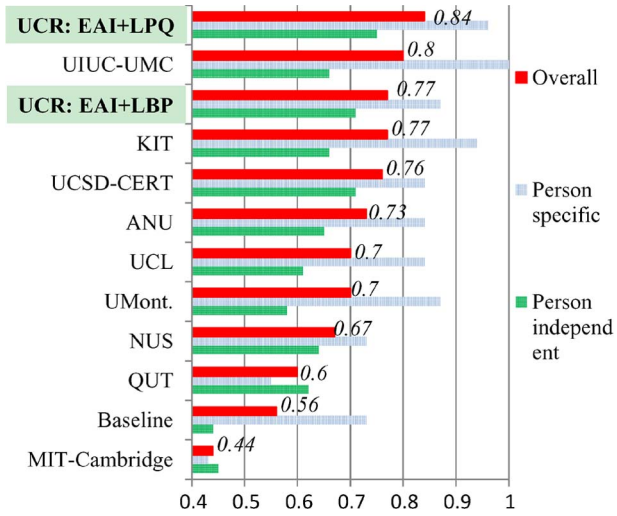

Fig. 11. Comparison of classification results in the *primary test* for person-specific, person-independent, and overall cases [36]. Teams are ranked based on the overall performance (numbers are labeled). Our (UCR) EAI methods ranked the first place and the third place for LPQ and LBP, respectively. **UCR**: University of California at Riverside; **UIUC-UMC**: University of Illinois at Urbana-Champaign; University of Missouri; **KIT**:Karlsruhe Institute of Technology; **UCSD-CERT**: University of California at San Diego; **ANU**: Australian National University; **UCL**: University College London; **UMont.**: University of Montreal; **NUS**: National University of Singapore; **QUT-CMU**: Queensland University of Technology; Carnegie Mellon University; **MIT-Cambridge**: Massachusetts Institute of Technology; University of Cambridge.

In Fig. 11, it is shown that our EAI representation combined with LPQ and LBP descriptors rank the first and third places, respectively, in the *primary test*. Our approach achieves the highest classification rate in the person-independent test (0.75 using EAI + LPQ). This is a positive evidence that our approach eliminates the person-specific information and captures the facial expression information. In addition, this demonstrates the desired ability of EAI for predicting the unseen data in real applications. In the person-specific test, our method achieves 96% classification accuracy. In the training data, each subject displays the same expression three to five times. The EAI representation achieves consistency when a subject displayed the same expressions in different videos [37].

Since the ground-truth label for each emotion video is easy to tell, the FERA2011 organizer required a *secondary test* where no participant can see the data. We submitted our facial expression recognition system program using EAI + LPQ to the organizer. Secondary test data are approximately half the size of the primary test set. Our approach achieves an 86% overall classification rate [36], which is consistent with the primary test.

The inherent characteristic of our approach is to eliminate facial dynamics while maintaining the emotion information. Unlike most of the other approaches [40], [48] which treat 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 method only considers them as 155 EAIs. Given more training videos, our system will most likely be improved since 155 videos of five emotions (approximately 30 videos/emotion on average) may not be sufficiently large to represent a single emotion across a large population.

TABLE IX
CONFUSION MATRIX FOR CK+ DATA SET.
(AN = ANGER, CO = CONTEMPT, DI = DISGUST, FE = FEAR, HA = HAPPY, SA = SADNESS, SU = SURPRISE)

		truth						
		An	Co	Di	Fe	Ha	Sa	Su
prediction	An	**81.8**	11.1	3.6	4.0	0.0	29.6	0.0
	Co	2.3	**55.6**	0.0	0.0	0.0	0.0	1.2
	Di	6.8	0.0	**85.7**	4.0	1.6	11.1	0.0
	Fe	0.0	5.6	1.8	**40.0**	0.0	0.0	0.0
	Ha	0.0	0.0	3.6	32.0	**98.4**	0.0	0.0
	Sa	2.3	5.6	1.8	0.0	0.0	**48.1**	0.0
	Su	6.8	22.2	3.6	20.0	0.0	11.1	**98.8**
average rate		**82.6**						

D. Evaluation on the CK+ Data Set

We also evaluated our system (implemented with the combination of level-2 EAI and LPQ) using 316 sequences from 123 subjects in the CK+ [14] data set. Seven emotion categories (Anger, Contempt, Disgust, Fear, Happy, Sadness, and Surprise) are included in this data set. No subject with the same facial expression has been collected more than once. We carry out leave-one-subject-out cross-validation experiment so that it belongs to the person-independent category. The confusion matrix is shown in Table IX. The average classification accuracy is 82.6%, which is consistent with our person-independent test result for the FERA data set in Table III.

Our algorithm performs not as good as in [4], [5], and [34] on this data set. The reasons are as follows. First, each sequence in the CK+ data set has only one apex, which reduces the intensity of the expression. The EAIs for CK+ look relatively neutral compared with the EAIs for the GEMEP-FERA data set. Second, the frontal view face images from the CK+ data set do not need sophisticated registration techniques. Thus, good dynamic facial features can be easily captured. Therefore, those approaches that use dynamic features outperform our approach that is based on simple features computed from the EAI representation. However, in a more realistic case where a good registration result is difficult to achieve (such as the GEMEP-FERA), our approach outperforms the approaches using complex dynamic features [42], [44]. Third, the training data might not be sufficient. We plot the relation between the number of training examples for each emotion category and the corresponding classification rate in Fig. 12. For classes Anger, Disgust, Happy, and Surprise, where training examples are greater than 40, the corresponding classification rate is significantly higher than that from the categories Contempt, Fear, and Sadness. We can expect an improvement of performance for a larger number of training instances.

E. Discussion

In more general cases such as spontaneous facial expression, facial feature motion is more subtle, and the temporal boundaries for expression are difficult to determine. As demonstrated in Section III, the registration process using SIFT flow can capture small changes in facial expressions if the changes are not extremely subtle. With respect to the temporal boundary issue, depending on application of the system, a facial expression can be segmented based on a single expression label or multiple labels. On the one hand, if a single-label assignment is acceptable for an application, it is possible to sample the data based on the appearance change and to learn the temporal boundary [52]. On the other hand, if the application needs to capture subtle information and multiple labels are required, one can consider learning the relation between different labels and the appearance feature.

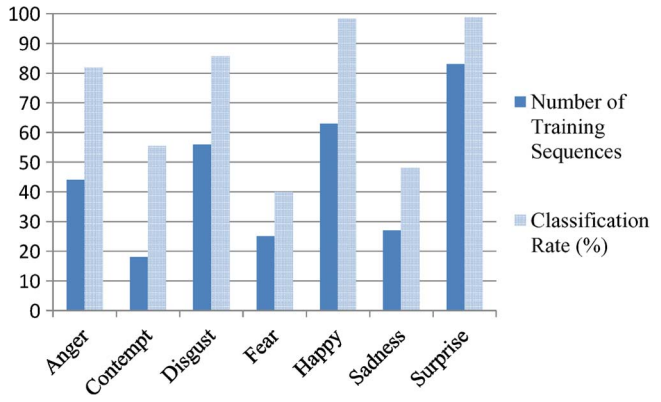

Fig. 12. Relation between the number of training images and the classification rate. The semantic meanings of the y-axis are different for the two classes. The classification rates for categories with more training example are significantly higher.

In the process of developing a real-time system, several issues need to be addressed. The avatar reference is created during the training phase. During the test phase, the detected faces are directly aligned with respect to the avatar reference using SIFT flow. As discussed in the previous paragraph, the EAIs can be computed given a temporal buffer resulting from the resampling process. The real question is that whether SIFT flow can be implemented in real time or not. The dense SIFT descriptor can be computed in a parallel fashion, whereas loopy belief propagation cannot. However, if we can lower the face resolution from 200×200 (as used in this system) and sacrifice a small amount of the recognition rate, it is possible to carry out SIFT flow in real time.

V. CONCLUSION

Given the temporal segmentation of a video, we explore the new idea of condensing a video sequence into a single EAI representation. We adopt SIFT flow for aligning the face images, which is able to compensate for large rigid head motion and maintain facial feature motion detail. Then, an iterative algorithm is used to generate an avatar reference face model onto which we align every face image. We experimentally demonstrate that the level-2 EAI has the potential to generate a higher classification rate. Our EAI representation combined with LPQ and LBP texture descriptors achieved excellent performance in both person-independent and person-specific cases when tested on the challenging facial expression recognition data set, i.e., the GEMEP-FERA data set. Given the consistency of our EAI representation, the performance of our approach is dramatically improved when compared with the baseline [47] and other approaches [38], [40]–[42], [44]–[46], [48]. In the future, we will incorporate larger data in our system. To generalize our system, we will also study on how to automatically segment a video in a meaningful manner.

ACKNOWLEDGMENT

The authors would like to thank the organizers of FERA2011 Grand Challenge for providing the training data, test data, and evaluating the results. The author also like to thank the provider of the CK+ data set at Carnegie Mellon University. All this effort has been highly valuable in standardizing the evaluation of different approaches and in advancing the field of facial expression recognition.

REFERENCES

[1] J. A. Russell and J. M. Fernández-Dols, *The Psychology of Facial Expression*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[2] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[3] J. Yu and B. Bhanu, "Evolutionary feature synthesis for facial expression recognition," *Pattern Recog. Lett.*, vol. 27, no. 11, pp. 1289–1298, Aug. 2006.

[4] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[5] T. Wu, M. Bartlett, and J. Movellan, "Facial expression recognition using Gabor motion energy filters," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshop Human Commun. Behav. Anal.*, Jun. 2010, pp. 42–47.

[6] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 258–273, Feb. 2010.

[7] P. Yang, Q. Liu, and D. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–6.

[8] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychol. Sci.*, vol. 16, no. 5, pp. 403–410, May 2005.

[9] J. N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Pers. Social Psychol.*, vol. 37, no. 11, pp. 2049–2058, Nov. 1979.

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "Computer expression recognition toolbox," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Mar. 2011, pp. 298–305.

[11] R. El Kaliouby and P. Robinson, "Mind reading machines: Automated inference of cognitive mental states from video," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2004, vol. 1, pp. 682–688.

[12] M. Pantic and L. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1449–1461, Jun. 2004.

[13] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005, p. 5.

[14] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2000, pp. 46–53.

[15] S. M. Boker, J. F. Cohn, B.-J. Theobald, I. Matthews, T. R. Brick, and J. R. Spiesaff, "Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars," *Philosoph. Trans. B Roy. Soc.*, vol. 364, no. 1535, pp. 3485–3495, Dec. 2009.

[16] Challenge Data. [Online]. Available: http://sspnet.eu/fera2011/fera2011data/

[17] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[18] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[19] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn, "AAM derived face representations for robust facial action recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2006, pp. 155–160.

[20] J. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 29–Oct. 2, 2009, pp. 1034–1041.

[21] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999.

[22] T. Ojala, M. Pietikäinen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[23] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *Image Vis. Comput.*, Jun. 2006, vol. 24, no. 6, pp. 615–625.

[24] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. ICISP*, Berlin, Germany, 2008, pp. 236–243.

[25] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[26] Y. Yacoob and L. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 636–642, Jun. 1996.

[27] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, Jul. 1997.

[28] M. Wang, Y. Iwai, and M. Yachida, "Expression recognition from time-sequential facial images by use of expression change model," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Apr. 1998, pp. 324–329.

[29] C. Hu, Y. Chang, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image Vis. Comput.*, vol. 24, no. 6, pp. 605–614, Jun. 2006.

[30] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshop Human Comput. Interact.*, Oct. 2007, pp. 118–127.

[31] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[32] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2006, pp. 223–230.

[33] Y.-I. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[34] F. Zhou, F. De la Torre, and J. Cohn, "Unsupervised discovery of facial events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 2574–2581.

[35] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[36] FERA2011 Challenge. [Online]. Available: http://sspnet.eu/fera2011/

[37] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 866–871.

[38] U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang, X. Lv, and T. X. Han, "Emotion recognition from an ensemble of features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 872–877.

[39] T. Gehrig and H. Ekenel, "A common framework for real-time emotion recognition and facial action unit eetection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshop Human Comput. Interact.*, Jun. 2011, pp. 1–6.

[40] G. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett, "The Motion in Emotion—A CERT based approach to the FERA emotion challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 897–902.

[41] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 878–883.

[42] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze, "Emotion recognition by two view SVM 2K classifier on dynamic facial expression features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 854–859.

[43] H. Meng, J. Shawe-Taylor, S. Szedmak, and J. Farquhar, "Support vector machine to synthesise kernels," in *Deterministic and Statistical Methods in Machine Learning*, J. Winkler, M. Niranjan, and N. Lawrence, Eds. Berlin/Heidelberg, Germany: Springer-Verlag, 2005, pp. 242–255.

[44] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 884–888.

[45] R. Srivastava, S. Roy, S. Yan, and T. Sim, "Accumulated motion images for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 903–908.

[46] S. W. Chew, P. J. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 915–920.

[47] M. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 921–926.

[48] T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. el Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshop Facial Express. Recog. Anal. Challenge*, Mar. 2011, pp. 909–914.

[49] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[50] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[51] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[52] A. Cruz, B. Bhanu, and S. Yang, "A psychologically-inspired match-score fusion model for video-based facial expression recognition," in *Proc. Int. Conf. HUMAINE Assoc. Affective Comput. Intell. Interact. Workshop 1st Int. Audio/Visual Emotion Challenge*, Oct. 2011, pp. 341–350.

Songfan Yang (S'10) received the B.S. degree in electrical engineering from Sichuan University, Chengdu, China, and the B.S. degrees in computer science and electronic engineering technology from Eastern New Mexico University, Portales, in 2009. He is currently working toward the Ph.D. degree in electrical engineering with the Center for Research in Intelligent Systems, University of California, Riverside.

His research interests include computer vision, pattern recognition, machine learning, human behavior understanding, and facial expression/emotion understanding.

Mr. Yang was a recipient of the Best Entry Award in the FG 2011 Facial Expression Recognition and Analysis Emotion Sub-Challenge competition.

Bir Bhanu (S'72–M'82–SM'87–F'95) received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge; the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles; and the M.B.A. degree from the University of California, Irvine.

He is the Distinguished Professor of electrical engineering; a Cooperative Professor of computer science and engineering, mechanical engineering, and bioengineering; and the Director of the Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory with the University of California, Riverside (UCR). In addition, he serves as the Director of National Science Foundation (NSF) The Integrative Graduate Education and Research Traineeship program on Video Bioinformatics, UCR. He has been the Principal Investigator of various programs for NSF, Defense Advanced Research Project Agency, National Aeronautics and Space Administration, Air Force Office of Scientific Research, Office of Naval Research, Army Research Office, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine–vision applications. He is the author or coauthor of 400 reviewed technical publications, including over 100 journal papers and 40 book chapters; seven published books; and three edited books. He is the holder of 18 (five pending) patents. His current research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human–computer interactions, and biological, medical, military and intelligence applications.

Dr. Bhanu is Fellow of American Association for the Advancement of Science, International Association of Pattern Recognition, and The International Society for Optics and Photonics.