Synthesis of Realistic Facial Expressions Using Expression Map

Swapna Agarwal and Dipti Prasad Mukherjee, *Senior Member, IEEE*

***Abstract*—For synthesis of realistic facial expressions displaying emotions, we need an efficient representation of pure (e.g., *surprise*) as well as mixed (e.g., *happily surprised*) emotional expressions. In this paper, we train an expression map (XM) that efficiently represents the emotional expressions. We propose an algorithm that utilizes the XM to synthesize emotional expressions, tailor-made for the facial structure of the target person. The proposed method can also control the proportions of different basic emotional expressions of those, when mixed together, to generate realistic emotional facial expressions. Unlike many existing methods, our expression synthesis model requires only one expression-neutral face image of the target person. Both qualitative and quantitative tests on four data sets show promising results. On average, we have achieved 92.4% correct validation of the expressions synthesized by our method. We also show that for both basic and mixed emotional expressions, our method generates finer expression details compared to existing state-of-the-art works.**

***Index Terms*—Expression map, mixed expression synthesis, facial structure, realistic facial expression synthesis.**

I. Introduction

THE field of facial expression synthesis has made significant progress. Successful models [1]–[4] have been developed to transfer facial expressions of emotions acted by an actor (source) to the face image of the target person. Such models are expensive as an actor is involved. Such models often assume that a wide variety of facial expressions of the target person is available. Collecting this set may not be always feasible. There also exist synthesis models [5], [6] that require only one expression-neutral face image of the target person. Such methods generally synthesize expressions of basic emotions. But in real life, faces show mix of basic emotions. For example, one may be *happily* or *fearfully surprised*. Therefore, we need a model that (a) can synthesize realistic expressions representing both basic as well as mixed emotions and (b) requires only one face image of the target person as input. In this paper we present such a model. In [7] we introduced *Expression Map* (XM). The XM was shown to be useful for estimating the percentage of different basic emotions in a given facial expression. For example, a facial expression may show 40% *happiness* and 60% *surprise*. The current proposed method utilizes the XM to synthesize a specified combination of basic emotional expressions on a given expression-neutral face image. The AMD features, as used in [7], being histogram in nature are not suitable for synthesis. For synthesis, we use a special set of features from [8] representing facial shape, texture, face structure and expression intensity. The main difference between the features used in [8] and here is that in the proposed method the texture feature is divided into low and high frequency components which are processed separately. This helps us in reducing the noise in the synthesized facial expressions. Using these features we train the XM following the learning process of Self Organizing Map (SOM) [7], [9]. We visualize expression as a change in facial appearance (shape and texture) from the expression-neutral face. This change gets stored as a pattern in the node of the trained XM. We propose an objective function that chooses a specific pattern from a node of the trained XM given an emotion (to be synthesized) and an expression-neutral target face image. For synthesizing expression representing the mix of two basic emotions, our method takes weighted combination of two patterns represented by the two nodes of the XM. Thus, the expressive face, synthesized by adding the chosen combination of patterns to the expression-neutral target face image, looks realistic and preserves subject identity.

Fig. 1 presents examples of basic and mixed expressions synthesized by the proposed approach. Observe (Fig. 1) that the synthesized expression looks natural on the target face. Different steps of synthesizing expression are depicted in the form of block diagram in Fig. 2 and Supplementary Fig. S1.

Our main contributions in this paper are as follows.

- We propose an algorithm that, using the XM, can synthesize basic and mixed (pre-specified combinations of basic emotions) emotional expressions on a given expression-neutral face image. The synthesized expressions are rich in details such as the appearance/disappearance of wrinkles, furrows, teeth *etc.* and look natural on the facial structure of the target face.

Manuscript received July 19, 2016; revised June 2, 2017 and November 6, 2017; accepted August 12, 2018. Date of publication September 19, 2018; date of current version March 22, 2019. The work of Swapna Agarwal was supported by DST, Government of India project no. SR/WOS-A/ET-53/2012(G). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qi Tian. *(Corresponding author: Swapna Agarwal.)*

The authors are with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: swapna_r@isical.ac.in; dipti@isical.ac.in).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes some supporting figures and tables which are referred to in the main paper. This material is 11.6 MB in size.

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/TMM.2018.2871417

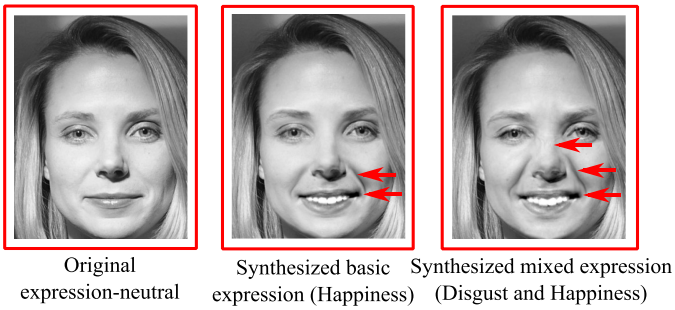


Fig. 1. Left: Expression-neutral target face, middle: Synthesized face displaying basic emotion *happiness*, right: Synthesized face displaying mixed emotion *disgust* with *happiness*. The arrows point to the changes in appearance due to imposition of expressions.

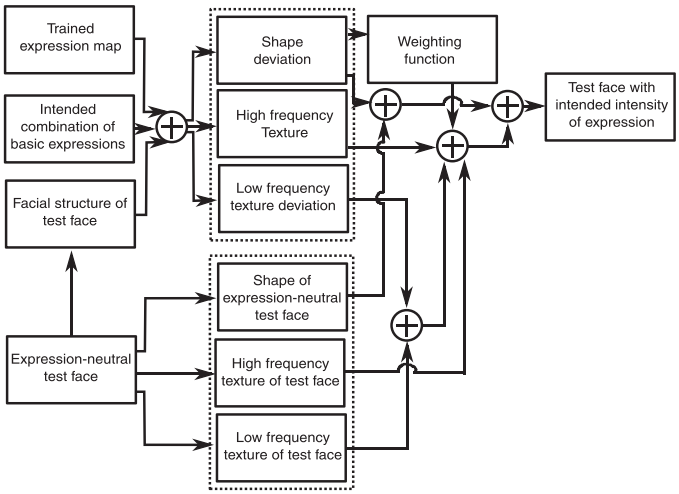


Fig. 2. Block diagram depicting different steps of synthesizing expression on a given expression-neutral face image using the trained expression map.

- The proposed algorithm needs only one expression-neutral face image of the target person.

Related works are presented in the next section. Section III describes the feature extraction procedure. The process of expression synthesis is described in Section IV. Section V presents the experimental results. The proposed method is compared with the state-of-the-art works in the same section. The conclusions are summarized in Section VI.

II. Related Work

For synthesis of realistic emotional facial expressions, we need an appropriate representation of expressions. Categorial models may not entirely represent mixed emotional expressions [10]. For example, categorical model may assign the label *happiness* to an expression showing the emotion *happily surprised*. Martinez and Yannakakis [11] concluded that ranking should be used rather than discrete classes of affect (emotion in this context) for a generalizable model. The work in [12] proposed a hierarchical Dirichlet process mixture model to represent the relationship between the real-world emotional states. Some researchers represented each emotion in a separate continuous space [13], [14]. Some others represented all the basic emotions in one single continuous space. Wang *et al.* [15] and Liu *et al.* [16] estimated the distribution of data in different emotion classes in the valence-arousal and PCA spaces respectively. All these works suggest that to represent realistic emotional expressions, it is better to assign continuous values to emotional expressions. We visualize the proposed Expression Map (XM) as representing one continuous emotion space for all the six basic emotions and their combinations. We use this XM for synthesis of realistic mixed emotional expressions.

Some papers on expression synthesis [2]–[4], [17], [18] assumed that a set of example images of the target person displaying different facial expressions are available. Zhang *et al.* [17] divided the face into a number of regions. Given a set of facial landmark points corresponding to the expression to be imposed, Zhang *et al.* [17] linearly combined the texture of the corresponding region of the example images to get the synthetic expressive texture. Malleson *et al.* [18] proposed a method to blend between multiple facial performances of an actor. They achieved the blending by non-linear temporal synchronization of the input performance videos. In some works [1]–[4], the expression imposed on the target face was driven by the expression on another face (let us call it source face). Asthana *et al.* [2] used the parametric correspondence between the Active Appearance Models (AAMs) of the source and the target faces to facilitate the transfer of facial performance. For facial performance transfer, from the example set of the target, Li *et al.* [3], [4] retrieved the face image displaying the expression which is most similar to that of the expressive source face. The retrieved expressive face image was further warped to the Expression Mapping Image (EMI) to produce final result. The EMI was obtained by multiplying the expression-neutral face of the target by the ratio of the expressive and expression-neutral images of the source face.

Some works on synthesis [5], [6] required a training set consisting of face images of different persons displaying basic emotional expressions and only one expression-neutral target face image. Huang and De la Torre [6] proposed to use 'bilinear kernel reduced rank regression' to learn a mapping between the appearance of a neutral face and a variety of expressions. Xiong *et al.* [5] learned a polynomial mapping from the training examples to get the shape (represented by landmark points) of the expressive target face given the shape of the expression-neutral target face. To synthesize the texture of the target, the EMI was obtained from the source person whose shape is similar to that of the target.

In recent years deep neural networks (deep nets) are successfully used for expression recognition [19] and facial animation (e.g., [20]–[22]). Susskind *et al.* [20] learned a deep belief net from examples for synthesizing facial expressions controlled by specifying Action Unit (AU) [23] labels. Olszewski *et al.* [21] trained a Convolutional Neural Network (CNN) to learn the mapping between the images of the source person's mouth region and the parameters of a digital avatar. Saito *et al.* [22] used CNN to segment facial region for expression transfer. Among different learning machines we choose SOM based XM due to its unique characteristics of topological ordering. This helps us in achieving our objective of synthesizing facial expressions of

mixed emotions. For training the XM, we use a set of handcrafted features. These features are specialized in generating face images with better minute expression details such as wrinkles, furrows etc. as compared to other deep net based methods such as Susskind *et al.* [20] that directly work on pixel values.

The methods in the literature generally consider shape and/or texture as features for synthesizing facial images of expressions. As shape feature, most of the works consider 2D landmark points ignoring the 3D structure of the face [5], [17]. As texture feature, pixel values normalized in certain ways, are considered [4], [20]. Some methods use either of shape [24] or texture feature [20] whereas others use both of them [17], [22]. In addition to landmark points, we also consider texture to derive facial structure. We decompose the texture feature into high and low frequency components that helps in producing less noisy expression images. In Section V, we compare our method with closely matching [5], [17], [20] and [4]. Whereas [5] and [20] report the results only on basic expressions and AUs, [4], [17] transfer facial expressions, our method efficiently synthesizes basic as well as mixed emotional realistic expressions.

III. Extraction of Feature

We assume that we have image sequences for training the XM. Each sequence displays one of the six basic emotional expressions. The first frame of each sequence shows an expression-neutral face. The intensity of the expression increases over the frames. The features to be extracted from each frame need to have two important characteristics: (a) the features should efficiently represent facial expression. (b) the features should be *reversible*. By reversible we mean that given a face image it should be possible to extract the features and on the other side, given the features it should also be possible to synthesize the required expression on a given face image. This means if there exists a function say $f(I) : F = f(I)$, there must also exist a function $f' : I' = f'(F)$ where, $I' \approx I$. A combination of features representing intensity of emotion, facial structure, facial shape and texture is used in our implementation.

Extraction of Shape Feature: We annotate each image of the training sequences with a fixed number (say, l) of landmark points [25]. For the test image the landmarks are detected automatically following Active Appearance Model (AAM) based approach [26]. One such example of landmark points is shown in Supplementary Fig. S2. A vector containing the x and y coordinates of the landmark points represents the shape of the training faces [8]. All these face shapes are aligned to a reference face shape using Procrustes analysis. The difference between the x and y coordinates of the landmark points of the kth frame and the first frame of an image sequence is denoted by the vector $\mathbf{d}_k^s$, where superscript s stands for shape feature. To model the domain of feasible shape deviations due to expression, we perform Principal Component Analysis (PCA) on $\mathbf{d}_k^s$ extracted from all the training sequences. Let P^s represent the matrix where each column represents a vector (principal component) that forms the basis of the PCA domain. Let $\boldsymbol{\mu}^s$ represent the mean of $\mathbf{d}_k^s$ $\forall k$, for all training sequences and $\mathbf{b}_k^s$ represent the weights of the bases of the PCA domain when $\mathbf{d}_k^s$ is projected to the PCA domain. Therefore, we can write,

$$\mathbf{b}_k^s = (P^s)^T(\mathbf{d}_k^s - \boldsymbol{\mu}^s). \tag{1}$$

The size of P^s, $\mathbf{b}_k^s$, $\mathbf{d}_k^s$ and $\boldsymbol{\mu}^s$ are $2l \times e^s$, $e^s \times 1$, $2l \times 1$ and $2l \times 1$ respectively, where e^s is the number of principal components chosen for 99% loading. The vector $\mathbf{b}_k^s$ acts as shape feature in our model.

Extraction of Intensity of Expression: The intensity of expression is correlated to the amount of movement of facial landmark points [27]. As in [8], we estimate the intensity of emotion in a frame by the amount of normalized shape deviations (from the first frame) averaged over all the landmark points in that frame. For normalization, deviations in the x and y directions are considered separately. Let i represent an index for the $2l$ elements of the vector $\mathbf{d}_k^s$. If $i <= l$, $d_{k,i}^s$ represents the deviation of the x-coordinate, else it represents the deviation of the y-coordinate of the ith landmark point. The normalization factor a_i for each element of $\mathbf{d}_k^s$ is set equal to maximum $d_{k,i}^s$ over all the training sequences. Let ϱ_k represent the intensity of expression. We find ϱ_k as follows.

$$\varrho_k = \frac{\sum_{i=1}^{2l}(d_{k,i}^s/a_i)}{2l}, \tag{2}$$

The variable ϱ acts a feature of the intensity of expression.

Extraction of Structure Feature : As structure feature, we use a vector of 21 elements represented by $\boldsymbol{\vartheta}$. Of these 21 features, 17 were proposed by Mao *et al.* [28]. Mao *et al.* [28] and other synthesis models [4], [5] consider only the 2D shape (represented by a set of landmark points) which is not enough to represent the facial structure. In [8], we proposed four new features for facial structure. The computation of these features considers facial texture in addition to face shape. To measure the structure feature, we consider the surface normal at each pixel position on the face. The full set of 21 structure features is listed in the Supplementary Table S1. For training the XM, the structure vector $\boldsymbol{\vartheta}$, extracted from the first frame of the sequence, is concatenated to each ϱ_k (intensity of emotion) calculated for each frame of the sequence. The concatenated feature vector is represented by $\boldsymbol{\varphi}_k$ for kth frame. To get the principal variances of $\boldsymbol{\varphi}_k$ we perform PCA on $\boldsymbol{\varphi}_k$ extracted from all the training sequences. So we have,

$$\mathbf{b}_k^{\varphi} = (P^{\varphi})^T(\boldsymbol{\varphi}_k - \boldsymbol{\mu}^{\varphi}), \tag{3}$$

where meaning of all the variables are the same as in (1). In (3) the superscript φ stands for feature representing both the intensity of emotion and facial structure. The vector $\mathbf{b}^{\varphi}$ is used as one of the features in our synthesis model.

Extraction of Texture Feature: Important changes in facial appearance like wrinkles and furrows, due to expression, are reflected through changes in texture. To eliminate the effect of shape from texture, we warp all the training face images to a reference shape using cubic polynomial function. Let us assume that the number of pixels within the convex hull defined by the landmark points of the warped image is n. To eliminate the effect of different skin tones and illuminations in our model, we decompose the shape-free texture using Total Variation regularization (TV) [29], [30]. This TV method decomposes the

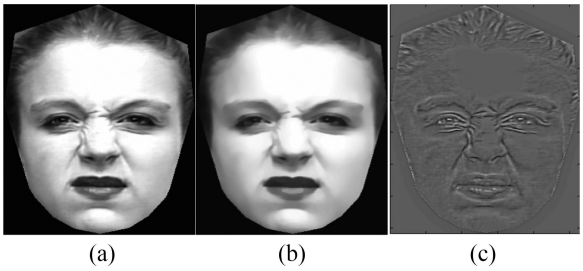


Fig. 3. TVL^1 decomposition of (a) an expressive face image into (b) low frequency and (c) high frequency components.

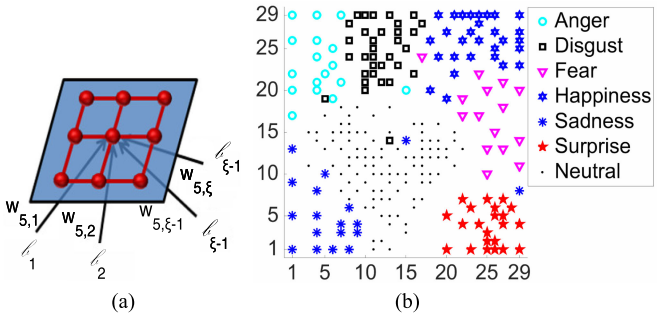


Fig. 4. (a) Schematic diagram depicting the architecture of a 2D grid of XM with $m = 9$ nodes. The synaptic weights connected to the fifth node are shown. All other nodes also have ξD synaptic weights which are not shown here. Each element of the weight ($\mathbf{w}$) is connected to corresponding element of the input ($\mathbf{b}$). (b) Expression map (2D grid of 29×29 neurons) after training with CK+ [23], [31] dataset. The x and y axes of the expression map represent the x and y coordinates of the neurons in the 2D grid.

texture of an image into u: the low frequency component and v: the high frequency component. Fig. 3 shows one example of this decomposition. The v component is less dependent on illumination and mainly represents expression [29]. The u component still depends on the illumination and tone of the face image. To counter this dependence, for each face image we divide each element of the u component by the mean value of u of that image. We represent the low and the high frequency components into $n \times 1$ vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ respectively. We find the deviation (represented by say, $\mathbf{d}_k^u$) from the first frame of the u component by $\mathbf{d}_k^u = \mathbf{u}_k - \mathbf{u}_1$. In the same manner as shape feature, to model the domain of feasible texture in expressions we perform PCA on $\mathbf{d}_k^u$ and $\mathbf{v}_k$ for all the training frames.

$$\mathbf{b}_k^u = (P^u)^T(\mathbf{d}_k^u - \boldsymbol{\mu}^u), \;\; \mathbf{b}_k^v = (P^v)^T(\mathbf{v}_k - \boldsymbol{\mu}^v). \quad (4)$$

In (4) the meaning of the variables are the same as in (1) except s is replaced by u and v.

We use $\mathbf{b}_k^u$ and $\mathbf{b}_k^v$ as features to represent the change in texture due to expression. We concatenate $\mathbf{b}_k^s$, $\mathbf{b}_k^\varphi$, $\mathbf{b}_k^u$ and $\mathbf{b}_k^v$ calculated using (1) (3) and (4) respectively to get the final feature vector represented by $\mathbf{b}$ for the kth frame of a video sequence. We use $\mathbf{b}_k$ as the input feature vector to train the XM for expression synthesis as detailed next.

IV. Expression Synthesis Using Expression Map

We design the Expression Map (XM) as a 2D grid of say, m number of neurons, one neuron per node of the grid. An example is shown in Fig. 4(a). We use the words 'node' and 'neuron' interchangeably in the rest of the paper. The training process of the XM is described next.

A. Training of Expression Map

The training dataset consists of images displaying the six basic expressions with different intensities of emotion. We extract the feature vector $\mathbf{b}$ from each training frame to train the XM. If $\mathbf{b}$ is ξ dimensional (ξD), then each neuron (say, j where $j = 1, \ldots, m$) of the XM gets connected to ξD input space through ξD synaptic weight vector represented by say, $\mathbf{w}_j$ (Fig. 4(a)). In the trained XM, each node (indexed by j) stores an expression pattern in the form of the vector $\mathbf{w}_j$. Prior to training, all the $\mathbf{w}_j$ of the XM are randomly initialized. We want that after training, the XM approximate the expressions space and also that similar emotional expressions be mapped to topologically close proximity in the XM. This objective is achieved by following the learning process of Self Organizing Map (SOM). In this learning, the XM gets trained through three steps of (a) *competition*, (b) *cooperation* and (c) *adaptation*. Given an input training vector $\mathbf{b}$, the m neurons of the XM *compete* among themselves to adapt to $\mathbf{b}$. The neuron say, $\imath(\mathbf{b})$ with the synaptic weight vector $\mathbf{w}$ that matches best to $\mathbf{b}$, wins. The winning neuron *cooperates* with the neighboring neurons for adaptation. The neighborhood function represented by say, $f_{j,\imath(\mathbf{b})}$ is a monotonically decreasing function of spatial distance (in the 2D grid) between the winning neuron $\imath(\mathbf{b})$ and the neighboring neuron j in the XM. The *adaptation* of the synaptic weight vectors (represented by $\mathbf{w}_j$) of all the neurons of the XM is governed by the following equation for $j = 1, 2, ..., m$.

$$\mathbf{w}_j(\mathtt{k}+1) = \mathbf{w}_j(\mathtt{k}) + \eta(\mathtt{k}) f_{j,\imath(\mathbf{b})}(\mathtt{k})(\mathbf{b} - \mathbf{w}_j(\mathtt{k})), \quad (5)$$

where iteration number is represented by $\mathtt{k}$ and η is the learning rate. The value of η and the spread of $f_{j,\imath(\mathbf{b})}$ decreases with iteration. Further details on training the XM is given in [7, Algorithm 1]. Fig. 4(b) shows the XM trained with the proposed combination of features (represented by $\mathbf{b}$) extracted from CK+ dataset [23], [31]. Similar XM trained with MUG [32] dataset is shown in Supplementary Fig. S3.

The training process of the XM (before labeling) is completely unsupervised. At the end of the training, for labeling the nodes representing the emotional expressions at intensity 1 or 0 (neutral expression), we consider the last and the first frames of each training sequence respectively. We map each training feature vector $\mathbf{b}$ (representing intensity 1 or 0) to a node which has a weight vector,s most similar to that training vector. We label any node of the XM with the emotion (six basic and one *neutral*) which is the emotion label of maximum number of training vectors, mapped to that node. If no training vector is mapped to a node, we do not label that node. Notice that in Fig. 4(b) neurons representing particular emotional expressions have formed clusters. Clusters representing similar looking expressions (*e.g.*, *anger* and *disgust*) are adjacent to each other. We use this trained and labeled XM for synthesis of mixed expressions.

B. Synthesis of Expression

For synthesizing an expression representing a specific combination of basic emotions (*e.g.*, *30% happy, 70% surprised*), we propose Algorithm 1. Algorithm 1 goes through a number of iterations where each iteration synthesizes one expression, say expression e, $e = 1, 2, \ldots, 6$ on the target face. Next we explain each step of Algorithm 1. Let S_e, $e = 1, ..., 6$ and S_7 represent the set of all the neurons (of the XM) labeled as basic expression e and *neutral* respectively. Since we train the XM with $\mathbf{b}$ extracted as explained in Section III, the pattern stored at each node of the trained XM represents a facial structure and the deviation in shape and texture of that particular facial structure. Let $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\vartheta}_\jmath$ represent the structure features extracted from the target face image and the one stored at the $\jmath$th neuron, $\jmath = 1, 2, \ldots, m$ in the XM respectively. The synthesized expression should suit the structure of the target face. So, for synthesizing an emotional expression say, e, from among S_e we choose the node say, N_e that stores the structure which is most similar to $\boldsymbol{\vartheta}_t$.

$$N_e = \arg\min_{\jmath}\{dist(\boldsymbol{\vartheta}_t, \boldsymbol{\vartheta}_\jmath) | \boldsymbol{\vartheta}_\jmath \in S_e\} \tag{6}$$

where $dist(.,.)$ is a suitable distance metric. We use Euclidean distance. This node selection forms step 2.1 of Algorithm 1. If we add the shape and texture deviations stored at node N_e to the shape and texture of the target expression-neutral face image, this will synthesize expression e at intensity 1. This is because the pattern stored at N_e represents an expression at intensity 1. We wish to synthesize intensity say, $\varrho_e \in [0, 1]$ of emotion e. So we choose a node, say N_7 representing neutral expression in similar manner (step 1 of Algorithm 1). Let the shape, low and high frequency texture patterns stored at N_e and N_7 be represented by $\{\mathbf{b}_e^s, \mathbf{b}_e^u, \mathbf{b}_e^v\}$ and $\{\mathbf{b}_7^s, \mathbf{b}_7^u, \mathbf{b}_7^v\}$ respectively. Let $\widetilde{\mathbf{b}}_e^s$, $\widetilde{\mathbf{b}}_e^u$ and $\widetilde{\mathbf{b}}_e^v$ represent the PCA projections of the shape deviation, low frequency texture deviation and high frequency texture respectively that are required for expression synthesis. To synthesize intensity ϱ of emotion e, we shall take linear convex combination of the expression patterns stored at N_e and N_7 as follows.

$$\begin{aligned}
\widetilde{\mathbf{b}}_e^s &= \varrho_e \mathbf{b}_e^s + (1 - \varrho_e)\mathbf{b}_7^s, \\
\widetilde{\mathbf{b}}_e^u &= \varrho_e \mathbf{b}_e^u + (1 - \varrho_e)\mathbf{b}_7^u, \\
\widetilde{\mathbf{b}}_e^v &= \varrho_e \mathbf{b}_e^v + (1 - \varrho_e)\mathbf{b}_7^v.
\end{aligned} \tag{7}$$

This combination of patterns forms steps 2.2, 2.3 and 2.4 of Algorithm 1. From these PCA projections, we find required vectors as follow.

$$\begin{aligned}
\widetilde{\mathbf{d}}_e^s &= P^s \widetilde{\mathbf{b}}_e^s + \boldsymbol{\mu}^s, \\
\widetilde{\mathbf{d}}_e^u &= P^u \widetilde{\mathbf{b}}_e^u + \boldsymbol{\mu}^u, \\
\widetilde{\mathbf{v}}_e &= P^v \widetilde{\mathbf{b}}_e^v + \boldsymbol{\mu}^v.
\end{aligned} \tag{8}$$

These equations form steps 2.5, 2.6 and 2.7 of Algorithm 1. The deviation vectors $\widetilde{\mathbf{d}}_e^s$ and $\widetilde{\mathbf{d}}_e^u$ respectively are added to the shape and low frequency textures ($\mathbf{s}_t$ and $\mathbf{u}_t$ respectively) of the target expression-neutral face in steps 2.5 and 2.6 of Algorithm 1. The resultant vectors are $\widetilde{\mathbf{s}}_e$ and $\widetilde{\mathbf{u}}_e$.

$$\widetilde{\mathbf{s}}_e = \mathbf{s}_t + \widetilde{\mathbf{d}}_e^s,\ \widetilde{\mathbf{u}}_e = \mathbf{u}_t + \widetilde{\mathbf{d}}_e^u. \tag{9}$$

Ideally if we add $\widetilde{\mathbf{u}}_e$ and $\widetilde{\mathbf{v}}_e$, we should get the texture of the intended synthesized expressive face in the reference shape. Due to such reasons as less than accurate warping of the training faces and the target face to the reference shape, the pixel to pixel correspondence among $\widetilde{\mathbf{u}}_e$ and $\widetilde{\mathbf{v}}_e$ remains imprecise. This may introduce noise to the synthesized face.

To discard the changes of texture due to noise, we introduce a weighting function of the texture. This function is designed based on the intuition that the change in facial texture should be dependent on the movement of facial components (represented by movement of landmark points). The weighting function is designed as a summation of weighted Gaussian functions. A Gaussian function is assumed to be centered at each landmark point. To define the effect of movement of each landmark point ι, $\iota = 1, ..., l$, the corresponding Gaussian function is weighted by the shape deviation magnitude (represented by $\| \widetilde{d}_{e,\iota}^s \|$) corresponding to that landmark point. The spread (represented by say, σ) of each Gaussian function represents the zone of influence of each landmark point. At each facial pixel position represented by say, i, we sum up the magnitudes of all (l) the weighted Gaussian functions to get the weight represented by $h_e(i)$ of that pixel (step 2.8.1 of Algorithm 1) as follows.

$$h_e(i) = \sum_{\iota=1}^{l} \big(\| \widetilde{d}_{e,\iota}^s \| \exp(dist(\iota, i)^2 / 2\sigma^2) \big). \tag{10}$$

In (10), the Euclidean distance between ith pixel and ιth landmark point is given by $dist(\iota, i)$. The process of deriving the weighting function h_e is explained in Fig. 5. To penalize the pixels with low weight and to reward the rest, we apply sigmoidal function (say, Ω) on $h_e(i)$.

To get the final low frequency texture of the synthesized expression, we take a combination of texture vectors $\mathbf{u}_t$ of the expression-neutral target face and $\widetilde{\mathbf{u}}_e$ which is derived from the XM in (9).

$$u_t(i) = \widetilde{u}_e(i)\Omega(h_e(i)) + u_t(i)\Omega(1 - h_e(i)), \tag{11}$$

$$\Omega(h_e(i)) = \frac{1}{1 + \exp(-\phi(h_e(i) - \tau))}. \tag{12}$$

Here $\Omega(h_e(i))$ is a sigmoidal function; ϕ and τ control the slope and translation of the sigmoidal function. Similar equation holds for the high frequency texture component v (steps 2.8.2 and 2.8.3 of Algorithm 1).

The equation (12) is designed such that the pixels with no or small movement are mainly influenced by the u component (u_t) and v component (v_t) of the input expression-neutral target image. Pixels with considerable movement are mainly influenced by the u component ($\widetilde{u}_e$) and v component ($\widetilde{v}_e$) derived from the XM. Finally (step 2.8.4 of Algorithm 1) we add the modified $u_t(i)$ and $v_t(i)$ at each pixel location $i = 1, \ldots, n$ to get the final texture $\widetilde{g}_e(i)$. The final texture displays the intended expression e with intensity ϱ_e. Using cubic polynomial function, the face in reference shape having the texture $\widetilde{g}_e(i)$ is warped to the

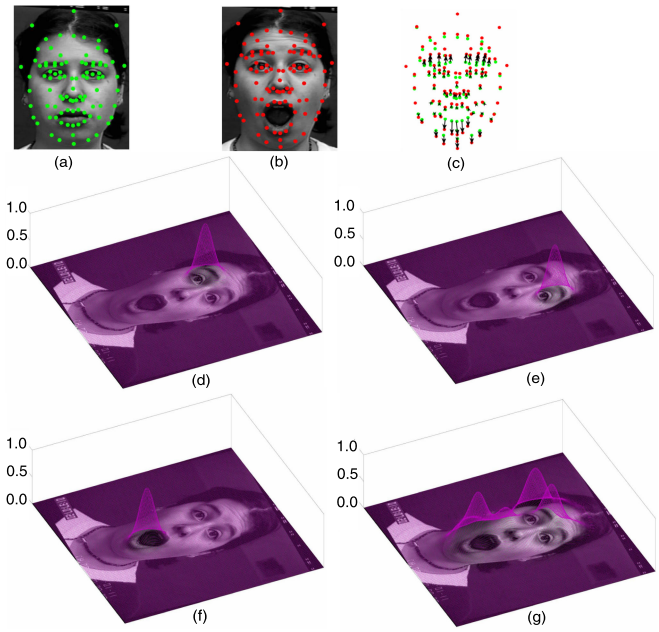


Fig. 5. Example showing derivation of weighting function h_e. (a) and (b) Landmarks plotted on face image displaying no expression and *surprise* respectively; (c) deviation of landmarks due to expression *surprise* shown by arrows; (d), (e), (f) some constituent Gaussian functions of h_e; (g) plot of h_e on the face image displaying *surprise*. The constituent Gaussian functions and the function h_e are shown by purple mesh. The height of the mesh shows the magnitude of the function at each pixel location. Observe that since *surprise* is mainly expressed through movement of eye-brows and lips, the function h_e forms peak at those locations.

expressive shape $\widetilde{\mathbf{s}}_e$. Thus, we get the final face image of the target face displaying expression e with intensity ϱ_e.

Step 2 of Algorithm 1 is iterated 6 times. Each iteration synthesizes one basic expression on the target face image. The unique combination of the sigmoidal and Gaussian functions in step 2.8 of Algorithm 1 ensures smooth blending of expressions of different emotions. The final image shows the required expression, synthesized (basic or mixed) on the given expression-neutral target face image. Next we present the results of our proposed model.

Algorithm 1: Algorithm for Finding the Shape Vector $\widetilde{\mathbf{s}}_e$ and Texture Vector $\widetilde{\mathbf{g}}_e$ for Synthesizing the Required Combination of Six Basic Expressions in the Target Expression-Neutral Face Image

INPUT: Trained and labeled expression map, expression-neutral target face image, set of basic expressions to be synthesized specified by e, $e = 1, 2, ..., 6$ and the set of intensities ($\varrho_e \in [0, 1]$) of each of the basic expression e to be synthesized.

OUTPUT: The shape vector $\widetilde{\mathbf{s}}_e$ and texture vector $\widetilde{\mathbf{g}}_e$ of the synthesized expression.

Let S_e, $e = 1, ..., 6$ represent the set of all the neurons labeled as basic expression e and S_7 represent the set of all neurons labeled as neutral. Let $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\vartheta}_\jmath$ represent the structure features extracted from the test image and stored at the $\jmath$th, $\jmath = 1, 2, \ldots, m$ neuron respectively. Also let $\mathbf{s}_t$, $\mathbf{u}_t$ and $\mathbf{v}_t$ represent the shape, low and high frequency texture features respectively, extracted from the expression-neutral target face image. Let $\mathbf{b}_e^s$ and $\mathbf{b}_e^u$ represent the weights of the shape and low frequency texture deviation features respectively and $\mathbf{b}_e^v$ represent the weight of the high frequency texture feature (projected onto the respective PCA domains represented by $\{P^s, \boldsymbol{\mu}^s\}$, $\{P^u, \boldsymbol{\mu}^u\}$ and $\{P^v, \boldsymbol{\mu}^v\}$) stored at the neuron N_e $e = 1, \ldots, 7$.

1. $N_7 = \arg\min_\jmath \{dist(\boldsymbol{\vartheta}_t, \boldsymbol{\vartheta}_\jmath) | \boldsymbol{\vartheta}_\jmath \in S_7\}$ where, $dist(.,.)$ is a suitable distance metric.
2. Repeat the following steps for $e = 1, ..., 6$.
   - 2.1. $N_e = \arg\min_\jmath \{dist(\boldsymbol{\vartheta}_t, \boldsymbol{\vartheta}_\jmath) | \boldsymbol{\vartheta}_\jmath \in S_e\}$.
   - 2.2. $\widetilde{\mathbf{b}}_e^s = \varrho_e \mathbf{b}_e^s + (1 - \varrho_e)\mathbf{b}_7^s$
   - 2.3. $\widetilde{\mathbf{b}}_e^u = \varrho_e \mathbf{b}_e^u + (1 - \varrho_e)\mathbf{b}_7^u$
   - 2.4. $\widetilde{\mathbf{b}}_e^v = \varrho_e \mathbf{b}_e^v + (1 - \varrho_e)\mathbf{b}_7^v$
   - 2.5. $\widetilde{\mathbf{s}}_e = \mathbf{s}_t + \widetilde{\mathbf{d}}_e^s$ where, $\widetilde{\mathbf{d}}_e^s = P^s(\widetilde{\mathbf{b}}_e^s) + \boldsymbol{\mu}^s$
   - 2.6. $\widetilde{\mathbf{u}}_e = \mathbf{u}_t + \widetilde{\mathbf{d}}_e^u$ where, $\widetilde{\mathbf{d}}_e^u = P^u(\widetilde{\mathbf{b}}_e^u) + \boldsymbol{\mu}^u$
   - 2.7. $\widetilde{\mathbf{v}}_e = P^v \widetilde{\mathbf{b}}_e^v + \boldsymbol{\mu}^v$
   - 2.8. Repeat the following steps for each facial pixel position $i = 1, \ldots, n$.
     - 2.8.1 $h_e(i) = \sum_{\iota=1}^{l} (\| \widetilde{d}_{e,\iota}^s \| \exp(dist(\iota, i)^2 / 2\sigma^2))$
     - 2.8.2 $u_t(i) = \widetilde{u}_e(i)\Omega(h_e(i)) + u_t(i)\Omega(1 - h_e(i))$
     - 2.8.3 $v_t(i) = \widetilde{v}_e(i)\Omega(h_e(i)) + v_t(i)\Omega(1 - h_e(i))$
     - 2.8.4 $\widetilde{g}_e(i) = u_t(i) + v_t(i)$

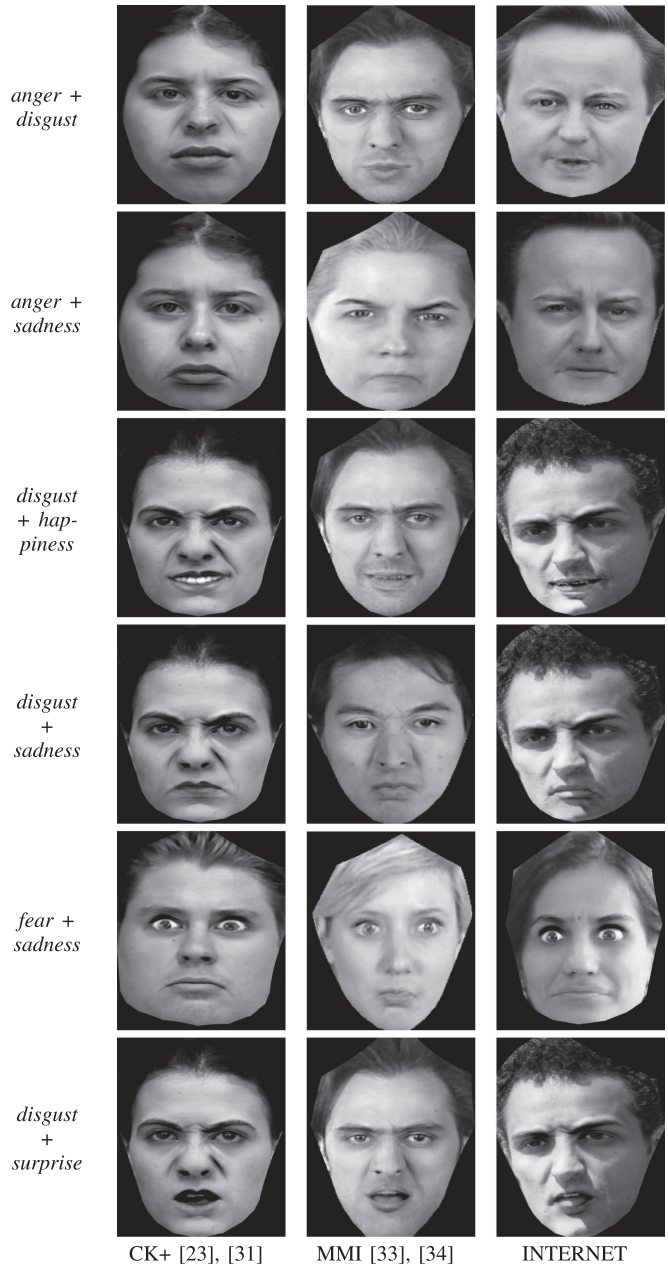
V. Results and Discussions

We have done extensive experiments (both qualitative and quantitative) to validate the synthesized expressions as detailed next.

A. Description of Dataset

We have used three publicly available datasets: CK+ [23], [31], MMI [33], [34] and MUG [32]. All the image sequences displaying one of the six basic emotional expressions on 118, 13 and 52 number of subjects from CK+, MMI and MUG datasets respectively are used for our experiments. While CK+ [23], [31] consists of a large number of data in different illumination conditions on faces with different tonicities, MMI [33], [34] data consists of a wide variety of expressions for the same basic emotion. MUG [32] also consists of a large collection of facial expressions with less ethnical variety. The image sequences from 98 and 31 subjects from CK+ and MUG datasets respectively are used for training the proposed XM, one XM for each dataset. To test the performance of the proposed algorithm on real-life (non-lab) face images, we have collected 50 faces (almost frontal) of celebrities displaying near-neutral expression under varying illumination condition from internet. We call this dataset INTERNET. Some faces from this dataset are shown in Fig. S4 of the supplementary material. For testing the XM trained using CK+ dataset, expressions are synthesized on expression-neutral face images of 20 subjects (which are not used for training the XM) from CK+ dataset and all the subjects from MUG, MMI and our INTERNET datasets. Similarly,

TABLE I
EXAMPLES OF SYNTHESIZED EMOTIONAL EXPRESSIONS WHERE TWO OF THE SIX BASIC EMOTIONAL EXPRESSIONS ARE PRESENT. EACH ROW SHOWS ONE MIXED EXPRESSION. THE EXPRESSION-NEUTRAL TARGET FACES ON WHICH THE EXPRESSIONS ARE SYNTHESIZED ARE TAKEN FROM CK+ [23], [31] (COLUMNS 2), MMI [33], [34] (COLUMNS 3) AND INTERNET (COLUMNS 4) DATASETS RESPECTIVELY. THE XM IS TRAINED WITH MUG DATASET [32]



for testing the XM generated using MUG dataset, 21 subjects (which are not used for training the XM) from MUG dataset and all the subjects from CK+, MMI and INTERNET datasets are used.

B. Results

Qualitative Evaluation: Table I shows some examples of mixed expressions synthesized by our method. The expressions can be explained with respect to Action Units (AUs) [35]. AU can be thought of as the unit of movement triggered by one or more facial muscles. A list of AUs and their corresponding visible facial changes are listed in [23]. From Table I, it can be noticed that two basic expressions have been mixed in a seamless manner to produce the intended mixed expression. For example, in Table I, first row, *the eyebrows drawn close together* (AU4) and *tightened lips* (AU23) carry the *anger* information. In the same images *deepened nasolabial furrow* along with *nasal root wrinkles* (AU9) carry *disgust* information. Together these characteristics express the emotions *anger* and *disgust* simultaneously. Some of the synthesized results showing basic emotions are shown in Supplementary Fig. S5. It can be seen that our method is able to produce realistic facial expressions enhanced by furrows and wrinkles. Also notice that our algorithm produces different types of expressions on people with different facial structures even for the same basic emotion category. For example, consider Table I, row 1. Notice the difference in the shape of the lips, nasolabial furrow and the shape of wrinkles between the eyebrows for each face. These examples demonstrate the efficiency of our approach in producing a wide variety of realistic mixed facial emotional expressions.

Note that for all our experiments, the set of target faces, on which the expressions are synthesized, is completely disjoint from the set of faces which are used to train the XMs.

Quantitative Evaluation of Synthesized Basic Expressions Using Automatic Tool: If an existing automatic facial expression recognition method (e.g., [37], [36]) can recognize emotions from the face images which are synthesized by our method, that establishes the success of our method. We utilize the method by [36] which uses LBP features. As classifier we use (a) Support Vector Machine (SVM) with linear and Radial Basis Function (RBF) kernels and (b) multiclass RBF classifier. We generate two XMs trained with CK+ and MUG datasets respectively. To validate the expressions synthesized using these two XMs, we train two sets of classifiers. The last three images with highest intensity of emotion per training sequence of CK+ and MUG datasets are used to train the two sets of classifiers. Table II shows the recognition accuracies of the six basic emotions from the synthesized images. The expressions are synthesized on faces taken from the four datasets. SVM being binary classifier, one *vs.* all classification strategy is used for classification. The parameters for the SVM and multiclass-RBF classifiers are chosen using 10-fold cross-validation of the training data.

Similar recognition accuracies can be observed irrespective of the test dataset or the classifier. From Table II it is seen that the expressions generated utilizing the XM trained using MUG dataset produce better expression recognition results. Note that SVM classifier and LBP feature may not give 100% recognition result even for original expressive face images. For example, classification accuracy using LBP in [36] is 92.6%. We also perform human perception based quantitative validation of our synthesized expressions as described next.

Quantitative Evaluation of Synthesized Basic Expressions Through Perception Test: Thirty volunteers have participated in this evaluation. These volunteers are normal people with no special training in facial expressions. Each volunteer is presented with randomly chosen 10 synthesized images per expression, per

TABLE II
RECOGNITION % OF SYNTHESIZED BASIC EXPRESSIONS USING AUTOMATIC TOOL. THE EXPRESSION-NEUTRAL INPUT FACE IMAGES, ON WHICH THE EXPRESSIONS ARE SYNTHESIZED, ARE TAKEN FROM CK+, MUG, MMI AND INTERNET. THE TWO XMS ARE TRAINED WITH CK+ AND MUG RESPECTIVELY. THE FEATURES FOR EXPRESSION CLASSIFICATION ARE GENERATED FOLLOWING [36]. TR: TRAINING DATASET, TS: TEST DATASET, RE: RECALL, PRE: PRECISION, INT: INTERNET, CL: CLASSIFIER, CL1: SVM WITH LINEAR KERNEL, CL2: SVM WITH RBF KERNEL, CL3: MULTI CLASS RBF

Tr	Ts	Cl	*Anger*		Disgust		Fear		Happiness		Sadness		Surprise	
			Re	Pre	Re	Pr	Re	Pre	Re	Pre	Re	Pre	Re	Pre
CK+	CK+	Cl1	80.95	62.92	85.37	76.54	51.35	71.04	96.77	95.24	60.71	71.34	84.62	84.06
		Cl2	57.14	80.34	90.24	78.21	40.54	96.17	93.55	98.41	83.93	59.76	65.5	90.5
		Cl3	47.62	92.70	85.37	79.23	32.43	97.81	90.32	98.94	62.50	87.20	61.54	97.58
	MUG	Cl1	55.56	71.37	69.84	73.68	74.55	49.58	77.78	81.01	70.27	68.50	76.09	84.08
		Cl2	49.21	81.58	39.68	84.65	69.09	62.29	48.15	97.89	80.5	78.5	39.13	94.69
		Cl3	66.67	89.41	80.95	80.70	85.45	71.61	75.93	98.73	21.62	99.21	93.48	92.65
	MMI	Cl1	72.50	64.22	76.32	83.93	67.50	56.89	50.54	94.44	62.26	76.52	65.82	79.80
		Cl2	50.21	81.58	51.32	93.11	62.50	67.74	49.21	81.50	83.02	73.48	63.29	84.11
		Cl3	66.67	89.41	82.89	81.97	67.50	70.09	41.94	79.86	62.26	75.51	96.20	97.02
	INT	Cl1	60.60	54.84	70.00	57.60	50.00	63.64	86.36	88.98	64.52	63.56	50.00	76.62
		Cl2	31.03	82.84	54.17	93.60	50.00	76.03	77.27	98.43	70.97	77.12	68.75	69.39
		Cl3	37.93	86.57	63.33	84.96	66.67	56.39	51.61	90.91	22.22	97.06	50.00	97.96
MUG	CK+	Cl1	68.15	66.24	72.73	59.09	61.54	60.77	71.76	89.41	67.92	84.91	66.67	63.96
		Cl2	71.34	71.97	84.85	86.36	66.15	60.00	68.24	87.06	69.81	67.92	61.26	59.46
		Cl3	72.90	78.21	78.79	81.02	52.31	78.19	67.24	86.21	66.67	58.29	59.21	69.45
	MUG	Cl1	88.89	72.84	100.00	74.29	76.67	71.67	57.58	86.87	83.78	71.62	88.24	97.06
		Cl2	81.48	94.44	92.86	98.57	90.00	71.67	69.70	100.0	70.27	70.27	73.53	85.29
		Cl3	96.30	85.89	92.86	99.80	80.00	89.75	78.79	99.13	59.46	95.88	76.47	93.90
	MMI+	Cl1	86.49	70.72	88.14	84.41	66.67	78.89	91.23	84.21	73.21	81.25	76.74	86.05
		Cl2	83.78	82.88	96.61	93.56	86.67	86.67	91.23	92.98	73.21	87.50	90.70	74.42
		Cl3	64.86	87.39	98.31	88.60	91.11	82.33	71.93	98.71	82.12	70.25	48.84	79.80
	INT+	Cl1	77.78	64.81	80.70	57.89	72.41	57.47	87.83	77.10	60.47	65.12	65.67	76.12
		Cl2	64.29	71.43	82.46	73.68	66.67	62.07	86.09	89.28	50.00	84.88	69.81	56.60
		Cl3	80.95	60.48	80.70	68.62	60.92	69.49	80.00	67.38	62.79	71.98	50.94	55.93

test dataset, considering four test datasets. This means, each volunteer is presented with $10 \times 6 \times 4 = 240$ synthesized images. These images are synthesized to display pure basic emotions. The volunteers are requested to identify which basic emotion out of the six, is expressed in a given synthesized face image.

Fig. 6 (a) and (b) compare the recognition accuracy (by human volunteers) of the six basic expressions from the synthesized images. The recognition accuracy of any particular emotion class is almost similar for all the test datasets independent of the dataset used to train the XM. The recognition accuracy for *disgust* is quite high (98.5% and 99.1%) for images synthesized from both the XMs; similarly for *happiness* (98.4% and 99.4%). Recognition rate of *fear* is comparatively low (87.7% and 89.9%) for both the XMs.

We wish to check how the recognition accuracy differs for the synthesized faces from the real ones. So, we repeat the same perception test on real expressive faces from three datasets (we do not have expressive faces from the INTERNET dataset). Fig. 6(c) shows the result. Even for real images recognition accuracy is not 100%. There is noticeable confusion between *anger* and *disgust*; *surprise* and *fear* especially for the MMI dataset. Let the variable A represent the recognition accuracies of the six basic expressions synthesized using the XM which is trained with the CK+ dataset. Let the variable B represent the same for the real expressions from the CK+ dataset. The correlation coefficient between A and B is 83.5%. The correlation coefficient calculated similarly for the MUG dataset is 76.6%. In MUG dataset, the *fear* expression is recognized more accurately as compared to synthesized faces. The high correlation between the accuracies of the synthesized and real expressions

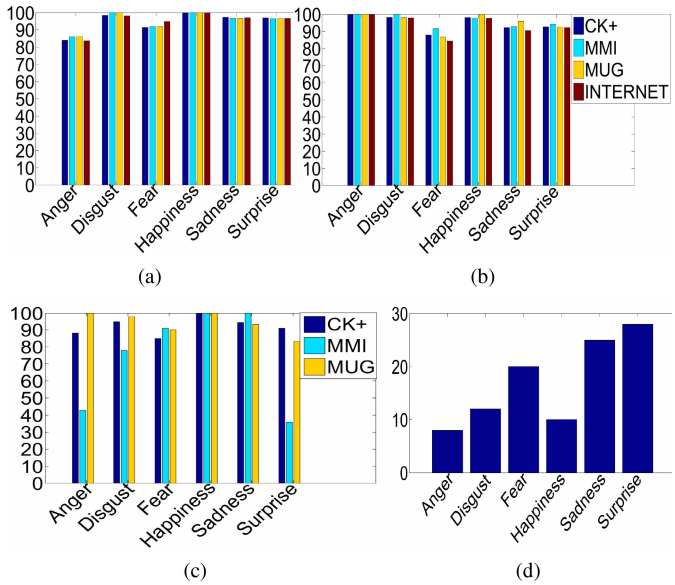


Fig. 6. (a), (b) Accuracy % in perception test for six basic emotional expressions synthesized on faces from four datasets. The XMs are trained with (a) CK+ [23], [31] and (b) MUG [32] datasets and are tested on the four datasets. (c) Accuracy % in perception test on real images from CK+, MMI and MUG datasets. (d) Recognition accuracy for identification of synthetic images from a mix of synthetic and real images. In each graph the y-axis represents the accuracy percentage.

show that the synthesized expressions are perceived accurately. To further test if the synthesized expressions look natural, we performed another experiment. We present the volunteers with randomly chosen 20 images per emotion. Of these 20 images,

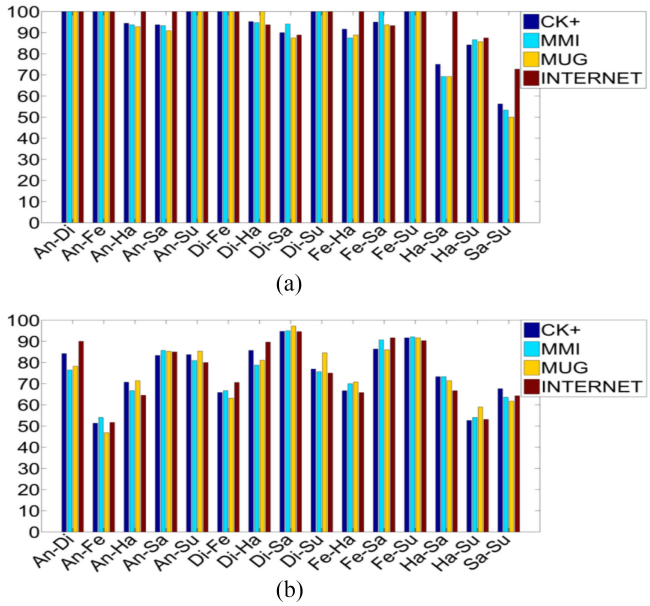


Fig. 7. Accuracy % in perception test for 15 synthesized mixed emotional expressions. The XMs are trained with (a) CK+ [23], [31] and (b) MUG [32]. Each group of bars is for one combination of two basic emotional expressions. Color of bars represents the dataset from which the test faces are taken. The y-axis represents the accuracy %.

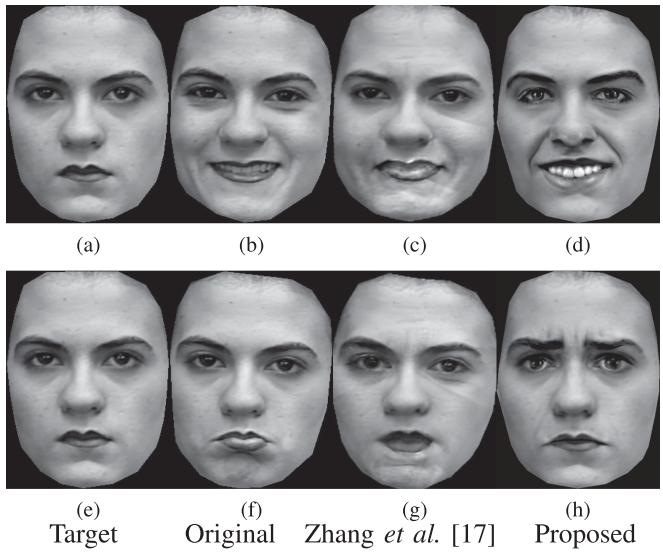


Fig. 8. Comparison of expressions synthesized by Zhang *et al.* [17] *vs.* our method. The four columns respectively show the expression-neutral target faces, the faces showing the emotions that are intended to be synthesized, the faces synthesized by [17] and the faces synthesized by our method.

10 are real and 10 are synthesized. The volunteers are requested to choose the synthetic faces. Fig. 6(d) shows the result. From Fig. 6(d) we see that 25% of the synthetic *sad* faces and 28% of the synthetic *surprise* faces are correctly identified as synthetic i.e., these images do not look realistic. This is because, to synthesize *surprise* (and *sorrow*) expression, the shape and the texture need to be changed considerably. For other emotions, synthetic expressions are incorrectly perceived as real by human volunteers. These experiments indicate that the proposed method is capable of synthesizing realistic facial expressions.

Quantitative Evaluation of Synthesized Mixed Expressions Through Perception Test: For validation of the mixed emotional expressions synthesized by the proposed method, we repeat similar perception test as described above. We generate a total of ${}^6C_2 = 15$ combinations of two basic emotional expressions for each of the target faces taken from all the four datasets. Volunteers are told that the images to be tested show mixed expressions where two of the six basic emotions are present. Volunteers are asked to select two basic emotions for each synthesized image. Fig. 7 shows the recognition accuracies for 15 mixed expressions. It should be noted that in the results presented here, a mixed expression is considered to be classified correctly only if both the constituent emotions are identified correctly by the volunteer. From Fig. 7(a) it can be noticed that for XM generated from CK+, 100% recognition accuracy could be found for a number of mixed expressions (*e.g.*, *anger-disgust*, *anger-fear*, *anger-surprise*, *disgust-fear*, *disgust-surprise* and *fear-surprise*). Relatively low accuracy could be observed for *happiness-sadness* and *sadness-surprise*. Further study into the corresponding confusion matrices suggests that *sadness* has been confused with *anger*. The confusion matrices for both basic and mixed expressions and the related analysis are presented in the supplementary Fig. S6 and Fig. S7. Thus, through qualitative and quantitative analysis presented above it is shown that the proposed method efficiently produces realistic basic as well as mixed emotional expressions. Synthesis of a mix of two emotional expressions on an image of size 140×230 require approximately 2 seconds in which the feature extraction time is 0.2 seconds.

C. Comparison With State-of-the-Art

We compare our method with five of the most relevant state-of-the-art methods [4], [5], [17], [20], [38].

Comparison with Zhang et al. [17]: Zhang *et al.* [17] proposed a method that generates a separate model for each target face using some example face images of that target face displaying a number of expressions. Fig. 8 displays some of the expressions synthesized following [17] and following our method. To synthesize a combination of two expressions, the technique in Zhang *et al.* [17] needs that each of the two expressions be present in the set of training example images of the target person. In Fig. 8, we present two examples of synthesis where one of the expressions (of the mixed expression) that we wish to synthesize is not present in the set of training example images.

Fig. 8(b) and Fig. 8(f) show two expressions (AU6+12+25) and (AU15+17) respectively [31] that we wish to synthesize on subject S130 of CK+ dataset (Fig. 8(a) and Fig. 8(e) respectively). To generate Fig. 8(c), we build a model following [17] using all the available 10 expressive images of subject S130. This set of training images do not include any face image displaying AU25. Notice that in Fig. 8(c) the furrow surrounding the lips (AU12) has been synthesized but the lips are not parted or teeth are not displayed (absence of AU25). Similarly, the synthesis model used to generate Fig. 8(g) includes all the available 10 images except the image displaying AU15. In Fig. 8(g) AU17

(*chin raiser*) has been synthesized but not AU15 (*lip corner depressor*). These examples show that [17] cannot synthesize an expression which is not present in the example images of the target face. This restriction necessitates collection of a wide variety of expressive images of the target person. This collection may not be always feasible. The method by [17] processes each region separately. Therefore, the illumination of texture of one region may vary from the neighboring regions resulting in visual oddities. This happens despite blending of region edges in the synthesized face image. This difference in texture can be seen in regions surrounding eyes, cheek regions and chin region in Fig. 8(c) and Fig. 8(g).

In Zhang *et al.* [17] the expression to be synthesized has to be specified by a set of landmark points. There is no such constraint in our approach. The input to our model is the labels of basic expressions and the proportions by which they are combined. For example, '100% *happiness*' (Fig. 8(d)). In order to compare our results with that of [17] (column 3 of Fig. 8), we look into FACS Manual Investigator's Guide [23]. We find which basic emotion(s) best describes the set of AUs present in each image in column 2 of Fig. 8. For example, AU6+12+25 of Fig. 8(b) represents *happiness* and AU15+17 represents *sadness*. We synthesize these emotions using our method on subject S130 of CK+ dataset and the results are shown in Fig. 8(d) and Fig. 8(h) respectively. Observe that the intended emotions, *happiness* (Fig. 8(d)) and *sadness* (Fig. 8(h)) are displayed nicely on the results generated by our method. The presence of wrinkles and furrows have given the synthesized expressions a realistic look. Unlike [17], our method requires only one expression-neutral image of the target face that too is not part of the training.

The expressions are generated from the XM which is pre-trained with different expressive examples of training faces. Collecting wide variety of expressive face images of different persons a-priori for training the model is much easier compared to collecting wide variety of expressive images of each target person. This makes our method more suitable for synthesis of a variety of expressions in real-time. Moreover, in our method, holistic approach for expression synthesis is taken rather than processing each region separately. This helps in generating smooth realistic looking expressive faces without discontinuity along the boundary (unlike [17]) of the face regions.

Comparison With Xiong et al. [5]: In [5], the texture representing an expression is synthesized using the ratio of expressive and neutral image (Expression Ratio Image or ERI) [39] of the selected (source) training face. Expression imposition using ERI is based on the preconditions that (a) the lighting directions for both the source and target faces are same, (b) the surface normals at each pixel position of the faces are same and (c) the geometrical shape of the target and the source faces are same [39]. ERI generates quality expressions when these conditions are satisfied [39]. Xiong *et al.* [5] used the shape represented in terms of landmark points to select the source face. This selection method ignores the illumination and the surface normals at each pixel (preconditions (a) and (b) for success of ERI) of the target and the selected source faces. Thus, the expressions synthesized by Xiong *et al.* [5] are noisy. Fig. 9 compares the expressions synthesized by [5] *vs.* our approach. Notice that when the illumination of the target face (column 2 of Fig. 9) and the training source face (expression-neutral: columns 3 of Fig. 9, expressive: columns 4 of Fig. 9) are different, the expressive images synthesized (column 6 of Fig. 9) following [5] are noisy. We also show an example where the illumination of the target (Fig. 9, column 1) and the selected training (Fig. 9, column 3 and 4) faces are roughly similar. Notice the noise on such regions as the corners of the eyes, lips *etc*. of the corresponding faces synthesized following Xiong *et al.* [5] in column 5 of Fig. 9.

To take care of the variation of illumination, we have divided the image pixel values into low and high frequency parts (Section III). The high frequency part is nearly illumination invariant [29] and represents expression. In our approach, the expression is synthesized mostly based on this high frequency part. To take care of the skin tone, we have divided the low frequency part of the face by the mean low frequency texture value of that face. Thus, our proposed approach synthesizes realistic expressions (columns 7 and 8 of Fig. 9) on faces with different illuminations and skin tones. Last two columns of Fig. 9 show *sadness* (first row) and *disgust* (second row) synthesized on target faces (columns 1 and 2 of Fig. 9) following our method. To represent the face structure, we not just depend on the 2D coordinates of the landmark points (as in [5]), but also on the surface normals (*e.g.*, structure feature number 19, 20 and 21 of the Supplementary Table S1). Thus, on different target faces, the proposed approach imposes the shape and texture (representing specified emotion) that better suits the structure of the face. Moreover, expression imposition by [5] involves multiplication operation (on the target face and the ERI). Our approach involves addition operation (on shape and textures and their deviations due to expression). Thus, our approach incorporates much less noise as compared to [5].

Comparison With Susskind et al. [20]: Susskind *et al.* [20] proposed an interesting hybrid (mix of unsupervised and supervised learning) approach based on Deep Belief Net (DBN) for expression synthesis. The DBN consists of three layers of Restricted Boltzmann Machine (RBM). The first two layers are trained in an unsupervised way. The output of the second layer concatenates to the vectors representing the face identity and AU labels to act as the input to the final penultimate layer. Once the DBN is trained, the input vectors corresponding to the AUs can be clamped to generate corresponding synthesized expressions. Fig. 10 presents some examples of expressions synthesized by [20] by clamping on one AU. The expressions generated by [20] lacks fine expression details such as wrinkles. Our method explicitly works on high frequency texture component to generate fine expression details. The method by [20] has no strict control over the particular AU to be imposed (see Fig. 10(f)) and on preserving the identity of the target (see Fig. 10(d)). Moreover, [20] has no measure to control the intensity of expression. A 92.4% recognition accuracy of the expressions synthesized by our method shows that our method has good control over the emotion to be synthesized. In our method, the user can also specify the percentage of different basic emotions to be mixed for synthesis.

Comparison With Li et al. [4]: Li *et al.* [4] proposed a technique that animates the face of the target person driven by the

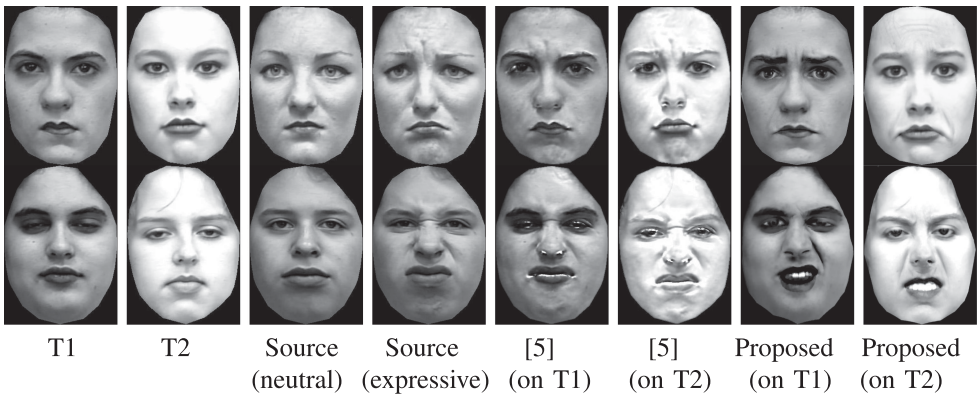


Fig. 9. Comparison of expressions synthesized by Xiong *et al.* [5] *vs.* our method. First two columns show the expression-neutral target faces (targets: T1 and T2). Each row of column 3 shows the expression-neutral face of the training source subject selected by Xiong *et al.* [5] as having similar shape as that of the target face. Column 4 shows the expressive face of the selected source training subject. Columns 5 and 6 show the expression displayed in column 4 synthesized on the faces in columns 1 and 2 respectively by [5]. Last two columns show *sadness* (first row) and *disgust* (second row) synthesized on target faces following the proposed method.

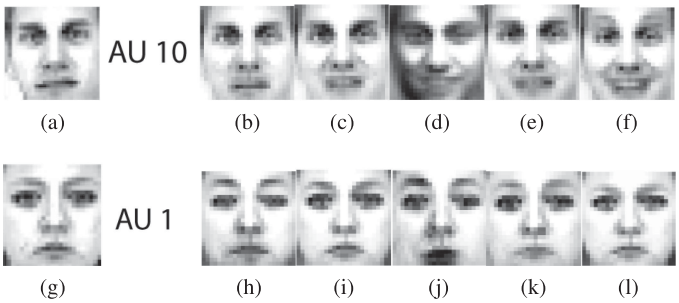


Fig. 10. Row 1 and row 2 show examples of expressions of AU10 (*upper lip raised producing square like furrow around nostrils*) and AU1 (*inner eyebrows raised*) respectively synthesized by Susskind *et al.* (a) and (g) show the target faces and (b) to (f) and (h) to (m) show the corresponding synthesized faces. In (d) the identity of the target has been changed. In (f) AU12 (*lip corners moved obliquely upward* as in *happiness*) has also been synthesized in addition to AU10. In (j) AU25 (*lips parted*) has also been synthesized in addition to AU1.

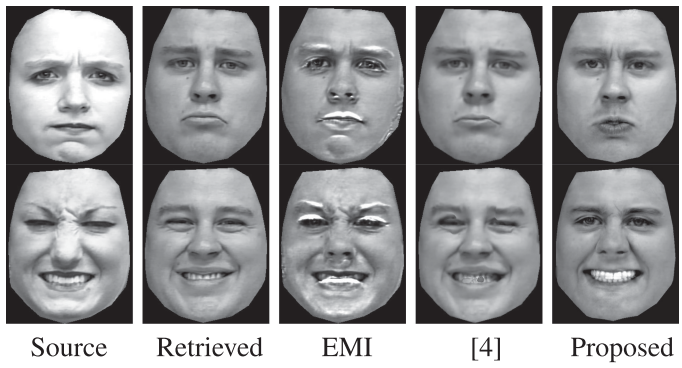


Fig. 11. Comparison of expressions synthesized by Li *et al.* [4] *vs.* our method. Column 1 shows the source persons' expression that we wish to synthesize. Column 2, to column 5 show the images retrieved from target database, the EMIs, the final images synthesized following [4] and the images synthesized by our method respectively.

facial expression of another person (source face). In [4], the synthesis process involves warping of the face image, retrieved from the expression database of the target, towards the Expression Mapping Image (EMI). Though the warping may change the shape of the facial components such as lips, but the transient facial features such as wrinkles and furrows, which are not present in the retrieved image, cannot be generated. In Fig. 11, row 1, column 4 displays *anger* synthesized by [4]. The lips appear pressed as in the corresponding retrieved image (Fig. 11, row 1, column 2). However the vertical wrinkles between the eyebrows as in the source image (Fig. 11, row 1, column 1) are not generated. This absence of vertical wrinkles between eyebrows in the synthesized images can also be seen in Fig. 7 of the article by Li *et al.* [4]. The first image in row 2 of Fig. 11 shows the source image displaying mix of *disgust* and *happiness*. The corresponding retrieved image and the final image (second and the fourth images respectively in row 2 of Fig. 11) synthesized following [4] display pure *happiness*. The wrinkles at the nasal root which is representative of *disgust* could not be synthesized following [4]. Earlier in the current section, we have mentioned the three conditions on which the success of ERI depends. The approach in [4] ignores all the three conditions. Therefore, when the surface normals at corresponding pixel positions are different for source and the target faces, the generated EMI contains noise (the third images in rows 1 and 2 of Fig. 11). When the retrieved face is warped towards the noisy EMI using optical flow, the final synthesized face naturally contains noise. For example, irregularity can be seen in the eyes and lips of the image (fourth image, row 2, Fig. 11) when mix of *disgust* and *happiness* is intended to be synthesized following [4].

Unlike Li *et al.* [4], our method learns the deviations in facial shape and texture due to an expression by *learning from examples*. Therefore, our method is able to generate transient facial features such as wrinkles, furrows *etc*. in the synthesized facial image (column 5 in Fig. 11). Our method does not depend on ERI (unlike [5] and [4]), but synthesizes the facial expression that best suits the facial structure of the target face. This helps in reducing noise in synthesized image. Also, Algorithm 1 mixes the basic expressions seamlessly to synthesize on the target face. The fifth image in the second row of Fig. 11 displays a mix of

Fig. 12. Comparison of expressions synthesized by Xie *et al.* [38] *vs.* our method. Column 1 and 2 show *happiness* synthesized by [38] and the proposed method respectively. Column 1 and 2 show *fear* synthesized by [38] and the proposed method respectively.

disgust and *happiness* synthesized by our method. This mixture expression could not be synthesized following [4].

Comparison With Xie et al. [38]: The basic method of [38] is similar to [5]. Both the methods first warp the target face to manipulate the shape and then modify the texture of the target face in accordance with the change in source face's texture. In [5] the shape representing expression on the target face comes from a non-linear model and in [38], the shape comes from the expressive face of the source person. The method by [5] modifies the texture of the target face by the ratio of the expressive and expression-neutral source face. The method by [38] proposes to modify the texture of the target face such that the lighting difference in a local region of the synthesized target face becomes similar to that of the source face. But like [5] and [4], Xie *et al.* [38] ignores the structure compatibility of the source and target faces. As a result, the synthesized expressions may not look realistic (Fig. 12). Expression synthesized by our method looks realistic as it considers the facial structure while synthesizing.

Quantitative Comparison With the State-of-the-Art: The qualitative comparison shows that our method produces realistic facial expression as compared to the five stated methods. To quantitatively compare the synthesized results, we perform another experiment. We employe 10 human observers. Each of them are presented with randomly chosen 10 sets of six synthesized faces, one face per competing approach. The human volunteers rank the faces in each set. The face looking most realistic should get rank 1 and the one looking most artificial should get rank 6. Thus, 100 synthesized expressive faces (10 sets for each of the 10 persons) per competing approach are used to estimate the rank-1 to rank-6 score of each of the six approaches. If for an approach, p and q number of faces out of 100 are given rank-1 and rank-2 respectively, then the cumulative rank-1 and rank-2 score for that approach become $(p \times 100/100)\% = p\%$ and $(p+q)\%$ respectively. The Cumulative Match Score (CMC) graph shown in Fig. 13 compares the cumulative ranks for all the six competing approaches. The greater the area under the curve, the better is the method. The rank of the proposed approach is similar to [4], [20] and [38] with our approach performing slightly better. The performance of [4] is same with our approach for rank-1 to rank3. For rank-4, our approach gives the best performance.

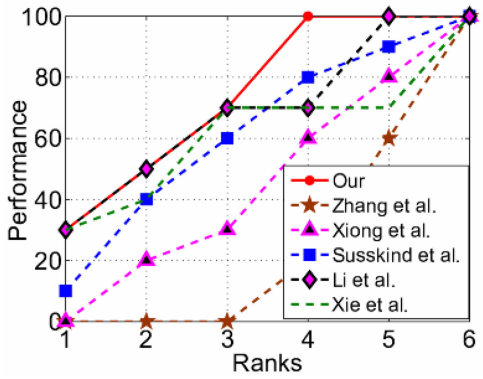


Fig. 13. Comparison of CMC curves for the six competing approaches: ours, Zhang *et al.* [17], Xiong *et al.* [5], Susskind *et al.* [20], Li *et al.* [4] and Xie *et al.* [38]. The maximum area is covered by the proposed approach followed by [4], [20] and [38].

VI. Conclusion

We have introduced a new method for synthesis of facial expressions representing a mix of basic emotions.

- To the best of our knowledge, our method is the first one to synthesize facial expressions using an example based pre-trained model where the expression to be synthesized is specified in terms of a combination of six basic emotions (*e.g.*, *30% happy and 70% surprised*).
- To synthesize expression, we need only one expression-neutral target face image.
- Both qualitative and quantitative analysis show that the proposed method synthesizes realistic and easily perceivable expressions.

We are now extending our synthesis model to render realistic animation.

References

[1] M. Song *et al.*, "A generic framework for efficient 2-d and 3-d facial expression analogy," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1384–1395, Nov. 2007.

[2] A. Asthana, M. de la Hunty, A. Dhall, and R. Goecke, "Facial performance transfer via deformable models and parametric correspondence," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1511–1519, Sep. 2012.

[3] K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu, "A data-driven approach for facial expression synthesis in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 57–64.

[4] K. Li *et al.*, "A data-driven approach for facial expression retargeting in video," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 299–310, Feb. 2014.

[5] L. Xiong, N. Zheng, Q. You, and J. Liu, "Facial expression sequence synthesis based on shape and texture fusion model," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2007, vol. 4, pp. IV-473–IV-476.

[6] D. Huang and F. De la Torre, "Bilinear kernel reduced rank regression for facial expression synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 364–377.

[7] S. Agarwal and D. P. Mukherjee, "Decoding mixed emotions from expression map of face images," in *Proc. Int. Conf. Workshops Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6, doi. 10.1109/FG.2013.6553731.

[8] S. Agarwal, M. Chatterjee, and D. P. Mukherjee, "Synthesis of emotional expressions specific to facial structure," in *Proc. 8th Indian Conf. Vis., Graph. Image Process.*, Dec. 2012, pp. 28:1–28:8, doi. 10.1145/2425333.2425361.

[9] S. Haykin, *Neural Networks A Comprehensive Foundation*. New Delhi, India: Dorling Kindersley (India) Pvt. Ltd., 2009.

[10] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse

regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.
[11] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 314–326, Jul./Sep. 2014.
[12] J.-C. Wang, Y.-S. Lee, Y.-H. Chin, Y.-R. Chen, and W.-C. Hsieh, "Hierarchical dirichlet process mixture model for music emotion recognition," *IEEE Trans. Affective Comput.*, vol. 6, no. 3, pp. 261–271, Jul./Sep. 2015.
[13] Y. Wu, H. Liu, and H. Zha, "Modeling facial expression space for recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 1968–1973.
[14] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans: Research overview and perspectives," *J. Mach. Learn. Res.*, vol. 13, pp. 1589–1608, 2012.
[15] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a gaussian mixture model," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 56–68, Jan./Mar. 2015.
[16] W. F. Liu, J. L. Lu, Z. F. Wang, and H. J. Song, "An expression space model for facial expression analysis," in *Proc. IEEE Congr. Image Signal Process.*, 2008, pp. 680–684.
[17] Q. Zhang, Z. Liu, G. Quo, D. Terzopoulos, and H.-Y. Shum, "Geometry-driven photorealistic facial expression synthesis," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 1, pp. 48–60, Jan./Feb. 2006.
[18] C. Malleson *et al.*, "Facedirector: Continuous control of facial performance in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3979–3987.
[19] T. Zhang *et al.*, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.
[20] J. M. Susskind, A. K. Anderson, G. E. Hinton, and J. R. Movellan, *Generating Facial Expressions With Deep Belief Nets*. London, U.K.: INTECH Open Access Publisher, 2008.
[21] K. Olszewski, J. J. Lim, S. Saito, and H. Li, "High-fidelity facial and speech animation for VR HMDs," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 221.
[22] S. Saito, T. Li, and H. Li, "Real-time facial segmentation and performance capture from RGB input," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 244–261.
[23] T. Kanade, J. Cohn, and Y.-L. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53, doi. 10.1109/AFGR.2000.840611.
[24] M. D. Zeiler, G. W. Taylor, L. Sigal, I. Matthews, and R. Fergus, "Facial expression transfer with input-output temporal restricted Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1629–1637.
[25] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Eur. Conf. Comput. Vis.*, May 2004, pp. 25–36.
[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[27] H. Bhaskar, D. La Torre, and M. Al-Mualla, "Exaggeration quantified: An intensity-based analysis of posed facial expressions," in *Proc. Adv. Face Detection Facial Image Anal.*, 2016, pp. 101–128.
[28] H. Mao, L. Jin, and M. Du, "Automatic classification of chinese female facial beauty using support vector machine," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2009, pp. 4842–4846, doi. 10.1109/FG.2013.6553750.
[29] W. Yin, D. Goldfarb, and S. Osher, "Image cartoon-texture decomposition and feature selection using the total variation regularized l^1 functional," in *Variational, Geometric, and Level Set Methods in Computer Vision*, vol. 3752. New York, NY, USA: Springer, 2005, pp. 73–84.
[30] H. Yu, O. G. Garrod, and P. G. Schyns, "Perception Deiven facial expression synthesis," *Comput. Graph.*, vol. 36, no. 3, pp. 152–162, Dec. 2011.
[31] P. Lucey *et al.*, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 94–101.
[32] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *Proc. IEEE Image Anal. Multimedia Interactive Services.*, 2010, pp. 1–4.
[33] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005, p. 5.
[34] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. Int. Lang. Resources Eval. Conf.*, May 2010, p. 65.
[35] P. Ekman and W. V. Friesen, *The facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press Inc., 1978.
[36] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.
[37] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[38] W. Xie, L. Shen, and J. Jiang, "A novel transient wrinkle detection algorithm and its application for expression synthesis," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 279–292, Feb. 2017.
[39] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 271–276.

Swapna Agarwal received the Bachelor's degree (rank: first in the university) from Vidyasagar University, Midnapore, India, and the Master's degree (rank: second in the university) from Visvabharati University, Santiniketan, India. She received the Ph.D. degree in computer science from Indian Statistical Institute, Kolkata, India. Her research interests includes computer vision, pattern recognition and affective computing.

Dipti Prasad Mukherjee is a Professor with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, India. He has written more than 100 peer-reviewed research papers. Prof. Mukherjee is the Fellow of the Computer Society of India and the Institution of Engineers, India. He is currently serving on the editorial boards of IEEE TRANSACTIONS ON IMAGE PROCESSING and *IET Image Processing*.