A Data-Driven Approach for Facial Expression Retargeting in Video

Kai Li, *Student Member, IEEE*, Qionghai Dai, *Senior Member, IEEE*, Ruiping Wang, *Member, IEEE*, Yebin Liu, Feng Xu, and Jue Wang, *Senior Member, IEEE*

***Abstract*—This paper presents a data-driven approach for facial expression retargeting in video, i.e., synthesizing a face video of a target subject that mimics the expressions of a source subject in the input video. Our approach takes advantage of a pre-existing facial expression database of the target subject to achieve realistic synthesis. First, for each frame of the input video, a new facial expression similarity metric is proposed for querying the expression database of the target person to select multiple candidate images that are most similar to the input. The similarity metric is developed using a metric learning approach to reliably handle appearance difference between different subjects. Secondly, we employ an optimization approach to choose the best candidate image for each frame, resulting in a retrieved sequence that is temporally coherent. Finally, a spatio-temporal expression mapping method is employed to further improve the synthesized sequence. Experimental results show that our system is capable of generating high quality facial expression videos that match well with the input sequences, even when the source and target subjects have big identity difference. In addition, extensive evaluations demonstrate the high accuracy of the learned expression similarity metric and the effectiveness of our retrieval strategy.**

***Index Terms*—Facial expression, expression retargeting, expression synthesis, expression similarity metric, data-driven.**

I. Introduction

FACIAL expression retargeting, or performance-driven facial animation, usually refers to the problem of synthesizing facial expressions of a target subject that exhibits the same expressions of a source subject. There are three criteria that such a successful retargeting system should meet: (1) *similarity*, meaning that the synthesized expressions should be perceptually close to those in the input performance, although the subjects are different; (2) *naturalness*, meaning that the synthesized expressions should look natural without noticeable artifacts; and (3) *efficiency*, the proposed system should require minimal user input and is general enough to handle various subjects.

This problem has drawn plenty of attention since the 1980s, yet it still largely remains unsolved. Previous methods often fail to meet all requirements mentioned above at the same time. For instance, many previous approaches [1], [2] focus on the similarity criterion, however they require too much user interaction for generating the output expressions. Other methods such as performance-based facial animation [3] focus on photo-realistic rendering of the synthesized expressions, however they require accurate 3D face models of the subjects, which are hard to obtain without using special devices and setups.

Recently, data-driven approaches have shown great potential in various synthesis problems such as creating human motions [4] and completing occluded faces [5]. Inspired by this methodology, we pre-capture a video database of the target subject to achieve photo-realistic expression retargeting. Our database includes some basic expressions such as neutral, angry, disgust, fear, happiness, sadness, and surprise. Since video frames in the database contain the ground-truth appearance of the target person under various expressions, they can be used as strong appearance priors for rendering new expressions. This allows us to develop an efficient facial expression retargetting system without using accurate 3D models which are hard to obtain.

However, developing such a data-driven expression retargeting system is a challenging task. One key problem is how to accurately measure the facial expression distance (or similarity) between different subjects from 2D images, as both identity and expression differences result in appearance differences. In our approach, we use the estimated motion (i.e., optical flow field) from the neutral face to an expression face of the same subject to characterize the latter. Based on this description, we derive a strict metric for measuring expression difference between different subjects. Furthermore, since the exact motion with respect to the same facial expression is different for different subjects, a metric learning approach is utilized to learn the subject-specific metric between the source and the target subject.

Another problem is how to query the database to generate a sequence that has similar expressions to the input video, while maintaining its temporal coherence. A straightforward solution

Manuscript received April 27, 2013; revised September 19, 2013; accepted October 11, 2013. Date of publication November 26, 2013; date of current version January 15, 2014. Part of this work was done when R. Wang was a postdoctoral researcher at Tsinghua University. F. Xu contributed to this work when he was a Ph.D. student at Tsinghua University. This work was supported by the National Basic Research Project (No. 2010CB731800) and the Project of NSFC (No. 61035002 & 61120106003). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris.

K. Li is with the Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China, and also with the Graduate School at Shenzhen, Tsinghua University (e-mail: l-k10@mails.tsinghua.edu.cn).

Q. Dai and Y. Liu are with the Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China (e-mail: qionghaidai@tsinghua.edu.cn; liuyebin@tsinghua.edu.cn).

R. Wang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangruiping@ict.ac.cn).

F. Xu is with the Microsoft Research Asia, Beijing 100080, China (e-mail: fengxu@microsoft.com).

J. Wang is with Adobe Systems Incorporated, Seattle, WA 98103 USA (e-mail: juewang@ieee.org).

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.2013.2293064

is to find the most similar expression in the database for each input frame, and then concatenate them into a sequence. However, this approach does not guarantee any temporal coherence of the synthesized video. In our system, instead of simply finding the most similar expressions, we select k nearest neighbors for each input frame, and optimize the retrieved sequence via the shortest path algorithm to balance the trade-off between temporal coherence and expression similarity.

Finally, given the limited size of the database, it is impossible to always find a good match for a given expression in the input video. To handle such cases, it is required to synthesize new expressions of the target subject that are not in the database in order to match with the input ones. Previous expression mapping methods [2], [6] can be used for this purpose, however they have difficulties to generate the correct facial texture, as we will demonstrate later. Our system employs a new spatio-temporal expression mapping method for synthesizing new expressions. We further combine the mapped expressions with the retrieved ones through a refinement step, resulting in a final synthesized video that meets both the similarity and naturalness requirements.

The preliminary version of this paper has appeared in [7]. Besides updating the current state of the arts and presenting more examples, we improve the core component of the system by redefining the expression distance measure as a strict metric and presenting a learning-based approach to improve its robustness and accuracy. We conduct a thorough evaluation on the new learning-based metric and show that it outperforms the original expression metric proposed in [7]. We also improved the expression mapping technique used in the previous work, by introducing a novel spatio-temporally coherent expression mapping method, which is specifically designed for video. Finally, we present more thorough experiments and evaluations to demonstrate the effectiveness of the proposed system.

A. Related Works

Early research on facial animation heavily relies on manually specified facial features. For example, Litwinowicz and Williams [1] animated the image with line drawings by texture mapping. Beier and Neely [2] presented a metamorphosis technique by providing line pairs on face image. Liu *et al.* [6] proposed the concept of expression ratio image, i.e., the illumination changes from the neutral image to an expression image, which is applied to another person's neutral face to achieve expression transfer purpose. Zhang *et al.* [8] generated the photorealistic expressions through a combination of examples images in each face subregion. The synthesized texture is inferred by applying the geometry relationship in each subregion to the texture of example images.

Using 3D face models for expression transfer has also been extensively studied. Williams [3] developed a puppetry system by tracking the expressions of a human face with markers and applying them to a textured head model. Pighin *et al.* [9] showed how to create a textured 3D model for facial animation from multiple images. Noh and Neumann [10] proposed a technique to clone the facial expression motion vectors from an input model to a target one. More advanced 3D face capture and animation systems, such as [11], were developed for high-resolution 3D model acquisition, but the cost of these systems is high.

Various statistical face models, e.g., PCA-based models, have been widely explored for generating new facial expressions. Active Appearance Model (AAM) [12] and Constrained Local Model (CLM) [13] are built upon training images with labeled landmarks to capture the facial appearance statistics. Similarly, the morphable model [14], [15] is formed by transforming the shape and texture of existing 3D textured models through linear combination. The multilinear model [16], [17] decomposes the training data into vertex, expression, and identity dimensions separately. These models are designed to be robust to model unseen faces using statistics from the training data. However, they often fail to take into account the high frequency expression details, since they only preserve the principal components. Furthermore, the base vectors of PCA-based models usually lack semantic meaning, thus using them for facial animation is non-intuitive. A region-based PCA model [18] was recently proposed to partially overcome these problems. Besides the PCA-based models, blendshape model [19]–[22] has proven its worth in industry. In this model the shape of a new expression is modeled as a person-specific blending of its neutral expression and the displacements of other expressions w.r.t the neutral pose. However, building this model for a target subject requires extensive skills.

Some new face video generation and editing systems emerged recently. Kemelmacher-Shlizerman *et al.* [23] developed a near-realtime puppetry system through image retrieval. However, the generated face video often lacks temporal coherence when the dataset of the target subject is small. Based on this work Kemelmacher-Shlizerman *et al.* [24] further reported an approach for creating face movies from large amount of photos, by seeking an optimal order of these photos and introducing smooth transitions between aligned face images. Dale *et al.* [17] presented a system to replace faces in videos that are of approximately the same appearances, expressions, and poses with a 3D multilinear model. Yang *et al.* [25] proposed a facial expression editing system by reconstructing the 3D geometry of each frame and factorizing the 3D models using a spatio-temporal multilinear model. This allows the user to manipulate the expression and identity dimensions separately.

Our work is also related to previous research on developing facial expression distance metric. Given the large volume of work on facial expression recognition, it is natural to try to use the facial features previously used for expression recognition for developing an expression distance metric. However, these features may not be able to yield a continuous distance function. For instance, the Facial Action Coding System (FACS) [26] used in the CERT system [27] only performs well when recognizing the large motions of action units. It is unclear how to distinguish subtle changes of action units. Furthermore, these features usually do not take identity difference into account. For example, the Gabor wavelet used in [28] and Local Binary Pattern (LBP) [29] used in [23], [24] may generate noticeably different scores if the same expression is performed by different subjects. In contrast, the metric proposed in our system takes identity difference into consideration and only measures the expression difference.

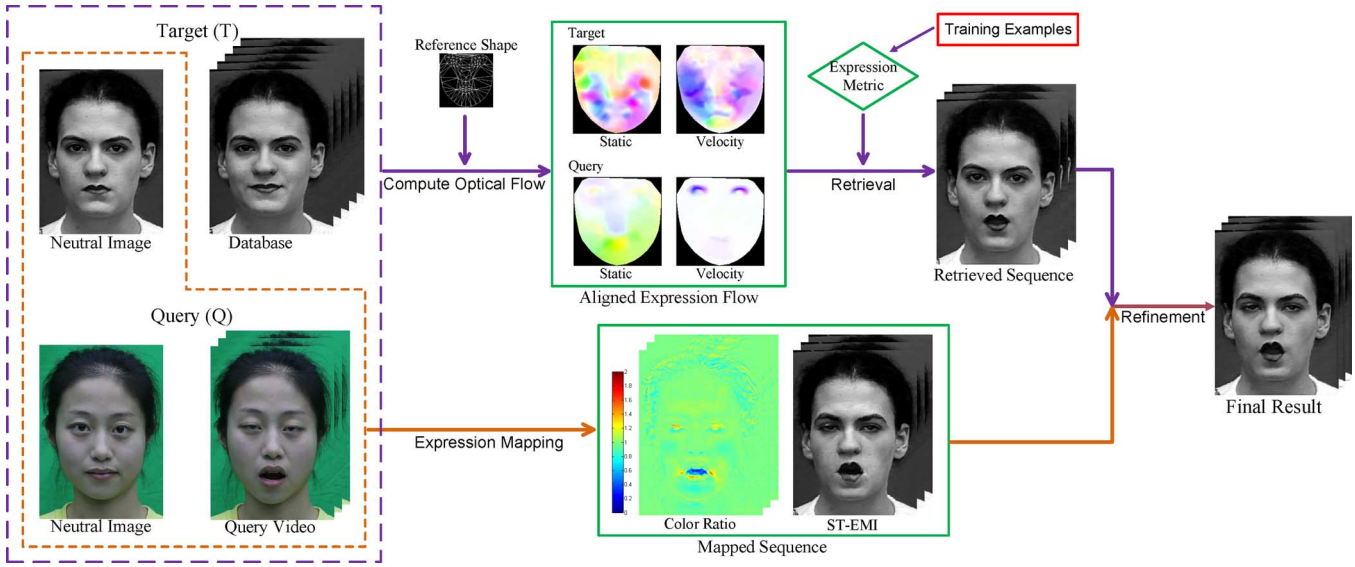


Fig. 1. System overview. To generate a new video of the target subject based on the input facial performance of the query subject, we query the pre-captured expression database of the target subject to obtain an initial retrieved sequence. We then generate another sequence using expression mapping techniques, and combine two sequences together to generate the final result.

B. System Overview

The system flowchart is shown in Fig. 1. Given the input sequence of the query subject, and the expression database of the target subject, we first identify the neutral expression for both subjects, which will be used in the expression metric for retrieval. The expression metric measures not only the motion from the neutral frame to an expression frame, but also the temporal motion velocity at the expression frame. The metric is further improved by a learning-based approach. Using the metric the system produces a retrieved sequence. The system also generates a synthesized sequence using expression mapping, and finally combine these two sequences together to produce the final result.

II. Learning-Based Expression Metric

In this section, we describe our expression similarity metric that takes into account the appearance difference between different subjects. Our metric is built upon an optical flow-based descriptor, which can capture subtle facial expression changes between two facial images, as described in Section II.A. We further show that for video frames, good expression matching requires not only their static facial expression distance, but also higher order statistics such as the velocity of expression changes at individual frames. By incorporating expression velocity, we construct an Euclidean facial expression similarity metric, as described in Section II.B. Finally, in Section II.C we present a metric learning approach which improves the accuracy and robustness of the proposed metric.

A. Optical Flow-Based Descriptor

Given an expression image Q_e of the query subject Q and another expression image T_e of the target subject T, our goal is to compute the expression distance (or similarity) between Q_e and T_e. We take advantage of the corresponding neutral face Q_n of person Q and T_n of T to achieve this goal, which are manually selected from the videos. This needs to be done only once for each subject. The motion filed between Q_n and Q_e (similarly between Q_n and Q_e) can well capture the facial difference caused by the expression shown in Q_e. We thus estimate the motion using an existing optical flow approach [30] and build the metric upon it.

Denote the optical flow between Q_n and Q_e as $\boldsymbol{F}_{Q_n \to Q_e}$, so is $\boldsymbol{F}_{T_n \to T_e}$ between T_n and T_e. However, directly comparing $\boldsymbol{F}_{Q_n \to Q_e}$ and $\boldsymbol{F}_{T_n \to T_e}$ to measure the expression similarity is not good, as two flow fields contain identity difference and are in different coordinates. Therefore, registration is required for a meaningful comparison.

To accurately align the optical flow fields, we first extract facial landmarks using the Active Shape Model (ASM) [31], [32] (Fig. 2(a)), then remove the non-deformation offset such as 2D translation, rotation, and scaling, by computing a similarity transform matrix between two models on the nose region (Fig. 2(b)), which is mostly invariant to expression changes. To account for the difference between two facial shapes, we build a piece-wise affine mapping function via Delaunay Triangulation, as shown in Fig. 2(c). Using this mapping function, we map both flow fields to a common reference face shape $\Re$, which is a pre-defined canonical shape. If such a canonical shape is not available, we use the mean shape of the neutral faces of both the query and the target subject as $\Re$, as shown in Fig. 2(c). Note that this is different from the mapping method used in [7], where the flow filed of the query subject is directly mapped onto the face region of the target subject. This one directional mapping leads to an asymmetric distance metric. We avoid this problem by mapping both flow fields to a canonical shape, ensuring that the derived distance is symmetric.

Specifically, given a pixel a in $\Re$'s face region and the piecewise mapping function $g : Q_n \to \Re$ which encodes the pixel correspondence between Q_n and $\Re$, the corresponding pixel in Q_n is given by $b = g^{-1}(a)$. Then, the optical offset Δb for pixel

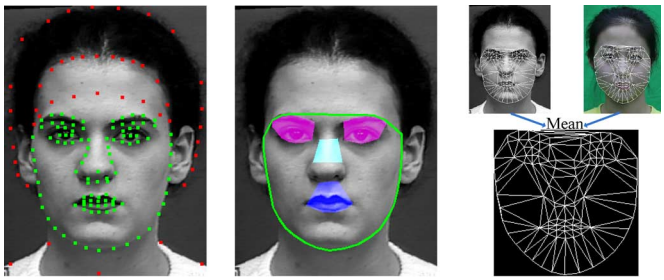


Fig. 2. Initial processing of the neutral face. (a) Green markers are automatically detected, and red ones are manually labeled used for expression mapping in Section IV.A. (b) The green contour shows the face region. The eye, nose and mouth regions are marked in magenta, cyan, and blue, respectively. (c) The mean shape of query neutral face and target neutral face can be regarded as a reference shape.

b can be interpolated from the optical flow matrix $\boldsymbol{F}_{Q_n \to Q_e}$. Therefore, the optical offset Δa for pixel a is computed as:

$$\Delta a = g(b + \Delta b) - a. \tag{1}$$

Equation (1) holds under a direct assumption that the mapping function from Q_e to the virtually deformed expression face of $\Re$ equals to g. Since facial landmarks change according to the facial expressions, the piece-wise correspondence roughly does not change, thus the assumption holds. In this way we can compute two mapped and aligned flow fields on the canonical shape as $\boldsymbol{M}_{Q_n \to Q_e}$, $\boldsymbol{M}_{T_n \to T_e} \in \mathbb{R}^{n \times 2}$, where n is the number of pixels in $\Re$'s face region, i.e., only the optical flow in the face region is considered.

With the aligned flow fields, the L2 norm can be used to measure their difference. For two optical offset vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, their distance is

$$d_b(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|_2, \tag{2}$$

where $d_b(\cdot, \cdot)$ is called the basic distance function. However, one more thing we need to consider is that physically, the optical offsets of pixels on face regions differ in their movement scales. For example, when expression changes from neutral to smile, the eye and mouth regions usually deform significantly, while the nose region remains roughly unchanged. To account for different motion scales of different facial components, we introduce a weight for each pixel:

$$d_e^2(Q_e, T_e) = \sum_i w_i d_b^2(M_{Q_n \to Q_e, i}, M_{T_n \to T_e, i}), \tag{3}$$

where $M_{Q_n \to Q_e, i}$ indicate the ith pixel's offset, i.e., the ith row of flow matrix $\boldsymbol{M}_{Q_n \to Q_e}$, and w_i is the weight for the ith pixel. In our system we set $w_i = \alpha_e = 1.0$ for the eye region (marked in magenta in Fig. 2(b)), $w_i = \alpha_m = 0.1$ for the mouth region (marked in blue in Fig. 2(b)), and $w_i = 0$ for pixels outside these two regions. This is because perceptually, the expression can be well captured by the movement in the mouth and eye regions. The ratio $\alpha_m / \alpha_e = 0.1$ is verified by the cross validation experiment shown in Fig. 6, which can be interpreted as that the movement in the mouse region is ten times as large as that in the eye region.

B. Incorporating Expression Velocity

The distance function derived in (3) only considers the static expression distance. When dealing with video frames, the velocity of expression changes at each moment also needs to be taken into account. In other words, when comparing a query frame and a database frame, we expect not only the static expression distance between them is minimized, but also the expression change momentum at these two times needs to be matched. This will greatly improve the temporal coherence of the retrieved frames.

We measure the expression velocity of a frame in a sequence by computing the optical flow between the current frame and the next frame. Let $Q_e^{(q)}$ be the qth frame of the query sequence, its expression velocity $\mathrm{d}\boldsymbol{F}_{Q_e^{(q)}}$ is given by:

$$\mathrm{d}\boldsymbol{F}_{Q_e^{(q)}} = \boldsymbol{F}_{Q_e^{(q)} \to Q_e^{(q+1)}}. \tag{4}$$

The expression velocity of the tth frame of the target database sequence, i.e., $\mathrm{d}\boldsymbol{F}_{T_e^{(t)}}$, is defined in the same way.

To compute the velocity difference, both $\mathrm{d}\boldsymbol{F}_{Q_e^{(q)}}$ and $\mathrm{d}\boldsymbol{F}_{T_e^{(t)}}$ need to be firstly aligned to the reference face shape $\Re$ to remove identity difference. Specifically, for $Q_e^{(q)}$, aligned flow $\boldsymbol{M}_{Q_n \to Q_e^{(q)}}$ implies that each pixel in the reference shape corresponds to a point in $Q_e^{(q)}$, thus optical flow between $Q_e^{(q)}$ and $Q_e^{(q+1)}$ can be easily mapped back to the reference shape. Let $\mathrm{d}\boldsymbol{M}_{Q_e^{(q)}}$ and $\mathrm{d}\boldsymbol{M}_{T_e^{(t)}}$ denote the mapped velocity flow of $\mathrm{d}\boldsymbol{F}_{Q_e^{(q)}}$ and $\mathrm{d}\boldsymbol{F}_{T_e^{(t)}}$, respectively. The velocity distance is computed as:

$$d_v^2\left(Q_e^{(q)}, T_e^{(t)}\right) = \sum_i w_i d_b^2\left(\mathrm{d}M_{Q_e^{(q)}, i}, \mathrm{d}M_{T_e^{(t)}, i}\right), \tag{5}$$

where w_i and $d_b(\cdot, \cdot)$ are the same as those in (3).

Combining the static expression distance in (3) and the expression velocity difference in (5) together, our complete expression distance measure for video frames is defined as:

$$\mathcal{D}_E^2\left(Q_e^{(q)}, T_e^{(t)}\right) = \gamma_e d_e^2\left(Q_e^{(q)}, T_e^{(t)}\right) + \gamma_v d_v^2\left(Q_e^{(q)}, T_e^{(t)}\right), \tag{6}$$

where γ_e and γ_v are balancing weights for two distance terms, subject to $\gamma_e + \gamma_v = 1$ ($\gamma_e = 0.9$ and $\gamma_v = 0.1$ in our system, which are obtained by the cross validation experiment shown in Fig. 6). A small weight for the expression velocity term helps to make the expression distance more accurate. The proposed distance function in (6) is a strict metric that satisfies non-negativity, symmetry, and triangle inequality.

It is easy to show that the distance metric in (6) can be formulated as linear algebra operations. We vectorize the optical flow matrix $\boldsymbol{M}_{Q_n \to Q_e^{(q)}}$ and $\mathrm{d}\boldsymbol{M}_{Q_e^{(q)}}$, and then integrate them into one single column vector $\boldsymbol{m}_{Q_e^{(q)}} \in \mathbb{R}^{4n}$, which is called the Expression Flow feature:

$$\boldsymbol{m}_{Q_e^{(q)}} \doteq \left[\mathrm{vec}\left(\boldsymbol{M}_{Q_n \to Q_e^{(q)}}\right); \mathrm{vec}\left(\mathrm{d}\boldsymbol{M}_{Q_e^{(q)}}\right)\right]. \tag{7}$$

The distance measure in (6) is rewritten as:

$$\begin{aligned} &\mathcal{D}_E^2\left(Q_e^{(q)}, T_e^{(t)}\right) \\ &\quad = \mathcal{D}_E^2\left(\boldsymbol{m}_{Q_e^{(q)}}, \boldsymbol{m}_{T_e^{(t)}}\right) \\ &\quad = \left(\boldsymbol{m}_{Q_e^{(q)}} - \boldsymbol{m}_{T_e^{(t)}}\right)^T \boldsymbol{W} \left(\boldsymbol{m}_{Q_e^{(q)}} - \boldsymbol{m}_{T_e^{(t)}}\right), \end{aligned} \tag{8}$$

where $\boldsymbol{W} \in \mathbb{R}^{4n\times 4n}$ is a diagonal matrix:

$$\boldsymbol{W} = \mathrm{diag}([w_1\gamma_e, \ldots, w_n\gamma_e, w_1\gamma_e, \ldots, w_n\gamma_e, w_1\gamma_v, \ldots, w_n\gamma_v, w_1\gamma_v, \ldots, w_n\gamma_v]). \quad (9)$$

We can see that the proposed metric in (8) is a weighted L2 norm distance function.

C. Metric Learning

The expression metric proposed in (8) is an Euclidean distance. However, the flow-based metric does not encode expression semantics, thus may not be consistent with the human perception of facial expressions. For instance, consider the case that one subject changes his expression from sad to a subtle smile, and then to a big smile. The distance from sad to a subtle smile may be the same as the distance from a subtle smile to a big one, according to the flow-based metric, however the former is usually considered as a bigger mode change for human perception. Under this observation, we propose to use a learning-based approach to make the expression metric be more consistent with human perception.

Metric learning itself is an extensively studied topic in computer vision. Following some previous works [33], [34] which argue that Mahalanobis distance has much better generalization performance than Euclidean distance, in this work we aim to learn a Mahalanobis distance metric for measuring expression similarity. The principle of our Mahalanobis distance learning approach is to select a positive definite matrix $\boldsymbol{A}$, which parameterizes the distance function as:

$$\begin{aligned}&\mathcal{D}_M^2\left(Q_e^{(q)}, T_e^{(t)}\right)\\ &= \mathcal{D}_M^2\left(\boldsymbol{m}_{Q_e^{(q)}}, \boldsymbol{m}_{T_e^{(t)}}\right)\\ &= \left(\boldsymbol{m}_{Q_e^{(q)}} - \boldsymbol{m}_{T_e^{(t)}}\right)^T \boldsymbol{A}\left(\boldsymbol{m}_{Q_e^{(q)}} - \boldsymbol{m}_{T_e^{(t)}}\right). \end{aligned} \quad (10)$$

When $\boldsymbol{A}$ equals to $\boldsymbol{W}$, the Mahalanobis distance degrades to the weighted Euclidean distance in (8). Davis *et al.* [34] solve this problem by minimizing the relative entropy between two multivariate Gaussians. It can handle various constraints, including similar or dissimilar constraints (the distance should be relatively small or large), and relative relations between pairs of distance. In addition, It is fast and robust over high dimensions. Therefore, we use the Information-Theoretic Metric Learning (ITML) approach [34] to select $\boldsymbol{A}$.

Specifically, we have a dataset $\{(Q_e^{(q)})_q, (T_e^{(t)})_t\}$, from which we can extract the expression flow feature set $\Omega = \{(\boldsymbol{m}_{Q_e^{(q)}})_q, (\boldsymbol{m}_{T_e^{(t)}})_t\} = \{(\boldsymbol{m}_i)_i\}$. Suppose the dataset contains some labeled image pairs, i.e., similar and dissimilar pairs. All pairs of samples that are labeled as similar (and dissimilar) form a pairwise similar constraint set S (and a dissimilar constraint set D). Then, the optimization problem becomes how to regularize $\boldsymbol{A}$ to be close to $\boldsymbol{W}$ given the similar and dissimilar constraints. This closeness is measured in an information-theoretic approach. There exists a multivariate Gaussian distribution $p(\boldsymbol{m}; \boldsymbol{A}) = (\frac{1}{Z})\exp(-\frac{1}{2}\mathcal{D}_M^2(\boldsymbol{m}, \boldsymbol{\mu}))$ which corresponds to the distance function in (10), where Z is the normalizing constant, $\boldsymbol{\mu}$ is the mean of all distances, and

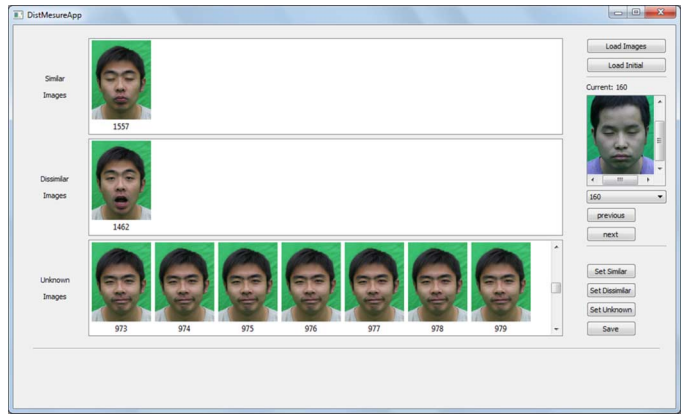


Fig. 3. Our user interface for labeling image pairs for metric learning. Given one expression of one subject shown on the right side of the board, the UI allows the user to specify an image with similar expression (left, top row) and another image with dissimilar expression (left, second row), from the candidates in the last row.

$\boldsymbol{A}$ is the inverse covariance. Consequently, the metric learning problem is formulated as a relative entropy minimization problem:

$$\begin{aligned}\text{minimize}_{\boldsymbol{A}} \quad & \mathrm{KL}(p(\boldsymbol{m}; \boldsymbol{W}) \| p(\boldsymbol{m}; \boldsymbol{A}))\\ \text{subject to} \quad & \mathcal{D}_M(\boldsymbol{m}_i, \boldsymbol{m}_j) \le u, (\boldsymbol{m}_i, \boldsymbol{m}_j) \in S\\ & \mathcal{D}_M(\boldsymbol{m}_i, \boldsymbol{m}_j) \ge l, (\boldsymbol{m}_i, \boldsymbol{m}_j) \in D \end{aligned} \quad (11)$$

where $p(\boldsymbol{m}; \boldsymbol{W})$ and $p(\boldsymbol{m}; \boldsymbol{A})$ are two multivariate Gaussian distributions parameterized by $\boldsymbol{W}$ and $\boldsymbol{A}$, respectively, and KL stands for the Kullback-Leibler divergence between two distributions. u and l are pre-defined upper and lower bounds (we use the 5th and 95th percentiles of the initial Euclidean distances), respectively. The information-theoretic formulation in (11) is solved via Bregman optimization [35]. Please refer to [34] for the detailed procedure.

The labeled dataset required for metric learning can be generated automatically in some cases, if some prior knowledge about the input data is available. For instance, the CK+ dataset [36] contains labeled expression images, thus we can use images with the same expression label to form the similar constraint set S, and those with different labels to form the dissimilar constraint set D. If for other datasets such prior knowledge is not available, we propose an efficient user interface, which allows the user to manually define such relationships, as shown in Fig. 3.

One may notice that in the preliminary version of this work [7], a direction difference term is incorporated in the basic expression distance function to characterize the movement direction of optical flow:

$$d_b^2(\boldsymbol{u}, \boldsymbol{v}) = \beta_m \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 + \beta_o(-\boldsymbol{u}\cdot\boldsymbol{v} + \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2), \quad (12)$$

where β_m and β_o are two balancing weights for the magnitude and orientation terms (they are both set to be 0.5 in this paper, according to the cross validation result shown in Fig. 6), respectively. However, the direction term is defined in an add-hoc way. It makes the resulting distance measure to be not a strict metric. In contrast, the learning-based metric proposed here is more

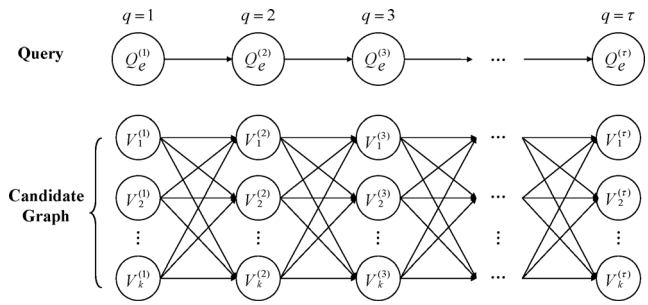


Fig. 4. The directed graph for shortest path optimization.

principled and effective. Compared to the distance measure defined in (12), the learned Mahalanobis distance discards the orientation term to form a strict metric, and makes the parameter matrix more flexible (i.e., not just diagonal). Furthermore, the empirical mixup of magnitude differences in (12) may have poor performance to unseen data. On the contrary, the learned Mahalanobis metric achieves better accuracy by automatically combining the differences from different dimensions, and obviating empirical parameter selection, both implying a better generalization performance. In addition, the parameter matrix $\boldsymbol{A}$ can be decomposed into $\boldsymbol{K}^T\boldsymbol{K}$, where $\boldsymbol{K}$ can be seen as a transform of the expression flow feature. This suggests that the expression flow features can be more distinguishable in the transformed space.

III. Optimization-Based Retrieval

In this section, we show how to query the database to obtain a retrieved sequence which closely matches with the expressions in the input sequence, while maintaining temporal coherence. The main idea is to retrieve multiple candidates for each input frame, and then formulate the retrieval problem as a shortest path energy minimization problem to find the optimal solution.

Specifically, using the facial expression metric, we query the database of the target person to obtain the k nearest neighbors ($k = 20$ in our system) for each frame in the input video, which are called the candidate frames. For the qth input frame $Q_e^{(q)}$, the jth candidate frame is denoted as $C_j^{(q)} \in \{(T_e^{(t)})_t\}$, where $j \in \{1, \ldots, k\}$ and $q \in \{1, \ldots, \tau\}$. Instead of directly lining up the most similar expression images to form the output sequence, we build a directed acyclic graph of the candidate frames, as shown in Fig. 4, to generate the optimal sequence.

In Fig. 4, a directed arc only exists between candidate frames at adjacent moments. For an edge $r_{ij}^{(q)}$ between $C_i^{(q)}$ and $C_j^{(q+1)}$, its length is defined as:

$$\mathcal{L}\left(r_{ij}^{(q)}\right) = \mathcal{D}\left(C_i^{(q)}, Q_e^{(q)}\right) + \mathcal{D}\left(C_j^{(q+1)}, Q_e^{(q+1)}\right) + \eta \exp\left(-\frac{\left(\left|\mathcal{T}\left(C_j^{(q+1)}\right) - \mathcal{T}\left(C_i^{(q)}\right)\right| - \mu\right)^2}{2\sigma^2}\right), \quad (13)$$

where $\mathcal{D}(\cdot,\cdot)$ is the distance function defined in (10), $\mathcal{T}(\cdot)$ is the timestamp function indicating the timestamp in the database sequence of the input frame, η is the balancing weight for the temporal term (an empirical value of 0.05 is used in our system), and μ and σ are the parameters of the exponential penalty function. Specifically, μ is a temporal scale variable which controls the preferred temporal spacing of the selected database frames for adjacent input frames. For instance, if in the input video the subject performs the expression changes much faster than the videos in the database, then we prefer a larger value for μ. The L2 norm in the temporal term allows small temporal shift, and penalizes on large shift. μ can potentially be set automatically at different times according to the motion speed. However, doing so does not result in significant changes in the results, according to our observation. Therefore, μ is fixed to 1 and σ is fixed to 2.

Once the graph is constructed, the retrieval problem can be formulated as a shortest path optimization problem, i.e., finding the path associated with the minimal cost from the start to the end moment. Let $P_{C_i^{(1)} \to C_j^{(\tau)}}$ denote a path from the start node $C_i^{(1)}$ to the end node $C_j^{(\tau)}$. The optimal path $\mathcal{P}_{ret}$ is minimized via:

$$\mathcal{P}_{ret} = \arg\min_{i,j} \arg\min_{P_{C_i^{(1)} \to C_j^{(\tau)}}} \sum_{r \in P_{C_i^{(1)} \to C_j^{(\tau)}}} \mathcal{L}(r), \quad (14)$$

where the inner minimization is to select the optimal path given the fixed start node $C_i^{(1)}$ to end node $C_j^{(\tau)}$, and the outer minimization is to choose the minimal path from all possible combinations of start and end nodes. The above shortest path problem can be solved efficiently via a number of algorithms, such as Dijkstra's algorithm with Fibonacci heaps, as described in detail in [37].

It is worth mentioning that our optimization-based temporal coherence strategy is inspired by, but different from some prior works on creating temporally coherent animation [11], [24], [38]. For instance, Kovar and Gleicher [38] mainly explored how to blend motions with smooth transitions to address the temporal coherence problem, while our goal is not only to create a temporally coherent sequence, but also to make sure each frame in the new sequence has the same expression as the input frame. To the best of our knowledge we are the first to introduce this optimization strategy to the problem of expression retargeting in video.

IV. Expression Refinement

As a common limitation of data-driven methods, the quality of the retrieved sequence generated in Section III is limited by the size of the database. Ideally, the size of the database should be large enough to capture a dense sampling of the entire expression space of the target subject. However in practice this is not always possible. With a limited database, the retrieved images may not exactly match with the input expressions. Furthermore, the temporal jitter will be inevitable when we retrieve expressions for which we do not have dense samples in the database. To deal with these issues, we propose a refinement step which improves the quality of the retrieved sequence by combining it with expression mapping techniques.

A. Expression Mapping

Liu *et al.* [6] proposed so called Expression Ratio Image (ERI) to map the facial expression from one individual to another on 2D images. Using this method, given a neutral image

X and an expression image X' of the same subject as example data, the expression, including its associated appearance details (e.g., winkles around mouth corners when smiling), can be transferred to the neutral face Y of another subject to create a new expression image Y'. The key idea of ERI is to utilize illumination changes to describe facial expression changes. Following this idea, in the preliminary version of the paper [7], we proposed a new method called Expression Mapping Image (EMI) to synthesize novel facial expressions of the target subject given the examples of the query subject. However, the EMI sequence just consists of independently-generated EMI images. Thus, the quality of the generated sequence may be affected by temporal inconsistency. To address this problem, we propose a spatio-temporally coherent expression mapping method (ST-EMI) for video, which can generate temporally coherent expression mapping sequences.

Before introducing the ST-EMI method, we first briefly explain the previous EMI approach. Given an expression face image of the query subject Q_e, with the help of two neutral faces (Q_n and T_n) of both the query and the target subjects, the corresponding expression face T_e of the target person is synthesized by the following steps. Let $a \in Q_n$ and $a' \in Q_e$ denote the same facial point before and after expression changes, and let $b \in T_n$ denote the corresponding point to a. We can see that b equals to $h(a)$, where h is the mapping function from Q_n to T_n. The synthesized expression image is computed via:

$$c_{b'} = c_b \frac{c_{a'}}{c_a}, \tag{15}$$

where b' is the shifted point b after the same expression changes, and $c_{\{\cdot\}}$ is the color value of its subscript symbol. The color ratio

$$p_{b'} = c_{a'}/c_a \tag{16}$$

is called the expression ratio.

The correspondence of neutral and expression face is built by manually labeled markers in [6], which is labor intensive when generating a sequence. Therefore, we compute the correspondence via optical flow as in [7]. To avoid interpolating color values in non-integer position b', which will cause blur, we start the mapping procedure by going through the pixels in T_e instead of Q_n. Given a pixel b' in the expected image T_e, we get the corresponding point a' in Q_e by $a' = h^{-1}(b')$. The original point a in Q_n of shifted a' is obtained via the optical flow between Q_e and Q_n. In this way we find the expression deformation offset of point b. Finally, the expected color value of pixel b' is determined by (15).

Now we demonstrate how to extend the EMI approach to be spatio-temporally coherent. Since the generated images of the target subject all refer to a common neutral image, the key is to make the expression ratio to be both spatially and temporally smooth. We can achieve this by solving the following energy minimization problem:

$$\text{minimize}_{\boldsymbol{R}^{(i)}} \left\| \boldsymbol{R}^{(i)} - \boldsymbol{P}^{(i)} \right\|_2^2 + \nu \left\| \nabla_3 \boldsymbol{R}^{(i)} \right\|_2^2, \tag{17}$$

where $\boldsymbol{P}^{(i)}, i \in \{1, \ldots, \tau\}$ is the original color ratio computed via (16), $\boldsymbol{R}^{(i)}$ is the expected expression ratio at moment i, $\nabla_3 = (\partial_x, \partial_y, \partial_t)^T$ is a three-dimensional gradient operation, and ν is the weight for the gradient-based term (an empirical value of 0.5 is used in our system). The temporal gradient $\partial_t \boldsymbol{R}^{(i)}$ is defined as $\boldsymbol{R}^{(i)} - \boldsymbol{R}^{(i-1)}$, and $\boldsymbol{R}^{(0)}$ is set to $\mathbf{1}$.

Since L2 norm imposes heavy penalty on outliers, robust functions can be embedded in (17) to allow discontinuity. Therefore, we re-formulate the energy function as:

$$\text{minimize}_{\mathbf{R}^{(i)}} \sum_{\mathbf{x}} \Psi\left(\left| R_{\mathbf{x}}^{(i)} - P_{\mathbf{x}}^{(i)} \right|^2 \right) + \nu \Psi\left(\left| \nabla_3 R_{\mathbf{x}}^{(i)} \right|^2 \right), \tag{18}$$

where $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}$. In our system we choose ϵ to be a small value 0.001, resulting in a L1 minimization. $\boldsymbol{x} = (x, y)$ is the pixel position, and $R_{\boldsymbol{x}}^{(i)}$ is the element of ratio matrix $\boldsymbol{R}^{(i)}$ at position $\boldsymbol{x}$. The solution of the above optimization formulation 18 must satisfy the Euler-Lagrange equation:

$$\Psi'\left(\left| R_{\boldsymbol{x}}^{(i)} - P_{\boldsymbol{x}}^{(i)} \right|^2 \right) \cdot R_{\boldsymbol{x}}^{(i)} \cdot \left| R_{\boldsymbol{x}}^{(i)} - P_{\boldsymbol{x}}^{(i)} \right| - \nu \text{div}\left(\Psi'\left(\left| \nabla_3 R_{\boldsymbol{x}}^{(i)} \right|^2 \right) \nabla_3 R_{\boldsymbol{x}}^{(i)} \right) = 0, \tag{19}$$

where $\text{div}(\cdot)$ is the divergence operator. The Euler-Lagrange Equation can be efficiently solved by various numerical approaches, such as successive over-relaxation (SOR) [39]. Finally, by multiplying the expression ratios on all frames with the neutral image of the target person, we obtain the final expression mapping sequence.

B. Combined Synthesis

So far we have obtained the retrieved sequence (Section III) and expression mapping sequence (Section IV.A). On one hand, the retrieved sequence cannot exactly match with the query video and there exists a small amount of temporal jitter. However, the facial appearances of the retrieved sequence come from real data. On the other hand, the synthesized sequence generated from expression mapping is temporally coherent, but the facial appearances on some frames may have artifacts since they are generated from a single neutral image. Given that two results are complimentary to each other, we therefore can combine them together to achieve the final optimal solution.

Our approach for combining two sequences is to compute the optical flow between their corresponding frames, and then use it to warp the retrieved frame to the expression mapping frame to generate a warped frame as the final result. In this way, the advantages of both the retrieved frame and the expression mapping frame are retained. As shown in Fig. 5, the final result not only matches well with the input performance, but also has realistic, artifact-free facial appearance. In addition, it is temporally coherent.

V. Experiments

In this section, we conduct extensive experiments to demonstrate: (1) the accuracy of our proposed facial expression metric, (2) the effectiveness of the temporal coherent retrieval approach, and (3) the effectiveness of the whole system for facial expression retargeting.

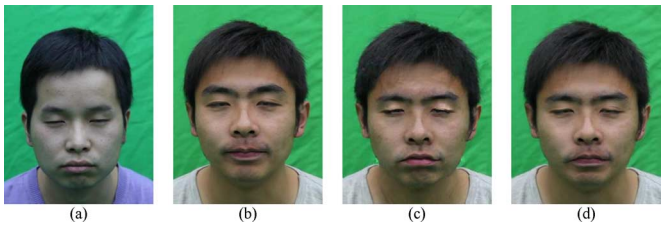


Fig. 5. An example of expression refinement. (a) The query frame. (b) The retrieved frame. (c) The frame generated by expression mapping. (d) The final combined result.

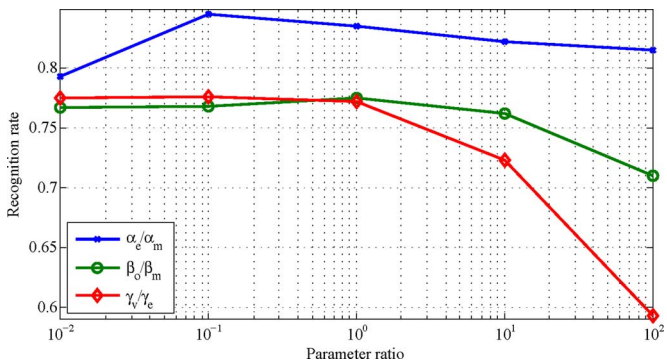


Fig. 6. Cross validation results for tuning parameters. α_e and α_m, β_m and β_o, and γ_e and γ_v are three pairs of parameters need to be tuned. We vary one variable each time. We can see that the best performance can be achieved when $\alpha_m/\alpha_e = 0.1$, $\beta_o/\beta_m = 1$, and $\gamma_v/\gamma_e = 0.1$.

A. Expression Metric Verification

We first evaluate the proposed expression metric on the task of facial expression recognition, which is conducted on the CK+ dataset [36]. The dataset consists of the facial expression sequences of 122 subjects performing from the neutral expression to one certain extreme expression. We pick out only the extreme expression and the neutral expression to construct our recognition dataset. Since only eight expressions are labeled, the recognition task is essentially an eight-category classification (578 neutral, 44 anger, 18 contempt, 58 disgust, 25 fear, 68 happiness, 27 sadness, 82 surprise). A naive k nearest neighbor classifier ($k = 3$ in this experiment) is used for expression classification. The parameters of each method are tuned via 5-fold cross validation, as shown in Fig. 6. We repeat the experiment twenty times and average the recognition rates as the final result.

There are a few options to generate the training set for both the kNN classifier and our learning-based metric. First, following previous multi-class classification methods, we randomly select an equal number of samples from each expression class for training, and use the rest of samples for testing. We use the training dataset to train both our metric and the kNN classifier. We compare the recognition results of the LBP method used in [23], the heuristic method proposed in [7], and our proposed metric in Section II.C, which are labeled as LBP, Unlearned, and Learned in Fig. 7, respectively. The results of this test case, using different number of training examples in each category (maximum number is 18 as the "contempt" category only has 18 examples), are shown in Fig. 7(a). It shows that our learned metric improves the accuracy by up to 20.0% compared with the other two metrics.

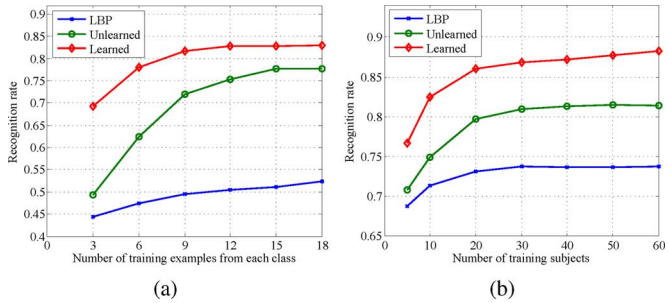


Fig. 7. Facial expression recognition results, whose training examples are obtained by randomly selecting (a) an equal number of samples from each class and (b) a certain number of subjects, respectively.

However, the training and test datasets created in this way may contain the expressions of same subjects. In our retargeting problem, we use the facial expression video of one subject as input to synthesize new facial expressions of another subject, therefore it is more reasonable to include the facial expression of different subjects in the training and test datasets. That is, two datasets should not contain the expressions of the same subject for our application. We then conducted the second round of experiments by using a certain number of subjects as the training set (ranging from 5 to 60), and the rest subjects as the test set. The results are shown in Fig. 7(b). We can see that the learned metric also improve the performance by up to 7.6% than our previous heuristic metric in [7]. More importantly, optical flow-based descriptor is demonstrated to significantly outperform the LBP feature, which fails to accurately describe facial expression changes for the weak kNN classifier.

B. Objective Evaluation on Retrieval

To show the superiority of our temporal coherent retrieval strategy, we conduct experiments for evaluating the quality of the retrieved sequence. First, a video of the query subject performing exactly the same expressions as one of the sequences in the database of the target subject, which we call the ground-truth sequence, is captured. We then remove that sequence from the database, and use different retrieval techniques to generate the retrieved results, using the video of the query subject as input. We evaluate each retrieved sequence by comparing it with the ground-truth sequence. The goal of the evaluation is twofold: first, we want to measure the match precision, which is the facial expression similarity between the retrieved and ground-truth frames; secondly, we want to measure the temporal coherence of the retrieved sequence, as human perception is quite sensitive to temporal jitter. We believe these two factors are the most important ones for measuring the quality of the retrieval results.

Specifically, in our experiments match precision is computed as the average feature distance between the retrieved and ground-truth frames. To accurately measure the feature distance, we manually label facial landmarks on both the retrieved and ground-truth frames, and compute the sum of Euclidean distances between corresponding feature points as the per-frame distance. Temporal coherence is measured by the maximum distance between consecutive frames, since even a single sudden expression jitter in the video can degrade its perceptual quality.

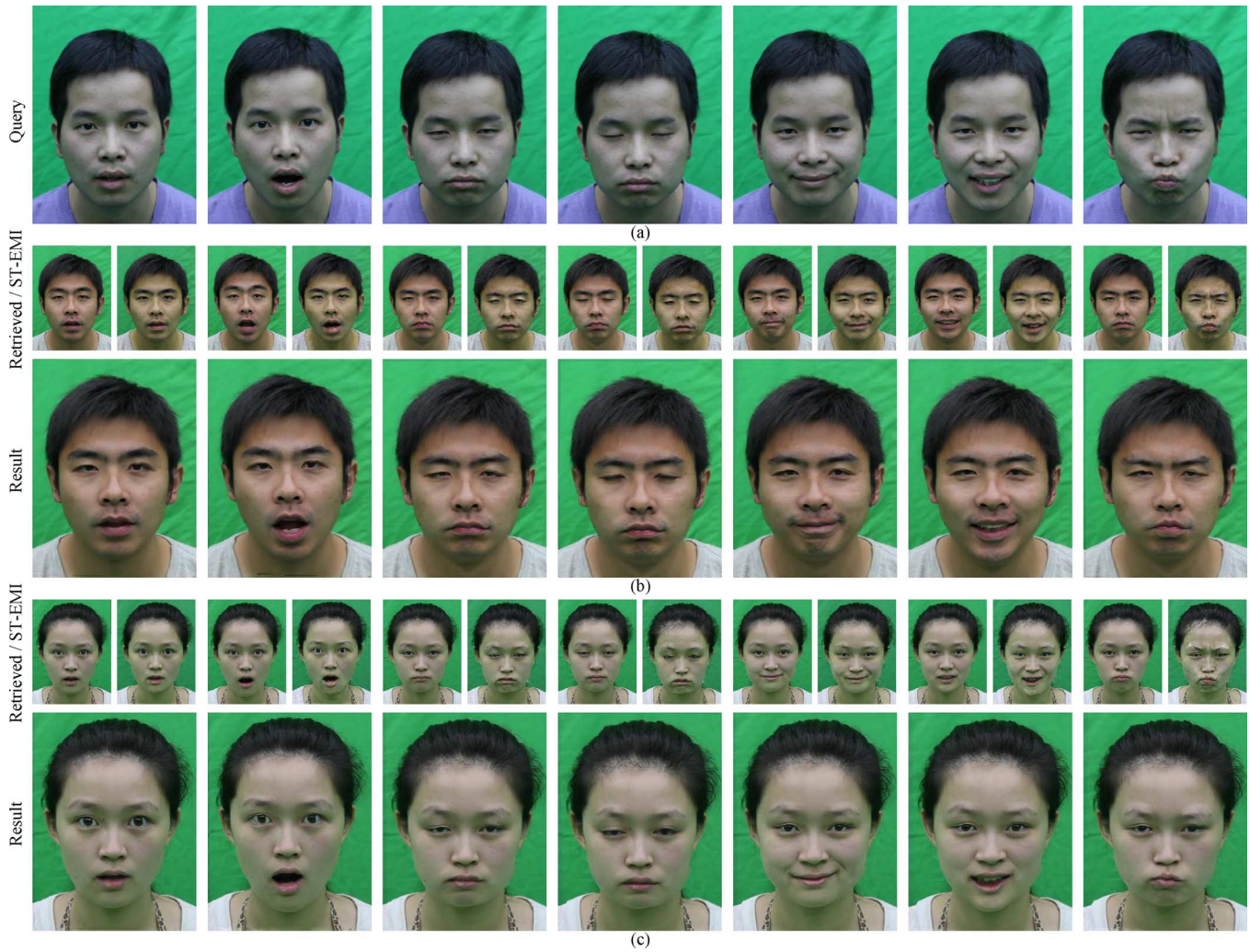


Fig. 8. Generated results of target person $T1$ and $T2$. (a) The common query frames of input person $Q1$. (b) The retrieved, ST-EMI, and final frames of target person $T1$. Frames in the first row are the retrieved and ST-EMI (left and right at each moment, respectively), and those in second row are the final results of $T1$. (c) The retrieved, ST-EMI, and final generated frames of target person $T2$.

We choose 100 pairs of query and target sequences in the CK+ dataset [36] to perform this evaluation. First, by using facial landmarks each pair of query and target sequence is normalized and aligned to a common shape. Since two sequences in each pair may not have exactly the same length, Dynamic Time Warping (DTW) [40] is then applied to align them in the time domain. To gather statistics from the results on different sequences, for each pair we treat the computed distance by the method proposed in [7] (shown in italic in Table I) as baseline. Specifically, the distances of other methods are normalized by those of [7] for each pair of sequences. Then the geometric mean of these ratios among all the test pairs is regarded as the final evaluation result. The training set of the learned metric is generated by randomly choosing 5 subjects that are not used as neither query nor target from the CK+ dataset.

The results of the objective evaluation are shown in Table I. From the results we have the following findings:

1) without the optimization-based retrieval strategy presented in Section III, that is, just by selecting the most similar image for each input frame to construct the sequence, the temporal coherence drops significantly (12.9% for the unlearned metric in [7], 24.0% for the learning-based metric). Meanwhile, the match precision does not change much. This suggests that the optimization-based retrieval strategy can effectively improve the temporal coherence of the retrieved sequence.

TABLE I
OBJECTIVE EVALUATION RESULTS ON DIFFERENT RETRIEVAL STRATEGIES WITH DIFFERENT SIMILARITY METRICS. (A) MATCH PRECISION INDICATOR. (B) TEMPORAL COHERENCE INDICATOR

	No optimization	Retrieval in [23]	Our retrieval
LBP in [23]	1.1785	1.1721	1.1387
Unlearned metric	0.9960	0.9975	*1.0000*
Learned metric	0.8978	0.9003	**0.8881**

(a)

	No optimization	Retrieval in [23]	Our retrieval
LBP in [23]	1.6039	1.5406	1.0917
Unlearned metric	1.1290	1.1077	*1.0000*
Learned metric	1.0382	1.0220	**0.8372**

(b)

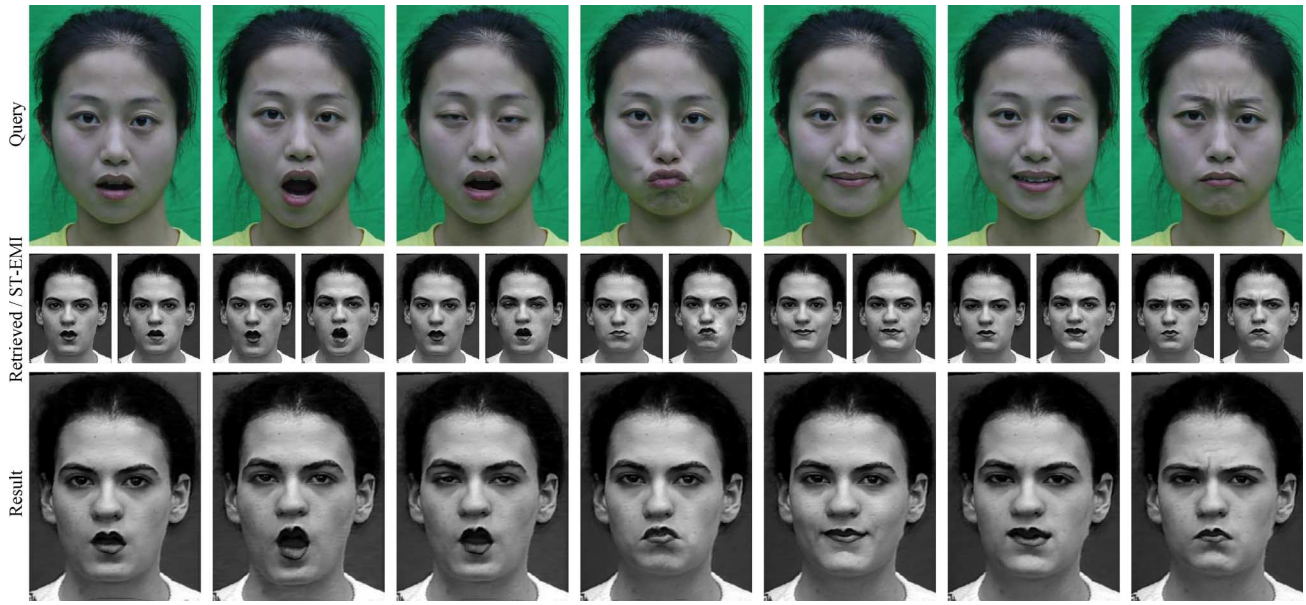


Fig. 9. Generated result of subject $S130$ from the CK+ dataset. Top row: the query frames of input person $Q2$. Middle row: the retrieved and ST-EMI frames of $S130$ (left and right at each moment, respectively). Bottom row: the final generated frames of $S130$.

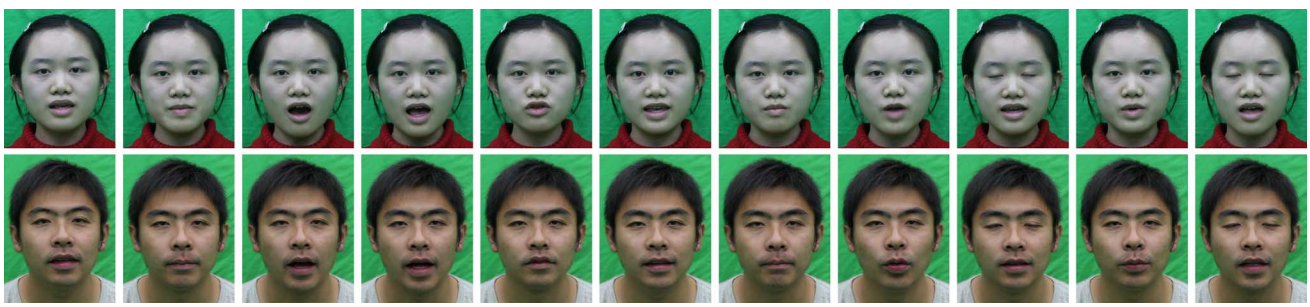


Fig. 10. A talking result of target person $T1$ driven by input person $Q3$. First row: query frames. Second row: synthesized frames.

2) Our optimization-based retrieval strategy outperforms the retrieval strategy proposed in [23]. In particular, with our retrieval method, the temporal coherence of the LBP feature is improved by 31.9%.
3) Our learned expression similarity metric outperforms the LBP feature and the unlearned metric [7] in every test.
4) Overall by combing the learned metric with the optimization-based retrieval strategy, our proposed method significantly outperforms the previous approaches presented in [23] and [7].

C. More Results and User Studies

To evaluate the system as a whole, we collect the databases of three target subjects. The videos of two of the target subjects, one male and one female, denoted as $T1$ and $T2$, respectively, are collected by ourselves. These two non-professionals are asked to perform some basic expressions for one minute. Thus each database video contains roughly 1500 frames. The third target subject is a female subject $S130$ from the CK+ dataset [36], whose database consists of 11 short sequences (220 frames). We also collect facial performance videos of other query subjects as input data to the system. Note that the query subjects are different from the target subjects.

Some generated results are shown in Fig. 8, Figs. 9 and 10. Fig. 8 is the result of target subject $T1$ and $T2$ driven by a common video of a male subject $Q1$. Fig. 9 is the result of subject $S130$ driven by a female subject $Q2$. Fig. 10 is the result of target subject $T1$ driven by a female subject $Q3$. From the results we can see that, first, our system is able to synthesize new expressions that are not performed in the database, such as the pouting expression shown in Fig. 8. This conclusion is also supported by the result with a small target database, as shown in Fig. 9. Secondly, our system can handle expressions performed at different speeds. For example, a fast talking sequence is used as input in Fig. 10, and our system can successfully synthesize the corresponding result for the target subject even though his database does not contain such fast motion examples. The whole set of generated videos is included in the supplementary material of this paper.

Moreover, we perform subjective evaluation experiments to demonstrate the effectiveness of our system. We prepare three pairs of input and output sequences, as shown in Fig. 8(b) (named as $T1$), Fig. 8(c) (named as $T2$) and Fig. 9 (named as S130), respectively. For each example we compare four

TABLE II
SUBJECTIVE EVALUATION RESULTS. THESE RESULTS ARE ALL STATISTICALLY SIGNIFICANT.

	$T1$	$T2$	$S130$
[23]	1.26	1.58	1.41
Our retrieved	2.48	2.97	2.58
Our ST-EMI	2.85	1.91	3.45
Our final	**4.02**	**4.52**	**4.06**

different results: result generated by the system proposed in [23], our retrieved sequence, our expression mapping sequence, and our final result.

We have gathered 30 volunteers in our user study, all of whom are undergraduate students in various departments in our university. For each subject, the query video and a result video are shown side-by-side three times. The subject is then asked to grade this result from 1 (not good at all) to 5 (very good) according to its visual quality compared with the input video. The average user scores for each method and each example are shown in Table II, which suggest that our system generates significantly better results than previous approaches.

D. Limitations

The presented data-driven framework for synthesizing facial expression videos also has some limitations. First, our system only focuses on frontal facial expression synthesis. Generating expressions under various poses is more challenging and is our future work. Secondly, although our system works well with a relatively small database, artifacts tend to appear more often when the size of the database is reduced, and it becomes hard to generate satisfactory results if only a few expressions are known in the database. How to generate better results with very small databases is still an open question. Thirdly, inaccurate optical flow estimation under extreme expressions can affect the performance of the similarity metric and the expression mapping, resulting in bad synthesis results. Finally, our system currently is not able to perform in real-time, which is the ultimate goal we would like to achieve.

VI. CONCLUSION

We have proposed a data-driven framework to transfer the expression performance in an input video to a target subject in the database. To achieve this we first present a learning-based expression similarity metric to measure facial expression similarity between different subjects. We then propose an optimization-based approach for generating a retrieved sequence which matches with the expression performance of the input video. Finally, we improve the retrieved sequence by combining it with the sequence generated by a sptatio-temporally coherent expression mapping method. Quantitative and qualitative evaluation results show that our system significantly outperforms previous approaches on achieving realistic, temporally coherent and accurate expression transfer results.

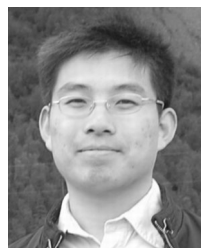
REFERENCES

[1] P. Litwinowicz and L. Williams, "Animating images with drawings," in *Proc. SIGGRAPH*, 1994, pp. 409–412.
[2] T. Beier and S. Neely, "Feature-based image metamorphosis," in *Proc. SIGGRAPH*, 1992, pp. 35–42.
[3] L. Williams, "Performance-driven facial animation," in *Proc. SIGGRAPH*, 1990, pp. 235–242.
[4] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt, "Video-based characters: Creating new human performances from a multi-view video database," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 32:1–32:10, 2011.
[5] Y. Deng, Q. Dai, and Z. Zhang, "Graph Laplace for occluded face completion and recognition," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2329–2338, Aug. 2011.
[6] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *Proc. SIGGRAPH*, 2001, pp. 271–276.
[7] 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. Computer Vision and Pattern Recognition*, 2012, pp. 57–64.
[8] Q. Zhang, Z. Liu, B. Quo, D. Terzopoulos, and H.-Y. Shum, "Geometrydriven photorealistic facial expression synthesis," *IEEE Trans. Visualiz. Comput. Graph.*, vol. 12, no. 1, pp. 48–60, 2006.
[9] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proc. SIGGRAPH*, 1998, pp. 75–84.
[10] J.-y. Noh and U. Neumann, "Expression cloning," in *Proc. SIGGRAPH*, 2001, pp. 277–288.
[11] Z. Li, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: High resolution capture for modeling and animation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 548–558, 2004.
[12] 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, 2001.
[13] J. M. Saragih, S. Lucey, and J. F. Cohn, "Real-time avatar animation from a single image," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2011, pp. 117–124.
[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proc. SIGGRAPH*, 1999, pp. 187–194.
[15] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, 2003.
[16] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 426–433, 2005.
[17] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister, "Video face replacement," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 130:1–130:10, 2011.
[18] J. R. Tena, F. De la Torre, and I. Matthews, "Interactive region-based linear 3d face models," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 76:1–76:10, 2011.
[19] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off: Live facial puppetry," in *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation*, 2009, pp. 7–16.
[20] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 32:1–32:6, 2010.
[21] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 77:1–77:10, 2011.
[22] Y. Seol, J. Lewis, J. Seo, B. Choi, K. Anjyo, and J. Noh, "Spacetime expression cloning for blendshapes," *ACM Trans. Graph.*, vol. 31, no. 2, pp. 14:1–14:12, 2012.
[23] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, and S. M. Seitz, "Being John Malkovich," in *Proc. Eur. Conf. Computer Vision*, 2010, pp. 341–353.
[24] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz, "Exploring photobios," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 61:1–61:10, 2011.
[25] F. Yang, L. Bourdev, E. Shechtman, J. Wang, and D. Metaxas, "Facial expression editing in video using a temporally-smooth factorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 861–868.
[26] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique For the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[27] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2011, pp. 298–305.
[28] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[29] T. Ojala, M. Pietikainen, 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, 2002.
[30] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[31] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vision Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.
[32] Z. Niu, S. Shan, and X. Chen, "Facial shape localization using probability gradient hints," *IEEE Signal Process. Lett.*, vol. 16, no. 10, pp. 897–900, 2009.
[33] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Adv. Neural Inf. Process. Syst.*, 2005, pp. 1473–1480.
[34] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information theoretic metric learning," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 209–216.
[35] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. New York, NY, USA: Oxford Univ. Press, 1997.
[36] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Workshop CVPR for Human Communicative Behavior Analysis*, 2010, pp. 94–101.
[37] T. H. Cormen, C. E. Leiserson, R. L. Riverst, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2009.
[38] L. Kovar and M. Gleicher, "Flexible automatic motion blending with registration curves," in *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation*, 2003, pp. 214–224.
[39] D. M. Young, *Iterative Solution of Large Linear Systems*. New York, NY, USA: Academic, 1971.
[40] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

Kai Li received the B.E. degree (with honor) in electrical engineering from Shanghai University, Shanghai, China, in 2010. He is now a fourth-year Ph.D. student of Department of Automation, Tsinghua University, Beijing, China. He worked as a Student Intern at Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, from September 2012 to May 2013. His research interests include image and video processing, computer vision and graphics. He is a student member of the IEEE.

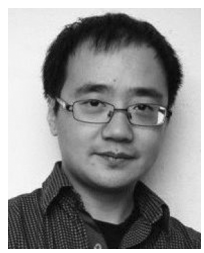
Qionghai Dai received the Ph.D. degree in automation from Northeastern University, Shenyang, China, in 1996. He has been a faculty member since 1997 and a Professor since 2005 of Department of Automation, Tsinghua University, Beijing, China. He has published more than 120 conference and journal papers, and holds 67 patents as well. His current research interests include the areas of computational photography, computational optical sensing and compressed sensing imaging and vision. His work is motivated by challenging applications in the fields of computer vision, computer graphics, and robotics. He is a senior member of the IEEE.



Ruiping Wang received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. He was a postdoctoral researcher with the Department of Automation, Tsinghua University, Beijing, from July 2010 to June 2012. He also spent one year working as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies (UMIACS), at the University of Maryland, College Park, from November 2010 to October 2011. He has been with the faculty of the Institute of Computing Technology, Chinese Academy of Sciences, since July 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning. He is a member of the IEEE.

Yebin Liu received the B.E. degree from Beijing University of Posts and Telecommunications, China, in 2002, and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor in Tsinghua University. His research areas include computer vision and computer graphics.

Feng Xu received a B.E. degree in physics from Tsinghua University, Beijing, China in 2007, and Ph.D. degree in automation from Tsinghua University, Beijing, China in 2012. He is currently working as an associate researcher in Microsoft Research Asia, Beijing, China. His research interests include image/video processing, computer vision, and computer graphics.



Jue Wang is a Senior Research Scientist at Adobe Research. He received his B.E. (2000) and M.Sc. (2003) from Department of Automation, Tsinghua University, Beijing, China, and his Ph.D. (2007) in electrical engineering from the University of Washington, Seattle, WA, USA. He received Microsoft Research Fellowship and Yang Research Award from University of Washington in 2006. He joined Adobe Research in 2007 as a research scientist. His research interests include image and video processing, computational photography, computer graphics and vision. He is a senior member of IEEE and a member of ACM.