Pattern Recognition 63 (2017) 279–290

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A dense flow-based framework for real-time object registration under compound motion

Songfan Yang[a,b], Le An[c], Yinjie Lei[a,*], Mingyang Li[d], Ninad Thakoor[e], Bir Bhanu[e], Yiguang Liu[f]

[a] College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, China
[b] FaceThink Inc., Beijing, China
[c] National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, Hubei, China
[d] Google Inc., Mountain View, USA
[e] Center for Research in Intelligent Systems, University of California, Riverside, USA
[f] School of Computer Science, Sichuan University, Chengdu, Sichuan, China

ARTICLE INFO

Keywords:
Object registration
Spontaneous facial expression
SIFT flow
Optical flow
Super-resolution

ABSTRACT

A moving object often has elastic and deformable surfaces (e.g., a human head). Tracking and measuring surface deformation while the object itself is also moving is a challenging, yet important problem in many video analysis tasks. For example, video-based facial expression recognition requires tracking non-rigid motions of facial features without being affected by any rigid motions of the head. In this paper, we present a generic video alignment framework to extract and characterize surface deformations accompanied by rigid-body motions with respect to a fixed reference (a canonical form). We propose a generic model for object alignment in a Bayesian framework, and rigorously show that a special case of the model results in a SIFT flow and optical flow based least-square problem. We demonstrate that dynamic programming can be used to speed up the computation of our algorithm. The proposed algorithm is evaluated on three applications, including the analysis of subtle facial muscle dynamics in spontaneous expressions, face image super-resolution, and generic object registration. Experimental results, in terms of both qualitative and quantitative measures, demonstrate the efficacy of the proposed algorithm, which can be executed in real time.

1. Introduction

Video registration is an important topic in video processing, computer vision and pattern recognition. It has various applications such as face recognition [1], facial expression recognition [2], image stitching [3], color demosaicking [4], etc. Depending upon different applications, there can be specific requirements for the registration techniques [5,6]. Broadly speaking, in the process of registration, most algorithms overlay objects spatially via motion estimation and compensation.

Video registration becomes a more challenging problem if there are object surface deformations which are further compounded by rigid-body motions or/and camera motion; in particular, if subtle surface non-rigid motions have to be detected and precisely characterized in applications such as medical imaging and facial expression. To appreciate the difficulties in precisely characterizing surface deformation amidst complex compound motion, let us examine a concrete example: the human facial expression analysis, in which the non-rigid muscle motion is of the central focus. Accurate facial expression analysis is hampered by the following complications:

1. Facial expression comprises non-rigid muscle motion and rigid head motion.
2. The head pose comprises both in-plane rotation and out-of-plane rotation.
3. The muscle motion is subtle in spontaneous expressions.
4. The data are streaming instead of being in a batch form.
5. The consecutive frames should comply with temporal smoothness constraint for micro-expression analysis.
6. The imaging condition varies, such as the illumination or resolution of the face region.

In this paper, we propose a new video registration approach, termed SIFT and Optical Flow Image Transform (SOFIT), that tackles the aforementioned challenges in aligning object features through video frames in the presence of compounded surface deformation

* Corresponding author.
E-mail address: yinjie@scu.edu.cn (Y. Lei).

http://dx.doi.org/10.1016/j.patcog.2016.10.015
Received 14 March 2016; Received in revised form 11 October 2016; Accepted 12 October 2016
Available online 14 October 2016

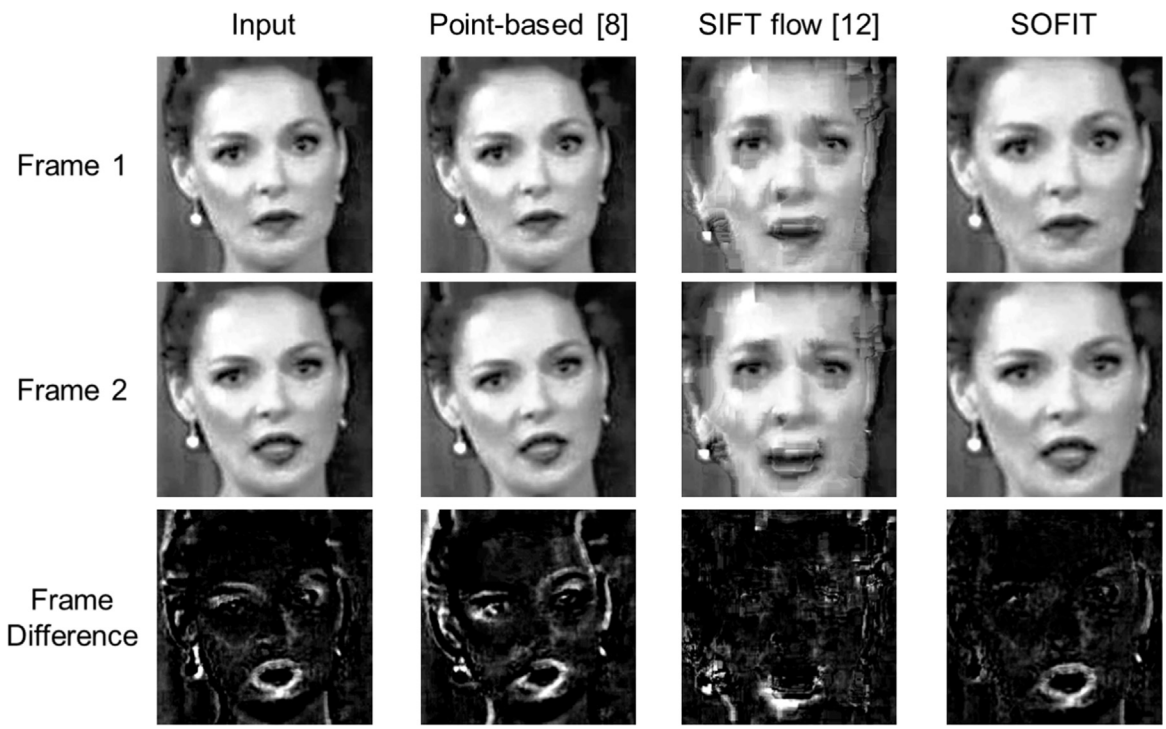

Fig. 1. Comparison of registration results. Row 3 is the absolute difference between frame 1 and frame 2. Column 2 is the point-based affine registration method used in [7–11], where affine (or piece-wise affine) transformation is computed from 83 facial feature points generated by the state-of-the-art detector [8]. Column 3 uses SIFT flow [12] to align with the Avatar Reference face model from [13]. Ideally, we would like the frame difference to show only at locations where the non-rigid motion is present (mouth area in this case). The proposed method, SOFIT, achieves the most plausible result.

and rigid motion.

In various tasks such as recognition, super-resolution, video compression, the deformed object should be aligned with respect to a canonical form or reference model. For instance, such a reference model is instrumental in facial expression analysis [13].

Facial muscle motion is similar for the same expression irrespective of the person [14], but the facial feature location (such as eyes, nose, mouth) of different people varies. Thus, finding a canonical reference feature location for all the faces is favorable for analyzing the dynamics of facial features across population. In other words, face registration is critical to facial expression recognition. In the proposed SOFIT approach, we need to transform every frame of the streaming video data into a canonical pose by neutralizing the effects of rigid body motion on the deformable object.

To further clarify the aforementioned design objective, let us examine, via Fig. 1, how different video registration methods behave when being applied to registering frame 2 with respect to frame 1. All methods in Fig. 1 are able to account for the in-plane head rotation. However, as illustrated by the frame difference images (row 3) for the point-based affine (or piece-wise affine) transformation (column 2) and the SIFT flow transformation (column 3), there is motion on most parts of the face. This is similar to the unaligned face image (column 1) where the image is the output of Viola-Jones face detector [15]. This suggests us to impose the temporal smoothness constraint so that the frame difference is small for areas with no motion; while for areas with motion (mouth area in this case), the frame difference should capture this change, as demonstrated by the results of the proposed method (column 4).

In this paper, we model the alignment-of-compound-motion problem in three steps. *First*, each frame is aligned with respect to a reference frame in a general distance measure, which is then instantiated to the SIFT flow criterion, thereafter. *Second*, our model enforces a smoothness constraint on adjacent frames. It is realistic for the consecutive frames to comply with the smoothness constraint. We realize this by depending this current transformation estimation on a number of previous frames in an optical flow criterion. *Third*, large transformation is penalized to prevent over-fitting. We also extend this approach to register many other types of objects and demonstrate applications in areas such as image super-resolution. More results can be found on our project website.[1]

The rest of the paper is organized as follows. After reviewing the related work and highlighting our contribution in Section 2, Section 3 presents our general model as well as the solution to the registration problem using the dense flow approximation to estimate the affine transformation parameters. The experimental results and discussions are provided in Section 4. Finally the conclusion is drawn in Section 5.

2. Related work and contributions

2.1. Related work

Video registration has been a fundamental topic in computer vision and image processing. Recent successful object retrieval and recognition methods, such as [16,17], have made progressive achievements, while accurate registration can further prompt the performance on these applications. As an object may undergo a complex motion (rigid and/or non-rigid), conventional video registration methods [5,6] attempt to correct both types of motion. On the contrary, we attempt to remove the rigid motion while retaining and characterizing the non-rigid motion. Such problem occurs when a moving object has deformable surface, which may contain crucial information (e.g. facial expression).

To analyze facial expressions, behavioral scientists have developed Facial Action Coding System (FACS) [14] as an objective standard to describe the muscle motion. According to FACS, human (coders) can decompose every possible facial behavior into Action Units (AU), which roughly correspond to the muscles that produce them. Automatic AU

[1] http://www.ee.ucr.edu/~syang/sofit/index.html.

recognition [9,18], has been quite successful for well-aligned, posed data, such as MMI [19] and CK+ [20]. For example, Fang [21] leveraged the salient information in expression video to select the peak expression. Li et al. [22] adopted Dynamic Bayesian Network(DBN) to model the dynamic and semantic relationships among multi-level AU intensities.

Unfortunately, AU recognition in an uncontrolled real-world environment remains a difficult problem, as shown in the Facial Expression Recognition and Analysis Challenge (FERA2011 [23] and FERA2015 [24]), due to the difficulties mentioned in Section 1. Existing face registration approaches attempt to solve different aspects of the aforementioned challenges. In the face recognition and image retrieval communities, researchers attempt to discard the non-rigid motion from facial data through registration using an ensemble of images [25–27]. These approaches are not suitable in the facial expression recognition domain, where the following three criteria should be met:

1. Non-rigid facial muscle motion, which carries essential information for expression inference, should be retained.
2. Facial features should be aligned under various muscle motions and pose variations.
3. Subtle facial muscle motion should be captured for spontaneous facial expression analysis.

To align faces with expressions, the state-of-the-art systems [7,9,28] track a set of anchor points on the face and estimate the affine transformation based on which the entire face is warped. Although the most recent facial point detection techniques [29–31,8] are able to achieve accurate detection results, there are two significant issues that need to be addressed. *First of all*, the affine estimation is sensitive to small perturbation of point detection results. Typically in point-based method, a number of facial fiducial points (e.g., 20 points in [9,29]) are detected. Each point carries much more weights in the estimation of the affine transformation matrix compared with methods that use corresponding information from the entire image, as demonstrated by Fig. 1. *Besides*, affine transform parameter estimation by a small set of points can be susceptible to detection errors. In a realistic case where the resolution of the face is not high enough, the accuracy of feature point detection will also degrade. Yang and Bhanu [13] adopted SIFT flow technique [12] to align every frame to a reference face. As shown in Fig. 1, column 3, the outcome of the SIFT flow transform displays a large amount of discontinuities and artifacts. Although they attempt to solve this issue by generating image-based face representations (i.e., Emotion Avatar Image) and a reference model (i.e., Avatar Reference), carrying out the double layer loopy-belief propagation in SIFT flow for every frame is computationally expensive and not suitable for real-time systems.

2.2. Contributions of this paper

The contributions of this work are summarized as follows:

1. Unlike methods in the registration literature that attempt to correct the motion, we attempt to solve a more challenging problem: aligning objects under compound motion, in the hope of compensating the rigid motion while retaining the non-rigid motion.
2. We propose a novel real-time streaming registration framework, SOFIT, that aligns the objects under compound motion. SOFIT is a holistic approach and no detection of local features (eyes, nose, mouth) is needed. Therefore, it is robust against noise, detection error, and low image resolution. The proposed method results in temporally smooth and aligned image sequences.
3. We quantitatively demonstrate the versatility of our registration method in the fields of spontaneous AU recognition and image super-resolution. We also show results for generic object alignment under various challenges.

3. Flow-based real-time object registration

The objective of this work is to align objects in video in an uncontrolled environment. Taking face images as specific examples, the original inputs to our system can be faces detected by the Viola-Jones detector [15] for the analysis and illustrations in expression analysis domain. We first formulate the generalized model in a Bayesian framework. A flow-based approximation results in an efficient closed-form solution. We also point out a dynamic programming implementation that will further optimize the registration algorithm.

3.1. The generalized model

Let p be the grid coordinate of I_i, the i-th frame in grayscale. For simplicity, we write the intensity of an image, I_i, as a shorthand for $I_i(p)$. Given a sequence of N unregistered frames of an object, our goal is to align individual frames with respect to a canonical representation of this object, denoted by I_c. Let w_i be the flow field to register frame i, then the i-th registered frame can be written as $I_i(p + w_i)$. To align the entire sequence, the objective is to recover $w_1, \ldots, w_N$ for each of the N images in the sequence. We model the distance measurement of $I_i(p + w_i)$ and I_c as a random variable Q_i corrupted by a Gaussian noise m_Q. Thus,

$$Q_i = Dist\,(I_i(p + w_i), I_c) + m_Q, \tag{1}$$

where $Dist\,(\cdot,\cdot)$ is a generic distance function. In this paper, we attempt to align every frame with respect to the canonical representation such that they share similar structure. However, in general, it is applicable to many other distance measures. m_Q is i.i.d. (i.e., independent and identically distributed) normally distributed zero-mean measurement noise. We model the measurement of the transformation, w_i, as a random variable Y_i. The difference between Y_i and w_i is modeled by an i.i.d zero-mean Gaussian distribution:

$$Y_i = w_i + m_Y, \tag{2}$$

where w_i is the underline true variable we intend to solve. This model penalizes excessive transformation due to over-fitting. The joint probability of all variables can be written as

$$L = P(w_{1:N}, Y_{1:N}, Q_{1:N}, I_c) = P(Q_{1:N}, Y_{1:N}|w_{1:N}, I_c,)P(w_{1:N}|I_c)P(I_c), \tag{3}$$

where $w_{1:N}$ is short for $w_1, \ldots, w_N$. Dropping the constant term and using the independence of our model definition in Eqs. (1) and (2), we obtain

$$L \propto P(Q_{1:N}|w_{1:N}, I_c)P(Y_{1:N}|w_{1:N})P(w_{1:N}) = \prod_{i=1}^{N} P(Q_i|w_i, I_c) \prod_{i=1}^{N} P(Y_i|w_i) \prod_{i=1}^{N} P(w_i|w_{1:i-1}), \tag{4}$$

where $\prod_{i=1}^{N} P(w_i|w_{1:i-1})$ can be viewed as the smoothness constraint. With the weakly coupled Markov assumption, we only take into account $H = \min(i, h)$ number of frames prior to frame i. The assumption is that the aligned frame I_i should have similar appearance with its previous h neighbors (if $h < i$). With the models in Eqs. (1) and (2) as the prior terms in Eq. (4), the joint probability can be written as

$$L \propto \prod_{i=1}^{N} P(Q_i|w_i, I_c) \prod_{i=1}^{N} P(w_i|w_{i-H:i-1}) \prod_{i=1}^{N} P(Y_i|w_i) \tag{5}$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma} Dist\,(I\,(p + w_i), I_c)\right\} \tag{6}$$

$$\times \prod_{i=1}^{N} \frac{1}{\epsilon\sqrt{2\pi}} \exp\left\{-\frac{1}{2\epsilon H} \sum_{j=1}^{H} \|I_i(p + w_i) - I_{i-j}(p + w_{i-j})\|_F^2\right\} \tag{7}$$

$$\times \prod_{i=1}^{N} \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-\frac{1}{2\tau}\|w_i\|_F^2\right\}, \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm; the smoothness constraint $P(w_i|w_{i-H:i-1})$ obeys zeros-mean Gaussian distribution; σ, ϵ, τ control the variance for the corresponding Gaussian distribution. Maximizing the likelihood, L, is equivalent to minimizing its negative log likelihood, E, where

$$E = \sum_{i=1}^{N} Dist\,(I_i(p + w_i), I_c) \tag{9}$$

$$+\frac{\alpha}{2H}\sum_{i=1}^{N}\sum_{j=1}^{H}\|I_i(p + w_i) - I_{i-j}(p + w_{i-j})\|_F^2 \tag{10}$$

$$+\frac{\beta}{2}\sum_{i=1}^{N}\|w_i\|_F^2, \tag{11}$$

where the constant terms are dropped. $\alpha = \sigma^2/\epsilon^2$ and $\beta = \sigma^2/\tau^2$ can be considered as two scaling parameters on the smoothness term and penalty term, respectively.

3.2. The flow-based instantiation

The *Dist* function in Eq. (9) measures the similarity between two images. One can define its form according to different applications. Since our objective is structural matching, we opt to use SIFT flow [12] for similarity matching under large, non-rigid transformation. SIFT flow [12] was originally designed to align an image to its plausible nearest neighbor which can have large variations. The SIFT flow algorithm matches dense SIFT features [32] between two images, resulting in a structural coherent image pairs. Although SIFT flow by itself generates block artifacts, it includes a structural matching constraint that allows non-rigid motion correspondence. Thus, the data matching term in Eq. (9) can be instantiated in the coordinate space as

$$\sum_{i=1}^{N} Dist\,(I_i(p + w_i), I_c) = \sum_{i=1}^{N}\|w_i - f_s^i\|_F^2, \tag{12}$$

where $f_s{}^i$ (shorthanded for $f_s(I_i, I_c)$) is the pixel-wise SIFT flow field given by matching I_i to canonical reference frame I_c. The reference frame of face images is chosen to be the Level-1 Avatar Reference (AR) image [13] generated from the FERA-GEMEP dataset [23]. AR is essentially a face model that reflects the expression and identity of the entire population in the dataset. It is computed offline by an iterative algorithm that estimates the reference model and the individual expression model simultaneously. It has been demonstrated to perform well across datasets [13]. For different classes of objects, the canonical image representation is generated in the same way.

Regarding the smoothness constraint in Eq. (10), we consider the optical flow between frames. Optical flow computes the motion between two frames by matching the corresponding intensity values. In the context of video processing, it is reasonable to assume that the frame rate is high enough for computing accurate optical flow between consecutive frames. With the pixel-level correspondence, we can approximate the current frame by its previous frame, i.e., $I_i \simeq I_{i-j}(p + f_o^{i-j,i})$, where $f_o^{i-j,i}$ is the optical flow field from frame $i-j$ to frame i. Thus, applying the corresponding registration transformation on both sides yields

$$I_i(p + w_i) \simeq I_{i-j}(p + f_o^{i-j,i} + w_{i-j}) \simeq I_{i-j}(p + f_o^{i-j,i} + f_s^{i-j}). \tag{13}$$

The approximation in the second line of Eq. (13) holds true according to the structural constraint in Eq. (12). Once again, due to the illumination invariant assumption under high frame rate for optical flow, Eq. (13) is equivalent to

$$w_i \simeq f_o^{i-j,i} + f_s^{i-j}. \tag{14}$$

Therefore, the smoothness constraint in Eq. (10) can be rewritten as

$$\frac{\alpha}{2H}\sum_{i=2}^{N}\sum_{j=1}^{H}\|I_i(p + w_i) - I_{i-j}(p + w_{i-j})\|_F^2 \simeq \frac{\alpha}{2H}\sum_{i=2}^{N}\sum_{j=1}^{H}\|w_i - (f_o^{i-j,i} + f_s^{i-j})\|_F^2. \tag{15}$$

Now, the cost function is written as the sum of three ℓ_2 norm terms in Eqs. (12), (15), and (11). Since the speed of the algorithm is a main concern for real-time applications in practice, we adopt ℓ_2 norm in our formulation, such that a closed-form solution for this optimization problem can be derived. We further assume that the transformation function is affine. The computation of the X, Y-component can be decomposed, which enables speedup using parallel computation. Thus, the cost function is instantiated as

$$E = \frac{1}{2}\sum_{i=1}^{N}\|T_i p - p - f_s^i\|_F^2 \tag{16}$$

$$+\frac{\alpha}{2H}\sum_{i=2}^{N}\sum_{j=1}^{H}\|T_i p - p + (f_o^{i-j,i} + f_s^{i-j})\|_F^2 \tag{17}$$

$$+\frac{\beta}{2}\sum_{i=1}^{N}\|T_i p - p\|_F^2, \tag{18}$$

where the flow w_i is written as $T_i p - p$; T_i is a 3×3 affine matrix. With minor abuse of notion, p is now a horizontal stacked coordinate location template of size $3 \times m$, where $m = r \times c$, assuming that r and c are the image height and width, respectively. Each column of p is a coordinate point $(x, y, 1)$ in homogeneous coordinates. Taking the derivative of E w.r.t. T_i and setting it to zero result in

$$T_i = (1 + \alpha + \beta)^{-1} \times \left(f_s^i + \frac{\alpha}{H}\sum_{j=1}^{H}(f_s^{i-j} + f_o^{i-j,i}) + (1 + \alpha + \beta)p\right)p^T (pp^T)^{-1}. \tag{19}$$

It is observed from Eq. (19) that for every input frame I_i, we have to compute its SIFT flow with respect to the canonical reference frame. Computing SIFT flow for every frame is time-consuming. However, given accurate optical flow estimation between frames (which is a practical assumption for video at high frame rate), we can approximate the SIFT flow computation of the i-th frame by the sum of the SIFT flow of the $(i-1)$-th frame and the optical flow from the $(i-1)$-th to the i-th frame, i.e.

$$f_s^i = f_s^{i-1} + f_o^{i-1,i}, \tag{20}$$

Therefore, the final closed-form solution can be written as

$$T_i = (1 + \alpha + \beta)^{-1} \times \left(f_s^{i-1} + f_o^{i-1,i} + \frac{\alpha}{H}\sum_{j=1}^{H}(f_s^{i-j} + f_o^{i-j,i})\right) + \mathbf{I}_3, \tag{21}$$

where $\mathbf{I}_3$ is a 3×3 identity matrix. Eq. (21) can be rewritten as

$$T_i = T_{i-1} + (1 + \alpha + \beta)^{-1}\left(f_o^{i-1,i} + \frac{\alpha}{H}g\right) + \mathbf{I}_3, \tag{22}$$

where $g = f_s^{i-1} - f_s^{i-H-1} + \sum_{j=1}^{H}(f_o^{i-j,i} - f_o^{i-j-1,i-1})$. It can be seen that Eq. (22) is written in a typical dynamic programming (DP) formulation, where we can cache the previously computed SIFT and optical flows. For the current frame i, we only carried out several optical flow computations, i.e., $f_o^{i-H,i}, \ldots, f_o^{i-1,i}$. When H is small, e.g., 3–5, the optical flow is accurate and the total amount of optical flow computation is small. The DP implementation dramatically reduces the computational cost and enables real-time execution of our algorithm.

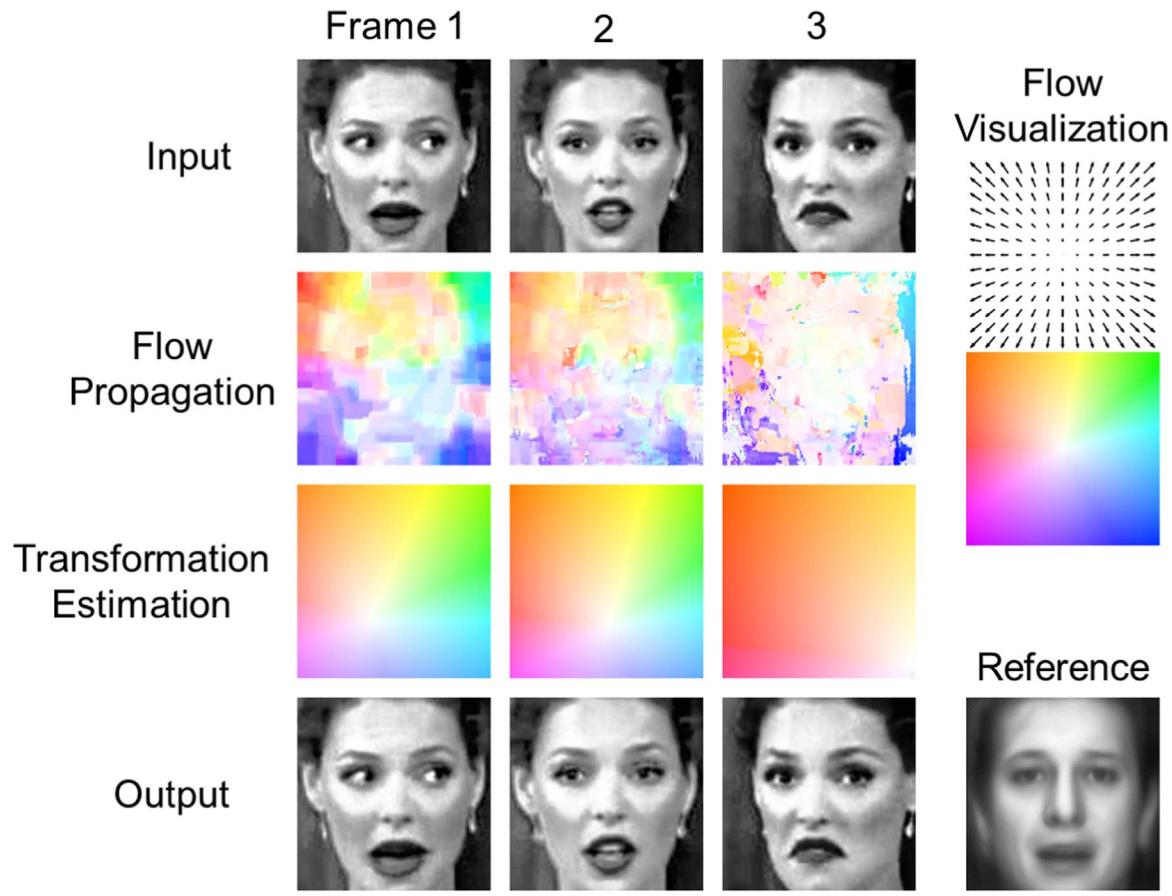

Fig. 2. SOFIT registration example. The input sequence is registered with respect to the reference frame shown on the bottom right. The flow visualization is coded as in [34]. Better viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Besides, individual optical flow can be computed in parallel to further speed up the algorithm. What's more, we use Iteratively Reweighted Least Squares (IRLS) [33] with a bisquare weighting function to robustly estimate the affine transformation matrix.

A registration example for face is visualized in Fig. 2. The flow propagation, computed by Eq. (21), is visualized in the second row for consecutive frames. The affine transformation is then robustly estimated for each frame. The output sequence is registered with respect to the reference frame and exemplar frames comply with smoothness constraint.

3.3. Error propagation and loop-closure rectification

In our model, we make a compromise between model optimality and computation efficiency. Therefore, the average registration error accumulates over time. The registration error is defined as the deviation from the canonical reference frame. Since we care about structural similarity, we compute the mean length of the SIFT flow from the current frame to the reference frame. For error analysis, we need videos with length of more than one minute (1800 frames in our case with 30 fps) to observe the noticeable cumulative error. Therefore, we register a long video sequence[2] and plot the error in Fig. 3. Although this error measurement consists of both global rigid head motion and local non-rigid muscle motion, we are still able to observe the error accumulation effect.

To solve this issue, we intend to update the global estimation at a certain rate without affecting the propagation computation. Inspired by the Loop-Closure (LC) strategy in robotics [35], we update f_s in Eq. (21) by recomputing the SIFT flow for every 300 frames. This update frequency is chosen because cumulative error is negligible within 300 frames based on empirical observations, and this also provides enough time for f_s to be updated in parallel and will not affect the overall flow propagation. As seen in Fig. 3, the error is rectified by the global flow estimation update.

The aforementioned procedures are summarized in Algorithm 1. The algorithm reinitalizes every K frames.

Algorithm 1. SIFT and Optical Flow Image Transformation

Input: image sequence to be aligned, $I_{0:N}$; reference frame, I_{ref}; reinitialize every K frame; windows size H; $init = False$

for i=0 to N **do**
  if $i\%K = 0$ **then**
    $init \leftarrow True$
  end if
  if *init* **then**
    $f_s^i \leftarrow \text{SIFT} - \text{flow}(I_i, I_{ref})$
    $T_i \leftarrow f_s^i p^T (pp^T)^{-1} + \mathbf{I}_3$ (Eq. (19))
  else
    $h \leftarrow \min(i, H)$
    for $j = 1$ to i **do**
      $f_o^{i-j,i} \leftarrow \text{optical} - \text{flow}(I_{i-j}, I_i)$
    end for
    $g \leftarrow f_s^{i-1} - f_s^{i-h-1} + \sum_{j=1}^{h} (f_o^{i-j,i} - f_o^{i-j-1,i-1})$
    $T_i \leftarrow T_{i-1} + (1 + \alpha + \beta)^{-1}(f_o^{i-1,i} + \frac{\alpha}{h} g) + \mathbf{I}_3$ (Eq. (22))
  end if

[2] source: https://www.youtube.com/watch?v=_aKNYRwb4-4.

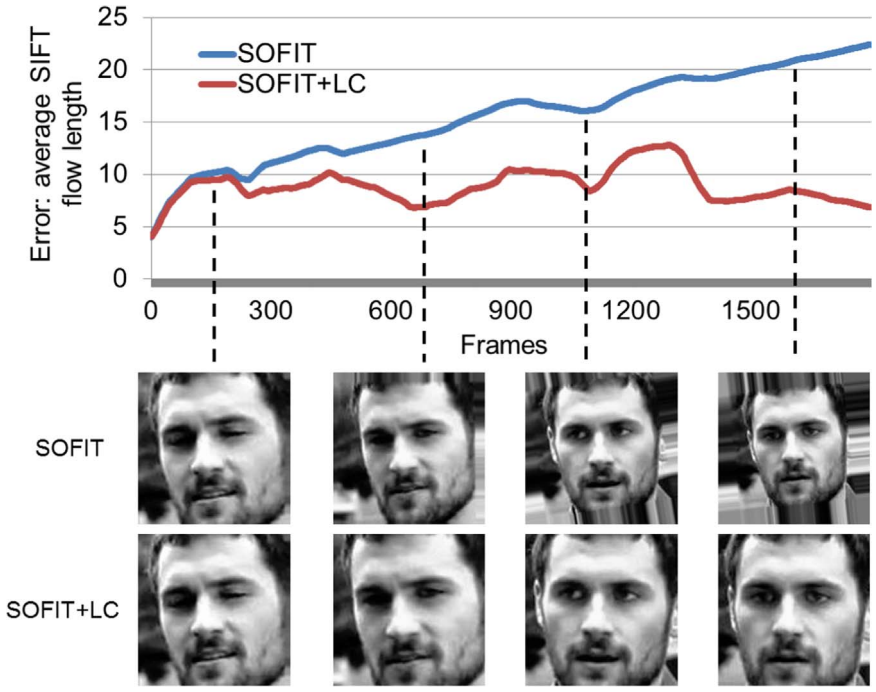

Fig. 3. The error accumulation over time. The error is defined as the SIFT flow of the current frame to the canonical reference frame. We use loop-closure (LC) to update the global flow estimation and rectify the error. The LC is carried out every 300 frames in this experiment.

Table 1
Person-independent AUC-score result on FERA-GEMEP AU training set.

		AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU17	AU18	AU25	AU26	Avg.
BaseLine [23]	LBP	0.69	0.69	0.58	0.68	0.61	0.68	0.68	0.52	0.61	0.57	0.53	0.52	0.61
	LBP-TOP	0.70	0.69	0.61	0.74	0.66	0.64	0.77	0.51	0.61	0.60	0.55	0.53	0.63
PA [8]	LBP	0.69	0.69	0.63	0.69	0.62	0.65	0.72	0.55	0.65	0.66	0.56	0.56	0.64
	LBP-TOP	0.67	0.69	0.73	0.73	**0.69**	0.64	0.70	0.66	0.59	**0.77**	0.53	0.51	0.66
EAI [13]	LBP	0.68	0.68	0.68	0.69	0.62	0.61	0.75	0.54	0.66	0.72	0.55	0.56	0.65
	LBP-TOP	0.71	**0.71**	0.61	0.66	0.63	0.66	0.78	0.67	0.67	0.69	0.51	0.52	0.65
SOFIT	LBP	**0.76**	0.70	0.62	0.78	0.67	**0.70**	0.74	**0.68**	**0.69**	0.74	**0.57**	0.58	0.69
	LBP-TOP	0.73	0.70	**0.75**	**0.80**	0.64	0.67	**0.82**	**0.68**	0.67	0.76	**0.57**	**0.63**	**0.70**

Table 2
Person-independent F1-score on BP4D development set (2D data only).

	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	Avg.
BaseLine [24]	0.35	0.26	0.40	0.77	0.74	0.78	0.81	0.60	0.25	0.52	0.26	0.52
PA [8]	0.36	0.26	0.42	0.70	0.74	0.81	0.84	0.60	**0.36**	0.58	0.35	0.55
EAI [13]	0.38	0.27	**0.47**	0.78	0.78	0.82	0.82	0.60	0.35	0.60	0.31	0.56
SOFIT	**0.41**	**0.32**	0.45	**0.80**	**0.80**	**0.86**	**0.89**	**0.62**	0.35	**0.61**	**0.36**	**0.59**

16: *init* ← *False*
17: **end for**
Output: $T_{0:N}$.

3.4. Computational cost analysis

We obtain considerable speedup using dynamic programming based implementation. In essence, the SIFT flow only needs to be computed for the initialization. The steps to be carried out for every subsequent frame are the following:

1. Compute the dense optical flow with respect to the previous H frame in parallel.
2. Estimate affine transformation matrix using Eq. (21) based on the updated coordinate value.

We adopt the optical flow implementation in OpenCV [36]. The aforementioned steps can be finished in approximately 47 ms on average for a 100 × 100 image on a quad-core Intel i7 4 GHz machine with 32 GB memory. In other words, the execution speed reaches 21 fps under this setting. Since the bottleneck of our method is optical flow computation, one can further speedup the algorithm by GPU implementation. The initialization takes on average about 1.2 s using the aforementioned settings. The computation of SIFT flow and optical flow are also in parallel. The LC re-initialization is also in parallel with the the optical flow computation, and it will not affect the speed of the registration procedure.

4. Experimental results

In this section, we show results on two applications of the proposed method, including facial AU recognition and image super-resolution.

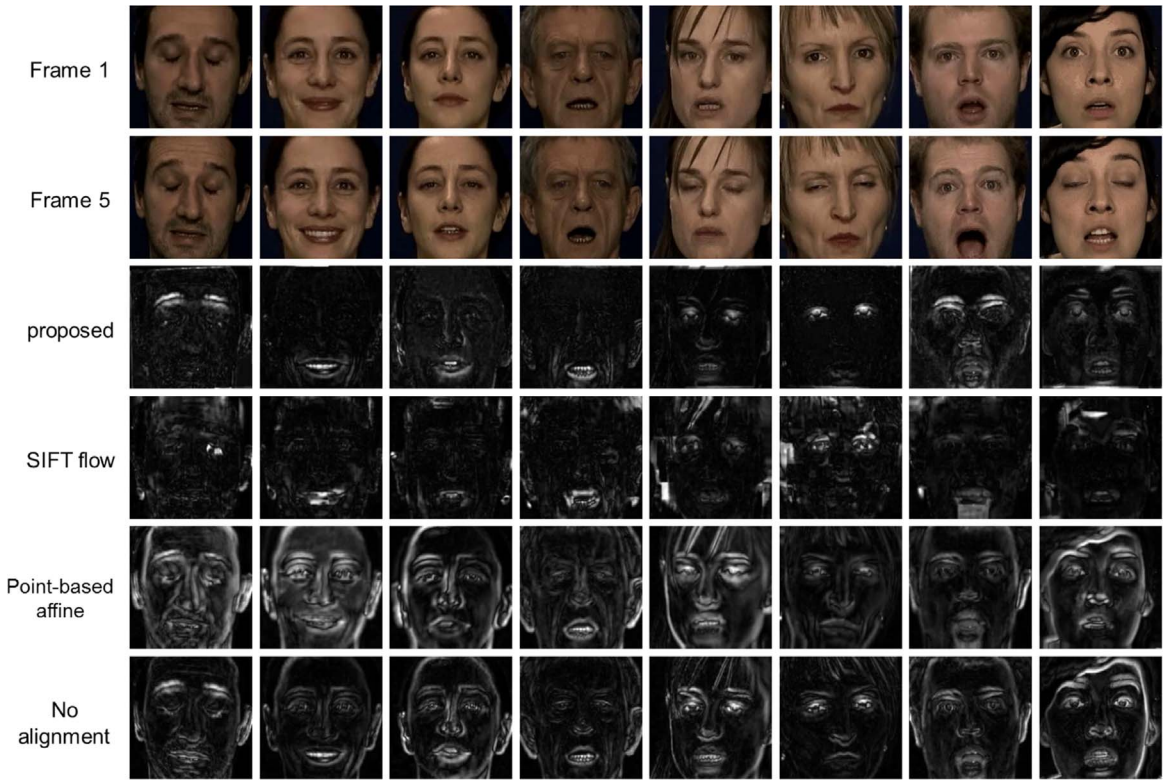

Fig. 4. Qualitative face registration results comparison. Row 1 and 2 are the first and fifth frame of a sequence. Row 3–5 are the cumulative absolute frame difference of 5 unaligned frames using method SOFIT, EAI [13], PA where points are detected using [8], respectively. The proposed alignment technique captures the correct non-rigid motion of face, for example, eyebrows raise for first column and mouth open for second column.

For AU recognition, we follow the challenge guideline of FERA2011 and FERA2015 for a thorough analysis and comparison. In addition, we have also included qualitative results on different types of objects to demonstrate the generalizability of our approach.

4.1. Facial action unit recognition

The goal of AU recognition is to detect a set of frequently occurring AUs on a per-frame basis. The main concerns here are two-fold:

1. Is registration really an important issue in real-world AU recognition in uncontrolled environment?
2. If yes, can we improve the recognition performance just by adopting a better registration algorithm, e.g., SOFIT?

Datasets: FERA-GEMEP. We first demonstrate SOFIT face registration technique by facial AU recognition on FERA Challenge 2011 dataset [37]. We use the same protocol as the FERA2011 [23] AU sub-challenge. The data we use for training is the GEMEP-FERA training dataset, which includes 87 sequences and around 5400 frames. The pose and gesture of the subjects in this dataset are uncontrolled, and therefore, this dataset is more realistic and complex compared to the legendary MMI [19] and CK+ [20] datasets.

To address the aforementioned issues, we use the exact same features as in the baseline approach for a fair comparison. The only variable in our experiment design is using different registration methods. The *baseline* registration method detects both eye locations on the faces, and then unifies their scales, and in-plane rotations. This registration belongs to in-plane image transformation category as summarized in [13]. A typical point-based registration method tracks facial landmarks and estimates the affine (or piece-wise affine) transformation based on a set of rigid landmarks such as eye corners or nose [7,10,11]. We term this family of methods as point-based affine (*PA*), and we use piece-wise affine in our experiment. Another registration technique in comparison is the Emotion Avatar Image (*EAI*) [13], which achieves the best performance in FERA Challenge 2011.

The feature extraction and classification are conducted in the same way as the baseline approach. After extracting the faces from Viola-Jones detector [15], we register all faces using the aforementioned methods. All registered face images are all of size 100×100. Subsequently, we divide the image into 10×10 blocks, where static features, Local Binary Pattern (LBP) [38], and dynamic feature, LBP in three orthogonal planes [18] (LBP-TOP), for each block are computed and concatenated separately. We then train 12 linear SVM binary classifiers based on the implementation of [39], each of which is trained independently regardless of the co-occurrence of AUs.

Concretely, the feature dimensions are the following:

1. LBP: the uniform LBP operator generates 59 basic patterns for a local patch. Given our region segmentation settings, the total feature dimension is $59 \times 10 \times 10 = 5900$.
2. TOP-LBP: since the feature dimension enlarges by a multiple of 3 compared with LBP, the total feature dimension is $3 \times 5900 = 17700$.

For registration methods with reference frame, i.e., EAI [13], and SOFIT, we use the Level-1 Avatar Reference [13] generated from the FERA Challenge training data [37]. To generate an EAI representation, we need to determine a temporal length parameter. In [13], this parameter is chosen as the length of a single video (around 2 s) for facial expression recognition on a per-video basis. To generalize this registration technique in AU recognition on a per-frame basis, we empirically determine the best value for the temporal length parameter. We carry out a leave-one-subject-out cross validation on the FERA-GEMEP AU training data, and find 0.56 s is a reasonable temporal length to achieve the best F1 score over all AUs. This means that for each frame in a video, approximately 14 consecutive frames are used to compute EAI representation. For the boundary frames, i.e. the

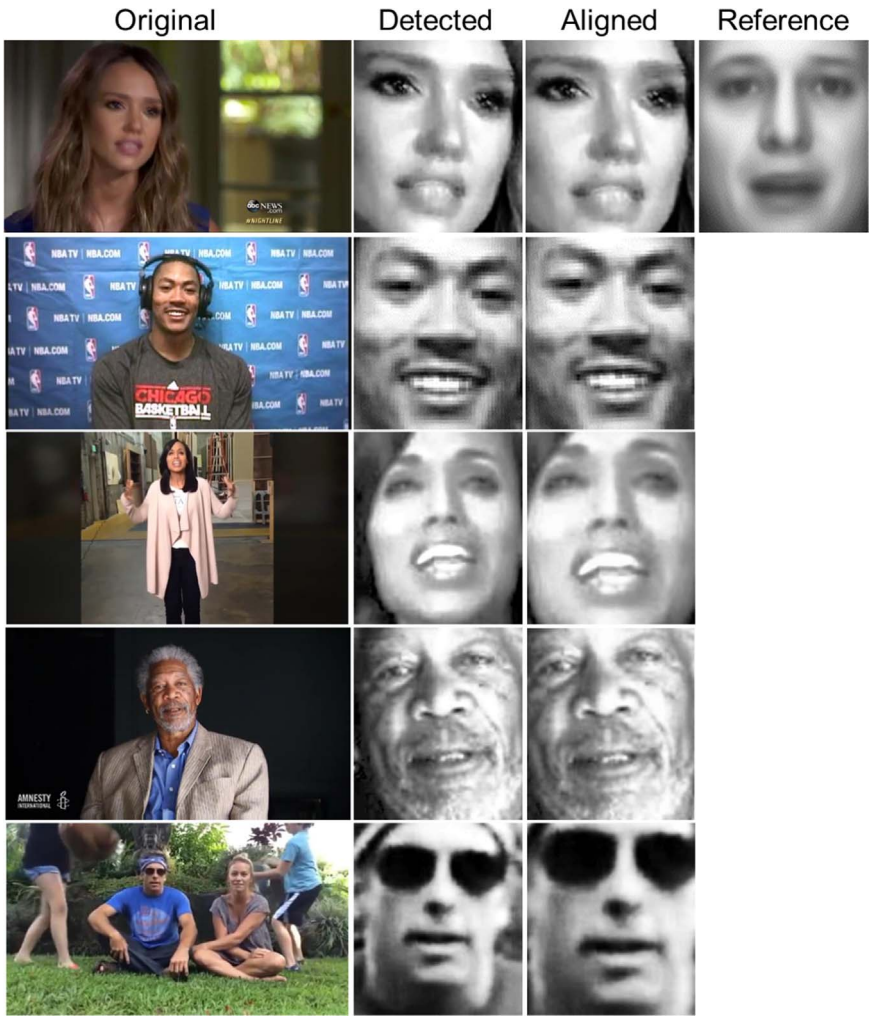

Fig. 5. Examples of face alignment results using SOFIT. By aligning with a canonical reference face, all pose rotations, translations, and scales are rectified.

starting and the ending 7 frames, we simply assign their values to be the 8th frame from the beginning and from the end, respectively. The aforementioned features are extracted from the EAI representations thereafter.

For PA method, 83 points are tracked using [8]. Each image is then registered by piece-wise affine warping estimated from 41 rigid points, including nose, eye corner, etc. We only include the texture in the face region and blacken the periphery of the entire region. Since the non-face region will not interfere the classification model training, for simplicity, we include all regions for feature extraction.

Since the ground-truth label is only available for the FERA-AU training set, we carry out a person-independent cross validation experiment, such that no test subject is used for training, and the average performance is reported. Due to the finite scale of the training exemplars, person-independent test is essential to demonstrate the generalization ability of an approach to unseen subjects. Table 1 shows the performance obtained using both LBP and LBP-TOP feature extractors, respectively. The area under curve (AUC) score of the receiver operating characteristic (ROC) curve is reported. As seen from Table 1, on average, SOFIT outperforms the other methods by a significant margin using both LBP and LBP-TOP features. SOFIT achieves the best performance in 9 out of 12 AU classification tasks. In general, LBP-TOP outperforms LBP likely due to the dynamic information extracted. EAI is on par with PA in terms of the average AUC score. EAI performs well in categorical emotion recognition [13]; in a per-frame based setting, however, EAI lacks the ability to maintain the subtle muscle motion. Moreover, dynamic feature is not advantageous in the EAI case. A plausible explanation is that the inherent dynamics are buried in the block artifacts caused by SIFT flow. We should point out that the video sequences in FERA-GEMEP is relatively short, i.e. around 2 s. Thus, the propagation error is negligible and thus we exclude the loop closure rectification procedure in this experiment.

Datasets: BP4D. We also carry out experiments on the BP4D [40] dataset, which is also used by the FERA2015 challenge [24]. It is a spontaneous facial expression dataset collected in an lab setting with uncontrolled pose and gesture. It contains 41 subjects participating in 8 tasks, which are designed to solicit expressions that are not deliberately posed. The subjects are aging from 18 to 29 covering various ethnicity groups. The original dataset includes both 2D and 3D videos, and we only use the 2D videos in our experiment. The dataset is partitioned into *Training* and *Development* sets with the ground truth label available. Similar to the FERA-GEMEP dataset, the ground truth label of each frame is obtained according to the FACS. Each subject has 8 sessions, and there are 168 sessions in the training and 160 sessions in the development partition, each of which has over 70k images.

In FERA2015, the performance is measured by the $F1$ score, calculated as:

$$F_1 = \frac{2PR}{P+R} \tag{23}$$

where P and R represents the precision and recall, respectively.

The challenge baseline system uses Viola-Jones face detector and directly extract the appearance feature, Local Binary Gabor Patterns (LGBP) [41]. LGBP is essentially the LBP features extract from the

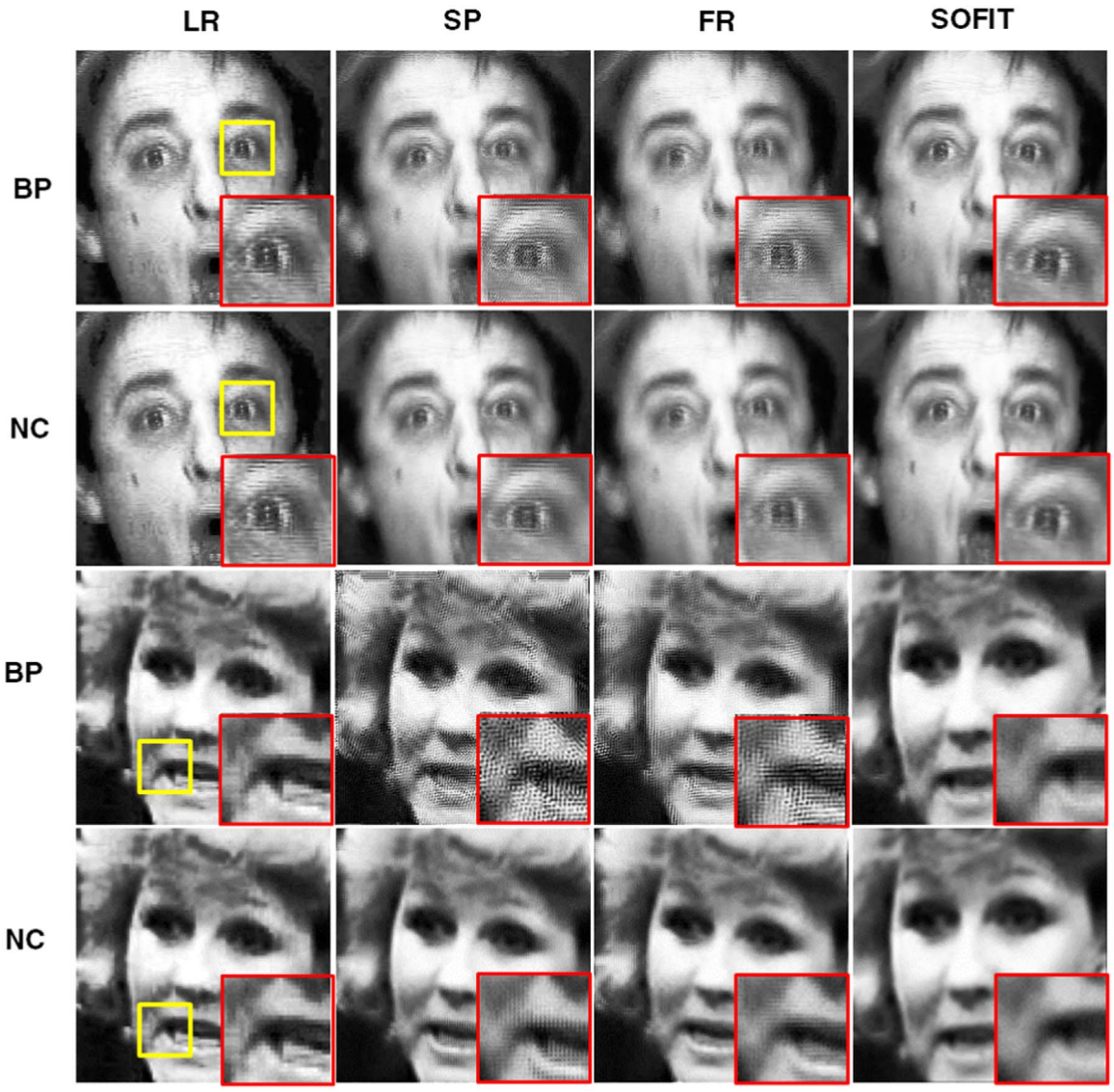

Fig. 6. The comparison of super-resolution results using different registration methods for 2 subjects. For each column from left to right: one of the LR inputs (enlarged by pixel replication), sub-pixel registration (SP) [42], frequency domain based registration (FR) [43], and the proposed registration method. We use two SR methods to reconstruct the high-resolution outputs: iterated back-projection (BP) [44] and normalized convolution (NC) [45]. The red blocks show the magnified parts from the yellow blocks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

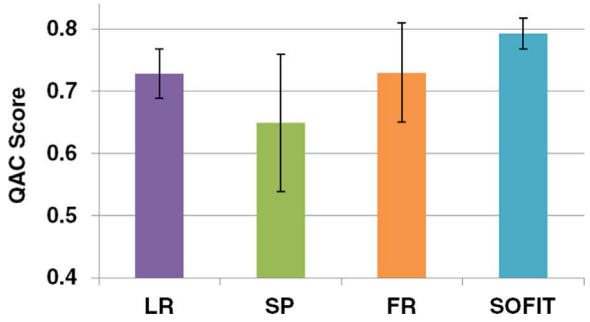

Fig. 7. Image quality comparison between video-based super-resolution results and the proposed method, SOFIT. A recent non-reference image quality assessment method [47] is used. The higher the score, the better the estimated visual quality is. LR stands for low-resolution input images. SP denotes the super-resolved results using [42]. FR is the super-resolved results using [43]. SOFIT outperform other benchmarks and has less variance, indicated by the black line.

Gabor filter magnitude response of the original image, resulting a 16, 992 feature dimension for each frame. The feature dimension is then reduced by Principle Component Analysis (PCA) while keep 98% of the energy.

We then carry out a person-independent experiment on the training set, similar to the ones in [24,40], where we follow a 10-fold cross-validation procedure and use 19 subjects for training and 2 for testing. The one-vs-all linear SVM is used to train each classifier. The best parameter combination is then used to the classification model on the entire training set, and the results for the development set is tabulated in Table 2. Similar observation can be made as in the FERA dataset. SOFIT achieves the best score in most AU detection tasks. By fix the other variables, it is clear that the performance improvement is due to the proposed SOFIT registration method.

Qualitative comparison. Fig. 4 shows the qualitative evaluation on explaining why our registration improves over the baseline and EAI approaches. We compute the absolute frame difference of the first 5 frames for both unaligned and aligned faces. As shown in the row 6 and 3 of Fig. 4, the unaligned frame difference reveals motion mainly caused by the *edge* feature of a face, while after alignment, the non-

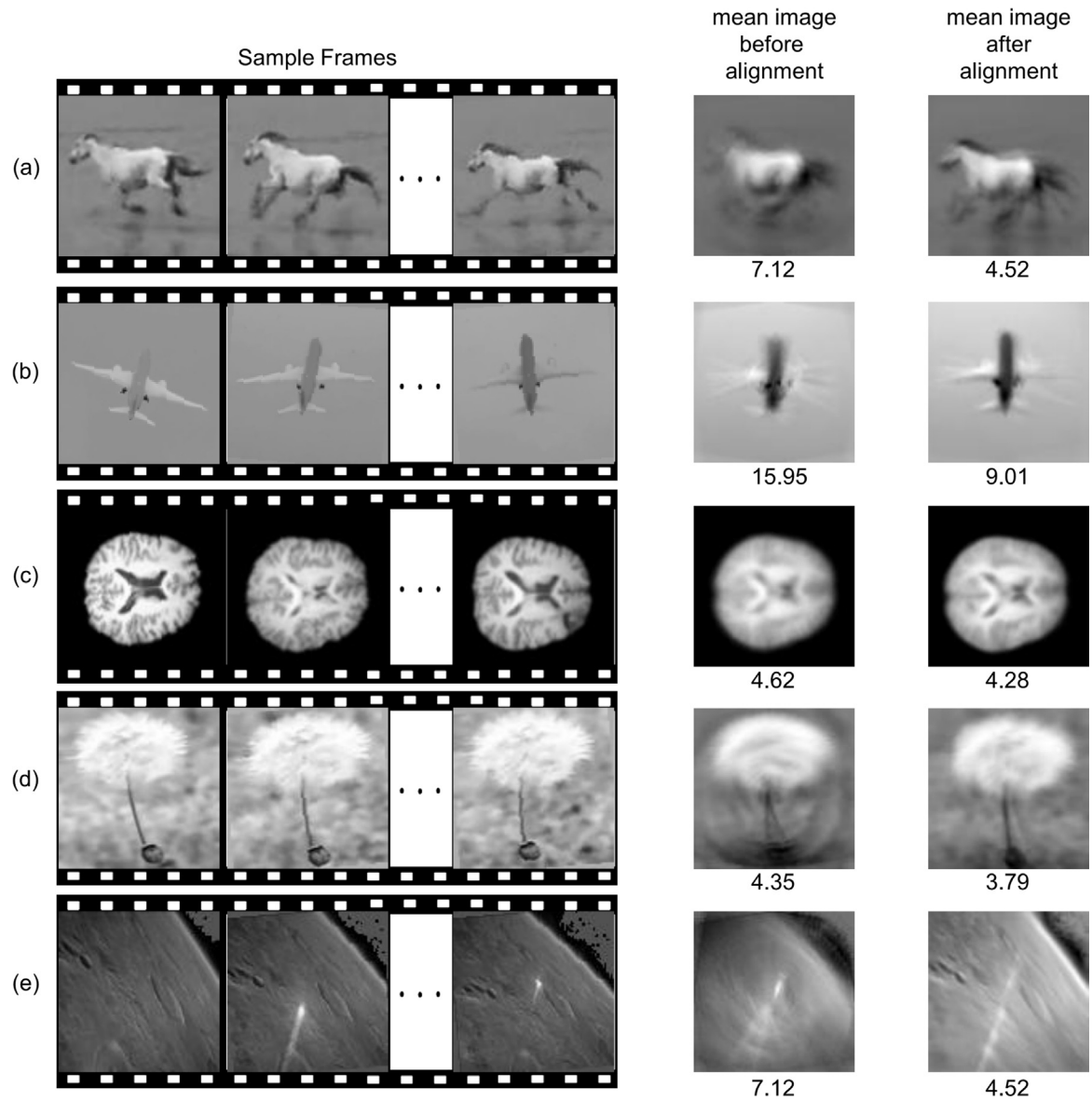

Fig. 8. Sample registration results using SOFIT for generic objects in video, each row of which represents challenges in different aspects. The reference frame is randomly selected from each sequence and the average representations before/after SOFIT alignment are shown in the last two columns. The average between-frame optical flow length is computed to evaluate the smoothness of a sequence. It is observed that the average optical flow length becomes smaller after alignment, demonstrating an image sequence with smoother transition generated by our algorithm. (a) Non-rigid motion, mainly caused by the poses of a horse. (b) In-plane and out-of-plane rotation. (c) Appearance variation. This row contains Magnetic Resonance Imaging (MRI) samples from different individuals. Though it is not strictly video per se, the results show that our method is robust against intensity variations. (d) Cluttered background. (e) Outliers. While the camera is non-stationary, there is a rocket moving against the Mars. Since we explicitly model the dominant motion of the scene, outliers (such as the rocket) have little impact on the alignment results. The number below each image is the average between-frame optical flow length of an entire sequence; lower number indicates a smoother sequence.

rigid muscle motion is retained. We provide more visual alignment results on faces in Fig. 5.

4.2. Multi-frame image super-resolution

In the imaging process, it is common to acquire images with low-resolution (LR) and/or certain artifacts such as blurriness, noise, etc. Image super-resolution (SR) is the process of generating a high-resolution (HR) image from one or more LR inputs. In the past few decades, there has been extensive work on super-resolution methods. Based on the inputs, the SR algorithms can be classified in two categories: single-image based [46] and multi-image based methods [44]. We apply SOFIT registration algorithm proposed in this paper to generate aligned images as inputs to different multi-image based SR methods. Here we compare our registration method with two other ones: frequency domain based method (FR) [43], and registration using sub-pixel displacement (SP) [42]. These registration methods are then used in two SR methods: iterated back-projection (BP) [44], and normalized convolution based method (NC) [45]. Fig. 6 shows the visual comparison of some sample results using different registration methods in different SR algorithms.

From Fig. 6 we can see that with our registration method, the SR results are significantly improved over the other SR methods in terms of visual quality. Despite the poor quality of the input frames, the results by our method are smooth with much fewer artifacts (e.g., noise and blockiness). The gain on the performance of SR directly comes from the accurate registration by the proposed method. The output images by our methods are also well rectified which would be desirable for post-processing or subsequent recognition tasks.

To quantitatively evaluate the image quality using our registration

method, we compute a recently proposed non-reference image quality index, Quality-aware clustering (QAC) [47], for output images using our method and the competing super-resolution methods. QAC is a general purpose blind image quality assessment method that has high correlation with human perception of image quality. Fig. 7 lists the average image quality scores on 87 sequences from GEMEP-FERA database [37]. Compared to the LR input and output using different registration methods, the output of SOFIT achieves the highest scores with the lowest standard deviation, which indicates better visual quality in terms of QAC image quality measure.

4.3. Generic object registration

In addition, we apply our method to other objects with challenges in various aspects, as shown in Fig. 8. Each video contains approximately 50 frames of a detected object with compound motions. For each video, we randomly select one frame to be the reference as the input of our method. The visual registration results are shown in the last two columns of Fig. 8, where the mean image results are sharper and reveals more local details. To quantitatively demonstrate that our algorithm generates image sequences with smoother transition, the ℓ_2 norm of the between-frame optical flow is computed and then averaged across the entire sequence. It is essentially the average optical flow length of a sequence, reported below the corresponding mean image in Fig. 8. It is observed that the average optical flow length is generally smaller, indicating a smoother sequence after our alignment algorithm.

4.4. Limitations

The proposed method assumes that the objects are already detected at similar scales. Such assumption holds in practice for some applications such as face analysis, which are indeed our main focus in this paper. Object detection in general is still a challenging topic. It still needs a great amount of effort to improve its performance even in predefined domains, such as face detection and pedestrian detection in the wild under low image resolution. For applications in which the detection itself is difficult or the object scale varies greatly, our method may fail to work.

5. Conclusions

We developed a real-time video-alignment technique, SOFIT, to register objects with compound motion (i.e. non-rigid surface motion accompanied by rigid body motion). We demonstrated its effectiveness in applications such as AU recognition and image super-resolution. This approach utilizes holistic dense flow-based information, and therefore, it is robust against detection error and noise. Minor out-of-plane rotation can also be corrected by employing structural correspondence from SIFT flow. More importantly, this method is able to generate temporally smooth registration results, which can improve the performance of various recognition and image super-resolution tasks.

Acknowledgment

This work (by S. Yang, L. An, N. Thakoor and B. Bhanu) was supported in part by the National Science Foundation under Grant 1330110 and in part by the Office of Naval Research, Arlington, VA, USA, under Grant N00014-12-1-1026. This work was supported in part by the National Natural Science Foundation of China under Grant Numbers 61501312, 61403265 and 61602193. This work was also supported in part by the Science and Technology Plan of Sichuan Province under Grant number 2015SZ0226.

References

[1] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, Y. Ma, Towards a practical face recognition system: robust registration and illumination by sparse representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 597–604.
[2] M. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, Meta-analysis of the first facial expression recognition challenge, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (4) (2012) 966–979.
[3] R. Szeliski, Image alignment and stitching: a tutorial, Found. Trends Comput. Graph. Vis. 2 (1) (2006) 1–104.
[4] X. Wu, D. Zhang, Improvement of color video demosaicking in temporal domain, IEEE Trans. Image Process 15 (10) (2006) 3138–3151.
[5] M. Uenohara, T. Kanade, Real-time vision based object registration for image overlay, in: Proceedings of the 1995 Conference on Computer Vision, Virtual Reality and Robotics in Medicine, 1995.
[6] Y. Caspi, M. Irani, Spatio-temporal alignment of sequences, IEEE Trans. Pattern Anal. Mach. Intell. 24 (11) (2002) 1409–1424.
[7] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, The computer expression recognition toolbox (CERT), in: Proceedings of IEEE International Conference on Automatic Face Gesture Recognition and Workshops, 2011, pp. 298–305.
[8] T. Baltruvsaitis, P. Robinson, L.-P. Morency, Continuous conditional neural fields for structured regression, in: European Conference on Computer Vision, 2014.
[9] M. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (1) (2012) 28–43. http://dx.doi.org/10.1109/TSMCB.2011.2163710.
[10] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, J. Cohn, Intraface, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015.
[11] D. McDuff, R. El Kaliouby, J. Cohn, R. Picard, Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads, IEEE Trans. Affect. Comput. 6 (3) (2015) 223–235.
[12] C. Liu, J. Yuen, A. Torralba, SIFT flow: dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 978–994.
[13] S. Yang, B. Bhanu, Understanding discrete facial expressions in video using an emotion avatar image, IEEE Trans. Syst. Man Cybern., Part B: Cybern. 42 (4) (2012) 980–992.
[14] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.
[15] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[16] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3D object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (9) (2012) 4290–4303.
[17] Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai, 3-d object retrieval with hausdorff distance learning, IEEE Trans. Ind. Electron. 61 (4) (2014) 2088–2098.
[18] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[19] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proceedings IEEE International Conference Multimedia and Expo., 2005, pp. 317–321. ⟨http://dx.doi.org/10.1109/ICME.2005.1521424⟩.
[20] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 46–53.
[21] H. Fang, N.M. Parthalin, A.J. Aubrey, G.K. Tam, R. Borgo, P.L. Rosin, P.W. Grant, D. Marshall, M. Chen, Facial expression recognition in dynamic sequences: an integrated approach, Pattern Recognit. 47 (3) (2014) 1271–1281.
[22] Y. Li, S.M. Mavadati, M.H. Mahoor, Y. Zhao, Q. Ji, Measuring the intensity of spontaneous facial action units with dynamic bayesian network, Pattern Recognit. 48 (11) (2015) 3417–3427.
[23] M. Valstar, B. Jiang, M. Méhu, M. Pantic, K. Scherer, The first facial expression recognition and analysis challenge, in: IEEE International Conference on Automatic Face Gesture Recognition and Workshops, 2011, pp. 921–926.
[24] M. Valstar, T. Almaev, J. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, J. Cohn, Fera 2015, second facial expression recognition and analysis challenge, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015.
[25] E. Learned-Miller, Data driven image models through continuous joint alignment, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 236–250.
[26] G. Huang, V. Jain, E. Learned-Miller, Unsupervised joint alignment of complex images, in: Proceedings ICCV, 2007.
[27] Y. Peng, A. Ganesh, J. Wright, W. Xu, Y. Ma, RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2233–2246.
[28] T. Baltrusaitis, M. Mahmoud, P. Robinson, Cross-dataset learning and person-specific normalisation for automatic action unit detection, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015.
[29] B. Martinez, M. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[30] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532–539.
[31] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in

the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886. ⟨http://dx.doi.org/10.1109/CVPR.2012.6248014⟩.
[32] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[33] P.J. Huber, Robust Statistics, John Wiley & Inc., Hoboken, NJ, 1981.
[34] S. Baker, S. Roth, D. Scharstein, M. Black, J. Lewis, R. Szeliski, A database and evaluation methodology for optical flow, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[35] M. Kaess, F. Dellaert, A markov chain monte carlo approach to closing the loop in SLAM, in: IEEE International Conference on Robotics and Automation, 2005, pp. 643–648.
[36] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools.
[37] FERA2011: Facial Expression Recognition and Analysis Challenge, ⟨http://sspnet.eu/fera2011/⟩.
[38] T. Ojala, M. Pietikäinen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[39] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[40] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) 692–706.
[41] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition, in: IEEE International Conference on Computer Vision, 2005.
[42] D. Keren, S. Peleg, R. Brada, Image sequence enhancement using sub-pixel displacements, in: Computer Society Conference on Computer Vision and Pattern Recognition, 1988, pp. 742–746.
[43] P. Vandewalle, S. Süsstrunk, M. Vetterli, A frequency domain approach to registration of aliased images with application to super-resolution, EURASIP J. Appl. Signal Process.
[44] M. Irani, S. Peleg, Improving resolution by image registration, Graph. Models Image Process. 53 (3) (1991) 231–239.
[45] T.Q. Pham, L.J. van Vliet, K. Schutte, Robust fusion of irregularly sampled data using adaptive normalized convolution, EURASIP J. Appl. Signal Process.
[46] J. Sun, Z. Xu, H.-Y. Shum, Image super-resolution using gradient profile prior, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[47] W. Xue, L. Zhang, X. Mou, Learning without human scores for blind image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 995–1002.

Songfan Yang received the B.S. degree in Electrical Engineering from Sichuan University, Chengdu, China, in 2009 and the M.S. and Ph.D. degree in Electrical Engineering from University of California, Riverside. He is currently an Associate Professor of College of Electronics and Information Engineering at Sichuan University. His research interests include computer vision, pattern recognition, and affective computing. He holds the Best Entry Award of the FG 2011 Facial Expression Recognition and Analysis emotion challenge (FERA) competition.

Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University in China in 2006, the M.Sc. degree in electrical engineering from Eindhoven University of Technology in Netherlands in 2008, and the PhD degree in electrical engineering from University of California, Riverside in USA in 2014. His research interests include image processing, computer vision, pattern recognition, and machine learning. He received the best paper award from the 2013 IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).

Yinjie Lei received his MS degree from Sichuan University, Chengdu, China in the area of image processing, and the Ph.D. degree in Computer Vision from University of Western Australia, Crawley, Australia. He is currently a Lecturer at Sichuan University, Chengdu, China. His research interests include image and text understanding, 3D face processing and recognition, 3D modeling, machine learning and statistical pattern recognition

Mingyang Li received B.Eng. degree in Automation Engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2009, and Ph.D. degree in Electrical Engineering from University of California, Riverside, in 2014. He is currently working as a research software engineer at Google Inc. His research interest lies primarily in the areas of robotics and computer vision. In particular, his research focuses sensor fusion, vision-aided inertial navigation, and multiple-view geometry in computer vision.

Ninad Thakoor received the B.E. degree in electronics and telecommunication engineering from the University of Mumbai, India, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Arlington, in 2004 and 2009, respectively. He is with Center for Research in Intelligent System at University of California at Riverside as a postdoctoral researcher.

His current research interests include vehicle recognition, stereo disparity segmentation, and structure-and-motion segmentation.

Bir Bhanu received the S.M. and E.E. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, the Ph.D. degree in Electrical Engineering, from the Image Processing Institute at University of Southern California and the M.B.A. degree from the University of California, Irvine. He is the Distinguished Professor of Electrical Engineering and Cooperative Professor of Computer Science and Engineering, Mechanical Engineering and Bioengineering, and Author Biography the Director of the Center for Research in Intelligent Systems (CRIS) and the Visualization and Intelligent Systems Laboratory (VISLab) at the University of California, Riverside (UCR). In addition, he serves as the director of NSF IGERT on Video Bioinformatics at UCR. Dr. Bhanu has been the Principal Investigator of various programs for NSF, DARPA, NASA, AFOSR, ONR, ARO, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine-vision applications. He has published seven coauthored and three edited books. He is the holder of 18 (3 pending) patents. He has published more than 500 reviewed technical publications, including over 140 journal papers and 44 book chapters. His current research interests are Computer Vision, Pattern Recognition and Data Mining, Machine Learning, Artificial Intelligence, Image Processing, Image and Video Database, Graphics and Visualization, Robotics, Human-Computer Interactions, Biological, Medical, Military and Intelligence applications. He is Fellow of IEEE, AAAS, IAPR, and SPIE.

Yiguang Liu received the M.S. degree from Peking University, Beijing, China, in 1998, and the Ph.D. degree from Sichuan University, Chengdu, China, in 2004. He was a Research Fellow, Visiting Professor, and Senior Research Scholar with the National University of Singapore, Singapore, Imperial College London, London, U.K., and Michigan State University, East Lansing, MI, USA, respectively. He was chosen into the program for new century excellent talents of MOE in 2008, and chosen as a Scientific and Technical Leader in Sichuan Province in 2010. He is currently the Director of the Vision and Image Processing Laboratory and a Professor with the School of Computer Science, Sichuan University. He has co-authored over 100 international journal and conference papers, and a chapter of the book entitled Computational Intelligence and Its Applications (H.K.Lam). His current research interests include computer vision and image processing, pattern recognition, and computational intelligence. Dr. Liu is a Reviewer of Mathematical Reviews of the American Mathematical Society.