Improving Action Units Recognition Using Dense Flow-based Face Registration in Video

Songfan Yang, Le An, Bir Bhanu and Ninad Thakoor

Abstract— **Aligning faces with non-rigid muscle motion in the real-world streaming video is a challenging problem. We propose a novel automatic video-based face registration architecture for facial expression recognition. The registration process is formulated as a dense SIFT-flow- and optical-flow-based affine warping problem. We start off by estimating the transformation of an arbitrary face to a generic reference face with canonical pose. This initialization in our framework establishes a head pose and person independent face model. The affine transformation computed from the initialization is then propagated by affine transformation estimated from the dense optical flow to guarantee the temporal smoothness of the non-rigid facial appearance. We call this method SIFT and optical flow affine image transform (SOFAIT). This real-time algorithm is designed for realistic streaming data, allowing us to analyze the facial muscle dynamics in a meaningful manner. Visual and quantitative results demonstrate that the proposed automatic video-based face registration technique captures the appearance changes in spontaneous expressions and outperforms the state-of-the-art technique.**

I. INTRODUCTION

Image registration is a classical topic in computer vision. It aims to overlay images of the same scene taken under different circumstances, such as time, viewpoint, and sensor [1]. In the context of human facial expression analysis, behavioral scientists have developed facial action coding system (FACS) [2] as an objective standard. According to FACS, human coders can decompose every possible facial behavior into action units (AU), which roughly correspond to the muscles that produce them, as shown in Fig. 1.

Automatic AU recognition [3], [4], has been quite successful for well-aligned, posed data, such as MMI [5] and CK+ [6] dataset. Unfortunately, AU recognition in an uncontrolled real-world environment remains a difficult problem, as shown in the Facial Expression Recognition and Analysis Challenge (FERA2011) [7]. This is due to the challenges in the automatic face registration for realistic data:

1) The facial muscle motion is non-rigid.
2) In the real data facial expressions are coupled with rigid motion of head pose.
3) The head pose comprises of both in-plane rotation and out-of-plane rotation.
4) The data are streaming instead of being in a batch form.
5) The consecutive frames should comply with temporal smoothness constraint.
6) The resolution of the face region is changing.

S. Yang, L. An, B. Bhanu, and N. Thakoor are with Center for Research in Intelligent Systems, University of California Riverside, Riverside, CA 92521, US. songfan.yang@email.ucr.edu, lan004@ucr.edu, bhanu@cris.ucr.edu, and ninadt@ucr.edu

Existing face registration approaches attempt to solve different aspects of the aforementioned challenges. In the face recognition and image retrieval community, researchers attempt to get rid of the non-rigid motion from facial data through registration using an ensemble of images [8], [9], [10]. These approaches are not suitable for the facial expression and AU recognition community, where registration is carried out so that the non-rigid facial muscle motion is retained and robustly recovered from streaming data in real-time. The state-of-the-art systems [4], [11] track a set of anchor points on the face and estimate the affine transformation based on which the entire face is warped. However, as demonstrated by Fig. 2, these methods do not address the temporal smoothness issue for proper alignment. 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 [12] adopt SIFT flow technique [13] to align every frame to a reference face. As shown in Fig. 2 column 3, the outcome of the SIFT flow transform has a large amount of discontinuities and artifacts. Although they solve this issue by generating image-based face representations (Emotion Avatar Image) and a reference model (Avatar Reference), carrying out the double layer loopy-belief propagation for every frame is computationally expensive and not suitable for real-time systems.

As illustrated in Fig. 2, we consider registration of frame 2 with respect to frame 1. All methods in this figure are able to account for the in-plane head rotation. However, as seen in the frame difference image (row 3) for the point-based affine transformation (column 2) and the SIFT flow transformation (column 3), there is a motion on most parts of the face. This is similar to the original face image (column 1) where the image is the output of Viola-Jones face detector [14] and is not registered. 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 our proposed method (column 4).

In this paper, we propose a video based face registration approach, namely SOFAIT, that tackles all the aforementioned challenges in real-world datasets for facial motion analysis. Inspired by the philosophy of [12] where the alignment is done with respect to a reference face model, we transform every frame of the streaming data onto a

Fig. 1. Examples of AUs defined in the FACS.

reference with canonical pose, expression, and illumination. This philosophy is significant in the facial expression analysis because facial muscle motion is similar for the same expression irrespective of the person [2], 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.

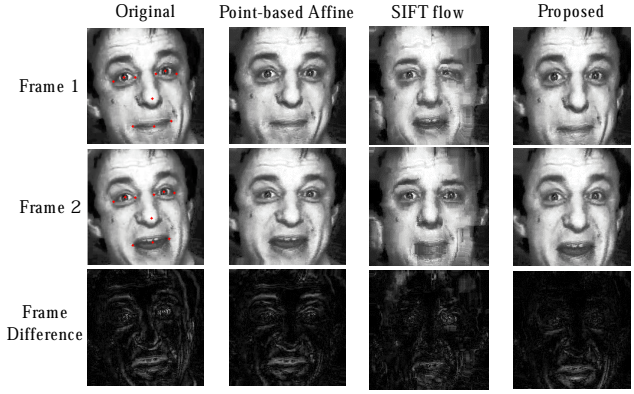

Fig. 2. Comparison of registration results. Row 3 is the absolute difference of frame 1 and frame 2. Column 2 follows the registration method used in [11], where affine transformation is computed from the points shown in column 1. Column 3 uses SIFT flow to align with the Avatar Reference face model from [12]. Ideally, we would like the frame difference to only show where the non-rigid motion is present (mouth area in this case). Our proposed method achieves the most plausible result.

The architecture of the SOFAIT is illustrated in Fig. 3. First, for faces extracted from Viola-Jones detector, we analyze their low-rank component of the starting frame by Transform Invariant Low-rank Texture (TILT) [15]. TILT algorithm can accurately rectify a face with in-plane rotation by recovering its intrinsic low-rank texture and domain transformation. Second, we compute the affine transformation from the dense SIFT flow of the rectified face to the reference face. Different from the point-based affine transformation, the dense flow-based affine is not only more robust to outliers but also able to maintain the non-rigid muscle motion. This solution is inspired by [16] in the image coding area where multiple affine models are estimated for motion segmentation of two images. We consider our face model subject to one dominant affine transformation which approximates the rigid head motion. Third, for consecutive frames thereafter, we apply a similar idea of dominant affine motion of the scene, and estimate the parameters for inter frame transformation from the dense optical flow. Fourth, we train a learning-based registration validation model to decide when to re-initialize the process once our system encounters any undesirable alignment result.

The contributions of this work are summarized as follows:

1) We propose a novel real-time video-based face registration technique, SOFAIT that aligns the real-world face data with non-rigid muscle motion for the facial expression analysis (Section II).
2) SOFAIT is an holistic approach and no detection of local features (eyes, nose, mouth) is needed. Therefore, it is tolerant to noise and low image resolution. The proposed method is also robust to minor out-of-plane head rotation, and it guarantees temporal smoothness for image sequences, as shown in Section II-B.
3) We define a learning-based model to validate the face registration results. The linear Support Vector Machine (SVM) is trained on Histogram of Oriented Gradients (HOG) features for classification (Section II-C).
4) We demonstrate that our registration method is applicable in spontaneous facial expressions analysis and existing AU recognition system performance enhancement. We carry out an *Independent* evaluation by a third party (the FERA2011 challenge organizer [7]) and show improvements over the state-of-the-art approach [12] and the baseline approach [7].

The rest of the paper is organized as follows: Section II presents our formulation and solution to the face registration problem using the dense flow information (both SIFT flow and optical flow) to estimate the affine transformation parameters. The experimental results are provided in Section III.

II. Real-time Video-based Face Registration

The objective of this work is to align real-world video-based face data in an uncontrolled environment. If not otherwise specified, the original input of our system is considered to be the detected faces by the Viola-Jones detector [14]. We use images from the Belfast Naturalistic dataset [17] to demonstrate that our approach is applicable in spontaneous expression recognition. We first introduce how TILT [15] is adopted in our system to accurately recover the in-plane head rotation. Subsequently, we use the structural information from SIFT flow to estimate the affine transformation for aligning faces with respect to a reference face model. This enables our system to tolerate an out-of-plane head rotation. The steps of TILT and SIFT flow based affine transformation are considered to be the initialization process. We then

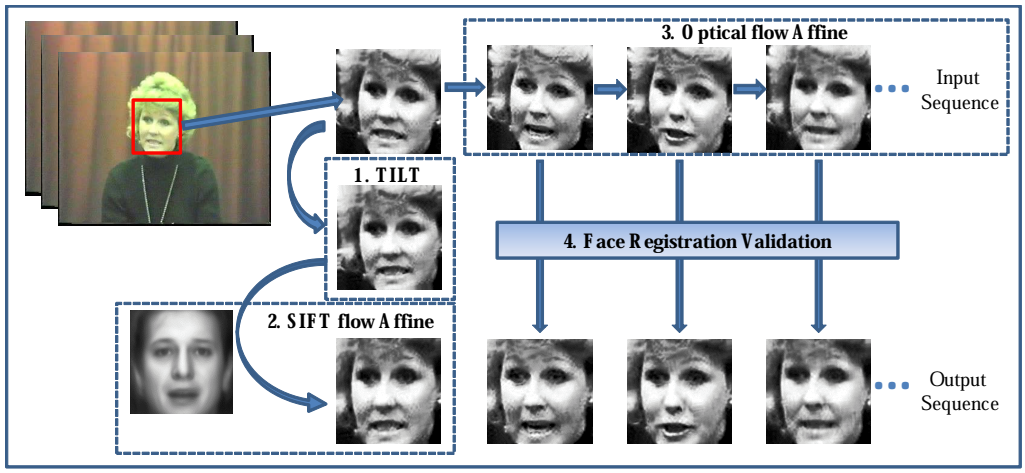

Fig. 3. The architecture of the proposed registration method SOFAIT .

impose temporal smoothness by computing the pairwise dense optical flow between the original consecutive frames. Finally, since the error is propagated during this process, we address the issue by training a SVM classifier to validate the registration results based on the HOG feature.

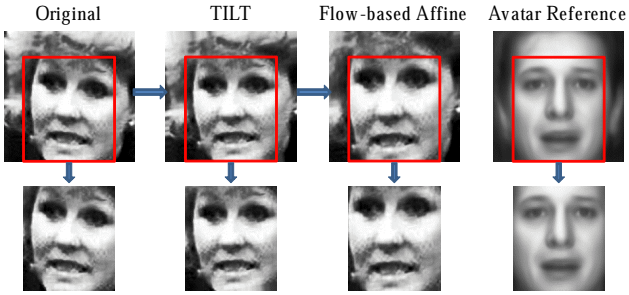

Fig. 4. The initialization procedure. We first use TILT [15] to eliminate the in-plane rotation of the face. Affine transformation is then estimated from SIFT flow to align the facial features. For comparison, we also show the level-1 Avatar Reference face model [12] (column 4) which is used for computing SIFT flow. The facial features of the initialization result (column 3) are aligned with respect to the reference face.

A. Transform Invariant Low-rank Texture (TILT)

In [15], TILT aims to extract the invariant structures in a 2D image that correspond to a class of patterns on a planar surface in 3D. One assumption is that the invariant structure has regular patterns, whose appearance can be modeled as a low-rank matrix, such as edges, corners, and symmetric patterns. In the context of face registration, we adopt TILT to recover the symmetric structure as well as the affine transformation from a given face image. The fundamental idea of TILT is to view the raw pixel values of an image as a matrix and discover a transformation that results in a low-rank matrix subject to sparse errors.

An $m \times n$ image I^0 of a texture is considered to be of low-rank if $rank(I^0) \ll min(m, n)$. For a given image observation I, the scene could be warped and corrupted by transformations, errors, or occlusions, such that I is not of low-rank. Thus, a transformed image, $I \circ \tau$, can be considered as of low-rank, subject to a sparse error matrix E that models the errors and occlusions. This formulation leads to the following optimization problem:

$$\arg \min_{I^0, E, \tau} \; rank(I^0) + \gamma \parallel E \parallel_0 \;\; \text{subject to } I \circ \tau = I^0 + E \tag{1}$$

where $\parallel E \parallel_0$ represents the number of non-zero entries in the matrix E; γ is a positive constant. However, this cost function is intractable as it is non-convex and the constraint is non-linear. Fortunately, recent work in compressive sensing [18] suggests that the cost function can be relaxed as:

$$\arg \min_{I^0, E, \tau} \; \parallel I^0 \parallel_* + \lambda \parallel E \parallel_1 \;\; \text{subject to } I \circ \tau = I^0 + E \tag{2}$$

where $\parallel . \parallel_*$ represents the matrix $nuclear\ norm$ (the sum of singular values); $\parallel . \parallel_1$ represents the matrix $1 - norm$ (the sum of absolute values of all entries); λ is a positive weighting factor. The cost function is now convex and continuous. The non-linear constraint is resolved by iteratively linearizing the constraint.

We use a 100×100 face image as the input to recover its low-rank component as well as the transformation itself. The algorithm converges in about 5 iterations. As the results of the TILT shown in Fig. 4, the in-plane head rotation is rectified and the facial expression is retained.

B. Dense Flow-based Affine Transformation

After recovering the in-plane rotation of the first frame, we align this frame (target face) to a reference face model by computing the dense SIFT flow, based on which the affine transformation parameters are estimated. The product of the previous two affine warpings (TILT and SIFT flow affine) is considered as the registration transformation for the first frame. For the subsequent frames, we compute the pairwise optical flow for the original input face sequence, and estimate the affine parameters of the current inter frame transformation.

1) Compute SIFT Flow for the Starting Frame: SIFT flow [13] was originally designed to align an image to its plausible nearest neighbor which can have large variations. The SIFT flow algorithm robustly matches dense SIFT features between two images, while maintaining spatial discontinuities. The local gradient descriptor, SIFT [19], is used to extract a pixel-wise feature component. For every pixel in an image, the neighborhood (e.g. 16×16) is divided into a 4×4 cell array. The orientation of each cell is quantized into 8 bins, generating a $4 \times 4 \times 8 = 128$-dimension vector as the SIFT representation for a pixel, or the so called SIFT image.

After obtaining the per-pixel SIFT descriptors for two images, a dense correspondence is built to match the two images. Similar to optical flow, the objective energy function is designed as:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} min(\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|_1, t) + \quad (3)$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) + \quad (4)$$

$$\sum_{(\mathbf{p},\mathbf{q}) \in \varepsilon} min(\alpha |u(\mathbf{p}) - u(\mathbf{q})|, d) + min(\alpha |v(\mathbf{p}) - v(\mathbf{q})|, d) \quad (5)$$

where $\mathbf{p} = (x, y)$ is the grid coordinates of the images, and $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at $\mathbf{p}$. $u(\mathbf{p}), v(\mathbf{p})$ are the flow vectors for x direction and y direction respectively. s_1 and s_2 are two SIFT images to be matched. ε contains all the spatial neighbors (a four-neighbor system is used). The *data term* in (3) is a SIFT descriptor match constraint that enforces the match along the flow vector $\mathbf{w}(\mathbf{p})$. The *small displacement constraint* in (4) allows the flow vector to be as small as possible when no other information is available. The *smoothness constraint* in (5) takes care of the similarity of flow vectors for adjacent pixels. In this objective function, the truncated $L1$ norm is used in both the data term and the smoothness term with t and d as the thresholds for matching outliers and flow discontinuities, respectively. η and α are scale factors for the small displacement and smoothness constraint, respectively. The dual-layer loopy belief propagation is used as the base algorithm to optimize the objective function. Then, a coarse-to-fine SIFT flow matching scheme is adopted to improve the speed and the matching result.

The SIFT flow computation from a target face with respect to a reference face is demonstrated in Fig. 5. The raw flow vectors (top row) are discontinuous, and the corresponding warping result has strong artifacts.

Remark. How to Choose a Reference Image: The reference frame is the level-1 Avatar Reference face model generated from the GEMEP-FERA training dataset [12] (Fig. 4 column 4). It can be viewed as the super-resolved version of the mean face of the entire dataset. This reference face model should be in canonical pose and expression. As demonstrated later in the experiment results in Section III, the choice of the Avatar Reference from different datasets will not significantly affect the registration and AU recognition results.

2) Flow-based Affine Transformation: Instead of detecting feature points (such as eye corners, nose tip etc.) and inferring affine transformation from them, we use the dense flow information to estimate the affine transformation, namely the dominant motion. In homogeneous coordinates, we represent the pixel location of a target frame and a reference frame by $\overline{\mathbf{p}} = (x, y, 1)$ and $\overline{\mathbf{p}}' = (x', y', 1)$, respectively. Given the target frame pixel location and its corresponding flow vectors, the reference frame pixel location can be written as $x' = x + u(\mathbf{p})$, $y' = y + v(\mathbf{p})$. Thus, we can model the affine transformation for all N pixels in a image as follows:

$$\begin{pmatrix} x'_1 & \cdots & x'_N \\ y'_1 & \cdots & y'_N \\ 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 & \cdots & x_N \\ y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{pmatrix} \quad (6)$$

where x and y components of the reference frame can be decomposed as:

$$\begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = \begin{pmatrix} x'_1 \\ \vdots \\ x'_N \end{pmatrix} \quad (7)$$

$$and \begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{pmatrix} \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} = \begin{pmatrix} y'_1 \\ \vdots \\ y'_N \end{pmatrix} \quad (8)$$

As flow vectors may suffer from outliers, we solve for the maximum likelihood estimates of this overdetermined system robustly by iteratively reweighted least squares (IRLS) [21].

Fig. 5 shows the affine transformation estimated from the SIFT flow vectors. The corresponding flow field captures the characteristics of the raw SIFT flow field in a continuous manner. With this, we finish the initialization process.

Consider the sequence of unregistered face images thereafter, their size may vary from Viola-Jones face detector even when the original images have the same scale. A naïve way of scale normalization cannot solve this problem. In order to align features under different scale and extract non-rigid facial muscle motion, we have to generate temporally aligned faces.

After registering the starting frame of the face sequence, we now consider processing the frame right after the starting frame. We compute the optical flow from the current frame to the unregistered previous frame (starting frame, in this case). As shown in Fig. 6(a), the warping based on optical flow almost perfectly aligns the entire face. Applying the aforementioned dominant motion philosophy, we compute the affine transformation from this dense optical flow. This transformation is considered as the warping of the current frame to the unaligned previous frame. Their results (Fig. 6(a) and (b)) are similar, but (b) is more smooth and able to retain the non-rigid motion in the mouth area. We then incorporate this

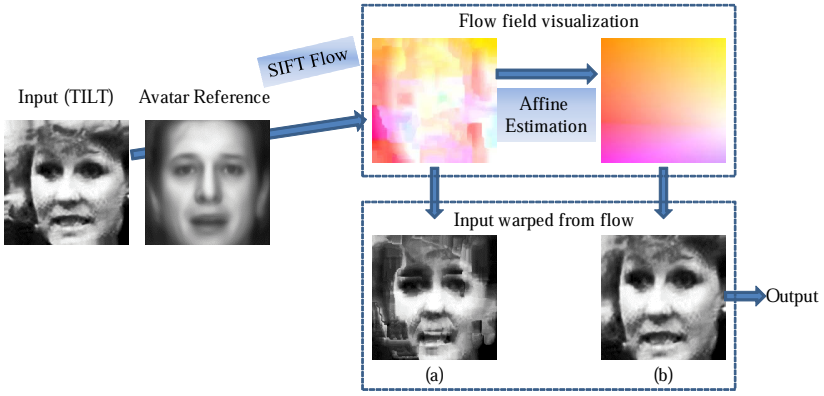

Fig. 5. The affine motion estimation from SIFT flow. The SIFT flow of the input is computed with respect to the level-1 Avatar Reference. The visualization of the flow field uses the color-coding scheme of [20]. For comparison, we also show the warped input image based on: (a) SIFT flow and (b) affine estimation of SIFT flow.

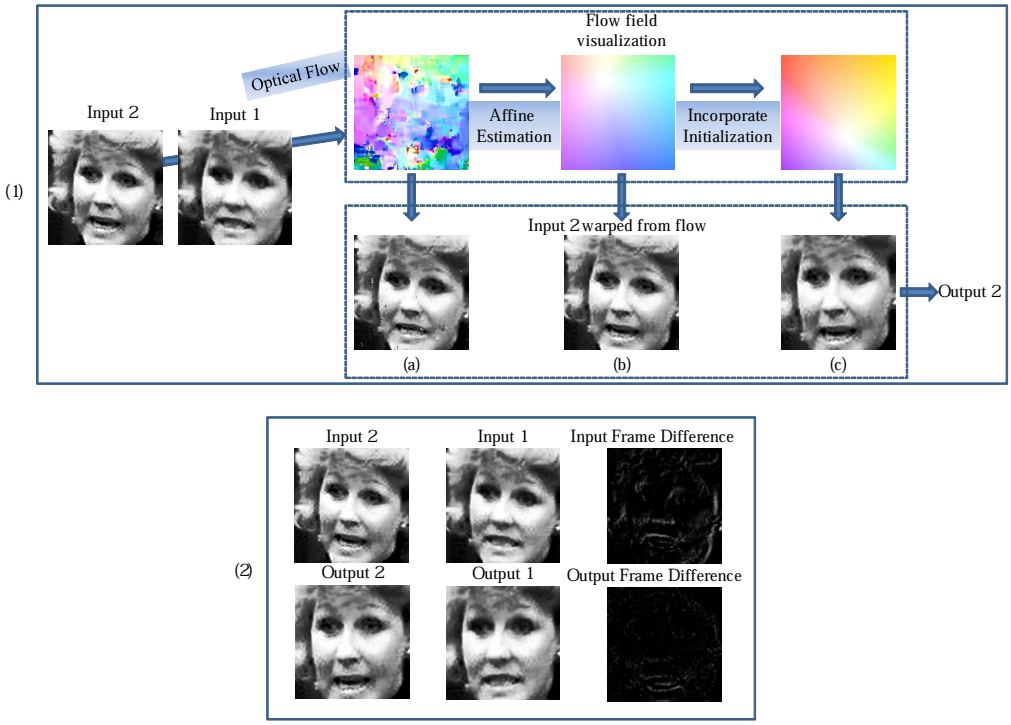

Fig. 6. (**1**) The affine motion estimation from optical flow. The optical flow of the input 2 is computed with respect to the the input 1. For comparison, we show the input 2 warped based on (a) optical flow; (b) affine estimation of optical flow; (c) incorporation of initialization transformation (from Fig. 5) and the flow from (b). (**2**) Temporal smoothness analysis. The output 1 is initialization result in Fig. 5 (b). The frame difference represents our registration method captures the non-rigid motion in mouth area.

affine transformation with the transformation of the previous frames (only the initialization in this case) by simple matrix multiplication, resulting in the registration transformation for the current frame. This process is repeated for the frames thereafter. To analyze the temporal smoothness, we also compare the frame difference of the original input image pair and the registered image pair in Fig. 6(2). The frame difference of the registered image pair represents the non-rigid motion in mouth area (similar to what the human visual system captures).

C. Evaluation of Registration Results - Learning-based Validation

Since the error is propagated by the affine matrix multiplication, we need an automatic approach to evaluate the current registration result. In this work, we treat this process as a binary classification problem. For 1200 positive and 2000 negative examples (samples are shown in Fig. 7), we compute Histogram of Oriented Gradients (HOG) [22] features. The edge and structure information captured by the HOG feature is used to train a linear SVM classifier. Since the true negative examples are limited (about 80 out

Fig. 7. Training examples for the registration validation model.

of 1100 face images from the GEMEP-FERA [7] dataset), we generate synthetic data by applying the affine transformation of misaligned examples to the entire dataset. This approach achieves 94.7% accuracy in predicting more than 5000 unseen examples.

Remark. Computational Cost Analysis: For SOFAIT registration architecture, we train our validation classifier in an offline phase. During the online phase, we only compute TILT and SIFT flow for the starting frame as the initialization step. The steps to be carried out for every frame are the following:

1) Compute the dense optical flow with respect to the previous frame
2) Estimate the affine transformation from the dense optical flow
3) Accumulate affine transformation by matrix multiplication
4) Apply affine transformation to register the current image
5) Calculate HOG feature for current image
6) Classify the feature using the trained validation model

We adopt both the optical flow and the LBP implementation in OpenCV [23]. The classification step is realized using the liblinear [24] library. The aforementioned steps can be finished in less than 20ms for a 100×100 image on a dual-core Intel 3.40GHz machine with 8GB memory, which means that we can carry out real time registration process at about 50 fps.

The initialization step takes on the average 1.69 second using the aforementioned settings. In the meantime, our system stores all the extracted face images in a buffer to prevent any information loss. After the system finishes initialization, we process the frames in the buffer. In our experiment, the video frame rate is 25 fps, our system takes another 1.72 second to empty the images in the buffer. Therefore, after 3.41 second, our system is able to process a face image right after it has been extracted from the face detector. Although the initialization delay will not affect processing a single frame of the video after initialization or causing any information loss, our system works in real-time after 3.41 seconds, literally.

III. Experimental Results

We demonstrate SOFAIT face registration technique by facial action unit (AU) recognition on FERA Challenge dataset [25]. We use the exact same feature as the baseline approach for better comparison and show superior performance. We also compare our approach with the state-of-the-art registration technique [12].

The goal is to detect 12 frequently occurring AUs (Fig. 1) on a per-frame basis. We use the same protocol as the Facial Expression Recognition and Analysis Challenge (FERA2011) [7] 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 MMI [5] and CK+ [6] datasets.

The baseline registration method detects both eye locations of the face, scale, and in-plane rotate the face. This registration belongs to in-plane image transformation category as summarized in [26]. The state-of-the-art registration technique is called Emotion Avatar Image (EAI) [12], [26]. It is another variation of SIFT flow method [13].

To demonstrate how our registration technique can improve the existing AU recognition system, we adopt exactly the same procedure and parameters as in the baseline approach [7] except the face registration step. After we extracted the face from Viola-Jones face detector [14], we resize them to 100×100 images and register them using the proposed method. Subsequently, we divide the image into 10×10 blocks, where the same type of Local Binary Pattern (LBP) [27] texture feature for each block is computed and concatenated as in the baseline approach. We then train 12 linear SVM binary classifiers based on the implementation of [28].

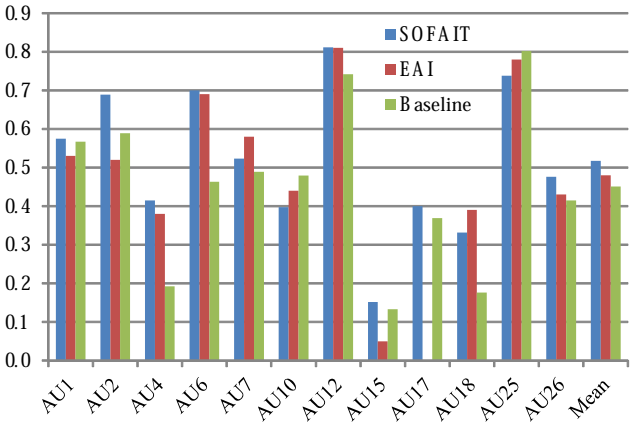

Fig. 8. (a) F1-score visualization between the proposed SOFAIT approach, EAI registration approach [12], and the baseline approach [7]. The corresponding number can be found in Table I. The performance of baseline algorithm is enhanced by our registration technique in all the AU classes except AU10 and AU25. The overall performance of the proposed registration method is also superior over the state-of-the-art EAI registration technique.

In order to compare our registration technique with EAI registration technique, we use the level-1 Avatar Reference generated from FERA Challenge training data [25]. In generating a EAI representation, we need to select a temporal

TABLE I

F1-SCORE COMPARISON WITH THE BASELINE APPROACH [7] AND EAI APPROACH [12] FOR THE COMBINATION OF PERSON SPECIFIC AND PERSON INDEPENDENT TEST.

AU	1	2	4	6	7	10	12	15	17	18	25	26	Mean
SOFAIT	0.58	0.69	0.41	0.70	0.52	0.40	0.81	0.15	0.40	0.33	0.74	0.48	0.52
EAI	0.53	0.52	0.38	0.69	0.58	0.44	0.81	0.05	0.00	0.39	0.78	0.43	0.48
Baseline	0.57	0.59	0.19	0.46	0.49	0.48	0.74	0.13	0.37	0.18	0.80	0.42	0.45

length parameter. The author of the original EAI paper [12] chooses the length of a single video (around 2 seconds) for this parameter for facial expression recognition on a per-video basis. To generalize this registration technique in AU recognition on a per-frame basis, we heuristically determine the best value for the temporal length parameter. We carry out a leave-one-subject-out cross validation on the FERA Challenge training data, and determine the parameter value to be 0.56 second for the best F1 score over all AUs. This means for each frame in a video, approximately 14 closest frames will be used to compute EAI representation. For the boundary frames, i.e. the starting and ending 7 frames, we simply assign their values to be the 8th frame from the beginning and the 8th from the end, respectively. Thereafter, the aforementioned features are extracted from the EAI representations.

We submitted the predictions on 71 testing sequences (around 4000 images) for both the SOFAIT and the EAI method to the FERA2011 organizer [7] for an independent evaluation. The organizer evaluated our predictions using the same FERA2011 challenge protocol and provided us the results. The result for baseline method can be found on FERA challenge website [25]. The overall performance (including person-independent and person-specific tests) is shown in Table I and Fig. 8. Our SOFAIT method outperforms the other two registration techniques in most AU classes. Fig. 9 illustrates that SOFAIT registration is able to enhance the baseline method in the majority of the AU classes (maximum improvement is 24%) as well as the overall performance (7% increase on average). SOFAIT did not help in AU10 and AU25 as for the 2 sequences which includes AU10 and AU25, our registration achieve poor results because of major out-of-plane head rotation.

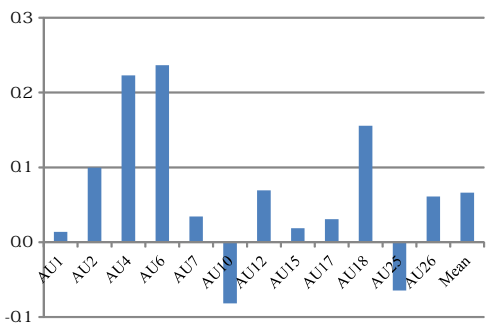

Fig. 9. F1-score percentage improvement over the baseline algorithm [7] using SOFAIT registration technique.

To demonstrate that choosing different reference face model will not significantly affect the AU recognition results, we carry out three experiments using level-1 Avatar Reference [12] computed from MMI, CK+, and GEMEP-FERA datasets. Their corresponding F1 scores of the leave-one-out cross validation are similar, as shown in Fig. 10.

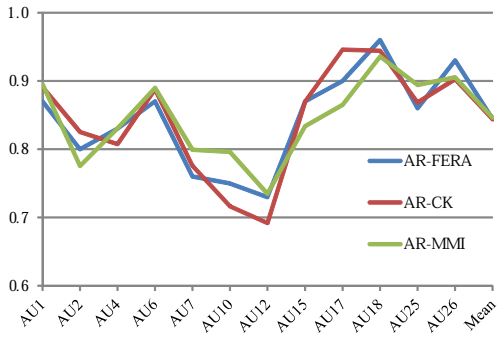

Fig. 10. F1-score comparison using Avatar Reference from different datasets. The variation in results is minor, which means the choice of the reference model will not affect the recognition result significantly.

IV. CONCLUSIONS

We developed a video-based real-time face registration technique, SOFAIT, and demonstrate its applications in AU recognition. This approach utilizes holistic dense flow-based information, and therefore, it is robust to detection error and noise. Minor out-of-plane head rotation can also be corrected by employing structural information from SIFT flow. Besides, this method is able to generate temporally smooth registration results which are essential for spontaneous facial expression analysis and super-resolution. Last but not the least, this method performs registration at 50 fps, and is suitable for real-time processing.

V. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0727129 and 0915270. The authors would like to thank Dr. Michel Valstar from the University of Nottingham, the organizer of FERA 2011 Challenge, for evaluating the AU testing results.

REFERENCES

[1] Zitová, B., Flusser, J.: Image Registration Methods: A Survey. Image and Vision Computing (2003)

Fig. 11. More results using SOFAIT registration technique.

[2] Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
[3] Zhao, G., Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE Trans. PAMI (2007)
[4] Valstar, M., Pantic, M.: Fully Automatic Recognition of the Temporal Phases of Facial Actions. IEEE Trans. SMC-B (2012)
[5] Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based Database for Facial Expression Analysis. In: IEEE Int. Conf. on Multimedia and Expo. (2005)
[6] Kanade, T., Cohn, J., Tian, Y.: Comprehensive Database for Facial Expression Analysis. In: Proc. FG. (2000)
[7] Valstar, M., Jiang, B., Méhu, M., Pantic, M., Scherer, K.: The First Facial Expression Recognition and Analysis Challenge. In: Proc. FG Workshop on FERA Challenge. (2011)
[8] Learned-Miller, E.: Data Driven Image Models Through Continuous Joint Alignment. IEEE Trans. PAMI (2006)
[9] Huang, G., Jain, V., Learned-Miller, E.: Unsupervised Joint Alignment of Complex Images. In: Proc. ICCV. (2007)
[10] Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images. In: Proc. CVPR. (2010)
[11] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The Computer Expression Recognition Toolbox (CERT). In: Proc. FG. (2011)
[12] Yang, S., Bhanu, B.: Facial Expression Recognition Using Emotion Avatar Image. In: Proc. FG Workshop on FERA Challenge. (2011)
[13] Liu, C., Yuen, J., Torralba, A.: SIFT Flow: Dense Correspondence across Scenes and its Applications. IEEE Trans. PAMI (2011)
[14] Viola, P., Jones, M.: Robust Real-time Face Detection. IJCV (2004)
[15] Zhang, Z., Liang, X., Ganesh, A., Ma, Y.: TILT: Transform Invariant Low-rank Textures. In: Proc. ACCV. (2010)
[16] Wang, J., Adelson, E.: Representing Moving Images with Layers. TIP (1994)
[17] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional Speech: Towards A New Generation of Databases. Speech Communication (2003)
[18] Candès, E., Li, X., Ma, Y., Wright, J.: Robust Principal Component Analysis? Journal of the ACM (2011)
[19] Lowe, D.: Object Recognition from Local Scale-invariant Features. In: Proc. ICCV. (1999)
[20] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A Database and Evaluation Methodology for Optical Flow. In: Proc. ICCV. (2007)
[21] Huber, P.J.: Robust Statistics. John Wiley & Sons, Inc., Hoboken, NJ (1981)
[22] Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proc. CVPR. (2005)
[23] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
[24] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A Library for Large Linear Classification. JMLR (2008)
[25] : (FERA2011: Facial Expression Recognition and Analysis Challenge) http://sspnet.eu/fera2011/.
[26] Yang, S., Bhanu, B.: Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image. IEEE Trans. SMC-B (2012)
[27] Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. TPAMI (2002)
[28] Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. (2001)