An Efficient Method Utilizing Temporal Correlationship for Real-time Face Alignment

Abstract— **This paper addresses the problem of real-time face alignment. Usually, this problem is tackled by alignment in each individual frame. However, it suffers from limited performance involving temporal continuity, which is concretely reflected in smooth alignments. The underlying reason is that the temporal correlationship between shapes of adjacent frames doesn't be utilized. Therefore, we propose a real-time face alignment technique utilizing the temporal correlationship, namely shape residuals between consecutive frames in this paper. Specifically, we postulate that the shape residuals have a sparse pattern in some domain. Under the hypothesis, They can be explictly depicted by an offline model trained within compressive sensing framework. Once the residuals are recovered, we can obtain an alignment of each current frame by adding them to shape of the corresponding last frame. Otherwise, to avoid shape drift or loss during the real-time face alignment, a reboot mechanism is adopted to effectively align a current shape. Experimental results demonstrate the superior performance of our approach compared with the state-of-the-arts in terms of smoothness and working efficiency.**

I. INTRODUCTION

Static face alignment, namely performing face alignment in a static images in large-scale and unconstrained conditions has been a most studied topic within the computer vision community during the last decades. As accurate localization of fiducials is a vital prerequisite task for variety of applications, e.g., face recognition [1], 3D face reconstruction [2], [3], face de-identification [4] and expression analysis [5], [6], [7].

The existing static face alignment approaches mainly start the alignment process from the mean facial shape [8], and deform the shape constrained by offline-trained facial deformable parameterized models (FDPMs) to minimize the shape residual by gradient descend optimization. Active Shape Model (ASM) based approaches [9] and Active Appearance Models (AAM) based approaches [10], [11] are typical representative subject to this category. Recently, regression based approaches [12], [13], [14], [15], [16], which attempt to infer a face shape through a discriminative regression function by directly mapping textual features to shape, have been proposed with excellent performance. Moreover, the tree based cascade regression approaches [8], [17], [14], [18] provide a super real-time performance. Although the methods of static face alignment have shown great success in standard benchmark datasets [19] with respect to the efficiency, many of them suffer from significant performance degradation especially in real-world scenarios under wild conditions [20], [21]. These methods are not appropriate for real-time face alignment, for they usually align each frame independently from a stable initial shape, e.g., a mean face in a tracking-by-detection manner by various facial landmark detectors [22] Therefore, this manner does not utilize the temporal correlationship between consecutive frames, and results in a un-smooth fitting. Otherwise, offline-trained static facial deformable models (FDMs) can not express some large shape residual precisely. Therefore, real-time face alignment, namely fitting facial landmarks on consecutive frames, is still very challenging and there is only a few published work in this direction. Therefore, this is a clear research gap that needs to be addressed, which is exactly the focus of this work.

To exploit the temporal correlationship, the existing approaches for real-time face alignment usually adopt a tracking manner [7], [23], in which the location of facial landmarks in preceding frames can be used to find the facial landmarks in the current frame. These approaches take advantage of location of facial landmarks in the last frame or multiple nearly-optimum initial estimated results to find the facial landmarks in a current frame by minimizing shape residual. It exploits the fact that the various changes of face shape over time are smooth in videos. Then, the offline trained FDMs can capture the shape residual if the previous landmarks were detected with acceptable accuracy, which make the initial estimated shape be close enough to the current landmarks. Hence, tracking based approaches are more likely to produce highly accurate, illumination-change-robust and partial-occlusion-robust fitting results than the aforementioned static approaches. As described above, the tracking based approaches still employ the trained FDMs built offline using a set of static annotated facial images, which does not include the subject being tracked [24] to recover the shape residual from a close initializations. This may inevitably result in shape drift when starting from poor initial estimations for the reason that the offline FDMs are weak in expressing shape residual. Hence how to enlarge expression to shape residual for offline FDMs is an important problem. Therefore, incremental learning based approaches [24], [21], [25], [26], [27] are proposed to address the problem in an online manner, which improve the generic offline FDMs into a person-specific one.

In summarize, the existed approaches for real-time face alignment are respectively:

- Static face alignment based approaches, as illustrated in Figure 1(a), fit the current shapes, represented by red points, from mean shape, represents by green point individually.
- Tracking based approaches, as illustrated in Figure 1(b), fit the current shape, represented by red point, by

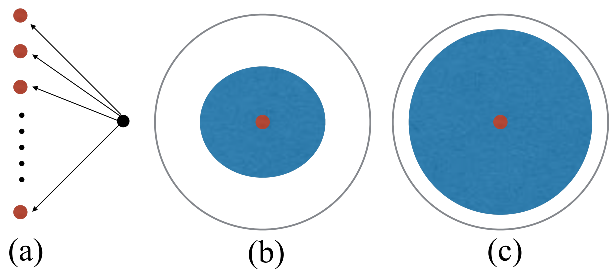

Fig. 1: Summary of the existed approaches for real-time face alignment. The red points represent the face shapes of current frames, the black point represents the mean face shape. The blue region, in which the face shapes of last frames may occur, indicates the ability that offline FDMs express the corresponding shape residuals. (a) represents the static face alignment based approaches. (b) represents the tracking based approaches, and the offline FDMs of these approaches can only tackle a part of the whole search region, in which the face shapes of last frames may occur. (c) represents the incremental learning based approaches and our proposed strategy, they all can enlarge offline FDM's in (b) expression to shape residuals. What the difference is, our strategy achieves that by an offline training task.

searching the inner green region which is smaller than the outer one, last frame shape may be in.

- Incremental learning based approaches, as illustrated in Figure 1(c), fit the current shape from a bigger searching region, represented by green color, relative to that of Figure 1(b).

However, incremental methods are nearly real-time [24], since incremental learning is costly, which severely impedes their applications on tasks need high real-time performance. Therefore, how to improve the offline FDMs with a better real-time performance is our concern.

A. Compressive sensing based approach for real-time face alignment

This problem can be addressed in an offline manner compared to incremental learning based approaches. Specifically, we hope to enlarge expression to shape residual for offline FDMs during an offline training task, as illustrated in Figure 1(c), which is same with the incremental learning based approaches but the enlarging manner of our proposed approach is offline. Motivated by the face recognition application [28], which proposes a general classification algorithm for object recognition based on sparse representation, we postulate that the shape residuals have a sparse pattern in some domain. Under the hypothesis, we propose a novel approach based on sparse representation to train an offline FDM which constructed from simulated sequential training data to perform real-time face alignment. To generate the simulated sequential training data, we randomly synthesis multiple shapes of last frame around the current truly shape. The offline FDM which captures residual between adjacent shapes is then achieved using sparse representation. During the fitting process, we start from the estimated result and inference the current shape through the trained offline FDM, while a reboot mechanism is adopted to alleviate model drifting. The major contributions of this paper are:

- To the best of our knowledge, this is the first work to perform real-time face alignment by an offline person-specific FDM.
- The proposed training strategy is radically different from existing methods with respect to the sparse representation.
- The proposed offline FDM is well designed for learning the temporal correlationship, and guarantees robust and real-time face alignment on the fly.

By conducting extensive experiments on multiple unconstrained databases, we show that our approach has significant real-time performance smoothment improvement compared with state- of-the-art, while constant computational cost w.r.t. both CPU time and memory usage is guaranteed. These merits make our approach very suitable for real-time and large scale applications.

II. Method

This section briefly reviews the theory of compressive sensing and sparse representation, and formulates our offline FDM under compressive sensing framework for real-time face alignment. Then, the details of training compressive sensing based offline FDM and fitting facial shape utilizing trained FDM are described.

A. Compressive Sensing and Sparse Representation

We use $x \in R^m$ to represent as a vector, the 1D dense signal to be reconstructed, let $D : R^n \to R^m$ represents sparse transform, which gives a sparse representation (mostly zero), $y \in R^n$, of x. Compressive sensing reveals that the dense signal x can be recovered if it is sparse under D. We formulate it as:

$$\min_{\alpha} \|\alpha\|_0 \qquad \text{s.t.} \quad x = D(\alpha) \tag{1}$$

Here $\|\cdot\|_0$ denotes l^0-norm, which counts the number of nonzero entries in a vector, $\alpha \in R^n$ is sparse representation of x. Usually, D can be pre-defined or data-driven sparse transform. The problem of finding the sparsest solution of an under-determined system of linear equations is NP-hard. However, there are greedy algorithms to solve this problem such as orthogonal matching pursuit (OMP) [29]. Alternatively, the l_0 quasi norm is replaced with its convex relaxation, the l_1 norm, at which the problem can be solved via linear programming.

B. Compressive Sensing Based Offline FDM (CS-OFDM)

At the beginning, we introduce some notation. Let I be a face image and X be its truth shape. For temporal scenario, we refer to I_t as the t frame image and the same to X_t. The recovery of shape residuals under the compressed sensing framework aims to solve the following optimization problem: Inspired by the theory of compressive sensing. Specifically, we presume the shape residuals have a sparse pattern in the domain concerned with face appearance, which could reflect the motion of face. Due to this hypothesis, our problem can be formulated under compressive sensing framework.

$$\begin{aligned} &\min_{\Gamma} \sum_{t} \|(X_t - X_{t-1}) - W\{D(I_t, X_{t-1})\}\|_2^2 \\ &\text{s.t.} \quad \|D(I_t, X_{t-1})\|_0 \le T_0 \quad \forall t \end{aligned} \tag{2}$$

The term $(X_t - X_{t-1})$ in Equation (2) represents the shape residuals to be recovered. Function $D(\cdot,\cdot)$ represents the sparse transform, which is our CS-OFDM. $D(I_t, X_{t-1})$ is the sparse representation of $(X_t - X_{t-1})$, and $W\{D(I_t, X_{t-1})\}$ is the estimation of $(X_t - X_{t-1})$. The l_0 quasi norm is used to encode the sparsity of the shape residual, and T_0 is the required sparsity level. Γ is used to denote the set of sparse representations of all shape residuals. This learning formulation minimizes the total fitting error of all shape residuals with respect to the sparse transform, $D(\cdot,\cdot)$, subject to sparsity constraints. Our sparse transform, $D(\cdot,\cdot)$, based on face appearance conducts sparse representation. In other words, we can index the sparse representation of the shape residuals by appearance of current frame and the last frame shape. Once that is indexed, the shape residuals can be estimated, and then, the estimation to current frame shape can be confirmed.

C. Training of Compressive Sensing Based Offline FDM

In this section, we give the explicit details of the training process, in which we obtain the data-driven CS-OFDM. In this paper, our CS-OFDM is trained with the static annotated dataset, $A : \{(I_0, X_0), (I_1, X_1), ..., (I_{n-1}, X_{n-1})\}$. Therefore, we should construct the sequential training samples. At first, we introduce some notation and re-formulate our problem. Then, how to simulate sequential training samples will be introduced. At last, we will depict the training details of our proposed CS-OFDM.

Let S be the set of training samples, in which the samples, S^j, concerned with image I_j, namely current frame image, will be $S^j : \{(I_j, X_{c,j}, \hat{X}^0_{l,j}), (I_j, X_{c,j}, \hat{X}^1_{l,j}), \ldots, (I_j, X_{c,j}, \hat{X}^{n-1}_{l,j})\}$. c represents the current frame, l represents the last frame. Hence, $X_{c,j}$ is the truly shape of the current frame and $\hat{X}^i_{l,j}$ is the ith simulated last frame shape. Then, Equation (2) can be re-formulated as:

$$\begin{aligned} &\min_{\Gamma} \sum_{i,j} \|(X_{c,j} - \hat{X}^i_{l,j}) - W\{D(I_j, \hat{X}^i_{l,j})\}\|_2^2 \\ &\text{s.t.} \quad \|D(I_j, \hat{X}^i_{l,j})\|_0 \le T_0 \quad \forall i,j \end{aligned} \tag{3}$$

1) Sequential training samples simulation: To generate the simulated sequential training samples, we randomly synthesis multiple last frame shapes around each annotated shape, which is regarded as current frame shape. The last facial shapes around the current one and the residual may occur in the scale, rotation and translation. The theory of compressive sensing implies that the precise choice of feature space is no longer critical: Even random features contain enough information to reflect the truth situation. Therefore, we can randomly sample multiple shapes around the current facial shape with various scale, rotation and translation settings. Let s denote scale random variable, r denote rotation random variable, t denote translation random variable. We suppose that the population distributions of these variables all are known to be normal, with mean μ_1, μ_2, and μ_3, and variance σ_1^2, σ_2^2, and σ_3^2, that are, $s \sim N(\mu_1, \sigma_1)$, $r \sim N(\mu_2, \sigma_2)$, and $t \sim N(\mu_3, \sigma_3)$, in which μ_1 is the scale of $X_{c,:}$, μ_2 is the rotation of $X_{c,:}$, and μ_3 is the translation of $X_{c,:}$. We obtain $\{s_1, s_2, ..., s_n\}$, $\{r_1, r_2, ..., r_n\}$, and $\{t_1, t_2, ..., t_n\}$ at the normal pattern. Once these parameters of deformation are sampled, we can apply them to $X_{c,\cdot}$ and obtain n simulated last facial shapes, $\{\hat{X}^0_{l,\cdot}, \hat{X}^1_{l,\cdot}, ..., \hat{X}^{n-1}_{l,\cdot}\}$.

2) Training CS-OFDM: We alternate between learning the sparse transform and sparse representations, and estimating the facial shape residual in a multi-stage coarse-to-fine manner to better fit the shape residuals. Otherwise, to obtain super real-time performance, we construct the sparse transform using the tree structure which has effective search. Specifically, after training stage d with training set S_d and before creating the regression tree based sparse transform (RTST) in stage $d+1$, we use the RTSTs, trained forest, to estimate the shape residual of each sample $s_i \in S_d$. Then, S_{d+1} is updated and is utilised to train the RTST in stage $d+1$. The forest is further grown in order to minimise the sum of square error loss as described in Equation (3). The alternative is repeated until a series of RTSTs are learnt, which will give a sufficient level of accuracy. The RTSTs constitute our final CS-OFDM. In the next part, we will introduce that how to learn the ith RTST.

The regression tree based sparse transform is trained using a standard regression random forest [30] with the aforementioned dataset S_j in stages, where each stage corresponds to a non-leaf vertex in the M. Starting with the root vertex, $i = 0$, of M we grow the tree with the objective of separating the samples in S_j into two cohesive sub-regions, $S(j,l)$ and $S(j,r)$, which correspond to the vertices $l(i)$ and $r(i)$, which are the children of the root node. The process is then repeated recursively on each split of the data, $S(j,l)$ and $S(j,r)$, until the information gain falls below a threshold. Once the RTST is constructed, for any leaf node, k, stores a 2D offset vector, C_k, that averages the samples, L_k, in the leaf, namely:

$$C_k = P_k^T * L_k \tag{4}$$

where, n_k is the number of samples in the kth leaf node, then P_k is, $[\frac{1}{n_k}, \frac{1}{n_k}, \cdot, \frac{1}{n_k}]^T \in R^{n_k}$.

This separation is achieved by using the pixel-difference appearance feature (PDAF). At each node, we randomly generate splitting candidates, $\Phi = \{(f_i, \tau_i)\}$, consisting of a PDF function,f_i and threshold, τ_i, and greedily pick the candidate (f_j, τ_j) from Φ that gives rise to maximum variance reduction.

For any two face images, I_i, I_j, with any estimated shapes, $\hat{X}_i$, $\hat{X}_j$, we would like to index the same positions of pixels in I_i relative to its shape, $\hat{X}_i$, as the positions in I_j to its shape, $\hat{X}_j$. To achieve this, each image can be warped to the mean shape based on the current estimated shape before extracting the features [17]. Otherwise, a very sparse representation of the image is much more efficient to warp the location of

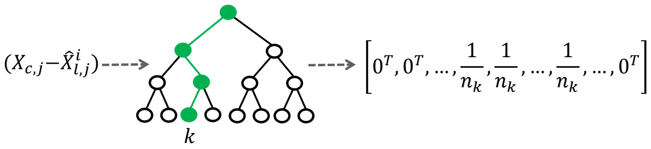
$(X_{c,j}-\hat{X}^i_{l,j})$
$\left[0^T,0^T,\dots,\frac{1}{n_k},\frac{1}{n_k},\dots,\frac{1}{n_k},\dots,0^T\right]$
k

Fig. 2: Sparse coding. The regression tree based sparse transform encodes the shape residual into a sparse representation.

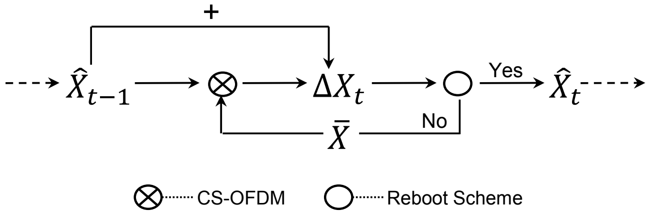

Fig. 3: Real-time face alignment utilizing the trained CS-OFDM. CS-OFDM uses last face shape $\hat{X}_{t-1}$ to estimate the shape residual ΔX_t between $\hat{X}_{t-1}$ and the current face shape $\hat{X}_t$ to be fitted. Then $(\hat{X}_{t-1}+\hat{X}_{t-1})$ is obtained to fit $\hat{X}_t$. if the fitting result is accepted by the reboot scheme, the real-time face alignment is going on. Otherwise, reboot is performed from the mean shape $\bar{X}$.

points as opposed to the whole image as suggested by [14], [17].

Once the RTST is learnt, sparse representation can be performed on each facial shape residual. Specifically, sparse representation, $D(I_j,\hat{X}^i_{l,j})$, of each image I_j with the last facial shape $\hat{X}^i_{l,j}$ can be solved based on the leaf, k, that shape residual reaches, i.e., the leaf at the end of the path in the tree the shape residual will follow, and for each node μ along this path we say that arrives at μ. Figure 2 illustrates the process of performing the sparse coding, and the sparse representation can be formulated as:

$$D(I_j,\hat{X}^i_{l,j}) = [0^T,0^T,\cdot,P_k,\cdot,0^T]^T \qquad (5)$$

where, we presume that the last frame shape reaches the leaf, k. Then, the estimation, $W\{D(I_j,\hat{X}^i_{l,j})\}$, of the facial shape residual $(X_{c,j}-\hat{X}^i_{l,j})$ will be:

$$W\{D(I_j,\hat{X}^i_{l,j})\} = [0^T,0^T,\cdot,P_k,\cdot,0^T]^T * [A_0,A_1,\cdot,A_k,\cdot,A_{2^n-1}] \qquad (6)$$

D. *Fitting Facial Shape Utilizing Trained CS-OFDM*

Our real-time fitting utilizing trained CS-OFDM is performed as illustrated in Figure 3. Because our approach fits current frame shape by recovering the shape residuals with last frame shape, the fitting of the first frame will be performed utilizing the mean shape of the annotated shapes. During acquisition, face image can be captured by camera in a real time manner. Then the facial shape of each current frame is fitted by the recovery to the shape residuals utilizing the trained CS-OFDM. To alleviate model drifting, we use the result of the face detection to judge the fitting result. If drift occurs, reboot is performed from mean shape.

III. Experiments

In this section we present qualitative and quantitative results on publicly available databases for the novel real-time face alignment approach presented in this paper, which serve to demonstrate the accurancy, the smoothness and the real-time performance of the proposed approach. We first presents the implementation details about training and experimental settings used in our experiments. Then we describe the publicly video datasets and evaluation protocols. We compare the proposed approach with the existing state-of-the-art tracking method on the discussed datasets.

A. *Implementation Details*

During the training process, we use image dataset to train our offline model. To simulate the video situation, Each image is regarded as current frame, and the last facial shapes are synthesized according to each truly shape corresponding to the current frame. Because our offline model recoveries current frame shape by reconstructing the shape residuals with the previous one. The capability that our offline model senses the various previous priors can ensure a reliable recovery. Therefore, we sample multiple last facial shapes using a Gaussian random pattern around the current truth shape. In our experiment, we generate 20 candidatets for each last facial shape. Once the training set is constructed, we train 5000 regression trees by evaluating 500 splitting candidates at each node and the shrinkage factor used to combat over-fitting is set to 0.01. The depth of the tree is set to 5.

B. *Datasets*

We use data from different datasets of static images to construct our training set which consists of: Helen [31], LFPW [32]. All training images are annotated with a 68-point configuration defined in 300-W challenge [33]. We compare our approach with increment approaches, Chehra tracker, on the 300VW [20]. The face videos from the dataset are categorized into Scenario 1, Scenario 2 and Scenario 3. Category 1 contains 31 videos recorded in controlled conditions, namely in laboratory and naturalistic well-lit conditions, whereas Category 2 includes 19 videos recorded under severe changes in illumination, namely in real-world human-computer interaction applications. Category 3 contains 14 videos captured in totally unconstrained scenarios, namely arbitrary conditions with challenge evaluation. All the videos are of a single-person, and have been annotated.

Fitting evaluation in all experiments is performed utilizing normalized Root Mean Square Error (RMSE), is computed for each frame by dividing the average point-to-point Euclidean error by the distance between two eye centers.

Our approaches are implemented in C++. All experiments are performed on an Intel 7 core 4 GHz CPU with 16 GB RAM. During detection, we use the face detector of OpenCV to detect the bounding box. This face detector is realized by Viola and Jones algorithm which uses Haar features and a cascade of classifiers.

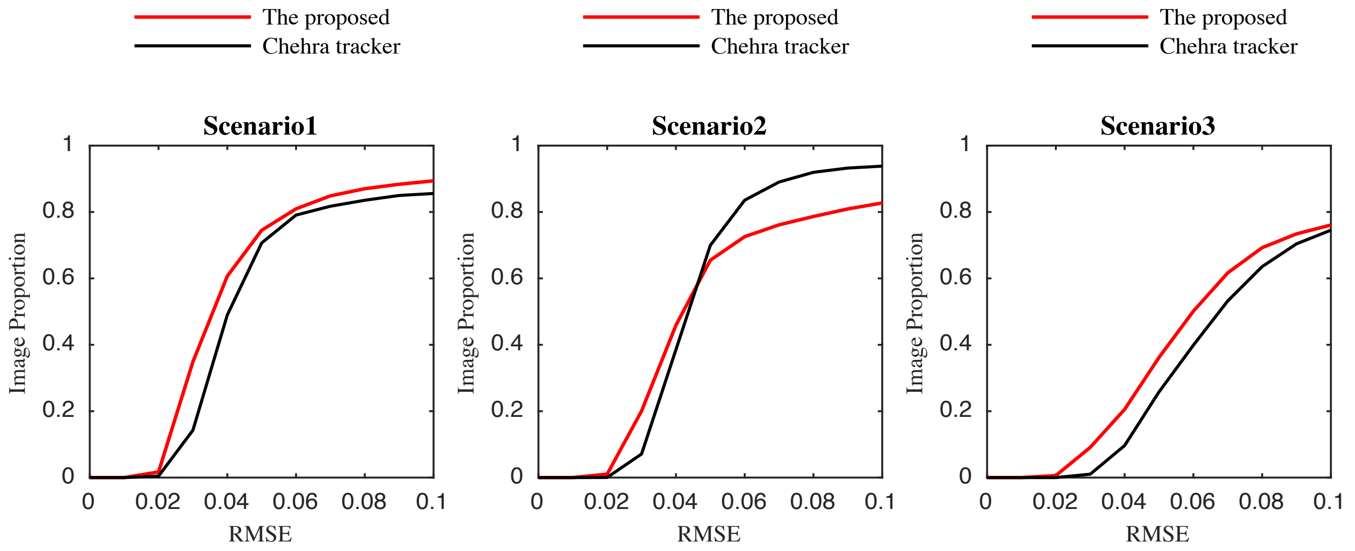

Fig. 4: Cumulative error curve of the proposed method (red) and the Chehra tracker (black) on 300-VW dataset (49 landmarks). The overview of delivery system.

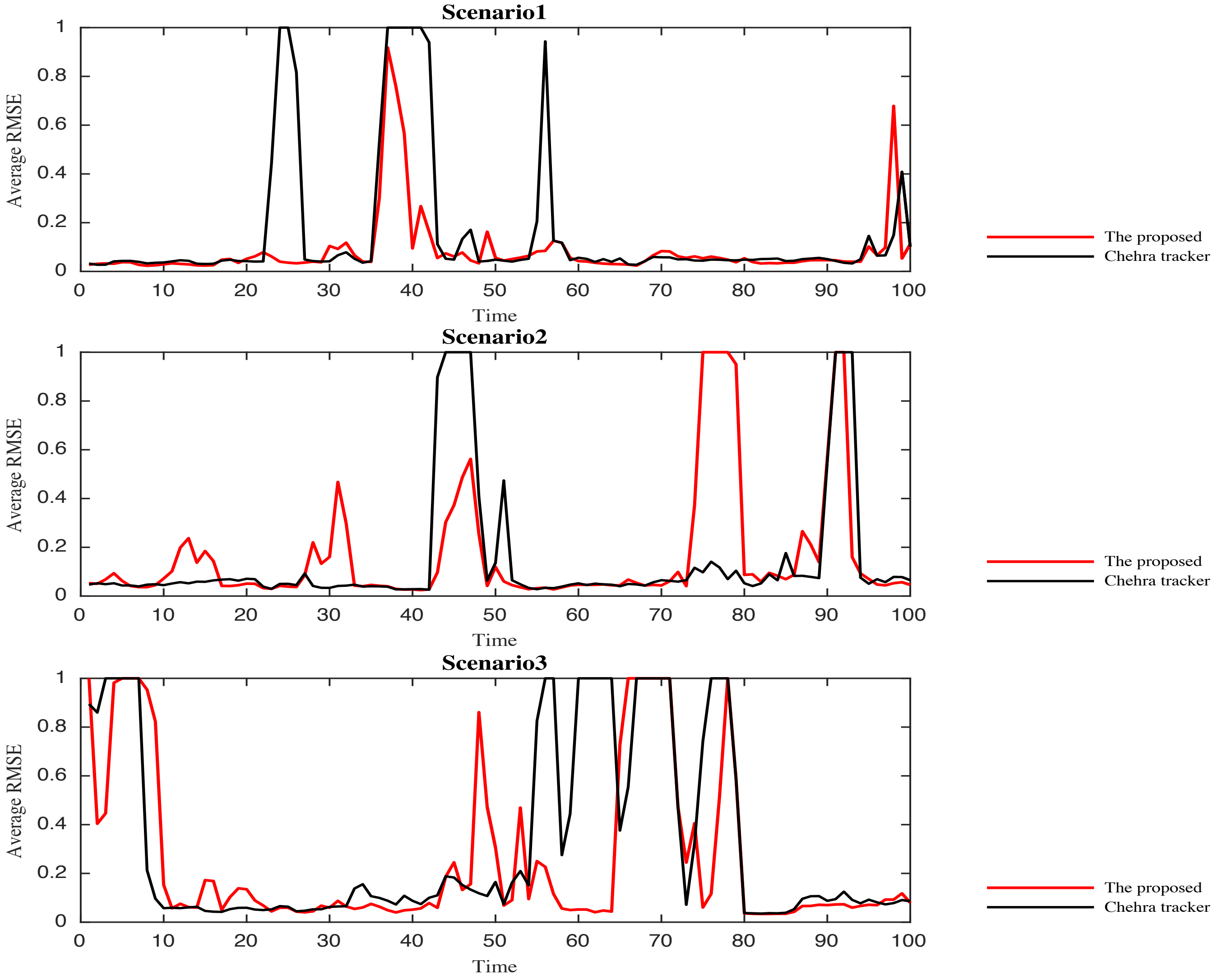

Fig. 5: The frame-by-frame average RMSE plot for proposed method (red) and the Chehra tracker (black) on 300-VW dataset(49 landmarks).

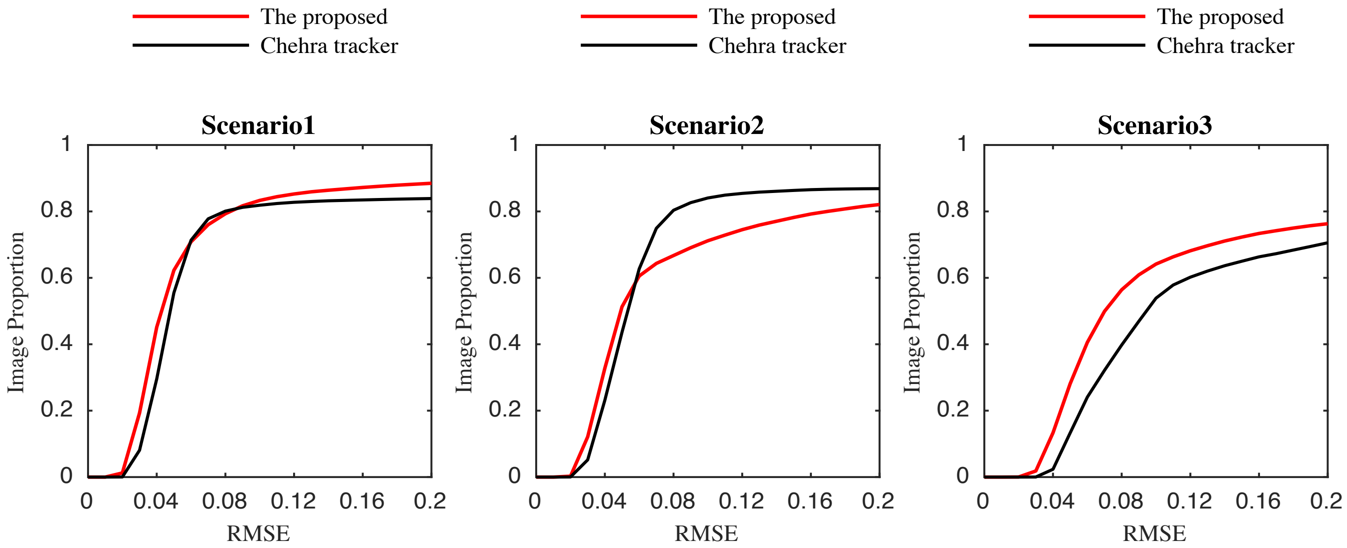

Fig. 6: Cumulative error curve of the proposed method (red) and the Chehra tracker (black) on image scale changed 300-VW dataset (49 landmarks). The overview of delivery system.

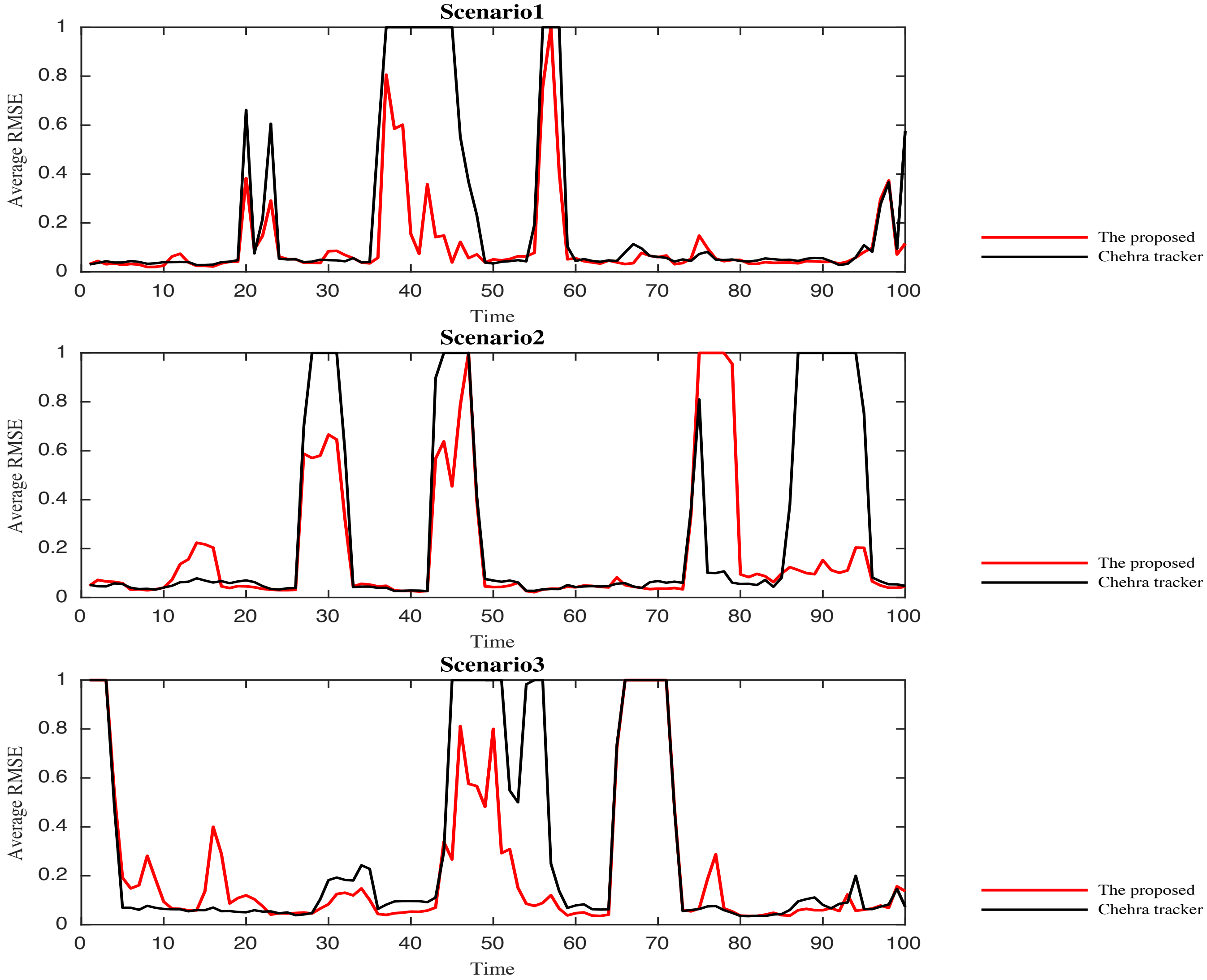

Fig. 7: The frame-by-frame average RMSE plot for proposed method (red) and the Chehra tracker (black) on image scale changed 300-VW dataset(49 landmarks).

C. Comparision

The fitting results of our proposed method and Chehra tracker on the 3 scenarios are reported using the RMSE evaluation metric. Comparison between the proposed method and participant is conducted by two types of curves. The first one is Cumulative Error Distribution curve (CED) , as shown in Figures 4 and 6, which examines the ability of our method to track face images in controlled and unconstrained scenarios. The second one plots average RMSE as function of time, as shown in Figures 5 and 7, which reflects the smoothness of tracking results from the two participants. In particular, we randomly sample a faction of frames, one hundred of frames in this paper, in all videos of each scenario. Then, we average all the RMSE at the same time. At last, we analyze the real-time performance.

1) Evaluation on 3 scenarios: Comparison is performed on three described scenarios, as shown in Figure 4. Compared to the participant, incremental method, the performance of the proposed method is much better on two of the three test sets, namely scenario 1 and scenario 2, as shown in Figure 4(a and c). which indicates that the proposed approach is robust under controlled conditions and arbitrary conditions. However, our approach does not show a better performance, as shown in Figure 4(b). One possible reason of this problem is that videos in scenario 2 are recorded under severe changes in illumination and the image feature, difference of two intensities, used by the proposed approach is not robust to the severe changes in illumination. Figure 5 shows three average RMSE curves on all scenarios. It can be seen that the proposed approach has a smaller average RMSE on scenario 1 and 2 at nearly the beginning and end of tracking, and shows a more smooth fitting results. Otherwise, the participant method exhibits more severe phenomenon. As discussed, our proposed approach does not provide a better fitting under the condition with severe changes in illumination. In summary, Our proposed approach provides a robust performance under many conditions expect the illumination.

2) Robust evaluation: In order to evaluate the comparison under severe changes in the size of image, each image in all scenarios is normalize into the size of 640×480, in which 640 represents the width of a image, and 480 is the height. From the other viewpoint, it also simulates the smart-phone environment. We achieve all the ground truth shapes by applying Procrustes Analysis to the original ground truth ones. Figures 6 and 7 show nearly same results as that in Figures 4 and 5. Moreover, our proposed method provide more robust results in scenario 2. Therefore, our proposed approach is robust to the changes in the size of image, and is more suitable to the smart-phone.

3) Analysis of real-time performance: The real-time performance of our proposed approach is similar to that of regression tree based static face alignment approaches for there similarity in efficiency of alignment to face shape. At the same time, the regression tree based static face alignment approaches are certificated to be of super real-time performance. Therefore, our approach can provide a better real-time performance than the existing state-of-the-art incremental learning based ones and achieves over 166 fps on a desktop.

IV. Conclusion

In this paper, we investigate the problem of smooth and accurate landmark-detection for real-time face alignment. We propose a novel compressive sensing based offline FDM for smoothly detect the location of the facial landmarks by recovering the shape residuals between consequent frames. Our trained offline model is composed of separate discriminative tree based sparse transforms, which are directly learned from samples of the facial shape residuals between the ground truth facial shape of the current frame and simulated facial shapes of last frame, which are randomly sampled around the current facial shape. Furthermore, we propose a strategy to oppose the drift during tracking. This allows our approach to keep a long period of tracking. We conduct experiments on the 300-VW challenge datasets. The results clearly demonstrate that our approach provides significant improvement over the baseline on the smooth performance. We also analysis the real-time performance of our approach. However, the intensity difference based image feature is not robust to the illumination changes. Therefore, We plan to investigate efficient feature fusion strategies to combine intensity and color information. Another research direction is to investigate more accurate sparse transform to produce better sparse approximation.

References

[1] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, “Toward a practical face recognition system: Robust alignment and illumination by sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.

[2] J. Roth, Y. Tong, and X. Liu, “Unconstrained 3d face reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2606–2615, 2015.

[3] J. Roth, Y. Tong, and X. Liu, “Adaptive 3d face reconstruction from unconstrained photo collections,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4197–4206, 2016.

[4] A. Jourabloo, X. Yin, and X. Liu, “Attribute preserved face de-identification,” in *Biometrics (ICB), 2015 International Conference on*, pp. 278–285, IEEE, 2015.

[5] Y. Guo, G. Zhao, and M. Pietikäinen, “Dynamic facial expression recognition with atlas construction and sparse representation,” *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 1977–1992, 2016.

[6] Q. Hu, X. Peng, P. Yang, F. Yang, and D. N. Metaxas, “Robust multi-pose facial expression recognition,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 1782–1787, IEEE, 2014.

[7] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, and D. Metaxas, “The best of both worlds: Combining 3d deformable models with active shape models,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–7, IEEE, 2007.

[8] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

[9] C. TECootes and A. Lanitis, “Active shape models: Evaluation of a multi-resolution method for improving image search,” in *Proc. British Machine Vision Conference*, pp. 327–338, 1994.

[10] G. J. Edwards, C. J. Taylor, and T. F. Cootes, “Interpreting face images using active appearance models,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 300–305, IEEE, 1998.

[11] I. Matthews and S. Baker, “Active appearance models revisited,” *International journal of computer vision*, vol. 60, no. 2, pp. 135–164, 2004.

[12] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, “Facial point detection using boosted regression and graph models,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2729–2736, IEEE, 2010.

[13] P. Dollár, P. Welinder, and P. Perona, “Cascaded pose regression,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1078–1085, IEEE, 2010.

[14] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[15] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539, 2013.

[16] G. Tzimiropoulos, “Project-out cascaded regression with an application to face alignment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3659–3667, 2015.

[17] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[18] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, “Robust face landmark estimation under occlusion,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1513–1520, 2013.

[19] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

[20] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, “The first facial landmark tracking in-the-wild challenge: Benchmark and results,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 50–58, 2015.

[21] X. Peng, J. Huang, and D. N. Metaxas, “Sequential face alignment via person-specific modeling in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 107–116, 2016.

[22] D. Decarlo and D. Metaxas, “Optical flow constraints on deformable models with applications to face tracking,” *International Journal of Computer Vision*, vol. 38, no. 2, pp. 99–127, 2000.

[23] Y. Wettum, “Facial landmark tracking on a mobile device,” B.S. thesis, University of Twente, 2017.

[24] E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar, “Cascaded continuous regression for real-time incremental face tracking,” in *European Conference on Computer Vision*, pp. 645–661, Springer, 2016.

[25] J. Sung and D. Kim, “Adaptive active appearance model with incremental learning,” *Pattern recognition letters*, vol. 30, no. 4, pp. 359–367, 2009.

[26] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, “Incremental face alignment in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1859–1866, 2014.

[27] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas, “Piefa: Personalized incremental and ensemble face alignment,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3880–3888, 2015.

[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[29] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, “3d facial landmark detection under large yaw and expression variations,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1552–1564, 2013.

[30] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, “Interactive facial feature localization,” in *European Conference on Computer Vision*, pp. 679–692, Springer, 2012.

[32] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, “Localizing parts of faces using a consensus of exemplars,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[33] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “A semi-automatic methodology for facial landmark annotation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 896–903, 2013.