Facial shape tracking via spatio-temporal cascade shape regression

Jing Yang
nuist_yj@126.com

Jiankang Deng
jiankangdeng@gmail.com

Kaihua Zhang
zhkhua@gmail.com

Qingshan Liu
qsliu@nuist.edu.cn

Nanjing University of Information Science and Technology
Nanjing, China

Abstract

In this paper, we develop a spatio-temporal cascade shape regression (STCSR) model for robust facial shape tracking. It is different from previous works in three aspects. Firstly, a multi-view cascade shape regression (M-CSR) model is employed to decrease the shape variance in shape regression model construction, which is able to make the learned regression model more robust to shape variances. Secondly, a time series regression (TSR) model is explored to enhance the temporal consecutiveness between adjacent frames. Finally, a novel re-initialization mechanism is adopted to effectively and accurately locate the face when it is misaligned or lost. Extensive experiments on the 300 Videos in the Wild (300-VW) demonstrate the superior performance of our algorithm.

1. Introduction

Face alignment is among the most popular and well-studied problem in the field of computer vision with a wide range of applications, such as facial attribute analysis [20], face verification [17], [28], and face recognition [31], [38], to name a few. In the past two decades, a lot of algorithms have been proposed [6], which can be roughly categorized as either generative or discriminative methods.

Generative methods typically optimize the shape parameters iteratively with the purpose of best approximately reconstructing input image by a facial deformable model. Active Shape Models (ASMs) [10] and Active Appearance Models (AAMs) [13], [9], [21] are typical representative subject to this category. In the ASMs, a global shape is constructed by applying the Principal Component Analysis (PCA) method to the aligned training shapes, and then the appearance is modeled locally via discriminatively learned templates. In the AAMs, the shape model has the same point distribution as that is in the ASMs, while the global appearance is modeled by PCA after removing shape variation in canonical coordinate frame. Discriminative methods attempt to infer a face shape through a discriminative regression function by directly mapping textual features to shape. In [12], a cascaded regression method built on pose-index feature has been proposed to pose estimation with excellent performance. Cao et al. [5] combine two-level boosted regression, shape indexed features and a correlation-based feature selection method to make the regression more effective and efficient. Xiong et al. [32] concatenate SIFT features of each landmark as the feature and obtain regression matrix via linear regression. In[29], a learning strategy is devised for a cascaded regression approach by considering the structure of the problem.

Although these methods have achieved much success for facial landmark localization, it remains an unsolved problem when applied to facial shape tracking in the real world video due to the challenging factors such as expression, illumination, occlusion, pose, image quality and so on. A successful facial shape tracking includes at least two characteristics. On the one hand, face alignment on images is supposed to perform well. On the other hand, face relationship between the consecutive frames should provide a solid transition. A typical work linking to face relationship between the consecutive frames is multi-view face tracking [8]. [11] demonstrates that a small number of view-based statistical models of appearance can represent the face from a wide range of viewing angles, in which constructed model is suitable to estimate head orientation and to track faces through wide angle changes. In [23], S. Romdhani et al. adopt a nonlinear PCA, i.e., the Kernel PCA [26], which is based on Support Vector Machines [30] for nonlinear model transformation to track profile-to-profile faces. In [14], an online linear predictor tracker without need for offline learning has been introduced for fast simultaneous modeling and tracking. [2] proposes an incremental parallel cascade linear regression (iPar-CLR) method for face shape tracking, which automatically tailor itself to the tracked face and become person-specific over time. [34] proposes a Global Supervised Descent Method (GSDM), an extension of SDM [32] by dividing the search space into regions of similar gradient directions.

In this paper, we construct a spatio-temporal cascade shape regression model for robust facial shape tracking, which aims at transferring spatial domain alignment into time-sequence alignment. A multi-view regression model is employed into robust face alignment, which greatly decreases the shape variance from face pose, thereby making the learned regression model more robust to shape variances. Futhermore, a time series regression model is explored to face alignment between the consecutive frames, thereby enhancing the temporal consecutiveness between alignment result in the former frame and initialization in the latter. In addition, a novel re-initialization mechanism is adopted to effectively and accurately locate the face when the face is misaligned or lost.

In summary, the main contributions are summarized as follows: (1) We improve the cascade shape regression model by constructing a multi-view cascade shape regression, making the learned regression model more view-specific, and better for generalization and robustness. (2) Our spatio-temporal cascade shape regression model is fully automatic and achieves fast speed for online facial shape tracking even on a CPU. (3) Extensive experiments on the 300 Videos in the Wild (300-VW) demonstrate the superior performance of our algorithm.

2. The proposed method

2.1. Overview

Figure 1 illustrates the proposed spatio-temporal cascade shape regression (STCSR) model for robust face shape tracking.

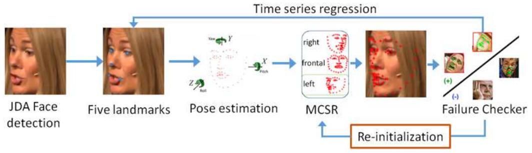


Figure 1. Overview of STCSR. MCSR denotes multi-view cascade shape regression. Re-initialization will be discussed in Section 2.4

In the first frame, the JDA [7] (Joint detection and alignment) face detector is utilized to initialize the system. Similarity transformation parameters (rotation, translation, and scale) are estimated from the five landmarks and the face view (left, front, and right) is also predicted by those five landmarks. Then a multi-view cascade shape regression is employed to predict face shape in the current frame, which will be discussed in section 2.2. When the score of the alignment result is larger than threshold, time series regression is performed for facial shape tracking, which will be discussed in section 2.3. When the score of the alignment result is smaller than a threshold, a re-initialization mechanism is adopted to avoid false convergence during facial shape tracking, which will be discussed in section 2.4.

Shape initialization from the JDA face detector and the alignment result of the previous frame are under a unified framework. On images, JDA is able to provide five facial landmarks to estimate face pose on images. Meanwhile, we assume that the face shape will not change abruptly between the consecutive frames on videos. So the parameters of similarity transformation and the yaw angle of the t-th shape are able to initialize the shape of the $t+1$-th frame. Based on the face pose, the algorithm selects the view-specific model and transforms the view-specific mean shape with similarity transformation parameters.

2.2. Multi-view cascade shape regression

The main idea of the cascade shape regression model is to combine a sequence of regressors in an additive manner in order to approximate an intricate nonlinear mapping between the initial shape and the ground truth. Specifically, Given a set of N images $\{I_i\}_{i=1}^{N}$ and their corresponding ground truth $\{X_i^*\}_{i=1}^{N}$. A linear cascade shape regression model [32] is formulated as:

$$\arg\min_{W^t}\sum_{i=1}^{N}\sum_{j}\left\|(X_i^*-X_{ij}^{t-1})-W^t\phi(I_i,X_{ij}^{t-1})\right\|^2, \tag{1}$$

where W^t is the linear regression matrix, which maps the shape-indexed features to the shape update. X_{ij}^{t-1} stands for the intermediate shape of image I_i, $t=1,\cdots,T$ is the iteration number, Φ is the shape-index feature descriptor, and j counts the perturbations. Usually, training data is augmented with multiple initializations for one image, which serves as an effective method for improving the generation capability of training. Inspired by the subspace regression [34] that splits the search space into regions of similar gradient directions and obtains better and more efficient convergence. We decrease shape variation by dividing the training data into three views (right, frontal, and left), then specific-view model is trained within each dataset. We estimate the face view with five landmarks (left eye center, right eye center, nose tip, left corner of mouth, right corner of mouth).

As shown in Figure 2, five facial landmarks indicate the face layout, so we use the locations of five landmarks to estimate the view status by

$$\arg\min_{R}\sum_{i=1}^{N}\|V_i-RP_i\|_2^2, \tag{2}$$

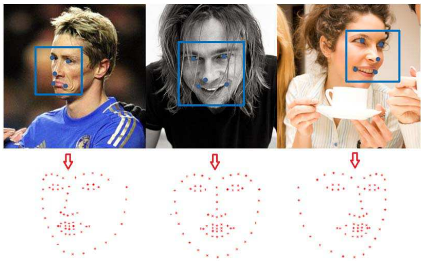


Figure 2. Illustrations of view specific shape initialization.

where V_i is the view status. $P_i \in \mathbb{R}^{10\times1}$ is the locations of the five facial landmarks. R is the regression matrix, which can be solved by least square method. In the experiments, we only categorize the face views into the frontal ($-15^\circ - 15^\circ$), left ($-30^\circ - 0^\circ$), and right ($0^\circ - 30^\circ$) views, which cover all of the face poses from the 300-W training dataset[1]. The overlaps between the frontal view and the profile views are used to make view estimation more robust.

The shape variance of each view set is much smaller than that of the whole set, and the mean shape of each view is much closer to the expected result, so the view-specific shape model is not only able to decrease the shape variance, but also it can accelerate the shape convergence.

2.3. Time series regression

Performing face detection on each frame for face alignment is time-consuming. Futhermore it tends to decrease the alignment accuracy on videos, because the initial mean shape is far from the ground truth shape under large face pose variation. So establishing a correlationship between the consecutive frames is of great importance. In this section, we propose three methods (box tracking, landmark tracking, and pose tracking) to link the consecutive frames.

Figure 3 shows the workflow of box tracking. In this method, we build a tracker based on face appearance model. Face location (x, y, w, h) at the current frame is estimated based on the tracker. Then a CSR is performed to predict the landmark locations from the mean shape based on the shape indexed features. This procedure is repeated until the last frame comes. The whole procedure combines the previous frame and the current frame with the face appearance information, and overlooks the relationship between two consecutive frames' landmarks. It is obvious that such a method is extremely time-consuming. Even worse, longtime tracking will cause tracking drift due to tremendous variation in the object appearance caused by illumination changes, partial occlusion, deformation and so on.

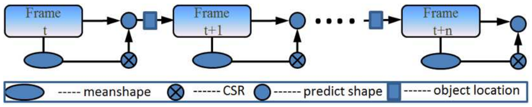


Figure 3. Box tracking. A visual tracker is employed to predict face location at present. Initial shape is the mean shape.

Figure 4 shows the workflow of landmark tracking. In this method, we deliver shape in previous frame directly to current frame as initial shape. Then MCSR is performed to predict the landmark locations from the alignment result of previous frame. For training CSR method in image datasets, the initial set of perturbations (ΔX) are obtained by Monte-Carlo sampling procedure [32], in that perturbations are randomly drawn within a *fixed pre-defined range* around the groundtruth shape X^*. Direct shape deliver approach cannot guarantee residual between previous shape and current shape within perturbation and might not converge to final shape due to cumulative error on videos.

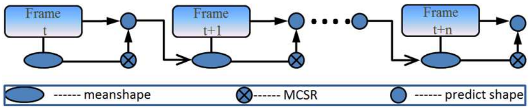


Figure 4. Landmark tracking. Shape in the previous frame is delivered directly to current frame as initial shape. Initial shape is previous shape.

Figure 5 shows the workflow of pose tracking. In this method, we deliver shape similarity transform parameters of previous frame to the current one. Parameters of face rigid changes from the previous shape is employed to adjust the mean shape, and the adjusted mean shape is taken as initial shape in current frame. MCSR is performed to predict the landmark locations from the transformed view-specific mean shape. Compared to landmark tracking, the noise of the initial shape from the previous frame is smoothed by pose tracking, thus making the facial shape tracking more stable.

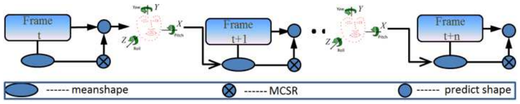


Figure 5. Pose tracking. Similarity transform parameters of the previous frame are delivered to the current frame. Initial shape is calculated based on the above information.

2.4. Re-initialization

As has been discussed above, MCSR is exploited to predict landmark location on each frame, while time series re-

[1] Face pose of each training image is estimated by http://www.humansensing.cs.cmu.edu/intraface/download.html.

Algorithm 1 Facial shape tracking via spatio-temporal cascade shape regression

```
Require: the $i$-th image frame in face video
if $i == 1$ then
    detect face location at the current frame $(x_t, y_t, w_t, h_t)$
    predict face shape $\widehat{X_i}$ via MCSR
else
    if $Score(i) > 0.7$ then
        Pose tracking is employed to predict the face shape.
    else
        detect face location at current frame
        if non face is detected then
            Adaptive compressive tracker is used to predict face location $(x_t, y_t, w_t, h_t)$.
            predict face shape $\widehat{X_i}$ via MCSR
        else
            predict face shape $\widehat{X_i}$ via MCSR
        end if
    end if
end if
Ensure: face shape $\widehat{X_i}$ at $i$-th image frame.
```

gression is employed to create a link between the consecutive frames. Both steps work when previous alignment is reliable to predict the current frame. If the previous alignment tends to drift, which will lead to face misaligned or lost, a novel re-initialization mechanism is adopted to effectively and accurately locate the face. In this work, we introduce the fitting score, which corresponds to the goodness of alignment. When fitting score is lower than the setted threshold (0.7), shape re-initialization is performed. For this purpose, we train an SVM classifier to differentiate between the aligned and misaligned images based on the last shape indexed features. We generate the positive samples from annotations and then randomly generate samples around the ground truth to generate the negative samples. The score from the trained SVM is used as the criteria to judge the goodness of alignment. In our experiments, confidence of face alignment above 0.7 is seen as a successful landmark location. Given a face video, if fitting score from the previous frame alignment is below 0.7, face detector embarks on face detection at the current frame. If non face is detected, adaptive compressive tracker [19] starts to locate the face with the appearance model built on the face appearance once alignment confidence is below 0.7.

The main steps of our facial shape tracking are summarized in Algorithm 1.

3. Experiments

We test our algorithm on two scenarios. One is face alignment on images, which is initialized with the output of a face detector. Another is face alignment on videos, which is initialized by the alignment result of the previous frame.

3.1. Experimental Data

Image datasets. A number of face image datasets [3, 18, 37] with different facial expression, pose, illumination and occlusion variations have been collected for evaluating face alignment algorithms. In [24], AFW [37], LFPW [3], and HELEN [18] are re-annotated[2] by the well established landmark configuration of Multi-PIE [16] using the semi-supervised methodology [25]. A new wild dataset called IBUG is also created by [24], which covers different variations like unseen subjects, pose, expression, illumination, background, occlusion, and image quality. IBUG aims to examine the ability of face alignment methods to handle naturalistic, unconstrained face images. In this paper, AFW, LFPW, HELEN and IBUG are used to train the multi-view cascade shape regression model.

Video datasets. Even though comprehensive benchmarks exist for localizing facial landmark in static images, very limited effort has been made towards benchmarking facial landmark tracking in videos [27]. 300-VW (300 Videos in the Wild) has collected a large number of long facial videos recorded in the wild. Each video has duration of about 1 minute (at 25-30 fps). All frames have been annotated with regards to the well-established landmark configuration of Multi-PIE [16]. 50 videos[3] are provided for validation, and 150 facial videos are selected for test. This dataset aims at testing the ability of current systems for fitting unseen subjects, independently of variations in pose, expression, illumination, background, occlusion, and image quality. There are three subsets for test with different difficulty:

Scenario 1: This scenario aims to evaluate algorithms that are suitable for facial motion analysis in laboratory and naturalistic well-lit conditions. There are 50 tested videos of people recorded in well-lit conditions displaying arbitrary expressions in various head poses but without large occlusions.

Scenario 2: This scenario aims to evaluate algorithms that are suitable for facial motion analysis in real-world human-computer interaction applications. There are 50 tested videos of people recorded in unconstrained conditions displaying arbitrary expressions in various head poses but without large occlusions.

Scenario 3: This scenario aims to assess the performance of facial landmark tracking in arbitrary conditions.

[2] http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/

[3] 300VW_Clips_2015_07_26.zip

There are 50 tested videos of people recorded in completely unconstrained conditions including the illumination conditions, occlusions, make-up, expression, head pose, etc.

3.2. Experimental setting

Data augmentation. Data augmentation serves as an effective method for improving the generation of training. We flip all of the training data and augment them with ten initializations for each image. We first get mean shape $\bar{X}$ from all ground truth shapes by Procrustes Analysis [15], then we train a linear regression to remove the translation and scale difference between the initial mean shape and the ground truth shape by the location of the face rectangle. Finally, the residual distribution between the initial mean shape and the ground truth shape is utilised to generate other initial shapes of identical distribution. Actually, the expectation of all of those initial shapes are the mean shape.

Shape initialization. Generally, the normalized mean shape is used as the initial shape during face alignment on images. The scale and the translation parameters of the initial shape are estimated from the output face rectangle of a face detector. The stability of the face detector is of great importance, because the drift from a face detector has more or less effect on the following face alignment. On videos, the initialization shape is generated from the alignment result of the previous frame, which makes face alignment more accurate due to the more accurate translation, scale, and face pose (yaw, pitch, roll) information inherited from the previous frame. However, in this paper we unify face alignment on images and videos by the proposed TSR model. Shape initialization is always from the five facial landmarks, which are utilized to remove rotation, translation and scale differences and select the view-specific models. The only difference is that the five facial landmarks are generated from JDA face detector on images and the previous alignment result on videos. We compare these different shape initialization methods and report the alignment result on IBUG dataset.

Regularization. To avoid overfitting, an additional L2 penalty term is added to the original least square objective function to regularize the linear projection. The regularization parameter is set as the number of the training example according to our experiment.

Evaluation metric. Fitting performance is usually assessed by the normalised mean error. In particular, the average Euclidean point-to-point error normalized distance is used. The error is calculated over (a) all landmarks, and (b) the facial feature landmarks (eyebrows, eyes, nose, and mouth).

The normalized mean error over all landmarks,

$$E_i = \frac{\frac{1}{M}\sum_{j=1}^{M} |p_{i,j} - g_{i,j}|_2}{|l_i - r_i|_2}, \tag{3}$$

where *M* is the number of landmarks, *p* is the prediction, *g* is the ground truth, *l* and *r* are the positions of the left eye corner and right eye corner.

The distance between eye corners is used to normalize the error as in [24]. The allowed error (localization threshold) is taken as some percentage of the inter-ocular distance (IOD), typically $\leq 10\%$. The normalization is able to make the performance measure independent of the actual face size or the camera zoom factor. Following the evaluation criteria of the 300-VW challenge, we use the cumulative error curve of the percentage of images against NME to evaluate the algorithms.

3.3. Evaluation of MCSR

We first investigate the influence of shape initialization on the face alignment. We compare the JDA face detector with the 300-W face detector[4] and the OpenCV face detector[5] on the IBUG dataset. For the JDA face detector and the OpenCV face detector, we select the rectangles with the largest overlap with the bounding box of the annotated landmarks. Moreover, for the OpenCV face detector, we drop the rectangles, which have smaller overlap than particular thresholds. We adopt two thresholds, 0.5 and 0.7, which are named as OpenCV ov0.5 and OpenCV ov0.7. For the JDA face detector, we give three kinds of alignment result. The first one is the shape initialization from face rectangle, which is named as JDA (box). The second one is the shape initialization from five landmarks of JDA, which helps to remove the rotation, translation and scale difference and is named as JDA (5 landmarks). The last one is the multiview shape initialization from five landmarks of JDA, which further reduces the shape variation from face pose and is named as JDA (5 landmarks, multiview).

We implement the linear cascade shape regression model, using HOG [33] as shape indexed feature, with seven steps of iteration. We train the cascade shape regression models based on these different kinds of face shape initialization, and the alignment results are shown in Table 1. Compared to the OpenCV ov0.5, the OpenCV ov0.7 is able to decrease the NME by 7.2%, which indicates that the drift of face detection generates great influence on the following face alignment and more accurate detection results can greatly improve the alignment accuracy. Compared with the performance of 300-W official detector, the JDA face detector improves alignment results by 14.12%, which indicates that the face rectangles generated from JDA are more semantically stable. Compared to shape initialization from normal face detectors, the face shape initialization generated from JDA five facial landmarks are better because the rotation, translation and scale difference are removed. Fi-

[4] rectangles marked as ”bb_detector” in http://ibug.doc.ic.ac.uk/media/uploads/competitions/bounding_boxes.zip

[5] haarcascade_frontalface_alt.xml

nally, we estimate face pose (yaw, pitch, and roll) from five landmarks, and the view-specific model is trained on each subset according to the yaw angle, which can further improve the alignment accuracy under different face poses and decrease the NME to 4.68%.

We further compare the proposed multi-view cascade shape regression with the other eight state-of-the-art methods reported in [22], including DRMF [1], RCPR [4], ES-R [5], CFAN [35], SDM [33], LBF [22], TCDCN [36], Linkface[6]. Please note that the distance used to normalize the mean error is the distance between eye center instead of eye corner. Table 2 lists the experimental results, and we can see that the proposed method outperforms the other eight methods by a large margin. Figure 6 illustrates some example results of the proposed method on the IBUG dataset. It can be seen that the proposed method is robust under various conditions.

3.4. Evaluation of TSR

We investigate the facial shape tracking on the videos by comparing the proposed three tracking strategies. As is shown in Figure 7, face box tracking is the worst method due to the shape initialization is always from the mean shape, and the alignment accuracy decreases dramatically under large face pose variation. Even worse, long time tracking will cause tracking drift, which will also affect the alignment performance. Shape initialization from the alignment result of the previous frame makes face alignment better due to the more accurate rotation, translation, and scale information inherited from the previous frame. However, cascade shape regression is an open-loop operation, and the expectation of the alignment result is not always the mean shape due to the noises in the shape indexed features, As a result, the cumulative bias tends to make the facial shape tracking not stable on videos. In order to utilize the alignment result of the previous frame and make the initial shape similar that of training data, we just use the similarity transformation parameters and the face pose from the previous frame. Compared to landmark tracking, the noise of the initial shape from the previous frame is smoothed by pose tracking, thus making the facial shape tracking more stable.

Another crucial component in the facial shape tracking scenario is the tracking failure checker. As is shown in Figure 8, we randomly select four videos from the validation set, and plot the normalized mean error as well as the corresponding score for each frame. The results show that our failure checker reasonably links the relationship between normalized mean error and corresponding score. When the failure occurs, the normalized mean error increases, meanwhile, corresponding score decreases.

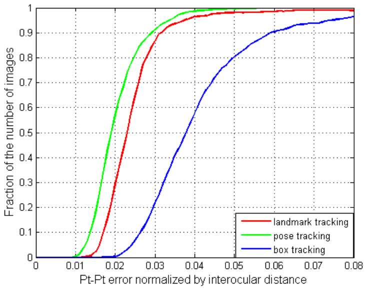


Figure 7. Cumulative error curves of different tracking methods on 300-VW validation set.

3.5. Results on 300-VW

Following the contest rule, the face videos from the 300-VW challenge dataset are categorized into "Scenario 1","S-cenario 2" and "Scenario 3". Figure 9 compares the experimental results (49 landmarks) of the proposed method and the Chehra tracker [2] on three kinds of videos respectively. Compared to the baseline method, incremental face alignment, the performance of the proposed method is much better on all of the three test sets. Besides the cumulative error curves provided by the contest, we also calculate the normalized mean error under certain thresholds in Table 3. For the 49 landmarks, which do not contain the landmarks on face contour, the proposed method achieves the mean error of 3.86% on the 95.91% frames of the hardest test set scenario 3, which indicates that the proposed facial landmark tracking method is robust under arbitrary conditions. On test set scenario 2, the normalized mean error of the proposed method is 3.16% on the 99.38% frames, which shows that the proposed method is quite suitable for facial motion analysis in real-world human-computer interaction applications. Although we obtain high accuracy on test set scenario 1, the returned results show that our face detector have lots of false positives on one single video. As we set the precision of our face detector at 99.8%, it is almost impossible that most of the detection results on one single video are false positive. One possible reason of this problem is that we only detect the largest face in each frame, which is suitable on the validation dataset, but may be not suitable on all of the test sets. We will investigate this problem after the videos are released.

As we can see from Figure 10 and Table 4, the alignment results of 68 landmarks are slightly worse than that of 49 landmarks. The proposed method obtains excellent performances on Scenario 1 and Scenario 2 test sets with the normalised mean error of 3.60% on 95.94% of frames and 3.83% on 98.56% of frames respectively, which indicates that the proposed method works well under laboratory and naturalistic well-lit conditions. However, uncon-

[6] http://www.linkface.cn/index.html

Detectors	OpenCV ov0.5	OpenCV ov0.7	300-W	JDA (box)	JDA (5 landmarks)	JDA (5 landmarks, multiview)
NME	8.50%	7.89%	7.72%	6.63%	5.59%	4.68%

Table 1. Face alignment results from different shape initializations

Algorithm	DRMF	RCPR	ESR	CFAN	SDM	LBF	TCDCN	Linkface	MCSR
NME68	19.79%	17.26%	17.00%	16.78%	15.40%	11.98%	9.15%	8.60%	6.74%

Table 2. **Eye center distance** normalized mean error on IBUG dataset

Dataset	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$	$\leq 20\%$
Scenario 1	$2.76\%(88.93\%)$	$3.06\%(96.66\%)$	$3.10\%(97.07\%)$	$3.12\%(97.24\%)$
Scenario 2	$2.83\%(89.90\%)$	$3.16\%(99.38\%)$	$3.19\%(99.82\%)$	$3.20\%(99.86\%)$
Scenario 3	$3.31\%(79.18\%)$	$3.86\%(95.91\%)$	$4.04\%(98.11\%)$	$4.11\%(98.60\%)$

Table 3. Normalized mean error of the proposed method under different error thresholds on 300-VW dataset (49 landmarks).

strained conditions such as occlusions, large pose variations still pose great challenges to face alignment on videos. As a result, the alignment results on Scenario 3 test set are slightly worse with the normalised mean error of 4.62% on 93.73% of frames.

4. Conclusion

In this paper, we have constructed a spatio-temporal cascade shape regression (STCSR) model for robust facial shape tracking. Firstly, we have presented a multi-view cascade shape regression model for robust face alignment, which greatly decreases the shape variance in regression model construction, making the learned regression model more robust to shape variances. Secondly, a time series regression model has been explored to face alignment between consecutive frames, thereby enhancing the temporal consecutiveness between alignment result in former frame and initialization in the latter. Finally, in order to increase the efficiency in videos, a novel re-initialization mechanism has been adopted to effectively and accurately predict face location when the face is misaligned or lost. Extensive experiments on the 300-VW dataset demonstrate the superior performance of STCSR.

Acknowledgments

This work is supported by the Natual Science Foudation of China (NSFC) under 61402233, by the Natural Science Foudation of Jiangsu under Grant BK20151529, by the Graduate Education Innovation Project of Jiangsu under Grant KYLX15_0881.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.

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

[3] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition*, pages 545–552. IEEE, 2011.

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, pages 1513–1520. IEEE, 2013.

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. IEEE, 2012.

[6] O. Çeliktutan, S. Ulukaya, and B. Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13, 2013.

[7] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.

[8] G. Chrysos, S. Zafeiriou, E. Antonakos, and P. Snape. Offline deformable face tracking in arbitrary videos. In *IEEE International Conference on Computer Vision Workshops (ICCVW), 2015*. IEEE.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, 2001.

[10] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[11] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and vision computing*, 20(9):657–664, 2002.

[12] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE, 2010.

[13] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. pages 300–305, 1998.

[14] L. Ellis, N. Dowson, J. Matas, and R. Bowden. Linear regression and adaptive appearance models for fast simultaneous

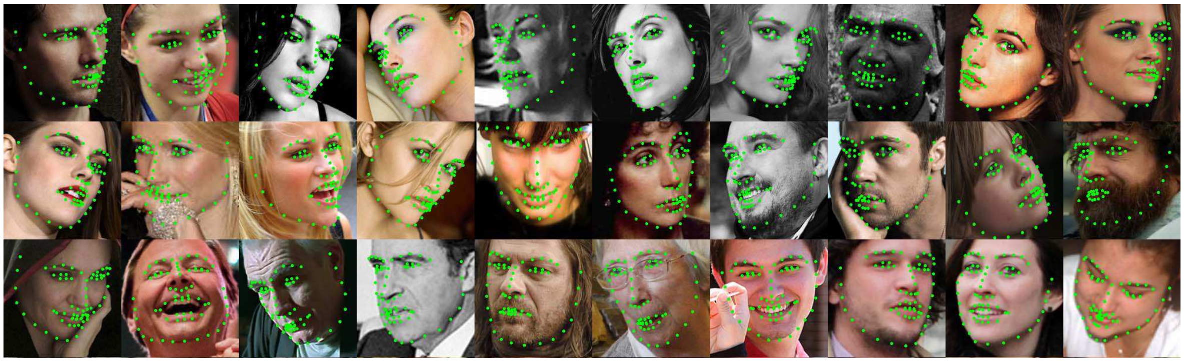
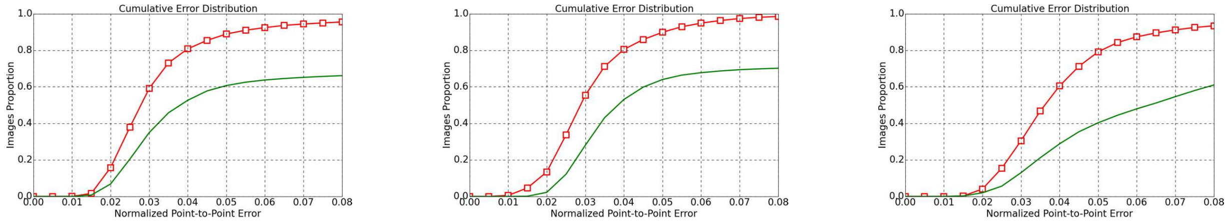
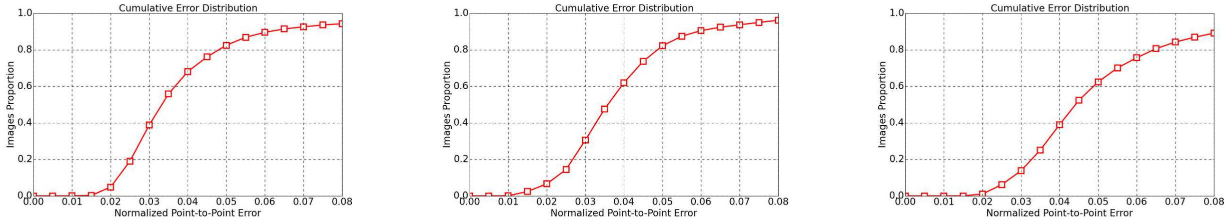


Figure 6. Example alignment results from IBUG.



(a) Cumulative Error Curve of Scenario 1 (b) Cumulative Error Curve of Scenario 2 (c) Cumulative Error Curve of Scenario 3

Figure 9. Cumulative error curve of the proposed method (red) and the Chehra tracker (green) on 300-VW dataset (49 landmarks).



(a) Cumulative Error Curve of Scenario 1 (b) Cumulative Error Curve of Scenario 2 (c) Cumulative Error Curve of Scenario 3

Figure 10. Cumulative error curve of the proposed method (red) on 300-VW dataset (68 landmarks).

Dataset	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$	$\leq 20\%$
Scenario 1	3.16%(82.50%)	3.60%(95.94%)	3.68%(96.93%)	3.71%(97.15%)
Scenario 2	3.31%(82.21%)	3.83%(98.56%)	3.90%(99.59%)	3.94%(99.82%)
Scenario 3	3.65%(62.42%)	4.62%(93.73%)	4.87%(97.17%)	5.00%(98.14%)

Table 4. Normalized mean error of the proposed method under different error thresholds on 300-VW dataset (68 landmarks).

modelling and tracking. *International journal of computer vision*, 95(2):154–179, 2011.

[15] J. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

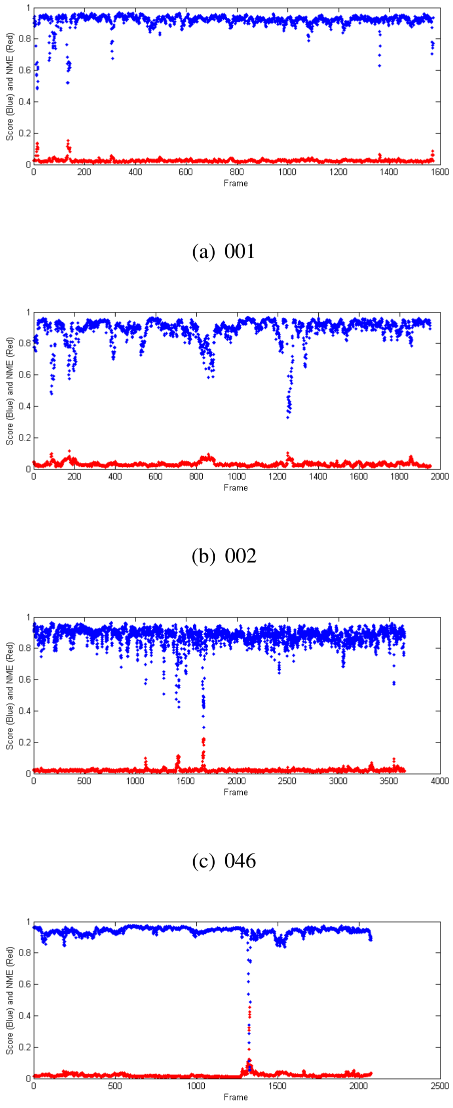
[16] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[18] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[19] Q. Liu, J. Yang, K. Zhang, and Y. Wu. Adaptive compressive tracking via online vector boosting feature selection. *arXiv preprint arXiv:1504.05451*, 2015.

[20] P. Luo, X. Wang, and X. Tang. A deep sum-product archi-



(a) 001

(b) 002

(c) 046

(d) 057

Figure 8. Score (Blue) and NME (Red) on four 300-VW validation videos

tecture for robust facial attributes analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2864–2871, 2013.

[21] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[22] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Computer Vision and Pattern Recognition*, pages 1685–1692. IEEE, 2014.

[23] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel pca. In *BMVC*, volume 10, pages 483–492, 1999.

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. pages 397–403, 2013.

[25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903. IEEE, 2013.

[26] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[27] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision Workshops (ICCVW), 2015*. IEEE.

[28] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.

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

[30] V. N. Vapnik. The nature of statistical learning theory. 1995.

[31] 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*, 34(2):372–386, 2012.

[32] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539. IEEE, 2013.

[33] X. Xiong and D. Fernando. Supervised descent method for solving nonlinear least squares problems in computer vision. *arXiv:1405.0601*, 2014.

[34] X. Xiong and F. D. la Torre. Global supervised descent method. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015.

[35] J. Zhang, S. Shan, K. Meina, and X. Chen. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[36] Z. Zhang, P. Luo, L. Chen, and X. Tang. Learning and transferring multi-task deep representation for face alignment. *arXiv:1408.3967*, 2014.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition*, pages 2879–2886. IEEE, 2012.

[38] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 113–120, 2013.