Topic-aware Deep Auto-encoders (TDA) for Face Alignment

Jie Zhang[1,2], Meina Kan[1], Shiguang Shan[1], Xiaowei Zhao[3], and Xilin Chen[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Imperial College London, London, UK

Abstract. Facial landmark localization plays an important role for many computer vision tasks, e.g., face recognition, face parsing, facial expression analysis, face animation, etc. However, it remains a challenging problem due to the diverse variations, such as head poses, facial expressions, occlusions and so on. In this work, we propose a topic-aware face alignment method to divide the difficult task of estimating the target shape into several much easier subtasks according to the topics. Specifically, topics are determined automatically by clustering according to the target shapes or shape deviations which are more compatible with the task of alignment. Then, within each topic, a deep auto-encoder network is employed to regress from the shape-indexed feature to the target shape. Deep model specific to each topic can capture more subtle variations in shape and appearance, and thus leading to better alignment results. This process is conducted in a cascade structure to further improve the performance. Experiments on three challenging databases demonstrate that our method significantly outperforms the state-of-the-art methods and performs in real-time.

1 Introduction

Face alignment or facial landmark localization is a vital problem in computer vision since many vision tasks depend on accurate face alignment results, including face recognition, facial expression analysis, face animation, etc. Although it has been studied for many years, facial landmark detection on the wild face images is still a challenging problem due to large shape variations, such as extreme head poses and facial expressions.

Typical *parametric methods*, such as Active Shape Model (ASM) [1, 2] and Active Appearance Model (AAM) [3, 4], employ the statistical model such as Principal Component Analysis (PCA) to capture the shape and appearance variations respectively. They perform well for face images with little pose variation, normal facial expression and good light conditions. However, they fail to get accurate shapes for those images with large head pose and exaggerated facial expressions since single linear model can hardly well capture the complex nonlinear variations in the wild data. To handle the large texture variations, van

et al. [5] extend the traditional AAM to MPPCA-AAM by using a mixture of probabilistic PCA [6] to model the complex appearance variations resulting in better performance. However, it is still sensitive to shape initializations as the traditional AAM.

Recently, *regression based methods* have achieved impressive results on both controlled and uncontrolled face images [7–10]. Instead of explicitly representing the shape or appearance variations with parametric models, these methods attempt to directly learn a mapping from appearance to face shape. As one of the most promising regression based method, SDM [9] employs a linear regression to estimate the shape deviation based on shape-indexed feature [7] under a cascade framework, and it achieves the state-of-the-art performance for facial landmark detection and tracking on the wild databases, e.g, LFPW [11], LFW-A&C [12], RU-FACS [13] and Youtube Celebrities [14]. To some extent, SDM is more robust to inaccurate shape initialization, but it may still get stuck on the images with extreme pose and exaggerated facial expressions when the initialization shape is far from the ground truth [15].

To relieve the influence of inaccurate initializations, [7, 8] use multiple initializations for testing and take the median result of all random fern regressors as the final estimation. Burgos-Artizzu et al. [16] propose a Robust Cascaded Pose Regression (RCPR) method to further improve the performance of CPR [7] under a novel restart scheme. Specifically, given an image, 10% of the cascade is applied for different initializations and then the variance of their predictions are checked. If the variance is low enough, the left 90% of the cascade is applied, otherwise restart with a different set of initializations.

Different from [7, 8, 16], Dantone et al. [17] employ a regression forest to estimate the head pose and then individually model the shape and appearance variations of facial landmarks for each head pose by using conditional regression forest. They argue that the exploiting of head pose provides a good shape prior for face alignment and conditional regression forests are easier to learn since the trees have no need to capture all shape and appearance variations. A good shape prior can provide better shape initialization even under extreme pose. Furthermore, Zhao et al. [18] propose an iterative Multi-Output Random Forests (IMOFR) algorithm to jointly estimate head pose, facial expressions and facial landmarks, which divides facial landmark detection into subtasks based on both head poses and facial expressions. It further achieves more accurate face alignment results than [17]. Zhu et al. [19] employ a mixture-of-trees model to capture the diverse variations of each viewpoint and partially address the initialization problem by evaluating the models of all viewpoints, which is thus accompanied by a high computation problem. Yu et al. [20] propose a group sparse learning method to select optimized salient facial landmarks for mixture-of-trees models and further refine the detection result by using two-step cascaded deformable shape model. This method can perform faster than [19]. However, it still cannot meet the real-time requirement and the performance degenerates when it fails to get accurate estimation of salient facial landmarks. In another interesting work [15], an exemplar-based approach is proposed to model the

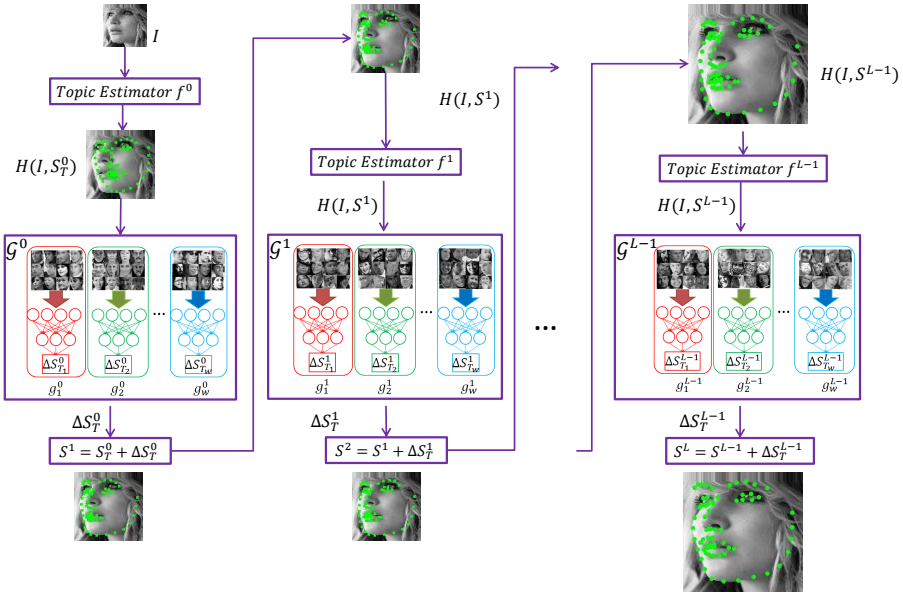


Fig. 1. Overview of our Topic-aware Face Alignment Algorithm with Deep Auto-encoders. f denotes a topic prediction function and $\mathcal{G}$ is the topic-specific deep models for face alignment. $H(I, S)$ is the joint shape-indexed features extracted around the landmarks of face shape S. ΔS is the shape difference between the ground truth and the current shape.

correlations between landmarks and their surrounding information and then a feature voting-based face alignment method is employed with non-parametric shape regularization. This method does not require initial face shape based on face detection result. Impressive results are achieved on two challenging data sets, i.e., AFW [19] and IBUG [21], but it is extremely time-consuming.

Auto-encoders and other deep models are widely applied for computer vision problem and achieve great success for image denoising, image classification, face analysis, etc [22–26]. Inspired by the success of deep network, some researches propose to employ it to solve the facial landmark detection problem. Sun et al. [24] design a deep convolutional neural network (DCNN) for facial landmark detection and achieve impressive results on two public datasets. However, the performance under extreme pose and exaggerated facial expressions may degenerates since it is hard to train a robust deep model to capture all facial variations without any prior knowledge. In [26], Wu et al. propose a deep model based on the Restricted Boltzmann Machines (RBM) for facial landmark tracking with shape prior in consideration of face pose and expressions. Specifically, deep belief network is employed to capture the shape variations due to facial expressions and a 3-way RBM is further used for modeling pose variations. Yet, it is still hard to handle the extreme variations of face poses and facial expressions simultaneously.

To deal with the facial landmark detection with large shape variations, we propose a topic-aware face alignment method to divide the difficult task of estimating the target shape into several much easier subtasks according to the topics, and an overview of the proposed method is shown in Fig. 1. Different from [17, 18], in which the topics are manually defined based on the appearance variations of head poses or facial expressions, we define the topics by automatically clustering the target shapes or shape deviations which are more compatible with the task of alignment. Then, within each topic, a deep auto-encoder network is exploited to detect the facial landmarks. Deep models specific to each topic can well capture the variations in shape and appearance even under extreme poses and facial expressions. This process is further conducted in a cascade structure to improve the performance. It is important to note that topic definitions in each cascaded stage are updated based on target shapes or shape deviations rather than the fixed manual definitions used in [17, 18]. As a result, the topics are closely related to the task, i.e., predicting the target shape.

The main contributions of this work are summarized as bellow:

1. By automatically discovering topics according to the target shapes/target shape deviations, the difficult face alignment task is divided into several much easier subtasks. As the defined topic is related to the shape, more compact subtasks can be achieved leading to better alignment results.
2. Deep model is employed as the alignment model for each topic. Benefited the great ability of modeling nonlinearity, deep model can well capture the diverse variations in shape and appearance leading to more accurate alignment results.
3. Our method outperforms the state-of-the-arts methods on three public data sets, i.e., XM2VTS, LFPW, IBUG, and performs in real time.

2 Topic-aware Deep Auto-encoder for Face Alignment

In this section, we will first give an overview of our topic-aware deep auto-encoder (TDA) method for face alignment. Then we will describe the technical details about each component of our approach.

2.1 Method Overview

Facial landmark detection on the wild face images is quite challenging mainly due to the large shape variations. To tackle this problem, we propose a topic-aware deep auto-encoders for the wild face alignment, which divides the difficult task of predicting the target shape into several much easier subtasks according to the topics, as illustrated in Fig. 1. To make the division of subtasks more compatible with the whole task, the topics are defined according to the target shape (or shape deviations). Furthermore, considering the great ability of capturing the nonlinearity, the deep auto-encoder is employed to solve the subtask within each topic to achieve better prediction.

Given an image I, the problem of facial landmark detection is generally formulated as learning a non-linear function D to predict the shape from the image:

$$D : S \leftarrow I, \tag{1}$$

where S is the face shape of input image I, i.e., the location of each landmark. In the wild condition, D is quite difficult to learn due to the large variations of shape and appearance. Therefore, we propose to divide D into several easier ones $\{D_1, D_2, \cdots, D_w\}$ according to the topics $\mathcal{T} = \{T_1, T_2, ..., T_w\}$. In this work, the topics $\mathcal{T}$ are defined by clustering the face images according to the shape (or shape deviation if the output of D_i is the deviation rather than the shape). To predict the topic of any input image, a deep auto-encoder f is used to model the regression from the input image to the topics:

$$T = f(I), T \in \mathcal{T}. \tag{2}$$

Within each topic, the variations is more compact than the overall topics, and a better shape prior S_T, i.e., the mean shape specific to each topic, can be achieved.

Then, for each topic $T \in \mathcal{T}$, we design another deep auto-encoder network, denoted as g_T, which attempts to infer the shape deviation $\Delta S = S_g - S_T$ as follows:

$$g_T : \Delta S \leftarrow H(I, S_T), \tag{3}$$

where S_g is the ground truth face shape (i.e., the target shape), H is the feature extraction function, and S_T is the shape prior of topic T or the shape from the previous stage.

After learning all topic-specific face alignment models $\mathcal{G} = \{g_{T_1}, g_{T_2}, ..., g_{T_w}\}$, the mapping function D from image I to face shape S can be reformulated as:

$$S = D(I) = S_T + \mathcal{G}(I, T), \tag{4}$$

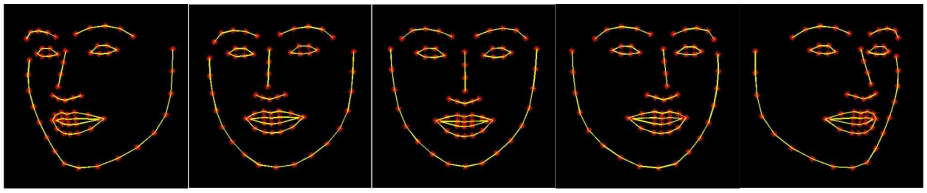
with $\mathcal{G}(I, T) = g_T$.

The above process is further conducted in a cascade procedure to improve the performance.

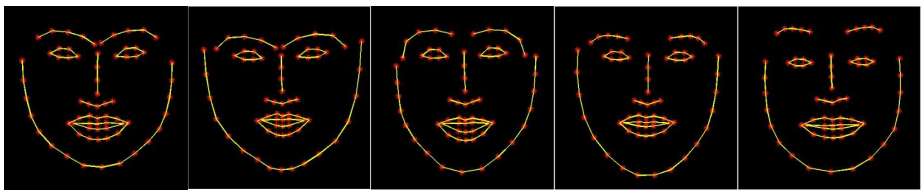
2.2 Topic Definition and Prediction

Topic Definition. In our method, the topics are defined by clustering the target shapes or shape deviations via k-means in each cascade stage. For the first stage, the topics are achieved by clustering according to the target shape, i.e., the ground truth shape S_g. As shown in Fig. 2(a), five topics are exploited by clustering which are roughly consistent with head pose variations since the head pose variations dominate the shape distribution of that dataset. A good shape prior can be easily achieved by taking mean face shape specific to each obtained topic.

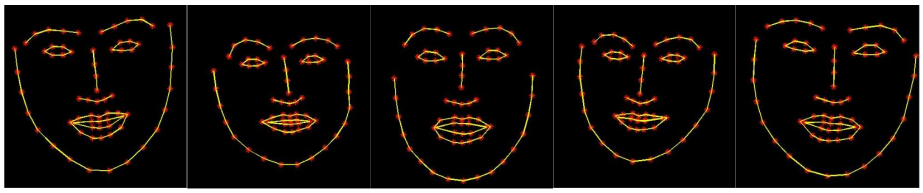
For the successive stages, the topics are achieved by clustering according to the shape deviations, i.e., the difference between the ground truth shape and the



(a) Topics of stage 1



(b) Topics of stage 2



(c) Topics of stage 3

Fig. 2. Topic discovery at each stage. Five topics are exploited by clustering the target shapes or shape deviations for each stage. For stage 1, we directly show the cluster centers of each topic, i.e., the mean shape of each topic. For stage 2 and 3, the cluster centers, i.e., the mean shape deviation, is added to the frontal face shape for better exhibitions.

shape from previous stage $\Delta S^{j-1} = S_g - S^{j-1}$, because the alignment model in the jth ($j \geqslant 2$) stage attempts to predict the shape deviations rather than the shape directly. In other words, the topics are defined according to the face shape deviations(i.e., the target of the task) rather than appearance. Finally, as shown in Fig. 2(b) and 2(c), the topics in each stage are different since the tasks, i.e., the deviations are different in each stage. Compared with the existing methods [17, 18] which also divide the overall tasks into several subtasks, our method is different in two-folds: 1) [17, 18] define the topics according to five head pose (profile left, left, front, right, profile right) or together with three facial expressions (neutral, happy and others), i.e., the characteristic of input image, while our method defines the topics according to the target shapes or shape deviations, i.e., the output of alignment task; 2) In [18], the topics are kept the same in all stages since characteristic of the input image is the same across all stages, while in our method, the topics in each stage are different since the task, i.e., the shape deviation, in each stage is different. Overall, the

definition of topics in our methods can make division of topics more compatible with the overall task, leading to better results.

Topic Prediction. After topics $\mathcal{T}$ are defined based on face shapes, a nonlinear function, i.e., f in Eq. (2), is designed to predict the topic $T \in \mathcal{T}$ of any input image I. Deep models like deep auto-encoder networks [22] is a good choice for its favorable ability of modeling the nonlinearity. Specifically, a deep network with $m-1$ hidden layers is designed. The prediction function f can be formulated as the following optimization problem:

$$f^* = \arg\min_f ||T - \psi_m(\psi_{m-1}(...\psi_1(I)))||_2^2 + \lambda \sum\nolimits_{i=1}^{m} ||W_i||_F^2, \tag{5}$$

$$\psi_i(a_{i-1}) = \sigma(W_i a_{i-1} + b_i) \triangleq a_i,\ i = 1, ..., m-1, \tag{6}$$

$$\psi_m(a_{m-1}) = W_m a_{m-1} + b_m \triangleq a_m,\ a_m \in \mathcal{T}, \tag{7}$$

where ψ_i is the nonlinear mapping of ith layer of deep auto-encoder networks parameterized with W_i and b_i, $\sum_{i=1}^{m} ||W_i||_F^2$ is a weight decay term to prevent over-fitting, σ is a sigmoid function which characterizes the nonlinearity mapping for feature representations $\{a_1, a_2, ..., a_{m-1}\}$ at the first $m-1$ layers. At the last layer, linear regression is employed to get the topic prediction T. Eq. (5) is iteratively optimized by L-BFGS [27]. After obtaining the solution f^*, the topic T of given image I can be achieved as $T = f^*(I) = \psi_m(\psi_{m-1}(...\psi_1(I))$.

2.3 Topic-specific Deep Auto-Encoder for Face Alignment

For a topic T, the face alignment task can be formulated as learning a regression function to predict the shape deviations ΔS between current shape S_T and the ground truth S_g. Considering that the regression function is a complex nonlinear mapping, a deep auto-encoder network denoted as g_T, $T \in \mathcal{T}$ is designed to infer the shape deviations, as shown in Fig. 1.

Specifically, within each topic T ($T \in \mathcal{T}$), the shape-indexed SIFT [28] features denoted as $H(I, S_T)$ are extracted around all facial points and further concatenated as the input for the deep network g_T. The deep network is optimized as follows:

$$g_T^* = \arg\min_{g_T} ||\Delta S_T - \phi_{T,n}(\phi_{T,(n-1)}(...\phi_{T,1}(H(I, S_T))))||_2^2 + \eta \sum\nolimits_{i=1}^{n} ||W_{T,i}||_F^2. \tag{8}$$

$$\phi_{T,i}(a_{T,(i-1)}) = \sigma(W_{T,i} a_{T,(i-1)} + b_{T,i}) \triangleq a_{T,i}, i = 1, ..., n-1, \tag{9}$$

$$\phi_{T,n}(a_{T,(n-1)}) = W_{T,n} a_{T,(n-1)} + b_{T,n} \triangleq \Delta S_T, \tag{10}$$

where $\phi_{T,i}$ is the nonlinear mapping of ith layer of g_T parameterized with $W_{T,i}$ and $b_{T,i}$. n is the number of layers in the deep network and $\sum_{i=1}^{n} ||W_{T,i}||_F^2$ is a weight decay term. After obtaining the solution g_T^*, the shape deviations can be achieved as $\Delta S_T = g_T^*(I) = \phi_{T,n}(\phi_{T,(n-1)}(...\phi_{T,1}(H(I, S_T))))$.

The deep network g_T with n layers has many parameters and is easier to get stuck in local minimum. To relieve this, we initialize the first $n-1$ layers through an unsupervised pre-train process. The objective function of the pre-train process for ith layer is:

$$\{\phi_{T,i}{}^*, \varphi_{T,i}{}^*\} = \arg\min_{\phi_{T,i},\varphi_{T,i}} ||a_{T,(i-1)} - \varphi_{T,i}(\phi_{T,i}(a_{T,(i-1)}))||^2 + \alpha(||W_{T,i}||_F^2 + ||W'_{T,i}||_F^2), \tag{11}$$

where $\phi_{T,i}(x) = \sigma(W_{T,i}x + b_{T,i})$ and $\varphi_{T,i}(x) = \sigma(W'_{T,i}x + b'_{T,i})$. For the first layer, we take the shape-indexed SIFT feature as input, e.g., $a_0 = H(I, S_T)$ and the output of this hidden layer is treated as the input of following layer. With the pre-trained parameters of the first $n-1$ layers and randomly initialized parameters of the last layer, the whole network is fine-tuned according to Eq. (8).

After learning all topic-specific face alignment models $\mathcal{G} = \{g_{T_1}, g_{T_2}, ..., g_{T_w}\}$, the face shape S of any image can be achieved by adding the predicted shape deviation ΔS_T from the corresponding model g_T to the shape prior $S_T : S = S_T + \Delta S_T$, where T is predicted topic from deep model f. The topic-specific deep models for face alignment can well capture the detailed variations in shape and appearance of each topic, which show favorable ability for handling extreme head pose and facial expression variations.

2.4 Cascade Topic-aware Face Alignment

Given an image I, we can get a shape prediction S from topic estimation model f and topic specific face alignment models $\mathcal{G}$. However, it is hardly to achieve accurate face alignment result with only one stage process as demonstrated above. So we perform topic-aware face alignment algorithm in a cascade structure.

After obtaining the shape estimation S^1 from the first stage, we further cascade $L-1$ stages to refine the face alignment result, where the jth stage attempts to predict the shape deviation $\Delta S^{j-1} = S_g - S^{j-1}$ based on shape-indexed feature $H(I, S^{j-1})$, $j = 2, 3, ..., L$. It is worth noting that topics $\mathcal{T}^j$ at each stage j is redefined by clustering with current target shape deviation ΔS^{j-1}. After defining the topics, we employ a deep auto-encoder network to predict the topic $T \in \mathcal{T}^j$based on shape-indexed feature $H(I, S^{j-1})$. For each stage j, the objective function of topic estimation model f^j is formulated as follows:

$$f^{j*} = \arg\min_{f^j} ||T^j - \psi^j_m(\psi^j_{m-1}(...\psi^j_1(H(I, S^{j-1}))))||_2^2 + \lambda \sum\nolimits_{i=1}^{m} ||W^j_i||_F^2, \tag{12}$$

After getting the topic estimation at stage j, we divide the whole training set into several subset based on the topic estimation result. Then a deep face alignment model g^j_T specific to each topic $T \in \mathcal{T}^j$ are trained with face images of the corresponding topic. The objective function of deep model g^j_T is shown below:

$$g^{j*}_T = \arg\min_{g^j_T} ||\Delta S^{j-1}_T - \phi^j_{T,n}(\phi^j_{T,(n-1)}(...\phi^j_{T,1}(H(I, S^{j-1}))))||_2^2 + \eta \sum\nolimits_{i=1}^{n} ||W^j_{T,i}||_F^2. \tag{13}$$

Finally, the overall model $\mathcal{G}^j = \{g^j_{T_1}, g^j_{T_2}, ..., g^j_{T_w}\}$ for the jth stage consists of the face alignment models specific to each topic.

After cascading L stages, the overall topic-aware face alignment model D can be represented as: $D = \{f^1, f^2, ..., f^L; \mathcal{G}^1, \mathcal{G}^2, ..., \mathcal{G}^L\}$. As a result, the face alignment performance is gradually improved stage by stage as shown in Sec.3.2. Our algorithm converged with 3 or 4 stages.

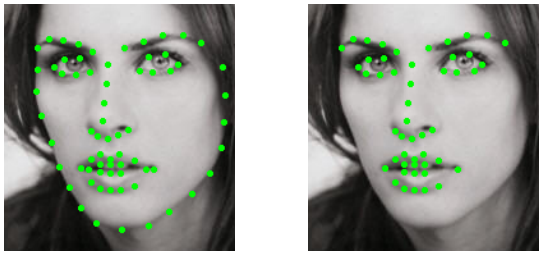
3 Experiments

In this section, the proposed topic-aware deep auto-encoder (TDA) method is evaluated on three public datasets. Firstly, the performance of each stage of TDA is investigated, and then the overall method is compared with the state-of-the-art methods.

3.1 Datasets and Methods for Comparison

To evaluate the effectiveness of the proposed TDA method, five public datasets are used, i.e., **XM2VTS** [29], **LFPW** [11], **HELEN** [30], **AFW** [19] and **IBUG** [31], among which three ones, i.e., IBUG, XM2VTS and LFPW test set, are used for testing, while the others are used for training. XM2VTS dataset contains 2360 face images of 295 individuals collected under laboratory conditions and the other datasets are collected from the internet, which contain more challenging images in the wild. LFPW contains 1432 images, including 1132 images for training and 300 images for testing. It is firstly published with 29 landmarks annotations by [11]. HELEN consists of 2330 high resolution images from *Flickr* with 194 annotated landmarks, which contains large variations such as head pose, facial expression, partially occlusion, etc. In AFW, 205 images with 468 faces are also collected from *Flickr*, containing complex backgrounds with large variations in head pose and facial expressions. However, only 6 landmarks(the center of eyes, tip of nose, the two corners and center of mouth) are released for this dataset [19]. Recently, the four datasets, i.e., XM2VTS, LFPW, HELEN and AFW, mentioned above, are relabeled with 68 landmarks and published in website [21]. Fig. 3(a) shows the definitions of 68 landmarks. The face detection results are also provided in website [21]. Besides these datasets, another 135 images with extreme head pose and facial expressions are released in website [21], denoted as IBUG dataset.

The proposed TDA model is trained with the images from LFPW training set, HELEN and AFW. For testing, the XM2VTS, LFPW test set and IBUG dataset are employed. XM2VTS dataset formulates a laboratory scenario, while LFPW test set and IBUG formulate the uncontrolled scenario which means much more challenging. Especially, IBUG dataset is even more challenging than LFPW due to the extreme head pose and exaggerated facial expressions. For all experiments, the number of stages is 3 and the number of topics is 3 for each stage.



(a) (b)

Fig. 3. Definition of facial landmarks: (a) 68 points mark-up. (b) 49 points mark-up.

The proposed TDA method is compared with a few state-of-the-art methods, e.g., Dantone et al. [17], Zhu et al. [19], Yu et al. [20], DRMF [10] and SDM [9]. For Dantone et al.'s method, the model released by authors can only detect 10 landmarks, we retrain it to detect 68 landmarks with the same training set for fair comparison. For Zhu et al.'s method, we use the model provided by Asthana et al., which shows better performance in [10]. The public code of SDM only predict 49 inner landmarks (as shown in Fig. 3(b)), so we retrain the SDM method to detect the 68 landmarks with the same training set as ours. Following the CMU 68 points mark-up, the methods, i.e., Dantone et al, Zhu et al., and SDM, are trained to detect 68 landmarks, while the original implementations of Yu et al.'s method and DRMF are directly used and they can only estimate 66 facial landmarks (as shown in Fig. 3(a) except two inner mouth corners). Therefore, in order to conduct a fair comparison, all methods are evaluated with the common 66 facial points.

Since all methods are initialized from face detection result, our TDA, SDM [9] and Dantone et al. [17] are conducted with face detection results from [21] and the face detectors for other methods [10, 20, 19] are kept the same as their papers.

To measure the performance of face alignment, the normalized root-mean-squared error (NRMSE) is employed. On XM2VTS and LFPW datasets, the NRMSE is normalized by the inter-ocular distance, while on IBUG, it is normalized by the face size for clear exhibition since this dataset is extremely difficult. Besides, the cumulative function (CDF) of NRMSE is used for performance evaluation.

3.2 Experimental Results

Performance of Each Stage. The proposed TDA is designed in a cascade structure, and thus we investigate the performance of facial landmark detection at each stage. The experiments are conducted on the most challenging IBUG dataset in terms of average detection accuracy of 66 facial landmarks. The experiment results are shown in Fig. 4. The "Mean Shape" denotes the alignment result by fitting a mean face shape to the face detection window. The "Top-aware Shape Prior" also takes the mean shape as the fitting results, but the

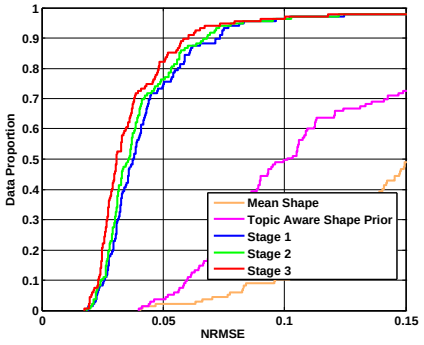


Fig. 4. Performance of each stage on IBUG.

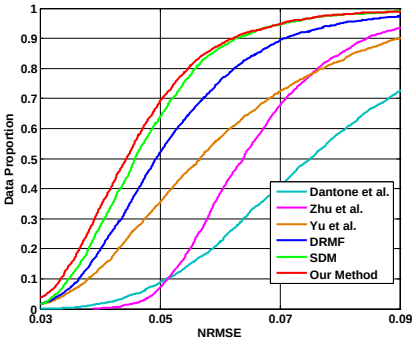


Fig. 5. Experiment on XM2VTS.

mean shape is from the corresponding topic at the first stage rather than the overall mean shape. "Stage 1,2,3" represent the facial detection result from the topic-aware deep face alignment model at each stage respectively.

As seen from Fig. 4, shape priors provided by topic discovery is more accurate than simply taking the mean shape as initialization, which demonstrates the effectiveness of topic-aware strategy for face alignment, especially under extreme head poses and facial expressions. Moreover, the deep auto-encoder network specific to each topic significantly improve the detection accuracy at stage 1. This improvement comes from two aspects, better shape prior from topic discovery and better capture of the detailed variations in shape and appearance from the deep model. Stage 1 handles the large variations and achieves a much better shape, but it is not accurate enough. To further handle the small shape variations and ensure a better shape, stage 2 and 3 are cascaded. As expected, the performance is improved progressively. It should be noted that, in order to well capture the subtle variations, a higher resolution image containing more subtle information is used in stage 3.

The experiments are conducted on a desktop (Intel i7-3770 3.4GHz CPU) with MATLAB implementation. The overall run time of TDA is about 150 milliseconds for one image, which means TDA is less time-consuming and can run in real-time.

Experiments on XM2VTS. We firstly compare the proposed TDA with the existing methods on the XM2VTS dataset. In XM2VTS, 2360 face images are collected over 4 sessions under the laboratory environment. It contains variations in shape and appearance due to identity, glasses, beard and so on. In DRMF [10], two face detectors are attached, and we choose the Viola-Jones face detector since all face images in this dataset are almost near frontal. For fair comparison, only the common face images returned by all face detectors are used for evaluation.

Fig. 5 shows the comparison results in terms of cumulative error distribution curves. Although Dantone et al. [17] divide the face alignment task as ours, it

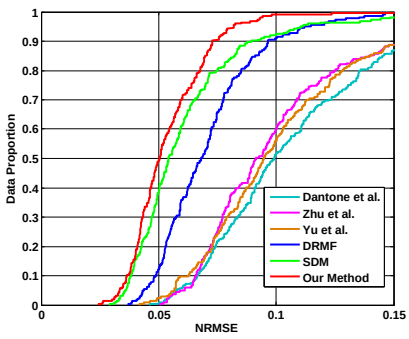


Fig. 6. Experiment on LFPW.

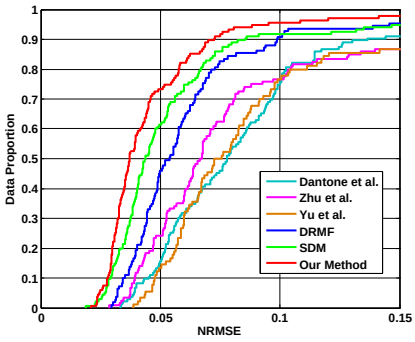


Fig. 7. Experiment on IBUG.

performs the worst as seen from Fig. 5, mainly due to the limitations of *manual* division of variations and its limited ability of capturing complex nonlinearity. Zhu et al. [19] performs a little better followed by Yu et al. [20], however both are worse than DRMF [10] and SDM [9], possibly because Zhu et al. does not model the correlation of nonadjacent nodes in the mixture-of-trees model, and Yu et al. suffer from the local minimum problem caused by Gauss-Newton optimization. Benefitted from the regression based framework, DRMF and SDM perform much better, and SDM performs even better than DRMF with finer shape-indexed feature. Moreover, our TDA method outperforms SDM, with an improvement up to 5% when NRMSE is 0.05, by taking the advantages of the topic-aware strategy and the deep models in a cascade structure.

Experiments on LFPW. To investigate the robustness to the large variations such as head pose and facial expression, all methods are further evaluated on the Labeled Face Parts in the Wild (LFPW) dataset, which contains large variations from pose, expression, occlusion, etc. The URLs of the 300 testing images are shared by [11], but some of them are no longer available. Recently, 224 testing images of LFPW are published as part of 300-W dataset [31]. So these 224 testing images from [31] are used for testing. For DRMF method, the tree-based face detector is employed for the wild scenario to achieve better face detection result.

The comparison results are shown in Fig. 6. As seen, all methods degenerate on this dataset as LFPW contains larger shape variations. On this dataset, Dantone et al. [17] also performs the worst as on XM2VTS. Yu et al. [20] is comparable to Zhu et al. [19] since Yu et al. degenerates a little due to inaccurate initializations from optimized part mixture models, especially in case of large variations. Similarly as on XM2VTS, DRMF and SDM perform better, and our TDA outperform all of them. Compared with the best performer SDM, the improvement of our TDA is even up to 10% when NRMSE is 0.05. Moreover, TDA achieves nearly perfect result, i.e., 100%, when NRMSE is 0.1. These comparisons demonstrate that our TDA is more robust to the large variations. On one hand, the improvement comes from the better shape prior from the topic-aware strategy, and on the other hand, the deep auto-encoder network can well

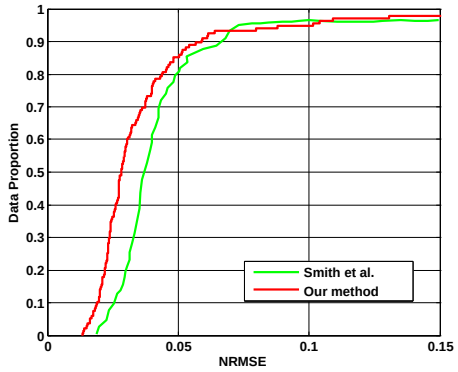


Fig. 8. Comparison with [15] on IBUG.

model the nonlinear mapping from the shape-indexed feature to shape, leading to further improvements.

Experiments on IBUG. IBUG, as another wild dataset, is more challenging than LFPW due to the extreme head poses and exaggerated facial expressions. The evaluation results of all methods are presented in Fig. 7. Considering the extreme challenges on this dataset, NRMSE is normalized by the face size rather than inter-ocular distance for clear exhibition.

As seen from Fig. 7, the similar conclusions can be obtained that SDM performs the best among the existing methods and our TDA method outperforms SDM. Even under the extreme shape variations, our algorithm outperforms all the other methods, demonstrating the effectiveness of our TDA, especially under the large variations.

Furthermore, we compare our TDA to method [15]. Smith et al. [15] propose a data-driven approach which is robust to extreme head pose and expressions and achieves state-of-the-art performance on IBUG dataset. Since only the detection result of 49 facial points (as shown in Fig. 3(b)) is published in [15], the common 49 landmarks are evaluated for fair comparison. As shown in Fig. 8, our method also outperforms [15] with more accurate detection result when NRMSE is below 0.07. Moreover, the MATLAB implementation of [15] requires 25.5 seconds for processing one image while the run time of our method is only 150 milliseconds per image, which demonstrates that our method performs more efficiently than [15]. We also compare our TDA with several deep learning methods, i.e., DCNN [24] and Zhou et al. [32] on IBUG: 1) The DCNN achieves an average error of 0.1052, while our TDA achieves better performance with an average error of 0.0848 in terms of five common landmarks. 2) The mean error of Zhou et al. is 0.1455, and our TDA achieves a much lower mean error as 0.1156 in terms of 19 common points. From these comparisons, our TDA also outperforms DCNN [24] and Zhou et al. [32] on the extremely challenging dataset benefited from automatic topic discovery. Fig. 9 shows the detection results of our TDA

Fig. 9. Exemplar results from IBUG dataset.

on some challenging images with simultaneous extreme poses, facial expressions and partial occlusions.

4 Conclusions

In this paper, we present a topic-aware deep auto-encoder network for face alignment. Instead of directly tackling the difficult alignment under large variations, we firstly divide it into several easier subtasks according to the topics, which are defined by clustering according to the target shapes or shape deviations. Then within each topic, a deep auto-encoder network is designed to regress from the shape-indexed feature to the shape or shape deviation specific to this topic. Benefitted from the better shape prior from the topic-aware strategy and the non-linear deep networks, our TDA method is robust to large shape variations, such as the head pose and facial expression. As evaluated on three challenging datasets, our method achieves the state-of-the-art performance, demonstrating the effectiveness of TDA. Moreover, our TDA can perform in real-time.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. 61025010, 61173065, and 61390511.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding (CVIU) **61** (1995) 38–59
2. Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: European Conference on Computer Vision (ECCV). (2008) 413–426
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **23** (2001) 681–685
4. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision (IJCV) **60** (2004) 135–164
5. van der Maaten, L., Hendriks, E.: Capturing appearance variation in active appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2010) 34–41
6. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Computation **11** (1999) 443–482
7. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010) 1078–1085
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 2887–2894
9. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
10. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 3444–3451
11. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 545–552
12. Saragih, J.: Principal regression analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 2881–2888
13. Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia **1** (2006) 22–35
14. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008) 1–8
15. Smith, B.M., Brandt, J., Lin, Z., Zhang, L.: Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
16. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision (ICCV). (2013)
17. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 2578–2585
18. Zhao, X., Kim, T.K., Luo, W.: Unified face analysis by iterative multi-output random forests. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)

19. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 2879–2886
20. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: IEEE International Conference on Computer Vision (ICCV). (2013)
21. : 300 faces in-the-wild challenge. (http://ibug.doc.ic.ac.uk/resources/300-W/)
22. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends® in Machine Learning **2** (2009) 1–127
23. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2012) 1106–1114
24. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 3476–3483
25. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 2480–2487
26. Wu, Y., Wang, Z., Ji, Q.: Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 3452–3459
27. Le, Q.V., Coates, A., Prochnow, B., Ng, A.Y.: On optimization methods for deep learning. In: International Conference on Machine Learning (ICML). (2011) 265–272
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60** (2004) 91–110
29. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA). Volume 964. (1999) 965–966
30. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision (ECCV). (2012) 679–692
31. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: The IEEE International Conference on Computer Vision Workshops (ICCVW). (2013)
32. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: The IEEE International Conference on Computer Vision Workshops (ICCVW). (2013)