Article:
Feng, Zhenhua, Huber, Patrik orcid.org/0000-0002-1474-1040, Kittler, Josef et al. (2 more authors) (2015) Random Cascaded-Regression Copse for Robust Facial Landmark Detection. IEEE Signal Processing Letters. pp. 76-80. ISSN 1070-9908

https://doi.org/10.1109/LSP.2014.2347011

Random Cascaded-Regression Copse for Robust Facial Landmark Detection

Zhen-Hua Feng, *Student Member, IEEE,* Patrik Huber, Josef Kittler, *Life Member, IEEE,* William Christmas, and Xiao-Jun Wu

***Abstract*—In this paper, we present a random cascaded-regression copse (R-CR-C) for robust facial landmark detection. Its key innovations include a new parallel cascade structure design, and an adaptive scheme for scale-invariant shape update and local feature extraction. Evaluation on two challenging benchmarks shows the superiority of the proposed algorithm to state-of-the-art methods.**

***Index Terms*—Facial landmark detection, cascaded regression, adaptive shape update.**

I. Introduction

Over the last few years, cascaded-regression (CR) based methods have shown impressive results in automatic facial landmark detection [1]–[6] in uncontrolled scenarios, as compared to the traditional ways of using Active Shape Models (ASM) [7], Active Appearance Models (AAM) [8], Constrained Local Models (CLM) [9] etc. Typically, a face shape is represented by the coordinates of P landmarks $\mathbf{s} = [x_1, y_1, \cdots, x_P, y_P]^T$. Given a facial image $\mathbf{I}$ and an initial face shape estimate, $\mathbf{s}_0$, the aim of facial landmark detection is to find a shape updater $\mathbf{U}$:

$$\begin{aligned} &\mathbf{U} : \mathbf{f}(\mathbf{I}, \mathbf{s}_0) \mapsto \delta\mathbf{s}, \\ &s.t. \quad \|\mathbf{s}_0 + \delta\mathbf{s} - \hat{\mathbf{s}}\|_2^2 = 0 \end{aligned} \tag{1}$$

where $\mathbf{f}(\mathbf{I}, \mathbf{s}_0)$ is a shape-related feature mapping function, $\delta\mathbf{s}$ is the shape update and $\hat{\mathbf{s}}$ is the ground truth shape.

The success of CR-based approaches emanates from four sources: 1) cascading a set of regressors greatly improves the representation capacity of a discriminative model; 2) local feature descriptors used in CR are much more robust than conventional pixel intensities; 3) the non-parametric shape model adopted in CR can express deformable objects, *e.g.* a human face, in more detail compared to a PCA-based parametric shape model; 4) the latent shape constraint of the coarse-to-fine cascade structure promotes the speed of convergence as well as accuracy of the detection result.

This work was supported by '111' Project (No. B12018) and Key Grant Project (No. 311024) of Chinese Ministry of Education, Fundamental Research Funds for the Central Universities (JUDCF09032), UK EPSRC project EP/K014307/1, European Commission project BEAT (No. 284989), National Natural Science Foundation of China (No. 61373055, 61103128), and Natural Science Foundation of Jiangsu Province (BK20140419, BK2012700).

Z.-H. Feng is with the School of IoT Engineering, Jiangnan University, Wuxi 214122, China, and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK. E-mail: Z.Feng@surrey.ac.uk

P. Huber, J. Kittler and W. Christmas are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK. E-mail: {P.Huber, J.Kittler, W.Christmas}@surrey.ac.uk

X.-J. Wu is with the School of IoT Engineering, Jiangnan University, Wuxi 214122, China. E-mail: xiaojun_wu_jnu@163.com

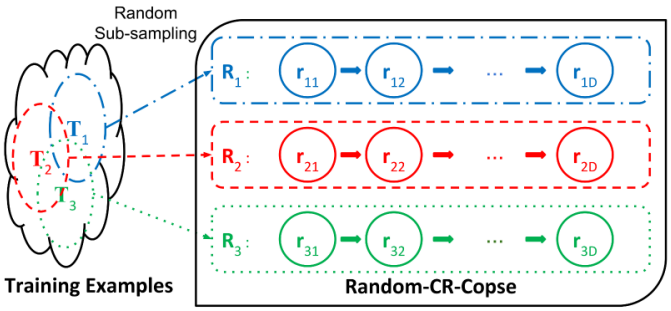

Fig. 1. A 3-wide and D-deep random CR copse.

In the development of a CR-based framework, there are two crucial design issues: 1) the cascade structure and 2) the method to extract local features. Perhaps the most widely adopted approach to the *first* issue is to simply concatenate a set of regressors in series [1], [3], [10]. Another successful cascade design is the two-layer structure used by [2] and [4], in which boosted regression was used for training a strong regressor with a sequence of weak regressors, each consisting of many sub-regressors. Regarding the *second* issue, both hand-crafted and learning-based feature extraction methods have been adopted. As an example of hand-crafted features, Xiong and De la Torre [3] used SIFT for facial landmark detection and tracking, and put forward a theoretical underpinning of cascaded regression as a supervised descent method (SDM). Yan *et al.* [10] have compared different hand-crafted local feature descriptors (HOG, SIFT, Gabor and LBP) and found that the HOG descriptor worked best. However, the hand-crafted feature extraction methods are not designed for the task of facial landmark detection specifically, whereas the learning-based feature extraction methods are self-adapting to the task [2], [4]. For example, cascaded Convolutional Neural Networks (CNN) have been successfully applied to facial landmark detection [5], [6]. The advantage of CNNs is that they fuse the tasks of feature extraction and network training in a unified framework. However, many free parameters need to be tuned when using CNNs. Subsequently, Ren *et. al.* [11] proposed a local binary feature learning approach that achieved great success both in accuracy and efficiency.

Through our early experiments, we found that simply using a strong regressor with a set of weak regressors in series performed badly in cases with occlusions and large-scale pose variations, confirming the observation made in [3]. Furthermore, it usually fails in the presence of deformation and scale variation of the human face. To counteract these problems, this

paper presents an adaptive Random-CR-Copse (R-CR-C) with two main contributions to the field: 1) We propose a new copse design with multiple CR threads in parallel. Each CR thread is trained on a subset generated by random sub-sampling from a pool of training examples. The proposed copse structure enhances the generalisation capacity of the trained strong regressor by fusing multiple experts. The independence among CR threads in the copse allows us to train them efficiently in parallel. 2) We propose an adaptive scheme for robust shape update and local feature extraction to counteract the deformation and scale variation of facial images. Compared to state-of-the-art algorithms, the proposed adaptive R-CR-C shows 15% improvement in accuracy on the newly released COFW benchmark [4].

II. Review of cascaded regression

Given a new image $\mathbf{I}'$ and an initial shape estimate $\mathbf{s}'_0$, the aim of a CR-based approach is to find a shape model updater to approach the true shape, as shown in equation (1). In a standard CR-based approach [1], [3], [4], the shape updater is a strong regressor formed by D weak regressors in series:

$$\mathbf{R} = \mathbf{r}_1 \circ \cdots \circ \mathbf{r}_D, \tag{2}$$

where $\mathbf{r}_d = \{\mathbf{A}_d, \mathbf{b}_d\}$ $(d = 1 \cdots D)$, $\mathbf{A}_d$ is the projection matrix and $\mathbf{b}_d$ is the offset of the dth regressor. Both $\mathbf{A}_d$ and $\mathbf{b}_d$ are learned recursively from a set of labelled facial images. This is discussed in detail in the next section. Assuming we have already trained a strong regressor $\mathbf{R}$, then, in the detection phase, we apply the first weak regressor to update the current shape $\mathbf{s}'_0$ to a new shape $\mathbf{s}'_1$ and then pass $\mathbf{s}'_1$ to the second weak regressor and so on, until the final shape estimate $\mathbf{s}'_D$ is obtained. More specifically, the dth shape is obtained by:

$$\mathbf{s}'_d = \mathbf{s}'_{d-1} + \mathbf{A}_d \cdot \mathbf{f}(\mathbf{I}', \mathbf{s}'_{d-1}) + \mathbf{b}_d. \tag{3}$$

Note that the shape-related feature $\mathbf{f}(\mathbf{I}', \mathbf{s}'_{d-1})$ is also updated after applying a new weak regressor to the current shape estimate. The process of facial landmark detection using a CR-based approach is schematically represented in Fig. 2.

Input: Test image $\mathbf{I}'$, initial shape estimate $\mathbf{s}'_0$ and a pre-trained cascaded strong regressor $\mathbf{R} = \{\mathbf{r}_1 \circ \cdots \circ \mathbf{r}_D\}$.
Output: Final facial shape estimate $\mathbf{s}'_D$.
Repeat:
for $d = 1 \cdots D$
  Obtain shape-related features $\mathbf{f}(\mathbf{I}', \mathbf{s}'_{d-1})$,
  Update current shape $\mathbf{s}'_{d-1}$ to $\mathbf{s}'_d$ using (3).
end

Fig. 2. CR-based facial landmark detection.

III. Adaptive random CR copse

In this section, we present the proposed R-CR-C structure design and the adaptive scheme. The key innovative idea is to design multiple cascaded regressors and fuse their estimates to obtain a better face shape estimate.

A. Random CR copse (R-CR-C)

We define the *width* W as the number of CR threads in a copse, and the *depth* D as the number of weak regressors in each CR thread. Fig. 1 illustrates a copse with three CR threads. Given a training dataset with N labelled facial images $\mathbf{T} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$, we generate W subsets $\{\mathbf{T}_1, \ldots, \mathbf{T}_W\}$ by applying random sub-sampling on $\mathbf{T}$. Each subset is used to train a single CR thread of the copse:

$$\mathbf{U} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_W\}, \tag{4}$$

where the wth CR thread $\mathbf{R}_w = \mathbf{r}_{w,1} \circ \cdots \circ \mathbf{r}_{w,D}$ contains D weak regressors trained on the wth subset. In contrast to training a single CR from all training examples, the procedure of random sub-sampling produces different experts (CR threads). This improves the generalisation capacity and achieves a better balance between over-fitting and reduced accuracy of the system by fusing the outputs of different experts. The proposed adaptive training of all the weak regressors in each CR thread will be described in the second part of the next subsection.

B. An adaptive scheme

Given a set of training images and their ground truth shapes, the initial shape estimates are obtained by putting a reference shape in the detected face bounding boxes. This is discussed in section IV-A. We can either use the mean shape [3] or a randomly selected shape [2] as the reference shape. To train the weak regressors, we need to obtain the extracted shape-related features of all initialised shapes and the differences between the initialised shapes and the ground truth shapes.

1) Adaptive local feature extraction: To extract the shape-related features, we could apply a local feature descriptor on a fixed-size neighbourhood of each landmark and then concatenate the extracted features into one vector. However, the local patches cropped from this fixed-size neighbourhood can be dramatically different in their content due to the deformations and scale variations of faces; *e.g.* we may crop the whole face part from a small face and only the nose part from a large face, as shown in Fig. 3. One solution of this problem is to resize all faces to a unified scale using the estimated face size from the face bounding box provided by a face detector [3], [10]. However, this strategy has two drawbacks: 1) the bounding box initialised by a face detector is too rough to accurately estimate the scale of a face; 2) resizing all images is computationally costly when we have a large number of images.

To meet the demands of scale-invariant local feature extraction, we propose an adaptive scheme. Rather than using a fixed neighbourhood, we set the patch size $S_p(d)$ of the dth weak regressor in a D-deep CR to:

$$S_p(d) = S_f / (K \cdot (1 + e^{d-D})), \tag{5}$$

where K is a fixed value for shrinking and S_f is the size of the face estimated from the previous updated shape $\mathbf{s}^{d-1}$. We can set S_f to either the distance between the pupils, or the distance between the mean of the two outer mouth corners and the mean of the two outer eye corners, or the maximum of these two distances. In this paper, we use the last of these

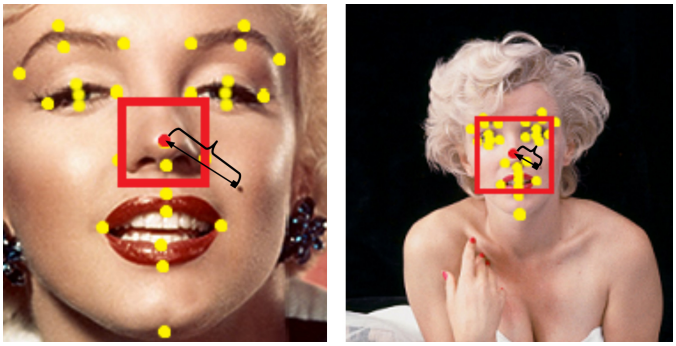

Fig. 3. Local patches as well as the shape updates between two facial images can be very different due to scale variations of the face.

three. As S_f is calculated from the previous updated shape directly, it is not very accurate after the first regressor, due to the rough initial shape estimate from the face bounding box. However, the estimate becomes more accurate as the current shape gets closer to the true value. Furthermore, it is worth noting that equation (5) involves a multi-scale technique, *i.e.* a bigger patch size for the first weak regressor and smaller patch size for the subsequent weak regressors, similar to [10]. For instance, when we set S_f to the pupil distance and pick the shrinking parameter $K = 2$ for a 5-deep CR copse, the patch size decreases from half size of the inter-ocular distance for the 1st weak regressor to a quarter for the last one. Finally, we resize these patches to a fixed size, 30×30 in our case, and then extract local features.

2) Adaptive shape update: The shape difference between the initial shape and the ground truth shape is also highly dependent on the face scale. For instance, the shape updates vary greatly when we set the initial shape estimate of the nose tip of each image in Fig. 3 at the centre of the left cheek. Rather than using an absolute shape difference $\delta\mathbf{s} = \hat{\mathbf{s}} - \mathbf{s}_0$, we propose to use a relative value $\delta\mathbf{s}/S_f$. Suppose the number of training examples in the wth training subset is M_w, we define the objective function of the *first* weak regressor in the wth CR thread as:

$$\frac{1}{2M_w}\sum_{i=1}^{M_w}\|\frac{\hat{\mathbf{s}}^i - \mathbf{s}_0^i}{S_f(\mathbf{s}_0^i)} - \mathbf{A}_{w,1}\cdot\mathbf{f}(\mathbf{I}^i,\mathbf{s}_0^i) - \mathbf{b}_{w,1}\|_2^2 + \lambda\sum\|\mathbf{A}_{w,1}\|_F^2, \quad (6)$$

where $\hat{\mathbf{s}}^i$ is the ground truth shape of the ith image, $\mathbf{s}_0^i$ is the initial shape estimate, $\mathbf{A}_{w,1}$ and $\mathbf{b}_{w,1}$ are the projection matrix and offset of the 1st weak regressor in the wth CR thread, and λ is the weight of the regularisation term. The minimum of this regularised cost function can be efficiently found by ridge regression fitting [12, p. 225]. The subsequent weak regressors in each CR thread are trained recursively using the updated shapes by applying previously trained regressors to the current shape estimates. It is worth noting that the classical CR is a special case of the proposed R-CR-C when W is set to 1 and S_f is set to a constant number.

The scale variation of human faces also affects the facial landmark detection phase. Thus, the output of the wth CR thread is obtained by modifying (3) to:

$$\mathbf{s}'_{w,d} = \mathbf{s}'_{w,d-1} + S_f(\mathbf{s}'_{w,d-1})\cdot(\mathbf{A}_{w,d}\cdot\mathbf{f}(\mathbf{I}', \mathbf{s}'_{w,d-1}) + \mathbf{b}_{w,d}). \quad (7)$$

The final estimated shape $\mathbf{s}'$ of the proposed R-CR-C is obtained by averaging the outputs of all the CR threads.

IV. Evaluation

The proposed algorithm has been evaluated on two challenging benchmarks: LFPW [13] and COFW [4]. Images in both are all 'faces in the wild', with 29 manually annotated landmarks, as shown in Fig. 4.

A. Implementation details

The shape initialisation and training data augmentation were performed in the same way as in [2] and [3]. Specifically, the initial shape estimate was obtained by putting the mean shape at the centre of the detected face bounding box. The training data was augmented by randomly perturbing the initialised shape estimates. The parameters of R-CR-C were tuned by cross validation, where we set the width W to 3, the depth D to 5 and 6 for LFPW and COFW respectively, and the weight of the regularisation term λ to 900. For each random subsampling on the original training dataset, we took 80% of all training examples to generate a random subset. Because Yan *et al.* reported that HOG worked better than SIFT, LBP and Gabor [10], we used two HOG descriptors [14]: Dalal-Triggs HOG (DT-HOG) [15] and Felzenszwalb HOG (F-HOG) [16]. We also used a learning-based 3-layer Sparse Auto-Encoder (SAE) [17] [18] to make a further comparison. For the SAE training, we set the sparsity to 0.025, the regularisation to 1×10^{-4} and the cost of the sparsity constraint to 5.

We measured the accuracy in terms of the average distance between the detected landmarks and the ground truth, normalised by the inter-ocular distance. It was calculated both on 17 and all 29 landmarks, where the former is the well-known '*me17*' measurement [9], shown in Fig. 4. We also measured the failure rate as the proportion of failed detected faces (*i.e.* whose average fitting error was larger than 10% of the inter-ocular distance), and the speed (fps). Our results were obtained using a single core 3.0 GHz CPU and MATLAB.

B. Comparison on LFPW

Although LFPW is a widely used benchmark for facial landmark detection, it only provides hyperlinks to the images. We were only able to download 797 training and 237 test images because some of the hyperlinks have expired. This is a common problem for experiments on LFPW. All results in [2]–[4], [11], [19], [20] are based on different numbers of training and test images. This is the main reason for also using the newly proposed COFW benchmark.

A summary of the performance obtained by state-of-the-art methods and the proposed algorithm using SAE, F-HOG and DT-HOG is shown in Table I. The proposed method beats the other algorithms both in accuracy and failure rate, at a competitive speed. Note that the speed of [11] does not include the time used for loading an image (around 20ms per image

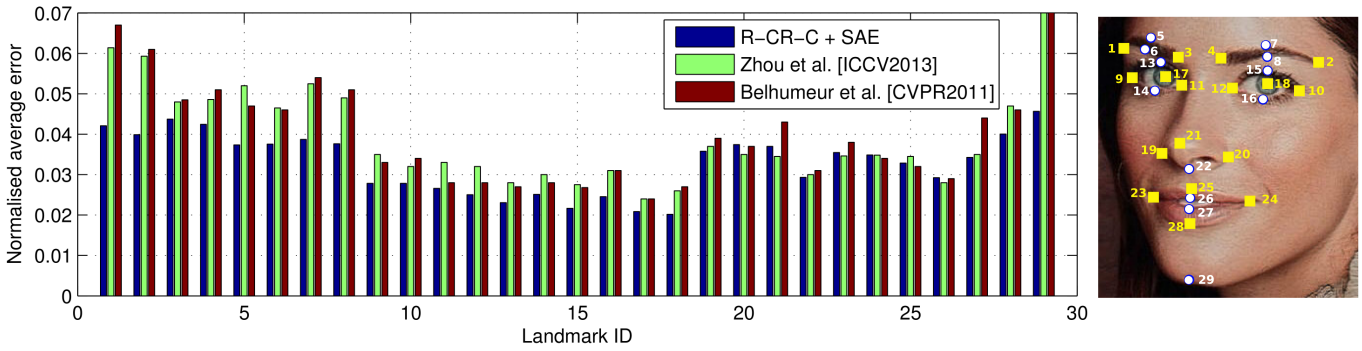

Fig. 4. *Left*: Comparison with Belhumeur *et al.* [13] and Zhou *et al.* [19] on all 29 landmarks of LFPW. *Right*: All 29 landmarks, and the 17 landmarks used for the *me17* measurement (squared landmarks).

TABLE I
COMPARISON ON LFPW.

Method	Error ($\times 10^{-2}$)		Failures	Speed(fps)
	me17	*me29*		
Asthana *et al.* [20]	6.50	-	5.74%	1
Belhumeur *et al.* [13]	3.96	3.99	≈6%	1
Zhou *et al.* [19]	3.89	3.92	-	25
Cao *et al.* [2]	-	3.43	-	20
Xiong and Torre [3]	-	3.47	-	30
Burgos-Artizzu *et al.* [4]	-	3.50	2.00%	12
Ren *et al.* [11]	-	3.35	-	**4200**
Results by Human [4]	-	*3.28*	*0.00%*	0.03
R-CR-C + SAE	**3.29**	**3.31**	**0.84%**	21
R-CR-C + F-HOG	3.37	3.35	1.27%	23
R-CR-C + DT-HOG	3.82	3.81	1.69%	26

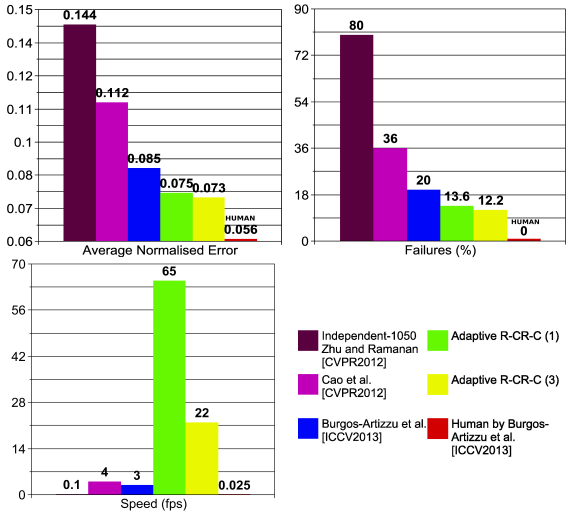

Fig. 5. Comparison of the proposed R-CR-C with 1 and 3 CR threads on COFW to Zhu and Ramanan [21], Cao *et al.* [2] and Burgos-Artizzu *et al.* [4].

for us on a 7200rpm hard disk) and it was measured on a more powerful CPU. At the same time, the use of an SAE shows competitive results compared to HOG descriptors. To the best of our knowledge, this is the first time that the use of an SAE has been explored in facial landmark detection. To gain a better understanding of the error distribution for different landmarks, we compare the detection error for all 29 landmarks in Fig. 4 with that of two state-of-the-art exemplar-based algorithms. It shows that the performance of the proposed approach is much more robust, especially for the landmarks at the eyebrows and chin (points 1, 2 and 29).

C. Comparison on COFW

The COFW benchmark consists of 1345 training images and 507 test images. It is much more challenging than LFPW due to strong pose variations and occlusions. As the performance of the SAE has been demonstrated to be better than HOG, we only present the results based on the SAE in this section. We first evaluate the proposed R-CR-C as a whole system on COFW. Comparisons on COFW with [21], [2] and [4] confirm the superiority of the proposed adaptive R-CR-C in accuracy, failure rate and speed (Fig. 5).

To examine the respective contributions of the proposed adaptive scheme and R-CR-R structure, we measured the performance of using only a single CR-based regressor trained on all training images, the proposed R-CR-C approach with 3 CR threads and their adaptive versions individually in Fig. 6. The results show that the use of our adaptive strategy and copse structure contribute to a similar extent. When both are used at the same time, the best performance is obtained.

Finally, to evaluate the accuracy and robustness of the proposed R-CR-C when using a different number of CR threads, we repeated the random sub-sampling several times to generate different adaptive R-CR-C regressors with different number of CR threads, and measured their accuracy in landmark detection with errorbars. Fig. 7 shows that the use of more CR threads improves both accuracy and robustness of the whole system.

V. CONCLUSIONS

In this paper, we proposed a novel R-CR-C structure with an adaptive scheme for robust facial landmark detection. We demonstrated that with multiple CR threads in parallel we are able to improve the generalisation capacity of the learning-based system. Also, we showed that the proposed adaptive scheme used for model training and local feature extraction makes the proposed R-CR-C approach more robust to scale

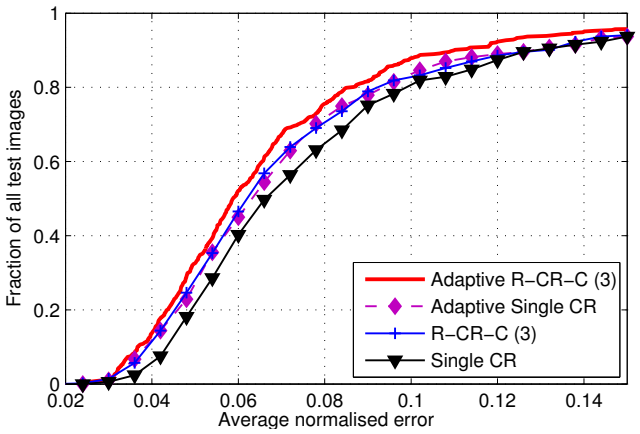

Fig. 6. Evaluation of the adaptive scheme and the proposed R-CR-C structure independently on COFW.

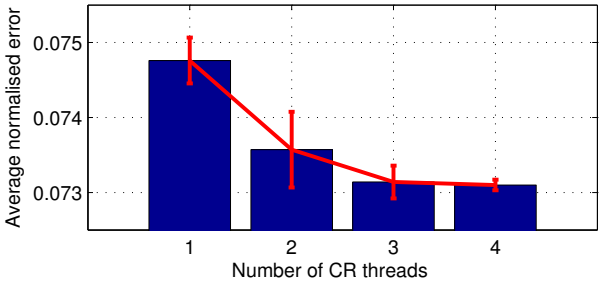

Fig. 7. Comparison on COFW using different number of CR threads.

variations and deformations of human faces. Moreover, the experimental results obtained on two challenging benchmarks using a sparse autoencoder demonstrate the superiority of the proposed algorithm compared to the state of the art.

REFERENCES

[1] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2010, pp. 1078–1085.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012, pp. 2887–2894.

[3] 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, CVPR*, 2013, pp. 532–539.

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the International Conference on Computer Vision, ICCV*, 2013.

[5] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade," in *Proceedings of the IEEE International Conference on Computer Vision Workshops on 300-W Challenge, ICCVW*, 2013, pp. 386–391.

[6] Y. Sun, X. Wang, and X. Tang, "Deep Convolutional Network Cascade for Facial Point Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2013, pp. 3476–3483.

[7] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[8] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision, ECCV*, 1998, pp. 484–498.

[9] D. Cristinacce and T. F. Cootes, "Feature Detection and Tracking with Constrained Local Models," in *Proceedings of the British Machine Vision Conference, BMVC*, 2006, pp. 929–938.

[10] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to Combine Multiple Hypotheses for Accurate Face Alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops on 300-W Challenge, ICCVW*, 2013.

[11] S. Ren, X. Cao, W. Wei, and J. Sun, "Face Alignment at 3000 FPS via Regressing Local Binary Features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR (Accepted)*, June 2014.

[12] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[13] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2011, pp. 545–552.

[14] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," http://www.vlfeat.org/, 2008.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 1, 2005, pp. 886–893.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[17] J. Ngiam, P. W. Koh, Z. Chen, S. A. Bhaskar, and A. Y. Ng, "Sparse Filtering." in *Proceedings of Neural Information Processing Systems, NIPS*, vol. 11, 2011, pp. 1125–1133.

[18] A. Ng, "Sparse autoencoder," Stanford CS294A Lecture notes - http://www.stanford.edu/class/cs294a/sparseAutoencoder.pdf.

[19] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based Graph Matching for Robust Facial Landmark Localization," in *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, December 2013.

[20] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust Discriminative Response Map Fitting with Constrained Local Models," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, June 2013, pp. 3444–3451.

[21] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012, pp. 2879–2886.