Correlation Filter Cascade for Facial Landmark Localization

Hamed Kiani Galoogahi
Pattern Analysis and Computer Vision Department
Istituto Italiano di Tecnologia, Genova, Italy
hamed.kiani@iit.it

Terence Sim
School of Computing
National University of Singapore, Singapore
tsim@comp.nus.edu.sg

Abstract

The application of correlation filters for the task of facial landmark detection has been studied by many vision works. Their success, however, is limited by the presence of large pose variations, expression and occlusion in face images. Moreover, existing correlation filters may suffer from poor discrimination to distinguish visually similar landmarks such as the right and left eyes. In this work, we present a new framework, referred to as Correlation Filter Cascade, to address the above limitations. The proposed framework consists of a set of correlation filters with different spatial supports (sizes) which are connected together in a cascade form. More specifically, the size of filters decreases from the lower to upper levels. Filters at lower levels implicitly code face shape information since they are trained using large patches stemmed from face images. This avoids ambiguous detections caused by landmarks with similar appearance. Detections in these levels, however, may not be accurate and suffer from small localization errors, mainly caused by face pose, expression and occlusion. Therefore, locations detected by lower levels will be further used by the higher levels to narrow down their search regions. Since the filters at higher levels have smaller size, they are less affected by pose, expression and occlusion, and thus, can perform more accurately. The evaluation on BioID and LFPW shows the superiority of our method compared to prior correlation filters and leading facial landmark detectors.

1. Introduction

Localizing facial landmarks has been widely studied by the vision community, due to its critical role for the task of face recognition and analysis [2, 24, 25, 8, 22, 28, 19]. This is a very challenging task when face images are captured under uncontrolled imaging conditions such as extreme pose, illumination, expression and occlusions. Existing approaches can be generally categorized in two classes: local regions [1, 2, 16, 29, 26] and global shape based approaches [20, 27, 17, 19, 21, 22].

Approaches in the first class involve training a part detector for each facial landmark, and localization is performed by classifying local patches in a face image as either landmark or non-landmark patch. Promising results of these approaches have been reported, especially in the presence of severe face pose, expression and occlusion [29, 2]. They, however, are not robust against ambiguous detections, in which a detector may return multiple visually similar candidate regions (e.g., the left and right eyes). In this case, spatial constraints can be applied (as post-processing) to find an optimal configuration of facial landmarks and remove the wrong candidates [10, 16].

The global shape based approaches, on the other hand, detect landmarks by exploiting visual information of the entire face or a set of local patches around the target landmark. This is typically followed by a post-processing step that refines the initial detection using face shape constrains. The Active Appearance Model [5], Active Shape Model [6] and their variations [2, 19, 20, 27] are typical methods which employ face global shape and geometric information for robust landmark detection. These approaches have shown to be capable of handling wrong local detections caused by visual similarity. They, however, are not robust against severe face pose, expression and occlusion. Moreover, optimization strategies (e.g. gradient descent) used by these approaches are very sensitive to the initialization and may suffer from local minimum. Furthermore, optimizing shape constraints might be time consuming and, thus, not appropriate for real time landmark localization.

Correlation filters, initially developed by Hester and Casasent [11], have been applied on many vision tasks [11, 15, 18, 4, 3, 13, 14, 9]. In particular, correlation filters aimed at learning a filter/template that returns corresponding desired outputs when correlated with a set of training images. Interest in correlation filters in the vision community has increased through the works of Bolme et al. on Minimum Output Sum of Squared Error (MOSSE) correlation filters [3]. This work addressed some of the classical problems with earlier correlation filters (e.g. overtraining and poor generalization) and was extremely effi-

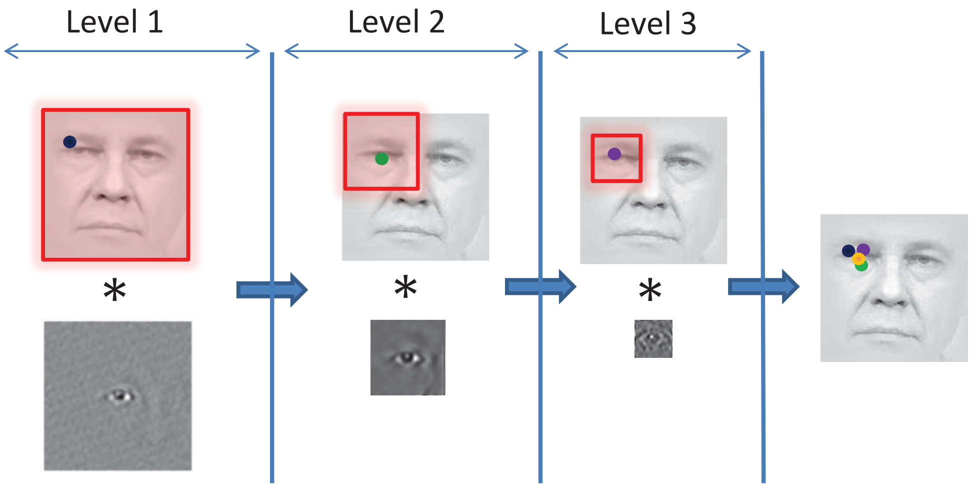


Figure 1. Correlation filter cascade for facial landmark detection. The size of filters decreases from the lower to upper levels. Detection at each level is used by the next level to limit its search region, shown by the red square. The color filled circles show the predicted landmark at each level. The final localization is performed by averaging all locations detected by all levels, the yellow filed circle. Best viewed in color.

cient in terms of computation and memory usage. Recently, Kiani et al. [13] introduced Multi-Channel Correlation Filters (MCCFs) to employ discriminative descriptors (e.g., HOG [7]) for learning multi-channel filters/detectors efficiently in the frequency domain. More recently, they addressed the problem of boundary effects during filter training using correlation filters with limited boundaries [9].

All of these correlation filter techniques were evaluated for landmark localization with very promising results. Almost all of these works (except [9]) used whole face images to train landmark correlation filters, which, as mentioned earlier, may not be robust against pose variations, expression and occlusion. In a different manner, the approach introduced by [9] used all possible patches within in a face image to train a correlation filter with much smaller size. Compared to other techniques, this approach was more robust against pose, expression and occlusion, but suffered from wrong detections caused by visually similar landmarks.

To deal with above limitations, we introduce a cascade framework for the task of facial landmark localization. Fig. 1 depicts the scheme of the proposed framework. In particular, our framework consists of a set of correlation filters with different spatial supports (sizes) which are hierarchically connected together in a cascade manner. The size of filters decreases from lower to higher levels, meaning that filters at lower levels have bigger size compared to those at higher levels. Filters at lower levels are trained using bigger patches (for instance, the filter at level 0 in Fig. 1 is trained using whole face images) and explicitly capture face shape information. This offers stability against ambiguous detections. Filters at higher levels, on the other hand, are trained using smaller patches, and, as a result, are more robust against uncontrolled face pose, occlusion and expression. The correlation outputs returned by each level is used to narrow down the search region for its upper level. The final landmark location is determined by averaging the locations estimated by all filters over the cascade framework.

2. Cascade Correlation Filters

The formulation of Multi-Channel Correlation Filters (MCCFs) in the spatial domain is defined as [13],

$$E(\mathbf{h}) \quad = \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D}||\mathbf{y}_i(j) - \sum_{k=1}^{K}\mathbf{h}^{(k)T}\mathbf{x}_i^{(k)}[\Delta\boldsymbol{\tau}_j]||_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K}||\mathbf{h}^{(k)}||_2^2 \qquad (1)$$

where $\mathbf{x}^{(k)}$ and $\mathbf{h}^{(k)}$ refers to the kth channel of the vectorized image/frame and filter respectively, and K represents the number of filters. D and N denote the length of vectorized image and number of training examples, respectively. $\mathbf{a}[\Delta\boldsymbol{\tau}_j]$ performs a j-step discrete circular shift to a vector $\mathbf{a}$, and λ is the regularization term. $\mathbf{y}_i(j)$ refers to the jth element of the vectorized correlation output $\mathbf{y}_i$. Any transpose operator ^{T} on a complex vector or matrix in this paper additionally takes the complex conjugate.

Solving this multi-channel objective in the spatial domain suffers from a high complexity of $\mathcal{O}(D^3K^3 + ND^2K^2)$ since one has to solve a $KD \times KD$ linear system [13]. Inspired by the efficiencies of posing single channel correlation filters in the frequency domain [3], Equation 1 can be expressed equivalently as,

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N} ||\hat{\mathbf{y}}_i - \sum_{k=1}^{K} \text{diag}(\hat{\mathbf{x}}_i^{(k)})^T \hat{\mathbf{h}}^{(k)}||_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K} ||\hat{\mathbf{h}}^{(k)}||_2^2 \quad (2)$$

where $\hat{\mathbf{h}} = [\hat{\mathbf{h}}^{(1)T}, \ldots, \hat{\mathbf{h}}^{(K)T}]^T$ is a KD dimensional supervector of the Fourier transforms of each channel. This can be simplified further,

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N} ||\hat{\mathbf{y}}_i - \hat{\mathbf{X}}_i\hat{\mathbf{h}}||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \quad (3)$$

where $\hat{\mathbf{X}}_i = [\text{diag}(\hat{\mathbf{x}}_i^{(1)})^T, \ldots, \text{diag}(\hat{\mathbf{x}}_i^{(K)})^T]$ and the solution for Equation 3 becomes,

$$\hat{\mathbf{h}}^* = (\lambda\mathbf{I} + \sum_{i=1}^{N} \hat{\mathbf{X}}_i^T \hat{\mathbf{X}}_i)^{-1} \sum_{i=1}^{N} \hat{\mathbf{X}}_i^T \hat{\mathbf{y}}_i \quad (4)$$

The cost of solving this linear system looks no different to the spatial domain as one still has to solve a $KD \times KD$ linear system. It is, however, shown in [13] that $\hat{\mathbf{X}}$ is sparse banded and Equation 3 can be efficiently solved through a variable re-ordering with smaller cost of $\mathcal{O}(DK^3 + NDK^2)$. Readers are encouraged to refer to [13, 3] for more details of multi/single channel CFs and their efficient memory usage and computational complexity.

Our proposed cascade framework contains a set of MCCFs arranged in a cascade form, denoted by $\{\hat{\mathbf{h}}_l\}_{l=0}^{L-1}$, where L is the number of cascade levels. Suppose that $\hat{\mathbf{h}}_l$ indicates the MCCF with the size of $h_l \times w_l \times K$ learned for level l (Eq 4)[1]. The size of filters decreases from the lower to higher levels, meaning that $h_i \leq h_j$ and $w_i \leq w_j$ if filter i is located at a higher level than filter j over the cascade.

Intuitively, the filter at the lowest level is trained using full face images. Thus, it implicitly encodes face shape and geometric information [13]. This significantly reduces ambiguous detections caused by visually similar patches. This advantage, however, comes at the high cost of inaccurate localization, especially when a face image displays severe pose, expression and occlusion. A localization is inaccurate if there is no ambiguous detection, but the landmark is detected with small spatial error [13]. The filter at the second level is trained using smaller patches with less face shape information. This increases the risk of ambiguous detection, while being more robust to face pose, expression and occlusion. To avoid ambiguous detection, the filter at the second level only explores the local region around the location predicted by the first level (red squers in Fig.1) to find the landmark. This procedure repeats till the last level. The final localization is determined by averaging all locations detected by all filters over the cascade.

Training cascade filters. Eq 4 states that the size of MCCF is same as the size of training samples. Therefore, we employ patches with different size to train filters in our cascade framework. To this end, we crop patches with size of $h_l \times w_l$ from face images centered upon the landmark of interest (e.g., the right eye) and use them to train filter $\hat{\mathbf{h}}_l$ at level l using Eq 4. This provides a set of MCCFs with different size which we further use to form our cascade framework.

Testing cascade filters. Given a test image, the MCCF at the first level is correlated over the whole face image and the maximum peak of the correlation output is selected as approximated landmark location. The filter at the second level is only correlated over the local region around the location provided by the first level. This procedure is performed for all levels of the cascade. The average of locations detected by all filters is considered as the final landmark localization.

3. Experimental Results

We evaluated the proposed framework on two publicly available datasets, BioID [12] and LFPW [2].

3.1. Datasets

LFPW (Labeled Face Parts in the Wild) All face images in this dataset are downloaded from the web and represent large variations in pose, illumination, expression and occlusion. The original version of this dataset contains 1100 training and 300 testing images. Since this dataset provides only web image URLs, some of URLs are not currently available. Therefore, we only downloaded 714 training and 214 testing images.

BioID dataset consists of 1521 near frontal face images of 23 subjects captured with various scales and face expressions in lab environment. This dataset is commonly used to evaluate most of the previous methods, allowing to compare our method to them. We used the evaluation procedure provided by [24], where 1000 images are randomly selected from the dataset for training and the rest for testing.

3.2. Implementation Details

Here, we explain implementation details and investigate the number of cascade levels using LFPW dataset.

Face Bounding Box. The face detector proposed by Viola and Jones [23] is applied to find the face bounding box, which is further enlarged by 20% in order to ensure that all facial landmarks are enclosed. All boxes are resized to have

[1] Please note that any type of CFs can be used in our framework (e.g., MOSSE [3] and ASEF [4]). We employ MCCF to exploit multi-channel features such as HOG in our framework.

the size of 128 × 128 pixels. We assumed that there is only one face in each image and all images are in gray scale.

Desired Correlation Outputs. A 2D Gaussian function with spatial variance of 2 is employed to define the desired correlation outputs whose peak is located upon the center of the target landmark, following [13, 3].

Landmark Detection and Evaluation: Following the previous works, we use the normalized inter-ocular distance, $d = \frac{\|\mathbf{p}_i - \mathbf{g}_i\|_2}{\|\mathbf{m}_l - \mathbf{m}_r\|_2}$, for evaluation, where $\mathbf{m}_r$ and $\mathbf{m}_l$ respectively indicate the true coordinates of the right and left eye, and $\mathbf{p}_i$ and $\mathbf{g}_i$ are the predicted and ground truth locations of the *i-th* landmark, respectively. A localization with distance $d < th$ is considered as a successful localization. The threshold *th* is set to a fraction of inter-ocular distance (0.10 in our experiments).

Feature Extraction. We extract 43 feature channels for each face image, including 40 Gabor features (eight different orientations and five scales), two Sobel features (horizontal and vertical gradient magnitudes) and one power-normalized intensity image. A Cosine-window is applied on all feature channels to reduce the frequency effects caused by opposite image borders in the Fourier domain [3, 13].

Number of Cascade Levels. We investigate the performance of our framework with respect to the number of cascade levels. For this purpose, we trained five different MCCFs of size 128 × 128, 64 × 64, 32 × 32, 16 × 16 and 8 × 8 using 128 × 128 LFPW training images. Then we employed these filters to form five cascade correlation filters with different number of levels, namely L0 (one level by 128 × 128 filter), L1 (two levels by 128 × 128 and 64 × 64 filters), L2 (three levels by 128 × 128, 64 × 64 and 32 × 32 filters), L3 (four levels by 128×128, 64×64, 32×32 and 16×16 filters) and L4 (five levels, including all the correlation filters).

The localization rate of these cascade filters for detecting 10 landmarks of LFPW is shown in Fig. 2. The lowest and highest localization rates belong to L0 and L4 detectors, respectively. Adding smaller filters in the cascade framework improves the detection performance, especially for mouth and lip landmarks which can be potentially more affected by pose and expression.

Figure 4 depicts how adding more levels over the cascade framework improves the detection performance for several examples with partial occlusions, pose variation and expression. Detection at the first level is not very accurate, particularly for occluded and under expression landmarks. The filter at the first level is trained using whole face images and, thus, is sensitive to pose, expression and occlusion. This results in inaccurate detections. These errors are improved over the next levels using smaller filters. Moreover, the search area proposed by the first level avoid smaller filters to detect ambiguous landmarks.

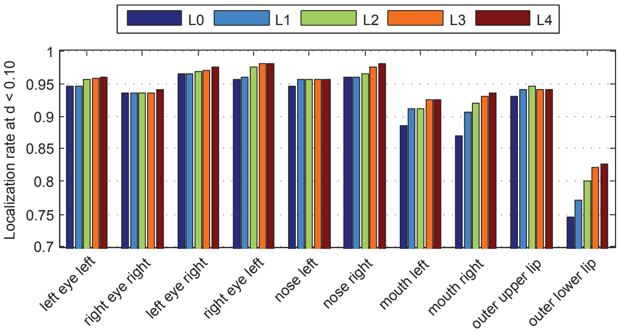


Figure 2. Evaluating the cascade framework with different number of levels on LFPW dataset.

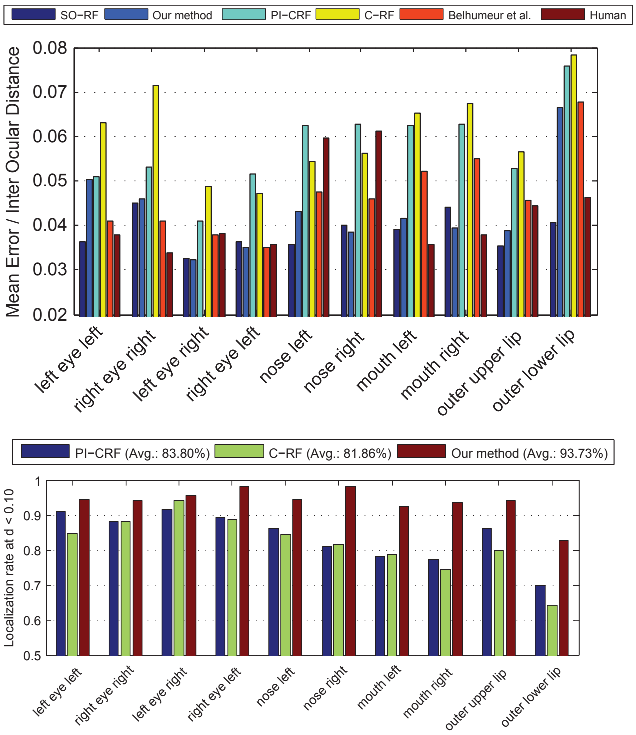


Figure 3. Comparing the proposed approach to state of the art methods on the LFPW dataset. (top) Mean normalized localization error, and (bottom) localization accuracy at threshold of $d < 0.10$.

3.3. Comparison with the-state-of-the-art

We compare our framework with the-start-of-the-art and leading landmark detectors in the literature on both LFPW and BioID datasets, including Structured-Output Regression Forests (SO-RF) [24], Conditional Regression Forests (C-RF) [8], Privileged Information-based Conditional Regression Forest (PI-CRF) [25], Belhumeur's part based detectors [2], Boosted Regression [22], Exemplar-based Graph Matching [28], Extended Active Shape Model [19].

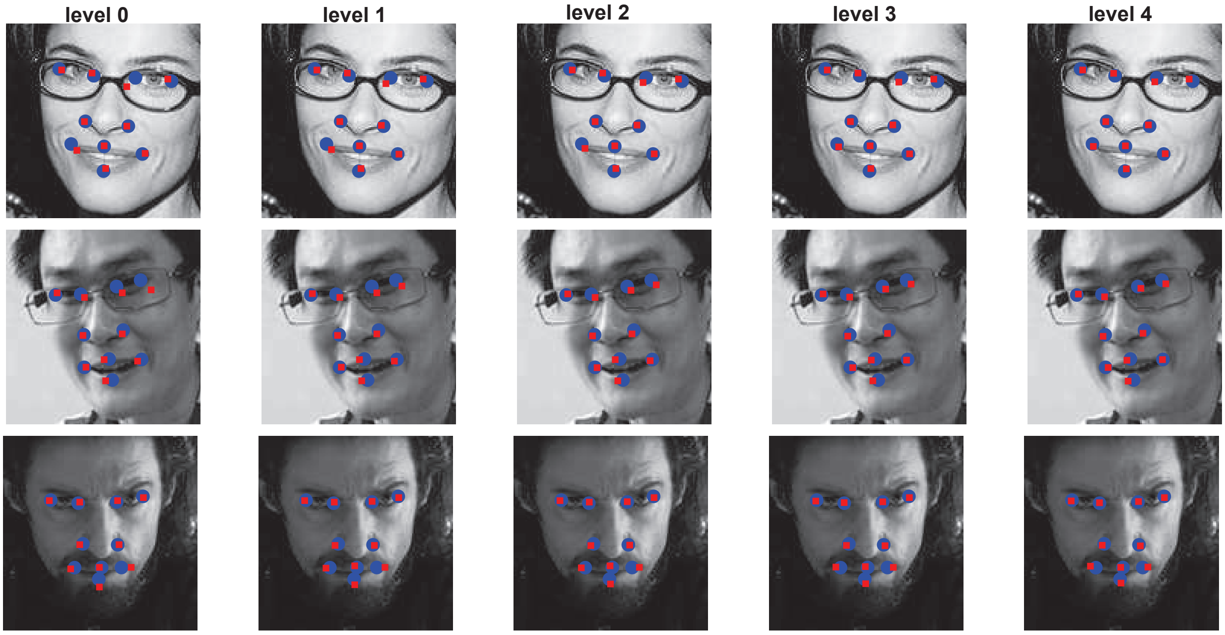


Figure 4. Accurate landmark detection over cascade levels. The ground truth and predicted locations are shown by blue dots and red squares, respectively. The images are selected from the LFPW testing set (best viewed in color).

Figure 3 illustrates the results of this comparison. Our method is trained using 714 available training and tested on 214 available testing examples. The results of other approaches are borrowed from their reference papers, where they evaluated their approach using 821-870 and 214 available training and testing examples, respectively. We also compared our approach to the performance of human annotators, borrowing from [24].

According to Figure 3, our approach outperformed all other approaches except SO-RF [24]. The main reason is that SO-RF considers shape constraints to remove wrong detections. This is, however, not the case in our method. Lacking direct shape constraints for landmark detection sometimes leads to wrong detections with large location errors. Even a very small number of these wrong detections will result in a very large average error. This rarely happens in approaches with shape constraints. Our method obtained superior performance compared to Conditional Regression Forests (C-RF) [8], Privileged Information-based Conditional Regression Forest (PI-CRF) [25], part based detector [2] for both mean detection error and localization rate. The average localization rate of our method for all landmarks is 93.73% compared to 81.86% of C-RF and 83.80% PI-CRF, Figure 3 (bottom). The accuracy of our method is very competitive to human annotators. We even achieved lower error for five landmarks, *left eye right*, *right eye left*, *nose left*, *nose right* and *outer upper lip*. Figure 6 depicts landmarks detected by our approach on LFPW images.

The result on BioID dataset is shown in Figure 5, comparing our method to the state of the art on this dataset. Figure 5(a) shows the cumulative error versus the fractions of inter-ocular distance d for m_e17 (for 17 facial landmarks of all 19 internal landmarks). These results are reported by [28], [2], [22] and [24]. This comparison shows the superiority of our method against all the other approaches for (almost) all fractions of inter-ocular distance. Figure 5(b) illustrates the mean detection error normalized by inter-ocular distance of our method, [22] and [24] for all 19 internal landmarks (chain landmark is discarded). We borrowed the parts ID from [24]. The mean error of P9 and P14 is not reported in [22]. The results shows that our method achieved the lowest errors for 15 of 19 landmarks. Figure 7 visualizes the landmarks of some BioID images localized by our method.

Detection speed. Aside from the competitive accuracy, our approach achieved superior detection speed by localizing each landmark within 400 face images in one second (2.5 ms to detect a landmark in a 128 × 128 face image), which is 16 times more faster than the state-of-the-art approaches with real-time detection speed (25 face images per second) reported by [24], [8] and [25]. This detection time includes all steps in the framework, including computing FFT/IFFT of filter and image channels, generating correlation outputs and evaluating the correlation outputs for final detection, excluding face detection in the image and feature extraction.

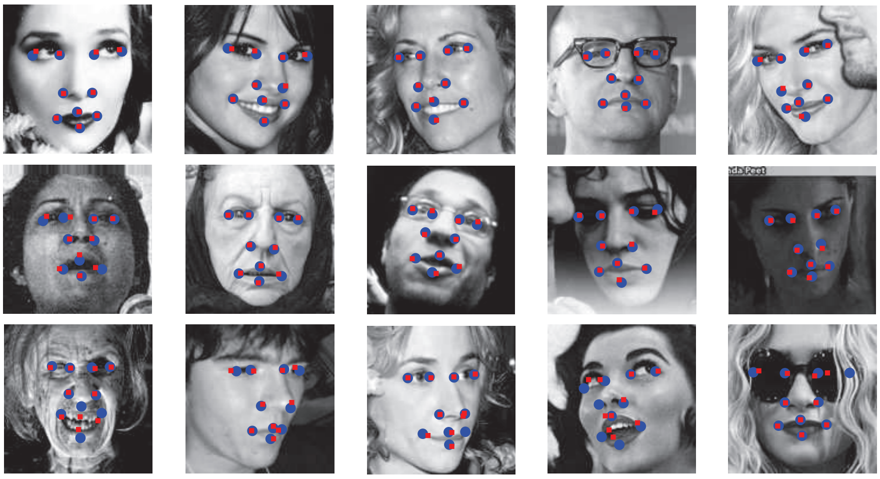


Figure 6. Detection examples of the LPFW dataset. The first two rows show the successful detections under challenging circumstances of expression, occlusion, pose, lighting and poor quality. The third row shows some failed cases. The red and blue marks respectively show the detected landmark and the ground truth (best viewed in color).

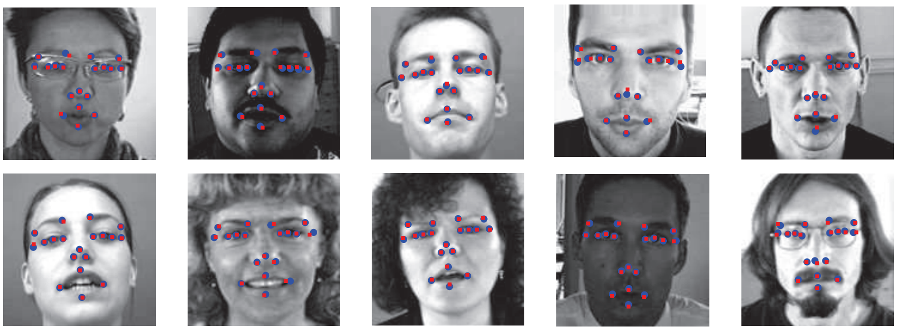


Figure 7. Detection examples of the BioID dataset. The red squares and blue dots represent the detected and ground truth landmarks, respectively (best viewed in color).

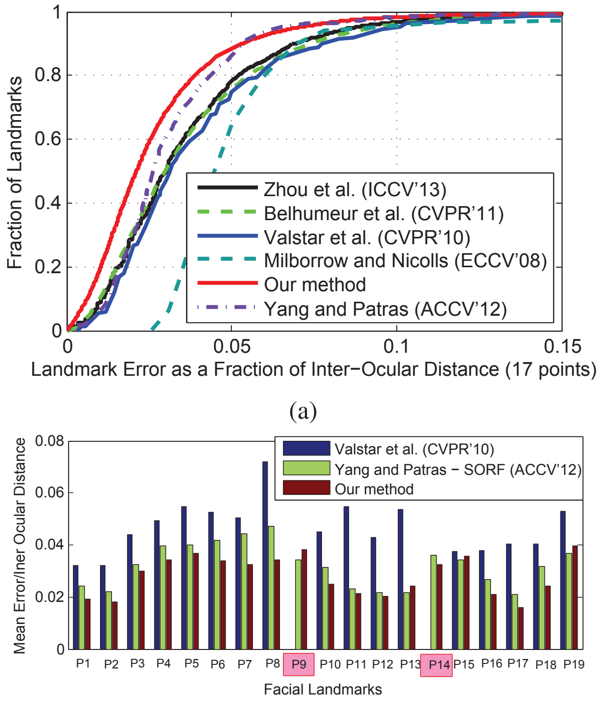
3.4. Comparison with Prior Correlation Filters

This experiment compares cascade correlation filters with the prior CF-based approaches, including single-channel correlation filters (*MOSSE*) [3], correlation filters with limited boundaries (*CF w LB*) [9], multi-channel correlation filters (*MCCF*) [13] and the cascade framework (*Cascade CF*) using the BioID dataset.

Similarly, 1000 face images are randomly selected for training and the rest for testing. We use normalized image intensities to train and test the single-channel correlation filters (MOSSE and CFwLB), and 43 feature channels (40 Gabor magnitudes, 2 Sobels and one normalized intensities) to train and test the multi-channel features (MCCF and our approach). We employ 64×64 landmark patches cropped from the face images (centered upon the landmark of interest) to train the *MOSSE* and *MCCF*. For correlation filters with limited boundaries, we employ full 128×128 face images to train 64×64 landmark filters.[2] Similarly, we use a five-levels cascade correlation filters. All these filters are applied on face images of size 128×128 for testing.

The result is illustrated in Figure 8, showing that the lowest and the best localization rate is obtained by MOSSE (77.45 %) and cascade CF (98.14 %), respectively. MOSSE obtained the worse accuracy since it employs image intensities which are not discriminative enough to distinguish patches with similar appearance (ambiguous detections). Moreover, as argued in [13], image intensities are not capable of handling low image quality and illumination. The localization rate of MCCF and CFwLB is much better than MOSSE. MCCF filters are trained using multi-channel features which are more discriminative to image quality and illumination, compared to image intensities [13]. During the

[2] we examined different sizes of MOSSE, MCCF and CFwLB including $16 \times 16, 32 \times 32, 64 \times 64$ and 128×128. Filters with size of 64×64 showed higher performance and, thus, selected for this experiment.



(a)

(b)

Figure 5. Comparison on BioID. (a) Average detection rate as a function of fraction of inter-ocular distance. (b) mean normalized error for each landmark. The parts (landmarks) ID are defined in [24]. The mean error of P9 and P14 is not reported in [22].

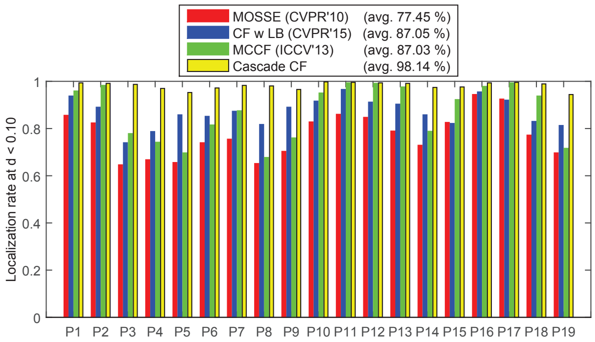


Figure 8. Comparing with prior correlation filters on BioID. Our approach significantly outperformed the other CFs techniques.

evaluation, we observed that most failure cases in MCCF and CFwLB are caused by face pose, expression and occlusion. Our approach significantly outperformed MOSSE, MCCF and CFwLB, since: (i) we use multi channel features which are robust to image quality and illumination, (ii) we use large size CFs which are able to handle ambiguous detections, and (iii) we use small size CFs to deal with face posing, expression and occlusion.

4. Conclusion

In this paper, we proposed cascade correlation filters for the task of facial landmark detection. In this framework, multiple correlation filters with different sizes are cascaded in a way that their sizes decrease upwards to the higher levels. The filter at the first level predicts the location of the target landmark. This predication is robust to ambiguous detections, while may suffer from smal localization errors originated from face pose, expression and occlusion. The position error in the first level is then refined by smaller filters at the higher levels of the cascade. The evaluation on the LPFW and BioID datasets demonstrated the superiority of our approach compared to the state of the art correlation filters and leading no-filter approaches.

References

[1] B. Amberg and T. Vetter. Optimal landmark detection using shape models and branch and bound. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 455–462. IEEE, 2011.

[2] P. N. Belhumeur, D. W. Jacobs, Kriegman, D., and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.

[4] D. S. Bolme, B. A. Draper, and J. R. Beveridge. Average of synthetic exact filters. In *CVPR*, 2009.

[5] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. volume 23, pages 681–685, 2001.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. volume 61, pages 38–59. Elsevier, 1995.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[8] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2578–2585. IEEE, 2012.

[9] H. K. Galoogahi, T. Sim, and S. Lucey. Correlation filters with limited boundaries. In *Computer Vision and Pattern Recognition, 2015. CVPR'15. IEEE Conference on*, 2015.

[10] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *Computer Vision–ECCV 2008*, pages 413–426. Springer, 2008.

[11] C. F. Hester and D. Casasent. Multivariant technique for multiclass pattern recognition. *Appl. Opt.*, 19(11):1758–1761, 1980.

[12] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[13] H. Kiani, T. Sim, and S. Lucey. Multi-channel correlation filters. In *ICCV*, 2013.

[14] H. Kiani, T. Sim, and S. Lucey. Multi-channel correlation filters for human action recognition. In *ICIP*, 2014.

[15] B. V. K. V. Kumar. Minimum-variance synthetic discriminant functions. *J. Opt. Soc. Am. A*, 3(10):1579–1584, 1986.

[16] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *Computer Vision–ECCV 2008*, pages 72–85. Springer, 2008.

[17] X. Liu. Generic face alignment using boosted appearance model. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[18] A. Mahalanobis, B. V. K. V. Kumar, and D. Casasent. Minimum average correlation energy filters. *Appl. Opt.*, 26(17):3633–3640, 1987.

[19] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *Computer Vision–ECCV 2008*, pages 504–513. Springer, 2008.

[20] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. volume 91, pages 200–215. Springer, 2011.

[21] P. Sauer, T. F. Cootes, and C. J. Taylor. Accurate regression procedures for active appearance models. In *BMVC*, pages 1–11, 2011.

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

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[24] H. Yang and I. Patras. Face parts localization using structured-output regression forests. In *Computer Vision–ACCV 2012*, pages 667–679. Springer, 2013.

[25] H. Yang and I. Patras. Privileged information-based conditional regression forest for facial feature detection. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[26] X. Zhao, X. Chai, Z. Niu, C. Heng, and S. Shan. Context modeling for facial landmark detection based on non-adjacent rectangle (nar) haar-like feature. volume 30, pages 136–146. Elsevier, 2012.

[27] X. Zhao, S. Shan, X. Chai, and X. Chen. Locality-constrained active appearance model. In *Computer Vision–ACCV 2012*, pages 636–647. Springer, 2013.

[28] F. Zhou, J. Brandt, and Z. Lin. Exemplar-based graph matching for robust facial landmark localization. In *ICCV*, pages 1–6. IEEE, 2013.

[29] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.