Image and Vision Computing 30 (2012) 136–146

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Context modeling for facial landmark detection based on Non-Adjacent Rectangle (NAR) Haar-like feature[☆]

Xiaowei Zhao [a,b], Xiujuan Chai [a], Zhiheng Niu [c], Cherkeng Heng [c], Shiguang Shan [a,*]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[c] Panasonic Singapore Laboratories Pte Ltd (PSL), Tai Seng Industrial Estate 534415, Singapore

ARTICLE INFO

Article history:
Received 6 July 2011
Received in revised form 30 November 2011
Accepted 4 December 2011

Keywords:
Context modeling
Face detection
Facial landmark detection
NAR Haar-like feature
Co-occurrence

ABSTRACT

Automatically locating facial landmarks in images is an important task in computer vision. This paper proposes a novel context modeling method for facial landmark detection, which integrates context constraints together with local texture model in the cascaded AdaBoost framework. The motivation of our method lies in the basic human psychology observation that not only the local texture information but also the global context information is used for human to locate facial landmarks in faces. Therefore, in our solution, a novel type of feature, called Non-Adjacent Rectangle (NAR) Haar-like feature, is proposed to characterize the co-occurrence between facial landmarks and its surroundings, i.e., the context information, in terms of low-level features. For the locating task, traditional Haar-like features (characterizing local texture information) and NAR Haar-like features (characterizing context constraints in global sense) are combined together to form more powerful representations. Through Real AdaBoost learning, the most discriminative feature set is selected automatically and used for facial landmark detection. To verify the effectiveness of the proposed method, we evaluate our facial landmark detection algorithm on BioID and Cohn-Kanade face databases. Experimental results convincingly show that the NAR Haar-like feature is effective to model the context and our proposed algorithm impressively outperforms the published state-of-the-art methods. In addition, the generalization capability of the NAR Haar-like feature is further validated by extended applications to face detection task on FDDB face database.

1. Introduction

Automatically locating facial landmarks in images is an important task in computer vision. It is useful for various practical applications, such as face recognition, facial attribute analysis, and intelligent human computer interface. Yet, facial landmark detection still remains challenging because the appearance of facial landmarks can vary a lot due to tremendous variations in pose, lighting, occlusion, and low quality imaging, etc.

To solve this problem, many methods have been proposed in recent years. One of the most widely used algorithms is the cascaded AdaBoost framework [1], which has remarkably improved the accuracy and efficiency of facial landmark detection. Typically, it is a local texture-based method, where Haar-like features are used to describe local textures around the target facial landmark (e.g., mouth corners, and the corners of the eyes). Based on the extracted Haar-like features, a cascaded classifier is built by exploiting AdaBoost learning and applied to exhaustively analyze the patches or windows of each testing image at all positions and at multiple scales. When a patch is extracted from the testing image, it is classified according to its local appearance and associated with a detection score. Similar AdaBoost-based methods can be found in [2,3]. In work [2], 2D cascaded AdaBoost framework is used for eye detection, which bootstraps both positive and negative samples in training stage. In work [3], a Gabor facial feature point detector (Gabor-ffpd) has been developed, where GentleBoost is used to learn classifiers and local textures are extracted by Gabor filters.

Although the local texture-based methods have been quite successful, one major drawback of these methods is that sometimes it is insufficient to distinguish facial landmarks from background by just exploring the local appearance information. Especially under complex environment, multiple similar modes in face region maybe undistinguished, or no matched mode can be detected in case of partial occlusions or image blur. In such situation, it is expected that the global facial information can be helpful for facial landmark detection. Actually, in the real world, objects tend to co-vary with other objects and particular environments (i.e., the context information) [4], and the visual systems of human and animal use these context constraints to improve their ability of object recognition. Specifically, taking the

[☆] This paper has been recommended for acceptance by special issue Guest Editors Rainer Stiefelhagen, Marian Stewart Bartlett and Kevin Bowyer.

[*] Corresponding author. Tel.: +86 10 6260 0548.
E-mail address: sgshan@ict.ac.cn (S. Shan).

0262-8856/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2011.12.004

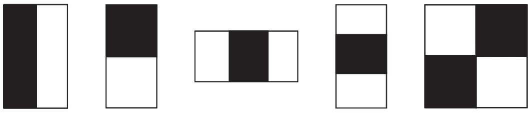

Fig. 1. Widely used Haar-like features.

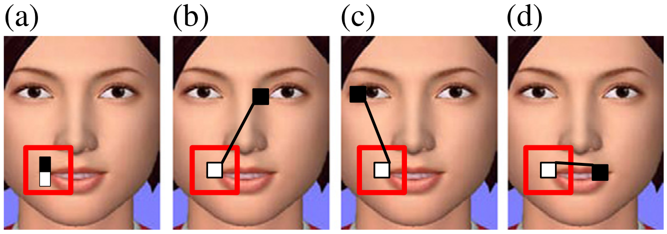

Fig. 2. Typical examples of traditional and NAR Haar-like features. Figure (a) shows a typical traditional Haar-like feature, and figures (b), (c), and (d) show typical NAR Haar-like features.

task of locating left mouth corner as an example, the position of left mouth corner is not only predicted by local textures around it but also predicted by other local texture patterns, such as nose and eyes.

Recently, the notion of exploiting context information to improve facial landmark detection has been increasingly realized [5–8]. For example, T. Kozakaya et al. present a weighted vector concentration approach [5]. They first place regular sampling points in the face region, and extract local pattern descriptors at each sampling point. Subsequently, both the local patterns and geometrical relationship between the sampling points and the target facial landmark are simultaneously learned. So, the location of the target facial point is predicted by local patterns extracted from the sampling points. Similarly, M. Valstar et al. propose a regression-based method for facial landmark detection [6]. In this method, each local texture pattern is related with a vector pointing to the target facial landmark. Regressors are learned based on these local patterns and vectors. So, the localization of the target facial landmark is voted by all regressors. In addition, Markov Random Filed model of all facial landmarks is constructed to prevent unfeasible predictions. Fink et al. also propose Mutual Boosting to incorporate context information to augment the detection of facial landmarks [7]. They train AdaBoost-based detectors for multiple facial landmarks in parallel. For each detector, the remaining intermediate detectors are used to enrich the weak learner set. Additionally, “auto-context” algorithm is proposed to learn an effective and efficient context model, together with an image appearance model [8]. It integrates the local image appearance model with the context information by learning a series of classifiers. The first classifier is trained to assign each pixel a confidence and produces a classification map for each image. Then, in the later stage, local texture features and context features extracted from the classification map are used to train classifiers.

In the above-mentioned context constrained methods, the context information is either modeled based on the classification map or densely extracted within the face region in detection stage. Different from these methods, this paper proposes a simple but discriminative feature to model the context. Essentially, it is a generalization of the widely used Haar-like features. We remove the constraint that the rectangles of Haar-like feature must be adjacent. In this case, it can model not only the short range but also the long range co-occurrence relationships of local texture patterns. To differentiate from the traditional Haar-like features, the proposed feature is named as Non-Adjacent Rectangle (NAR) Haar-like feature. The most significant advantage of the NAR Haar-like feature is that it can model the co-variation relationships of local textures implicitly by the co-occurrence of local texture features. For example, if we want to represent the co-occurrence relationship of left mouth corner and nose tip, it is convenient by using two separate rectangles which centered on left mouth corner and nose tip respectively. Although it's obvious that this feature represents the co-variation relationship of left mouth corner and nose tip in semantic level, it just represents the lighter or darker pattern of local textures in feature level. It must be pointed out that similar features were mentioned in [15] for face detection, but without details and experimental validations. More importantly, it is not designed for context modeling as in this work.

In our method, we try to combine the traditional Haar-like feature, which can characterize the local texture information, with the proposed NAR Haar-like feature to form a more representative feature set. Among this huge feature set, the most discriminative features are selected automatically through Real AdaBoost learning. In addition, with the computation advantage of integral image, the Non-Adjacent Rectangle (NAR) Haar-like feature can also be extracted efficiently. We evaluate the effectiveness of the proposed algorithm and the context modeling capability of the NAR Haar-like feature on BioID [9] and Cohn-Kanade [10] face datasets with respect to facial

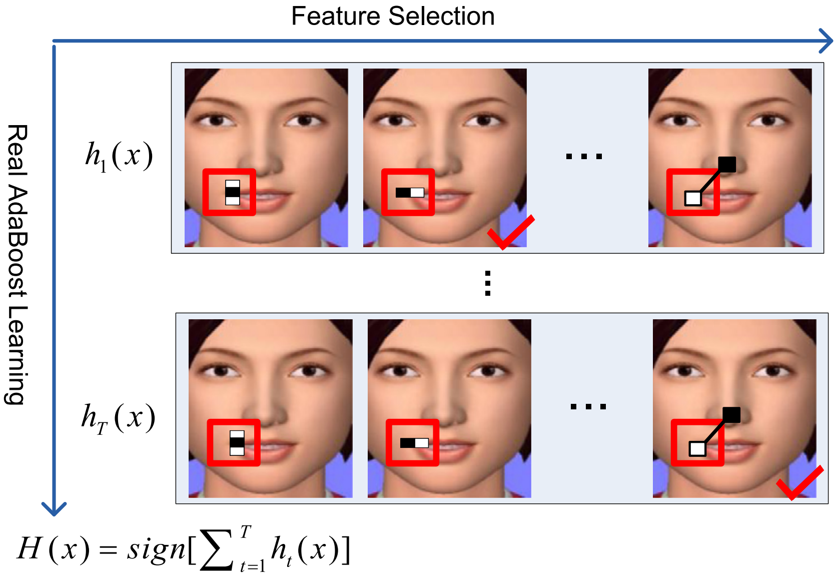

Fig. 3. Sketch map of classifier learning procedure with both traditional and NAR Haar-like features.

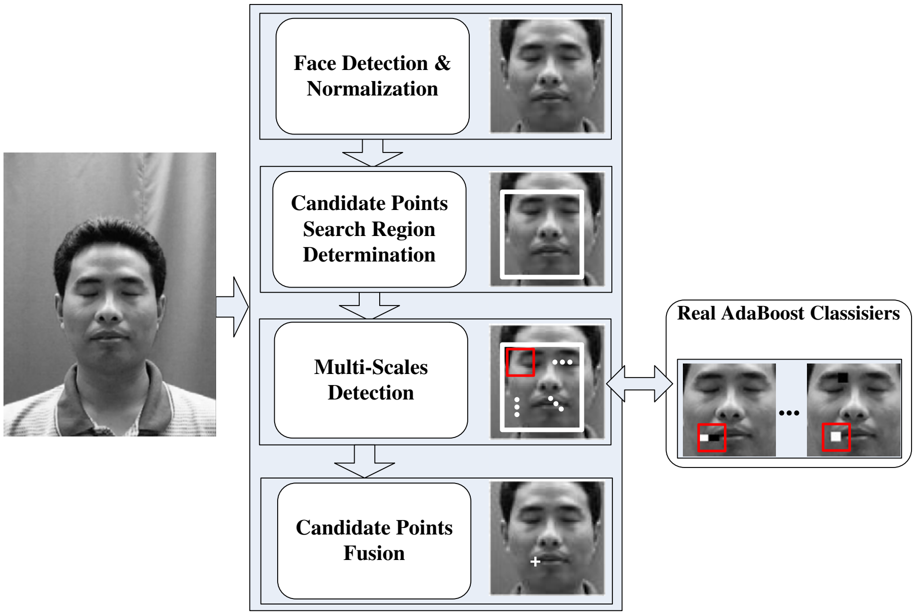

Fig. 4. Sketch map of our facial landmark detection procedure.

landmark detection task. Experimental results demonstrate that the proposed method is promising and outperforms many published state-of-the-art methods.

To verify the generalization capability of the NAR Haar-like feature, we further apply it to model context for face detection task. In addition, we deeply investigate the capability of the NAR Haar-like feature to characterize the local texture for face detection. Experiments conducted on the FDDB face dataset [11] convincingly show that the proposed NAR Haar-like feature generalizes well to model context for face detection, but just a slight improvement is obtained when the NAR Haar-like feature is used to model the local texture information for face detection. So, the proposed NAR Haar-like feature is more suitable to model context information.

This paper is an extension of our previous conference work [12]. In this extension work, we further validate the generalization capability of the NAR Haar-like feature to model the context for face detection. The remaining part of the paper is organized as follows. Section 2 introduces the proposed NAR Haar-like feature and the procedure of classifier learning. Sections 3 and 4 are the evaluations of our algorithm on facial landmark detection and face detection tasks respectively. Detailed experimental results and analysis are also given in these two sections. Section 5 concludes the paper.

2. Non-Adjacent Rectangle (NAR) Haar-like feature for context modeling

2.1. Haar-like feature and other local features

Haar-like feature [1] is a typical local texture descriptor, which measures the average intensity differences of the adjacent rectangle regions. Fig. 1 illustrates the widely used Haar-like features. Wealth of information, such as the intensity gradient at different locations, spatial frequencies and directions, can be captured by Haar-like features when we change the position, size, shape, and arrangement of rectangular regions. In literature [1], Viola and Jones use three kinds of Haar-like features to detect faces with very high efficiency. However, traditional Haar-like features are too simple and show some limits [13]. To enhance the capability of Haar-like features, many kinds of variations have been proposed, such as joint Haar-like feature [13], rotated Haar-like feature [14], and block difference feature [15].

Besides, many other types of local features are proposed to characterize the local appearance information of image patches, such as Gabor wavelet [16], SIFT [17], HOG [18], LBP [19], WLD [20], LAB [21], etc. In some specific applications, these features show good performance. For example, SIFT feature performs well in object matching and recognition area due to its invariance to scaling and rotations [22]. And the Gabor wavelet is widely used in image analysis applications, including texture classification, image registration and face recognition, etc.

However, the above-mentioned features are all local texture descriptors in essence. They can only characterize local texture information, while ignoring the facial context information.

2.2. NAR Haar-like feature

In local texture-based methods, features are usually extracted within a local image patch. As for the context constrained method, a larger region should be defined to include the context information. In order to characterize the co-variation relationship of the local image patch and its context, we generalize the traditional Haar-like feature by

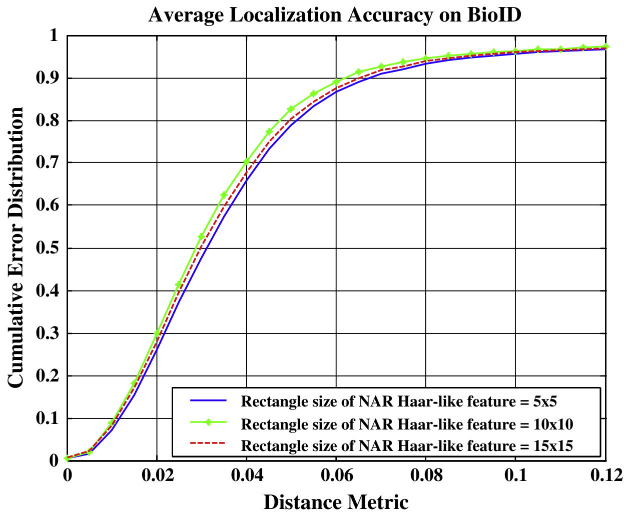

Fig. 5. Performance comparison on different rectangle sizes of the NAR Haar-like feature on BioID database.

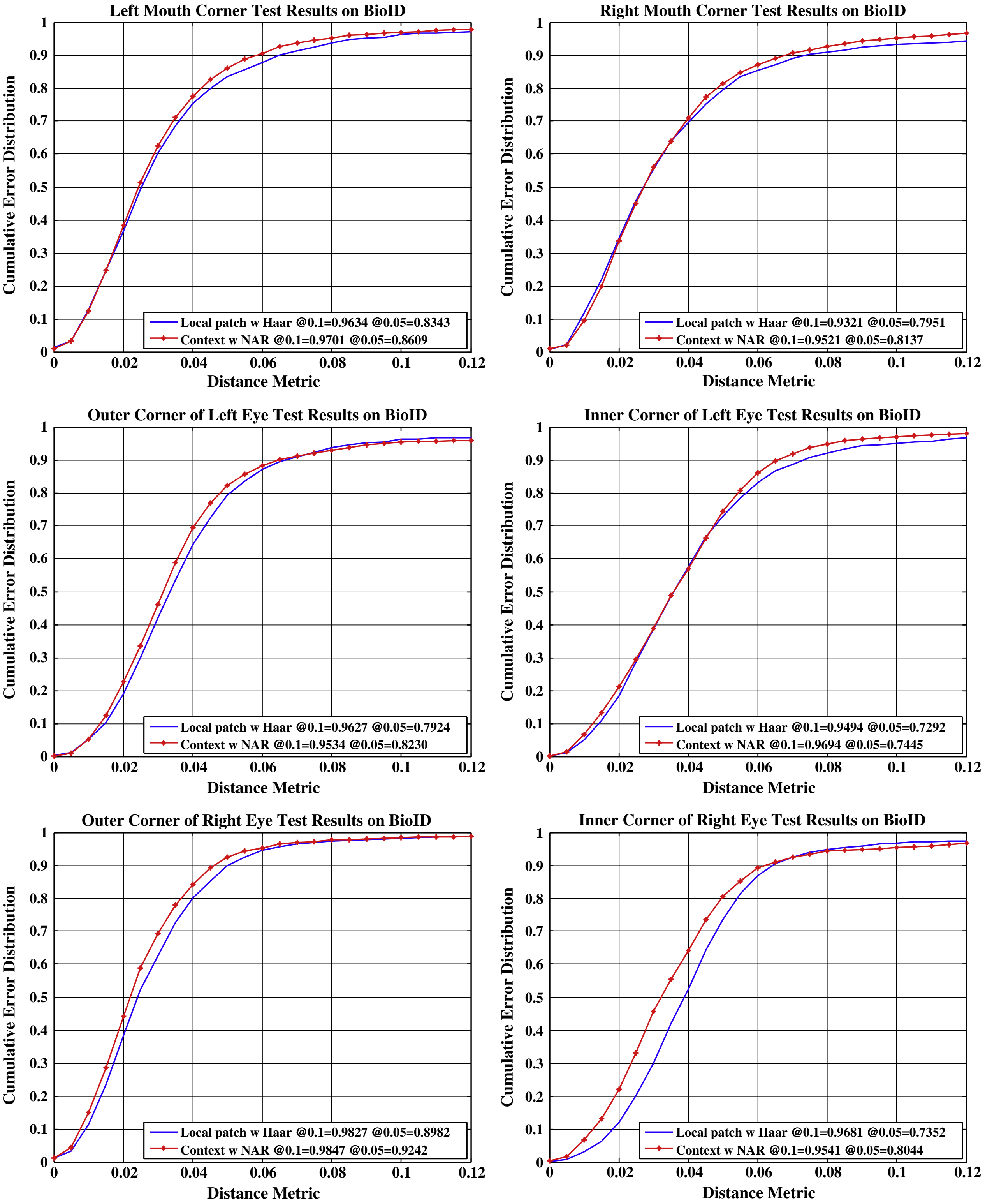

Fig. 6. Facial landmark detection results on BioID database, including four eye corners and two mouth corners.

removing the constraint that the rectangles of Haar-like feature must be adjacent. In our following experiments, we constrain one rectangle of the NAR Haar-like feature within the local image patch and the other rectangle lies in the whole context region. So, it can model the long range co-occurrences of local texture features within the local image patch and the context region. For example, if we want to locate left mouth corner in the face region, usually we perform "sliding-window-search" detection within the face region. For each window, traditional Haar-like features are extracted within it, as shown in Fig. 2 (a).

As an extension of the traditional Haar-like feature, the NAR Haar-like feature is characterized by separated rectangles in the face region, where one rectangle is extracted within the sliding window and the other lies in the normalized face region (i.e., the context region), as illustrated in Fig. 2 (b), (c), and (d). The constructed

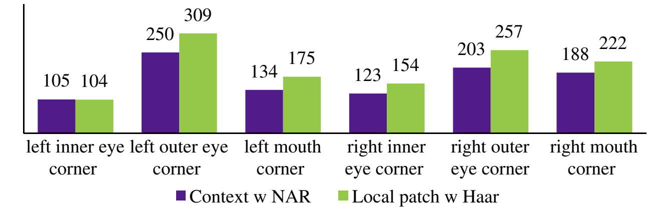

Fig. 7. A comparison of the feature number selected by our algorithm and the Haar-like feature-based method.

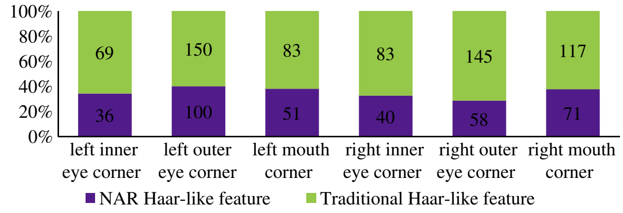

Fig. 8. Percentage of two different features in "Context w NAR" algorithm.

Table 1
Localization results on BioID database.

Method	Average of left eye corners	Average of right eye corners	Average of mouth corners
STASM [33]	0.927	0.962	0.911
ASAM [34]	0.935	0.978	0.957
Ours	0.961	0.969	0.963

nonadjacent rectangles describe the co-variation relationships between the target facial landmark and other facial textures.

The computation of the NAR Haar-like feature is similar to traditional Haar-like feature. The output value of the NAR Haar-like feature can be computed by the following formula:

$$output = \sum_{i=1}^{n} brec_i(j) - \sum_{i=1}^{n} wrec_i(j), \tag{1}$$

where $\sum_{i=1}^{n} brec_i(j)$ is the intensity sum of the pixels within the black rectangle, $\sum_{i=1}^{n} wrec_i(j)$ is the intensity sum of the pixels within the white rectangle. In addition, the output value of the NAR Haar-like feature can be calculated rapidly through integral image. Note that the rectangles must have the same size.

There are two parameters to control the size of the NAR Haar-like feature space: rectangle size of the NAR Haar-like feature and step size of feature sampling. Dense sampling or various rectangle size combinations will generate a huge feature set, which will burden the training process. So, in real applications, we just allow several possible combinations of rectangle sizes.

2.3. Classifier learning with the NAR Haar-like feature

It is known that Real AdaBoost is powerful for feature selection and classifier assembling [23]. So, in our implementation, Real AdaBoost is used to automatically select features and learn classifiers for the facial landmark detection task. Besides describing the local texture pattern by traditional Haar-like features, the proposed NAR Haar-like features are introduced to describe the context information. Different from "auto-context" algorithm, which extracts context features from the classification map produced by previous stage classifiers, we directly model the context by NAR Haar-like features. In the training stage, traditional Haar-like features and the proposed NAR Haar-like features are combined to train classifiers. It's up to the learning algorithm to select the most discriminative features, either the traditional Haar-like features or the NAR Haar-like features.

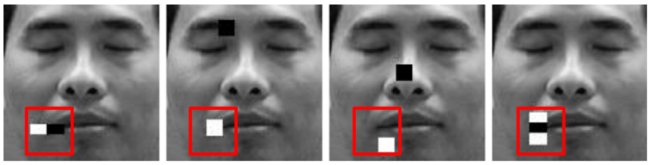

Fig. 9. Top 4 features of the proposed method (i.e., "Context w NAR"), which are selected automatically by Real AdaBoost learning.

As Fig. 3 illustrated, the most discriminative feature is selected automatically in each round of the Real AdaBoost learning, which is noted by a check mark. With all the selected features $h_t(x)$, we construct a strong classifier

$$H(x) = sign\left(\left[\sum_{t=1}^{T} h_t(x)\right]\right). \tag{2}$$

Similar to Viola–Jones [1], strong classifiers learned by Real AdaBoost algorithm are cascaded for facial landmark detection. To speed up the process of cascade learning, feature-inheriting technique [24] is adopted in our method. Using this technique, features learned by the previous stage are inherited by the current stage, which can greatly reduce the computation cost for feature selection. In addition, Look-Up-Table (LUT) type weak classifier proposed by literature [25] is used in our implementation to fit more complex distribution of samples. It's a real-valued weak classifier, which gives samples real-valued confidence instead of Boolean prediction.

3. Application to facial landmark detection

To evaluate the context modeling capability of the NAR Haar-like feature, we apply it to a facial landmark detection problem and compare our method with other state-of-the-art facial landmark detection methods.

3.1. Locating procedure

For the facial landmark detection problem, facial organs or smoothing skin regions can be regarded as context information for a specific target facial point. To model the co-occurrence relationship of these facial textures patterns, we restrict one rectangle of the NAR Haar-like feature to be located within the local image patch around the target point, the other rectangle should be located within the whole face region.

With the cascaded classifier learned in Section 2.3, our facial landmark locating procedure is shown in Fig. 4.

Here, left mouth corner is taken as the example target point. Given a testing image, we first detect face in it, and obtain the normalized face region accordingly. Secondly, a prior location model is applied to determine the search region for this target facial landmark. Note

Table 2
Localization results on Cohn-Kanade database.

Method	Left inner eye corner	Left outer eye corner	Left mouth corner	Right inner eye corner	Right outer eye corner	Right mouth corner	Average of six points
Gabor-ffpd [3]	0.96	0.92	0.97	0.99	0.96	0.91	0.952
STASM [33]	0.959	0.888	0.862	0.942	0.895	0.835	0.897
Ours	0.988	0.965	0.973	0.977	0.976	0.975	0.976

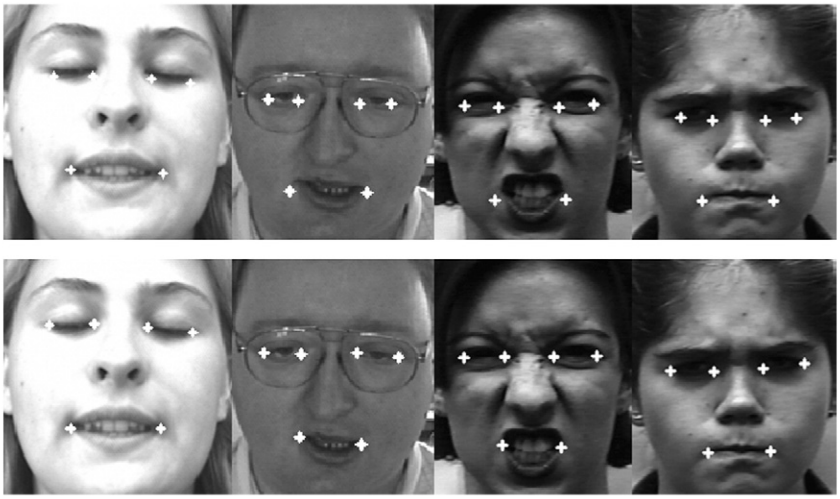

Fig. 10. Some example images from comparison of our method with STASM (v2.4). Top row: localization results of STASM. Bottom row: localization results of our method.

that the prior location model is learned as a bivariate Gaussian model on images with manually labeled landmarks, relative to the coordinate system of the normalized face. Then multi-scales detection is applied to each location in the search region. For each location, a confidence value that evaluates how well it represents the target facial landmark is assigned. Finally, the center of the k × k neighborhood with the maximal confidence is selected as the final position of the target facial landmark.

3.2. Data set and evaluation protocol

We collect about 7000 near-frontal face images from various sources, such as CAS-PEAL [26], CMU PIE [27], FRGC v1 [28], and FG-NET Aging [29], to train our facial landmark detectors. More synthesized positive samples are generated by transformations, such as shifting by ±1 pixel, in-plane rotation within 3°. Negative samples are image patches shifted 5–8 pixels away from the manually labeled ground truth position. Totally, a set of roughly 80,000 positive samples and 3,800,000 negative samples are taken as our training set.

In our experiments, six facial landmarks are localized and evaluated, which include four eye corners and two mouth corners. For each facial landmark, one cascaded classifier is trained separately. To reduce the size of feature space, we assume that the rectangle of the NAR Haar-like feature is square.

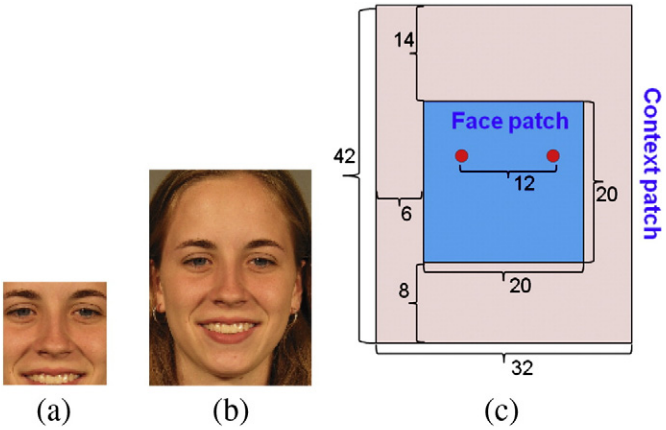

Fig. 11. Examples of the face patch pattern and context patch pattern, (a) face patch, (b) context patch, (c) specific definition of face patch and context patch. In our experiments, we set the distance of two eyes to 12 pixels. The sizes of face patch and context patch are set to 20 × 20 and 32 × 42 pixels respectively. The example image is from FRGC v2 [28] face database.

We evaluate our algorithm on two public databases, BioID and Cohn-Kanade. The BioID database consists of 1521 images of frontal faces taken in uncontrolled conditions using a web camera within

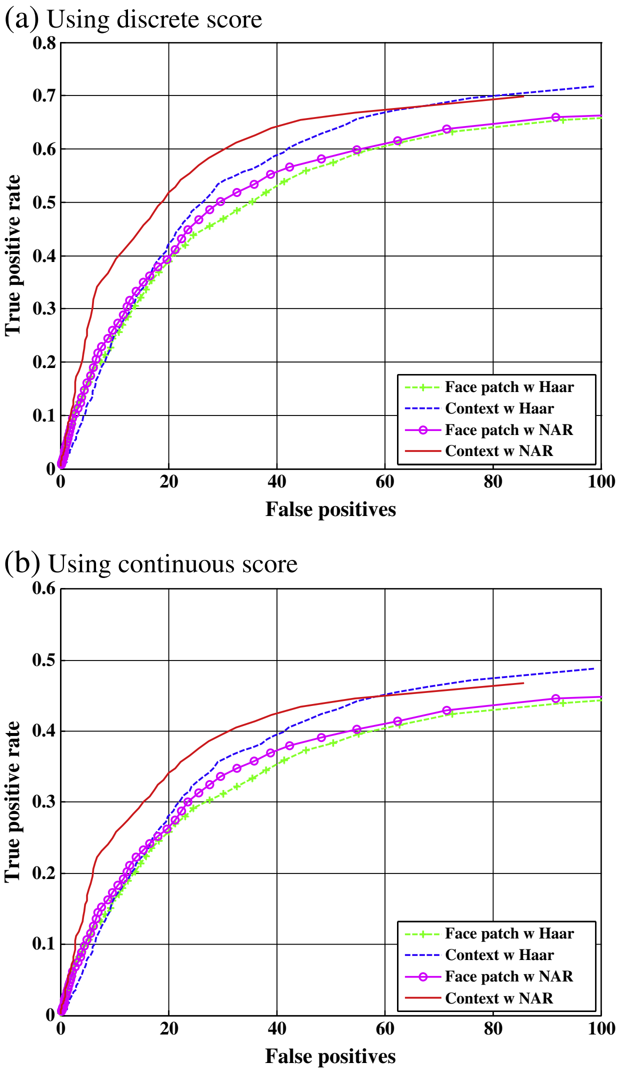

Fig. 12. Performance curves for different methods.

Table 3
Feature distribution statistics for different methods.

Method	No. of selected features	No. of NAR Haar-like features	Average time cost (ms)
Face patch w Haar	2292	–	1017
Face patch w NAR	2040	351	981
Context w Haar	433	–	692
Context w NAR	948	558	663

an office environment. It features a large variety of illuminations, backgrounds and face sizes. The Cohn-Kanade database includes 486 image sequences (8796 static images) in nearly frontal view from 97 subjects. Each sequence begins with a neutral expression and proceeds to a peak expression. Note that all of the databases we test on are definitely excluded from our training set.

In our experiments, we adopt the test criterion proposed in work [9]. The error measure for each individual landmark is defined according to the following formula:

$$e = \frac{||L-G||}{d_{IOD}}, \tag{3}$$

where d_{IOD} is defined as the Inter-Ocular Distance, i.e., the distance between the two eye centers. L represents the auto-located facial landmark location, G represents the location of manually labeled ground truth.

Thus, the cumulative error distribution (y) of the error (x) is used to evaluate the performance of the localization algorithm, as shown in the following formula:

$$y = \frac{Num(e \le x)}{N}, \tag{4}$$

where N is the number of all testing images.

3.3. Experimental results and analysis

In this section, we investigate the effectiveness of the proposed NAR Haar-like feature through evaluation experiments with respect to facial landmark detection.

3.3.1. Experiment on different rectangle sizes of the NAR Haar-like feature

In this subsection, we evaluate the performance of our method according to the rectangle size of the NAR Haar-like feature. More specifically, three different rectangle sizes are evaluated, which are 5×5, 10×10, and 15×15 pixels. In our experiments, the average localization accuracy of six landmarks (i.e., mouth corners, corners of the eyes, etc.) is calculated. The faces are all normalized into 105×105 pixels and the distance between two eyes is set to 50 pixels. Experimental results on BioID database show that with our current experimental configuration, the rectangle size of 10×10 pixels is the most appropriate parameter than the others for the best performance, as shown in Fig. 5. So, in the following experiments, the size of the NAR Haar-like feature is fixed to 10×10 pixels.

3.3.2. Evaluation on context-modeling effectiveness of the NAR Haar-like feature

In this subsection, we evaluate the effectiveness of the proposed NAR Haar-like feature on BioID database. Two facial landmark detectors are trained separately by using different Haar-like feature sets. One method just extracts traditional Haar-like features within local image patch of samples, briefly noted by “Local patch w Haar”. The other method uses both traditional and NAR Haar-like features, briefly noted by “Context w NAR” (i.e., the proposed method). The comparison results of these two methods on six facial landmarks are shown in Fig. 6.

It is obvious that better localization results are obtained by our proposed method. Especially, for the inner corner of right eye, the localization accuracy of “Context w NAR” algorithm is about 12% higher than the “Local patch w Haar” algorithm when the error is less than 4%. Besides the higher localization accuracy, distinctive feature set selected by “Context w NAR” algorithm is usually smaller than the “Local patch w Haar” algorithm, as Fig. 7 illustrated. It is also interesting to mention that about 1/3 distinctive features are selected from the NAR Haar-like feature set, shown in Fig. 8. As for the time cost of facial landmark detection, the “Context w NAR” algorithm costs 49 ms averagely when locating six points in images with 384×286 pixels, comparing with 51 ms of the “Local patch w Haar” algorithm. Fig. 9 shows the top four features selected by Real AdaBoost of the proposed method. Here, take the left mouth corner as the example target landmark.

From these experiments, it can be derived that: 1) the facial context information is useful for facial landmark detection; 2) the proposed NAR Haar-like feature can well model the facial context information.

3.3.3. Comparisons with other methods

Besides the cascaded AdaBoost-based methods, Active Shape Models (ASM) [30] and Active Appearance Models (AAM) [31] are two popular global shape-based methods for facial landmark detection. In these kinds of methods, statistical face models are constructed to prevent locating inappropriate facial landmarks. Also, some variants are proposed to get more robust and efficient performance [32–34].

In this subsection, we compare our method with current state-of-the-art methods: Gabor facial feature point detector (Gabor-ffpd) [3], Stacked Active Shape Model (STASM) [33], and Active Structure Appearance Model (ASAM) [34].

In the first comparison on BioID database, the implementation of STASM is available from the internet and the results of ASAM are obtained from their published paper. However, only average localization results of left eye corners, right eye corners, and mouth corners are given by the literature of ASAM. So, here we just compare the average localization results on BioID. Table 1 gives the localization results when the error is less than 10%. It shows clearly that we achieve better localization accuracy except on “average of right eye corners”.

In the second comparison on Cohn-Kanade database, besides the STASM method, Gabor-ffpd method is also taken as a compared

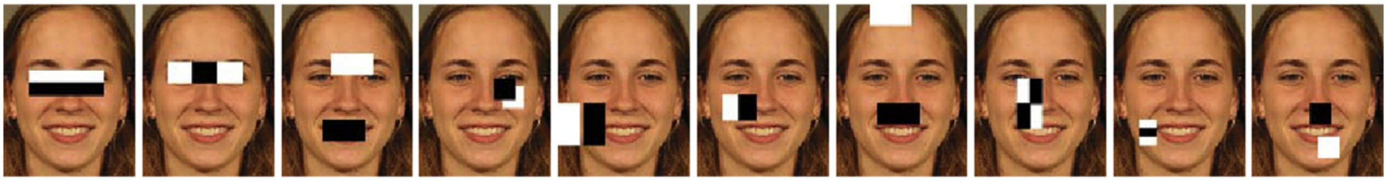

Fig. 13. Top 10 features of the proposed method (i.e., “Context w NAR”), which are selected automatically by real AdaBoost learning.

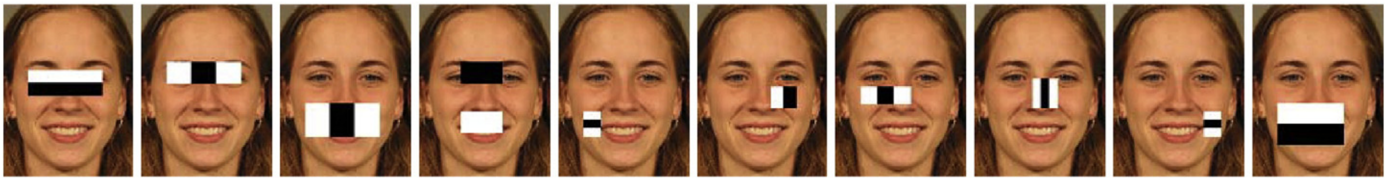

Fig. 14. Top 10 features of method “Face patch w NAR”, which are selected automatically by real AdaBoost learning.

benchmark for its high accuracy on this database. In their paper, localization results of 20 facial landmarks are published. They evaluate their method on the first frames of 300 video sequences, which only contain neutral expression. These 300 images are divided equally into 3 subsets. One of the 3 subsets is used as test set and the other two subsets are used as training set. A 3-fold cross-validation is conducted on the image set.

Because the partition of the testing set cannot be known precisely from the related publications, we just evaluate our method on all of the images (8796 images), including images with extreme facial expressions. Experimental results show that our method gets better localization results. Table 2 shows the location accuracy of six facial landmarks when location error is less than 10%. Averagely, our performance achieves 2.4% higher than Gabor-ffpd method and 7.8% higher than STASM method. Some facial landmark detection results compared with STASM are shown in Fig. 10.

4. Extended application to face detection

To further verify the generalization capability of the NAR Haar-like feature, we apply our algorithm again to face detection problem.

For face detection, information such as hair, neck, ear, etc., can be used as context. So, similar to work [35], a larger rectangle region surrounding the face is considered to contain the context information for face detection. The specific definition of “face patch” and “context patch” is shown in Fig. 11.

To model the co-occurrence relationship of face and the context, it is constrained that one rectangle of the NAR Haar-like feature should be located within the face patch region and the other rectangle is extracted within the whole context patch region. In the process of face detection, we use traditional Haar-like features (extracted from the face patch region) to characterize the local texture and NAR Haar-like features to model the context information.

4.1. Experimental setup and evaluation protocol

We collect 34,612 near-frontal face images from many public databases to train a near-fontal face detector, including AR [36], BANCA [37], FRGC v1 [28], CMU PIE [27], XM2VTS [38], ORL [39], Yale [40], CAS-PEAL [26], and FERET [41], etc. Most faces in the sample set have the variation of pitch rotation within range of [−20°, 20°]. Totally, a set of roughly 200,000 grayscale face samples with size of 32 × 42 are generated from the original 34,612 face images with manually labeled eyes by following transformations: mirroring, in plane rotation of −12°, −6°, 0°, and 6°, 12°. As for the negative samples, 6000 images without faces are downloaded from the internet for generating negative samples. In order to accelerate the training process, MSL [24] is adopted to train with an enormous sample set in our experiments. In each training stage, the training non-face negative samples are fixed to 10,000. For the positive bootstrap in MSL, the starting face sample set size is 3000. At each positive bootstrap, maximally 500 new samples are added. The minimum detection and maximum false alarm rate of each stage are set to 0.9999 and 0.4 respectively. The training process terminates automatically when there are no enough negative samples to train a new stage. To detect faces with various scales, test images are down-sampled with a coefficient of 0.9.

We evaluate the effectiveness of our method on the FDDB face database, which has been developed recently for evaluating the performance of face detection algorithms. This dataset contains photographs from several news sources, and includes images of face under very challenging, unconstrained environments. There are totally 5171 faces in 2845 images. Jain et al. also specified an evaluation scheme based on computing two ROC curves using (a)

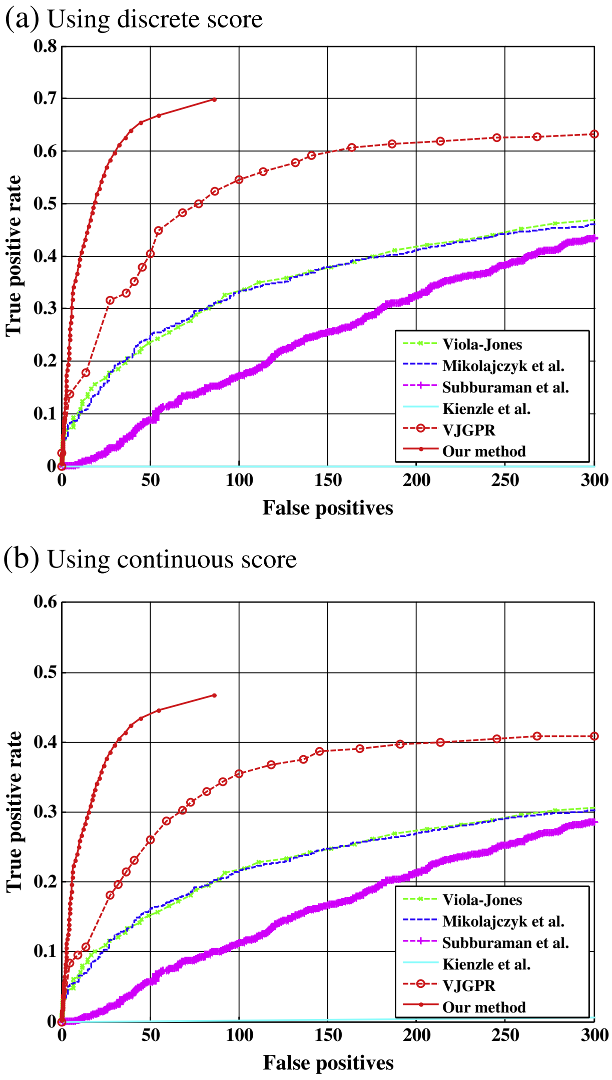

Fig. 15. A comparison with state-of-the-art methods published on the FDDB face database.

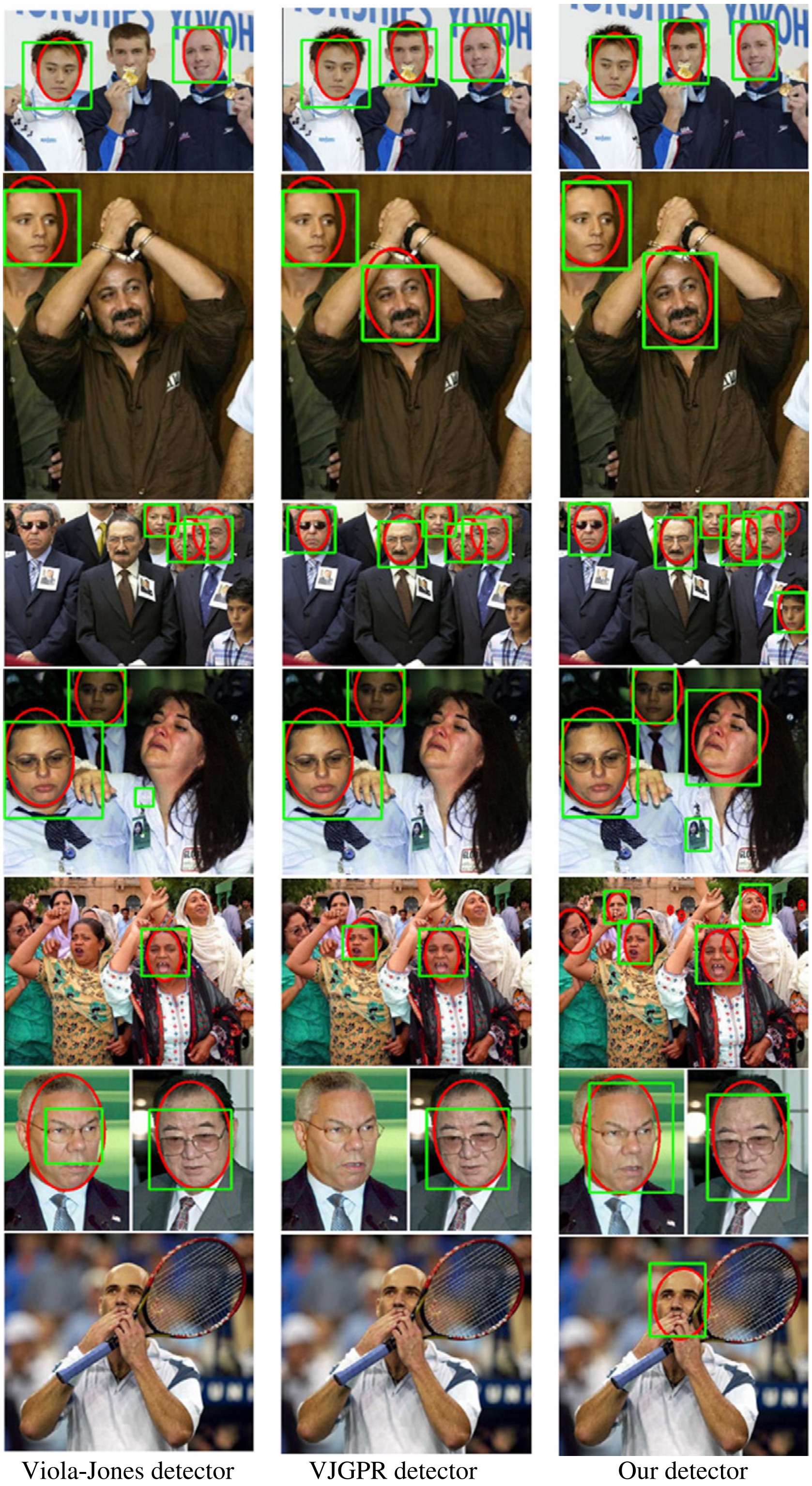

Fig. 16. Some example images from comparisons of our detector with Viola–Jones and VJGPR detectors.

discrete score, and (b) continuous score. The discrete score gives each detection result a binary match/non-match label. The continuous score associates a real-valued score with each detection result based on the overlap between the detected and annotated regions [42]. According to the test and evaluation protocols, our following experiments are carried out on ten folds of the data set separately and the cumulative performance is reported as the average curve of ten ROC curves.

4.2. Experimental results and analysis

4.2.1. Evaluation on context-modeling effectiveness of the NAR Haar-like feature

In this subsection, we comprehensively evaluate the effectiveness of the NAR Haar-like feature through comparing our method with other three kinds of methods. To differentiate from other methods, we briefly note our method as "Context w NAR", which extracts traditional Haar-like features within the face patch region and NAR Haar-like features with one rectangle in the face patch region and the other rectangle within the whole context patch region. In order to perform a fair comparison, for all these methods, the rectangle size of the NAR Haar-like feature is set to 5×5, 5×10, 10×5, 10×10. The step sizes of traditional and NAR Haar-like features are both set to 4 pixels in horizontal direction and 2 pixels in vertical direction.

The performance curves and statistics of feature number for all these methods are shown in Fig. 12 and Table 3 respectively.

Comparison 1. Context vs. local texture

To evaluate the effectiveness of the NAR Haar-like feature on context modeling for face detection, we just extract traditional Haar-like features within the face patch region (briefly noted by "Face patch w Haar") and compare it with our "Context w NAR" method. Experimental results show that the proposed method impressively outperforms "Face patch w Haar" method, which convincingly shows that the proposed NAR Haar-like feature generalizes well to model the context for face detection problem. The top 10 automatically selected features of our method are shown in Fig. 13.

Comparison 2. Context modeling with the NAR Haar-like feature vs. with traditional Haar-like feature

In comparison 1, we just extract NAR Haar-like features within the face patch region. So, to perform a more fair comparison, we extract traditional Haar-like features within the context patch region (briefly noted by "Context w Haar") and compare it with the proposed method. Experimental results show that the proposed method (i.e., "Context w NAR") outperforms "Context w Haar" method with a lower false alarm rate. And, in the proposed method, about 58% (558/948) features are NAR Haar-like features. It can be observed from the experimental results that the NAR Haar-like feature is more effective to model the context than just using traditional Haar-like features.

Comparison 3. Context modeling vs. local texture modeling with the NAR Haar-like feature

To deeply investigate the power of the NAR Haar-like feature, we exploit it to characterize the local texture of faces. In our experiments, we extract both traditional and NAR Haar-like features within the face patch region, called "Face patch w NAR", and compare it with "Face patch w Haar" method and the proposed "Context w NAR" method. As shown in Fig. 12, the performance of "Face patch w NAR" is just slightly higher than those of "Face patch w Haar". But the proposed "Context w NAR" method outperforms these two methods evidently. Based on these experimental results, it is observed that: the proposed NAR Haar-like feature is more suitable to model the context information rather than model the local texture information. The top 10 automatically selected features of "Face patch w NAR" method are shown in Fig. 14.

To evaluate the detection speed of each method, we conduct face detection experiments on 100 images with 384×286 pixels and the average detection speed is shown in Table 3.

4.2.2. A comparison with other methods

In this part, we compare our method with other published state-of-the-art methods [1,42–45] on the FDDB face database. As shown in Fig. 15, our method outperforms these methods on both discrete and continuous score measurements.

Fig. 16 gives some typical comparison examples among our face detector, Viola–Jones detector [1] and VJGPR detector [42]. It's important to note that these images are not manually selected by us, they are the examples used in VJGPR. We use these images here to expect a fair comparison. Quantitative and visualized results convincingly show that our method performs better than other methods.

5. Conclusions

Through introducing the context constraints, a novel solution is proposed for face and facial landmark detection by integrating the local texture information and the global context information. In our method, a novel type of the NAR Haar-like feature is designed to characterize the co-variation relationships between face or facial landmarks and its surrounding context. Essentially, it models these co-variation relationships implicitly through the co-occurrence of low-level features. Combined with the traditional Haar-like features, a powerful representation is formed to realize more robust face and facial landmark detection under Real AdaBoost framework. Experimental results convincingly show that the NAR Haar-like feature is effective to model the context for face and facial landmark detection. In addition, our proposed method obtains state-of-the-art performance on both face detection and facial landmark detection tasks for multiple databases.

Similar to the widely used rectangle Haar-like feature, the NAR Haar-like feature can also be designed to many different modes. Therefore, it would be our direct future work to extend the NAR Haar-like feature to more complicated patterns, such as the features characterized by three- or four-rectangle mode. And experiments on general object detection, such as car detection, horse detection, etc., will be further explored.

Acknowledgments

This paper is partially supported by the Natural Science Foundation of China under contracts No. 61025010, No. 60833013, and No. 61001193; and the National Basic Research Program of China (973 Program) under contract 2009CB320902.

References

[1] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proc. of Computer Vision and Pattern Recognition, 2001, pp. 511–518.
[2] Z. Niu, S. Shan, S. Yan, X. Chen, W. Gao, 2D cascaded AdaBoost for eye localization, Proc. of International Conference on Pattern Recognition, 2006, pp. 1216–1219.
[3] D. Vukadinovic, M. Pantic, Fully automatic facial feature point detection using Gabor feature based boosted classifiers, Proc. of Systems, Man and Cybernetics, 2005, pp. 1692–1698.
[4] A. Torralba, K.P. Murphy, W.T. Freeman, Using the forest to see the trees: exploiting context for visual object detection and localization, Commun. ACM 53 (3) (2010) 107–114.
[5] T. Kozakaya, T. Shibata, M. Yuasa, O. Yamaguchi, Facial feature localization using weighted vector concentration approach, Image Vision Comput. 28 (5) (2010) 772–780.
[6] M. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, Proc. of Computer Vision and Pattern Recognition, 2010, pp. 2729–2736.
[7] M. Fink, P. Perona, Mutual boosting for contextual inference, Proc. of Neural Information Processing Systems, 2003.
[8] Z. Tu, X. Bai, Auto-context and its application to high-level vision tasks and 3D brain image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1744–1757.
[9] O. Jesorsky, K. Kirchberg, R. Frischholz, Robust face detection using the Hausdorff distance, Proc. of Audio and Video-Based Biometric Personm Authentication, 2001, pp. 91–95.
[10] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, Proc. of Automatic Face and Gesture Recognition, 2000, pp. 46–53.
[11] V. Jain, E. Learned Miller, FDDB: a benchmark for face detection in unconstrained settings, technical Report, University of Massachusetts Amherst, 2010.
[12] X. Zhao, X. Chai, Z. Niu, C. Heng, S. Shan, Context constrained facial landmark localization based on discontinuous Haar-like feature, Proc. of Automatic Face and Gesture Recognition, 2011, pp. 673–678.
[13] T. Mita, T. Kaneko, O. Hori, Joint Haar-like features for face detection, Proc. of International Conference on Computer Vision, 2005, pp. 1619–1626.
[14] R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection, Proc. of International Conference on Pattern Recognition, 2002, pp. 900–903.
[15] S.Z. Li, Z. Zhang, FloatBoost learning and statistical face detection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1112–1123.
[16] B.S. Manjunath, W.Y. Ma, Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 837–842.
[17] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[18] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proc. of Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[19] T. Ojala, M. Pietik, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[20] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, WLD: a robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1705–1720.

[21] S. Yan, S. Shan, X. Chen, W. Gao, Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection, Proc. of Computer Vision and Pattern Recognition, 2008.

[22] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.

[23] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2) (2000) 337–407.

[24] S. Yan, S. Shan, X. Chen, W. Gao, J. Chen, Matrix-Structural Learning (MSL) of cascaded classifier from enormous training set, Proc. of Computer Vision and Pattern Recognition, 2007.

[25] B. Wu, H. Ai, C. Huang, S. Lao, Fast rotation invariant multi-view face detection based on real AdaBoost, Proc. of Automatic Face and Gesture Recognition, 2004, pp. 79–84.

[26] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale Chinese face database and baseline evaluations, IEEE Trans. Syst. Man Cybern. Part A Syst. Humans 38 (1) (2008) 149–161.

[27] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.

[28] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, Proc. of Computer Vision and Pattern Recognition, 2005, pp. 947–954.

[29] FG-NET Aging Database, http://www-prima.inrialpes.fr/FGnet/ 2002.

[30] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models: their training and application, Comput. Vision Image Underst. 61 (1) (1995) 38–59.

[31] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.

[32] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models, Proc. of British Machine Vision Conference, 2006, pp. 929–938.

[33] S. Milborrow, F. Nicolls, Locating facial features with an extended active shape model, Proc. of European Conference on Computer Vision, 2008, pp. 504–513.

[34] K. Kinoshita, Y. Konishi, S. Lao, M. Kawade, A fast and robust facial feature detection and 3D head pose estimation based on 3D model fitting, Proc. of MIRU, 2008.

[35] H. Kruppa, M.C. Santana, B. SchieFranz, B. Schiele, Fast and robust face finding via local context, Proc. of Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2003, pp. 157–164.

[36] A.M. Martinez, R. Benavente, The AR face database, CVC Technical Report #24, 1998.

[37] E.B. Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Pore, B. Ruiz, J. Thiran, The BANCA database and evaluation protocol, Proc. of Audio- and Video-Based Biometric Person Authentication, 2003, pp. 625–638.

[38] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Proc. of Audio Video-based Biometric Person Authentication, 1999, pp. 72–77.

[39] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, Proc. of the Second IEEE Workshop on Applications of Computer Vision, 1994, pp. 138–142.

[40] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[41] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face recognition algorithms, Image Vision Comput. 16 (5) (1998) 295–306.

[42] V. Jain, E. Learned-Miller, Online domain adaptation of a pre-trained cascade of classifiers, Proc. of Computer Vision and Pattern Recognition, 2011.

[43] V.B. Subburaman, S. Marcel, Fast bounding box estimation based face detection, Proc. of European Conference on Computer Vision Workshop on Face Detection: Where we are, and what next?, 2010.

[44] K. Mikolajczyk, C. Schmid, A. Zisserman, Human detection based on a probabilistic assembly of robust part detectors, Proc. of European Conference on Computer Vision, 2004, pp. 69–82.

[45] W. Kienzle, G. Bakir, M. Franz, B. Scholkopf, Face detection efficient and rank deficient, Proc. of Neural Information Processing Systems, 2005, pp. 673–680.