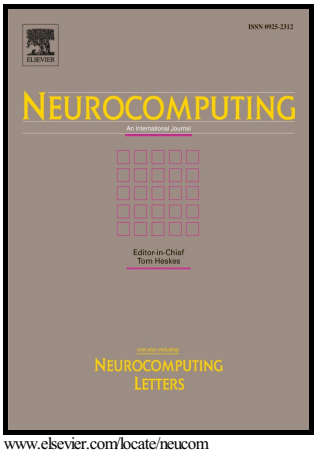
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

Robust face detection using local CNN and SVM based on kernel combination

Qin-Qin Tao, Shu Zhan, Xiao-Hong Li, Toru Kurihara

PII: S0925-2312(16)30566-5
DOI: http://dx.doi.org/10.1016/j.neucom.2015.10.139
Reference: NEUCOM17178

To appear in: *Neurocomputing*

Received date: 1 August 2015
Revised date: 27 September 2015
Accepted date: 21 October 2015

Cite this article as: Qin-Qin Tao, Shu Zhan, Xiao-Hong Li and Toru Kurihara Robust face detection using local CNN and SVM based on kernel combinatior *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2015.10.139

This is a PDF file of an unedited manuscript that has been accepted fo publication. As a service to our customers we are providing this early version o the manuscript. The manuscript will undergo copyediting, typesetting, an review of the resulting galley proof before it is published in its final citable form Please note that during the production process errors may be discovered whic could affect the content, and all legal disclaimers that apply to the journal pertain

ACCEPTED MANUSCRIPT

Robust face detection using local CNN and SVM based on kernel combination

Qin-Qin Tao[1], Shu Zhan[1,*], Xiao-Hong Li[1], Toru Kurihara[2]

[1] *Hefei University of Technology, Hefei, 230009, China*
[2] *Kochi University of Technology, Kochi, Japan*

Abstract

One key challenge of face detection is the large appearance variations due to some real-world factors, such as viewpoint, extreme illuminations and expression changes, which lead to the large intra-class variations and making the detection algorithm is not robust enough. In this paper, we propose a locality-sensitive support vector machine using kernel combination (LS-KC-SVM) algorithm to solve the above two problems. First, we employ the locality-sensitive SVM (LSSVM) to construct a local model on each local region, which can handle the classification task easier due to smaller within-class variation. Second, motivated by the idea that local features are more robust compared with global features, we use multiple local CNNs to jointly learn local facial features because of the powerful strength of CNN learning characteristic. In order to use this property of local features effectively, we apply the global and local kernels to the features and introduce the combination kernel to the LSSVM. Extensive experiments demonstrate the robustness and efficiency of our algorithm by comparing it with several popular face detection algorithms on the widely used CMU+MIT dataset and FDDB dataset.

Keywords: Face detection, Convolutional Neural Network, Kernel Combination, Support Vector Machine, Local Classifier

1. Introduction

Face detection is the foundation of computer vision and pattern recognition technology [1, 2, 3]. Recent research in this area focuses more on the uncontrolled face detection problem, where a number of factors such as pose changes, exaggerated expressions and extreme illuminations can lead to large visual variations in face appearance, so the difference between the face images may be large, and the difference between the background and face image may be fuzzy, which makes the classification more difficult. In other words, feature vectors extracted from faces are usually scattered into a feature spaces diversely and the classification boundary between face and background is usually highly complex due to large intra-class variances and vague inter-class discrimination. Instead

Accepted manuscript

ACCEPTED MANUSCRIPT

of learning such a complex global model, an alternative strategy is to build simple models on a set of locality-sensitive regions.

A few localized classifiers exist in literatures [4, 5], which leverage locality property to enhance the performance of classifiers. For example, Qi et al. [5] proposed a locality sensitive support vector machine (LSSVM) algorithm. Based on the local classifier [4], it imposed a global regularizer across local regions to avoid these local models from overfitting into locality-sensitive structures. They gain great success and achieve state-of-the-art performance. However, their performances are still limited because they did not study the kernel function further which is very important for object detection.

Kernel-based methods can represent non-linear variations easily, and therefore such methods have been used to cope with visual variations. However, because conventional methods apply one kernel to global features extracted from one image and global features are influenced easily by noise or occlusion, conventional methods are not robust to occlusion. Effectiveness of SVM based on local features has also been reported in recent years, since partial occlusion affects only specific local features [6, 7]. For example, Kazuhiro Hotta [6] arranged local kernels at all local regions of recognition target and used in SVM to realize robust face recognition under partial occlusion.

In this paper, we propose to exploit the kernel to improve the performance of LSSVM. Specific, we combine the global and local kernels and apply the combination kernel to the LSSVM algorithm, putting forward an improved local sensitive support vector machine using kernel combination (LS-KC-SVM) algorithm. Our algorithm has the advantages of LSSVM: make the classification task easier. Since each obtained locality region only contains images with relatively smaller variances, the local classifier for each cluster can easily handle them. At the same time, our algorithm can measure the detailed and rough similarity comprehensively through the combination of global and local kernels used in SVM, which improving the generalization ability of the algorithm.

On the other hand, the representation of face image is very important. Recently, CNN model is proven to be very powerful in learning discriminate visual features [8]. For instance, Zhang et al. [8] built a CNN that can simultaneously learn face/non-face decision, the face pose estimation problem, and the facial landmark localization problem. However, the features extracted by CNN are global features which are influenced easily by noise or occlusion. So in this paper we propose to learn expressive and robust face detection features with multiple local CNN models, each processes a local facial region. These learned local convolutional features are then combined together as the whole facial features. Thus, the obtained features can be more effective and robust compared with the global features extracted by the conventional CNN.

In summary, we propose a face representation learning framework with multiple local CNN models to obtain the robust local facial features. And then the local features are fed to our LS-KC-SVM to classify as face or non-face. The contributions of the paper include: 1) propose LS-KC-SVM algorithm, which can solve the problem of large intra-class variations, and is more robust to partial occlusion; 2) propose a multiple local CNNs fusion hierarchy, which learns

Accepted manuscript

ACCEPTED MANUSCRIPT

discriminate and occlusion invariance features that are suitable for LS-KC-SVM.

2. Related Works

In recent years, various classification algorithms have been emerged, such as optimum-path forest classifier [9, 10, 11, 12] and pose classifier [13, 14]. They are widely used in image classification, search and rerank fields [15, 16, 17, 18, 19, 20]. Among them, SVM is a famous machine learning algorithm and has been widely studied and applied to face detection [21, 22, 23, 24, 25, 26, 27, 28, 29]. One of the main focuses of the research is the study of kernel function [6, 7, 30]. Kernel method is an effective approach to solve the nonlinear pattern recognition problems in the field of machine learning. At present, multiple kernel method has become a new research focus [30]. Compared with the traditional single kernel method, multiple kernel method is more flexible, more interpretable and has better generalization performance when dealing with heterogeneous, irregular and non-flat distribution samples. For example, Hu et. al. [30] presented a multi-kernel SVM optimization model based on p-norm constraint. Besides multiple kernel method, SVM with local kernels has been proposed in recent years [6, 7]. Because conventional methods apply one kernel to global features and global features are influenced easily by noise or occlusion, the conventional methods are not robust to occlusion. The recognition method based on local features, however, is robust to occlusion because partial occlusion affects only specific local features. For example, Kazuhiro Hotta [6, 7]showed that the summation of local kernels is more effective and robust compared with global kernel based SVM. Kernel-based method is effective for object detection and recognition.

Compared with the general hand-crafted feature extracting methods [31, 32, 33, 34, 35, 36, 37], deep learning methods show notable potential in visual feature learning [38, 39, 40]. One of the main focuses of these methods is designing suitable deep network structure to accomplish some specific tasks [41, 42]. In [41], three levels of deep convolutional networks are cascaded to detect the face in a coarse-to-fine manner. In [42], multiple local convolutional sub-networks learn convolutional features from multiple local regions and fuse these local features by a tree-structured network. Differently, in our work we fuse these local features in a simple way to obtain the representation of the whole face, which mainly because the representation is more robust since partial occlusion affects only specific local features, the other local features are still same.

3. Background

In this section, we present a brief review of LSSVM which forms the basis of the proposed algorithm in this paper.

Accepted manuscript

ACCEPTED MANUSCRIPT

3.1. LSSVM

Give a set of training examples $S = (x_i, y_i)|i = 1, ..., N$. The whole training examples are clustered into several clusters by the clustering algorithm [43], and then a local classifier is trained for each cluster. Suppose that the local classifier $f_l = w_l x$ on the lth local region is a generalized hyperplane classifier.

Denote the classiier for region X_l by f_l, then for an arbitrary sample $x \epsilon X$ the classifier is

$$f(x) = \sum_{l=1}^{L} f_l(x) I(x \epsilon X_l). \tag{1}$$

Here I(E) is the indicator function taking value 1 if the event E occurs or 0 otherwise.

Since individual local learner is limited on local region, the overall classifier (1) combining these local learners cannot guarantee the regularity on the whole space even though local learners have regularization performance on their own regions. Thus, based on it, literature [5] assumed that exists a global reference classifier $f(x) = w \cdot x$ on the whole space which is regularized to control its complexity as well as best approximates the local learners on each local region. Such a criterion can be implemented by introducing the following regularization term on each local region X_l as

$$\Omega(w, w_l) = (w - w_l)^T X_l X_l^T (w - w_l) = (w - w_l)^T S_l (w - w_l), \tag{2}$$

where X_l is the matrix with x_{jl}, $j = 1, , N$ as its columns, x_{jl} is the jth example in the lth region, and $S_l = \frac{1}{N_l} \sum_{j=1}^{N_l} x_{jl} \cdot x_{jl}^T$ is sample correlation matrix for the lth region. N_l is the number of training examples on the lth region. With the above regularizer on each region, the learning problem for locality-sensitive classiier can be formulated as

$$\underset{w, w_l}{min} \ \frac{1}{2}\lambda \|w\|_2^2 + \frac{1}{2} \sum_{l=1}^{L} (w - w_l)^T S_l (w - w_l) + C \sum_{l=1}^{L} \sum_{j=1}^{N_l} \zeta_{jl}, \tag{3}$$

s.t., $y_{jl} \cdot w_l^T S_l x_{jl} \geq 1 - \zeta_{jl}, \zeta_{jl} \geq 0, j = 1, ..., N_l, l = 1, ..., l.$

Where $y_{jl} \epsilon (-1, 1)$ is the corresponding class label for each x_{jl}, and λ and C are the balance parameters.

Equation (3) can be solved by the Lagrange multiplier method. Specific derivation process is not repeated herein. At last, when the coefficients ∂_{jl} are solved the local classifiers are given as:

$$f_t(x) = \sum_{l=1}^{L} \sum_{j=1}^{N_l} \partial_{jl} \left\{ \frac{1}{\lambda N_l N_t} K_{jl}^T K_{lt} K_t(x) + \frac{\delta_{lt}}{N_l} K_{jl}^T K_t(x) \right\}, \tag{4}$$

s.t. $1 \leq t \leq L$.

Where $K_{jl} = [k(x_{jl}, x_{1l}), ..., k(x_{jl}, x_{N_l l})]$ is kernel matrix for samples in the lth region. $K_{lt} = [k(x_{jl}, x_{it})]_{N_l \times N_t}$ is the kernel matrix between all the samples in the lth and the tth regions. $K_t(x) = [k(x_{lt}, x), ..., k(x_{N_l t}, x)]$.

Accepted manuscript

ACCEPTED MANUSCRIPT

3.2. discussion

According to (4), we can acquire the local learners on each local region that is not sensitive to noise which can prevent the overfitting phenomenon. At the same time it can best handle the large intra-class problem, making the classification simple. However, in the literature [5], it does not consider the selection of kernel functions which is a very effective method for object detection.

As we know, the most commonly used method applies one kernel to global features. This method is useful in some cases since it can also provide some useful information. However, face detection in real-world situation is still challenging, with obstacles such as occlusion, illumination changes and pose changes to be overcome. With such obstacles, the global feature of face maybe destroyed. So the conventional method is not robust to such obstacles. Whereas local feature is more robust compare with global feature since partial occlusion affects only specific local feature. So in this paper, in order to use local features effectively, we combine the global and local kernel to measure the rough and detailed similarity together, proposing a robust and effective face detection algorithm. In particular, the global kernel is applied to the whole feature extracted from face, and the local kernel is applied to the local region of the feature. The use of local kernels in SVM requires local kernel integration. The summation of local kernels is used as the integration method in this study.

4. The Proposed Method

In this paper, we train a background filter in order to filter away the simple background as soon as possible. For this purpose, Adaboost is used since it is very effective and real time since Viola and Jones's [44] work. Section 4.1 describes the Adaboost background filter briefly and section 4.2 describes the local features we proposed based on CNN. At last, section 4.3 depicts the LS-KC-SVM algorithm using kernel combination, which is a new attempt. Fig. 1 shows the basic process of training and detecting human face respectively. In the training stage, we used training examples to train the three models respectively: Adaboost, local CNNs, and LS-KC-SVM. In the test stage, the input patches passed the Adaboost classifier first. If Adaboost judged its a face, then the patch would be passed to the next stage for further classification. Local CNNs extracted local features of the received patches, and then the local features would be passed to our LS-KC-SVM to make the final classification. If the judgment is no, the patch would be discarded directly.

4.1. The Adaboost based background filter

The boosting cascade framework by Viola and Jones [44] is a great breakthrough in the field of face detection. So we adopt the detector as our background filter in the first stage to filter out those obvious backgrounds, which ensuring the detection rate and increasing the speed of system at the same time.

It is important to set the threshold of the Adaboost appropriately. Setting the threshold too high may cause too many positive examples be rejected, reducing the overall detection rate, and setting it too low may lead to too many

Accepted manuscript

ACCEPTED MANUSCRIPT

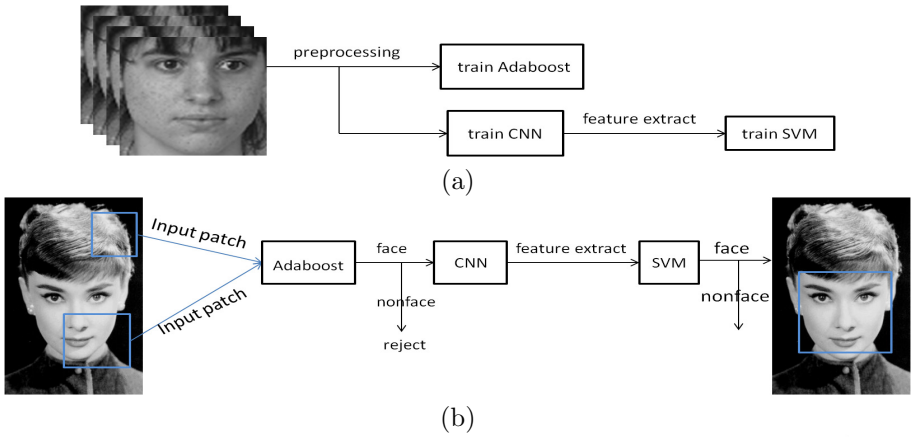


Figure 1: (a)Training of face detection framework and (b)Testing of face detection framework

pass-through patches that needs to be classified further by the SVM, slowing down the overall detection process. So a minimum detection rate of 99.8% and a maximum false positive rate of 50% were set as the training parameters.

4.2. Representation learning with multiple local CNNs

In our work, we apply CNN model to learn and extract discriminate visual features. However, instead of extract the whole facial feature in a single CNN model, we employ multiple local CNNs to learn features from multiple local regions and combine these local features into the final feature.

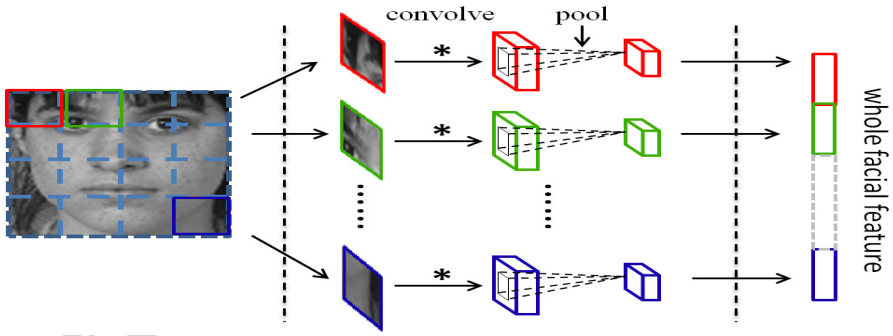


Figure 2: Local visual features learning by mutiple local CNNs

As shown in fig.2, we divided the face into 16 patches, each size is 7×7 pixels. These parameters are the results of our comprehensive consideration of the size of training face and the structure of CNN. We train a local CNN for each patch, getting the correspondence convolutional features. Then we assemble the convolutional features of each region in the correct order, getting the

ACCEPTED MANUSCRIPT

final representation of the whole face. As shown in fig.2, CNN models contain a series of planes where successive convolutions and subsampling operations are performed. These planes are called feature maps as they are in charge of extracting and combining a set of appropriate features through the convolutions and subsampling operations.

Convolution process: Each feature map unit computes a weighted sum of its input x by a 4×4 convolution kernel that can be learned, adds a trainable bias b_x , and then passes the results through Rectified Linear Units (ReLU), obtaining a convolution layer C_x. We adopt the ReLU as the activation function here. CNN with ReLU trains several times faster than the traditional CNN with tanh units, which has a great influence on the performance of large models trained on large datasets.

$$C_x = max\left(0, K * x + b_x\right). \tag{5}$$

Sampling process: Each unit computes the average of its four inputs, multiplies it by a trainable coefficient w_{x+1}, adds a trainable bias b_{x+1}, and passes the results through the ReLU. Thus produce a feature map S_{x+1}.

$$S_{x+1} = max(0, \Sigma C_x \times w_{x+1} + b_{x+1}). \tag{6}$$

The features extracted by convolutional layer are fed to full connected layer. The units in full connected layer compute the dot product between their input vector and weight vector, to which a bias is added. Then the result is passed to the ReLU function for nonlinear transformation. We take the output of the full connected layer as the extracted features of a facial patch. At last, we combine the local features extracted by the 16 CNN models into one, acquiring the final facial features.

We do this mainly because on one hand if a portion of face is occluded, the whole facial features extracted by the traditional CNN will be affected, and the characteristics of the not keep out face region will change too. This leads to that it is not robust enough. On the other hand, in order to use local kernels in LS-KC-SVM effectively, we want to use local appearance features. So we adopt the CNN model of such structure to get the local characteristics of human face, and apply it to our LS-KC-SVM algorithm, which both develops strength of CNN learning characteristics, and avoids the drawback of global characteristics are influenced easily by noise or occlusion.

4.3. locality-sensitive SVM using kernel combination

In the proposed method, local kernels are arranged at different face region. Each local kernel likes a visual cell specialized for local features. Therefore, we choose the Gaussian kernel as the basic kernel function whose stimulus selectivity is suited to develop the visual cells. Then the local kernels are integrated, and the integration kernel is then used for SVM. The summation of local kernels is considered as the integration method which satisfy Mercer's theorem and is more robust to occlusion compared with the product integration [6]. The local

Accepted manuscript

ACCEPTED MANUSCRIPT

Gaussian kernel is defined by

$$K_l\left(x\left(p\right),y\left(p\right)\right)=exp\left(\frac{-\left\|x\left(p\right)-y\left(p\right)\right\|^2}{\sigma_p^2}\right). \tag{7}$$

where p is the label of position, and $x\left(p\right)$ and $y\left(p\right)$ are the local features centered at position p. σ_p is the local variance at position p.

So the summation of local kernels output is defined as follows:

$$K_l\left(x,y\right)=\frac{1}{N_p}\sum_{p=1}^{N_p}K_l\left(x\left(p\right),y\left(p\right)\right)=\frac{1}{N_p}\sum_{p=1}^{N_p}exp\left(\frac{-\left\|x\left(p\right)-y\left(p\right)\right\|^2}{\sigma_p^2}\right), \tag{8}$$

where N_p is the number of local kernels. x and y are the global feature of the whole face.

Local kernel measures detailed similarity and global kernel measures rough similarity. Therefore, there is the case where a local features-based method misclassifies samples which are classified easily by global features. So, in this paper, we combine the global and local kernels to measure both detailed and rough similarity, thus improving the accuracy. The global kernel and local kernels are combined by the summation and then used as a kernel in locality-sensitive SVM. When x and y are the feature of the whole face, we denote the global kernel as $K_g\left(x,y\right)$, which is applied to the whole facial feature. The kernel function is still the Gaussian kernel. The combination kernel of the global and local kernels is defined as:

$$K_{com}\left(x,y\right)=K_g\left(x,y\right)+\frac{1}{N_p}\sum_{p=1}^{N_p}K_l\left(x\left(p\right),y\left(p\right)\right). \tag{9}$$

After the derivation above, we get the form of the combination kernel. This kernel also satisfies Mercer's theorem [6], the proof does not repeat herein. Next we will introduce how to apply the combination kernel to locality-sensitive SVM.

First, we use the combination kernel to measure the similarity between samples and get a correlation matrix K_t.

$$K_t=\left[K_{com}\left(x_i,x_j\right)\right]_{N\times N}, \tag{10}$$

where x_i and x_j denote the ith and jth sample in the feature space respectively. N is the number of training examples. The category is not considered here.

Then we calculate the kernel matrix K across the whole feature space through the correlation matrix K_t.

$$K\left(X_l,X_m\right)=\frac{1}{\lambda}K_t\left(X_l,X_l\right)\times K_t\left(X_l,X_m\right)\times K_t\left(X_m,X_m\right)+\delta_{lm}\times K_t\left(X_l,X_l\right)\times K_t\left(X_m,X_m\right), \tag{11}$$

where δ_{lm} takes the value 1 if $l=m$ or 0 otherwise. X_l and X_m are the samples in the lth and mth regions respectively. Correspondingly, $K_t\left(X_l,X_l\right)$

Accepted manuscript

ACCEPTED MANUSCRIPT

and $K_t(X_m, X_m)$ are sample correlation matrixes for the lth and mth regions respectively. $K_t(X_l, X_m)$ is sample correlation matrix between all the samples in the lth and mth regions. The kernel matrix in equation (11) measures the correlation between the samples within the same region and across the different region.

Therefore the final object function can be rewritten as:

$$\max_{\alpha_{jl}} \sum_{l=1}^{L} \sum_{j=1}^{N_l} \alpha_{jl} - \frac{1}{2} \sum_{l=1}^{L} \sum_{m=1}^{L} \sum_{j=1}^{N_l} \sum_{i=1}^{N_m} \alpha_{jl} \alpha_{im} K(X_l, X_m). \tag{12}$$

Finally, when the coefficient α_{jl} are solved the local classifiers are given as

$$f_t(x) = \sum_{l=1}^{L} \sum_{j=1}^{N_l} \alpha_{jl} K(x, X_l), \tag{13}$$

where $1 \leq t \leq L$.

The calculation of $K(x, X_l)$ is the same as $K(X_l, X_m)$, but we explain it here again. In particular, give a test example x. We first calculate the distance between the test example and the cluster center of the training examples, and determine which cluster it belongs to. Then we calculate the correlation matrix K_t between test example and training examples.

$$K_t(x, X) = [K_{com}(x, X)]_{1 \times N}, \tag{14}$$

where x is the test example and X are the training examples in the whole feature space.

Similarity, we calculate the kernel matrix K between the test example and training examples using the correlation matrix K_t.

$$K(x, X_m) = \frac{1}{\lambda} K_t(x, X_t) \times K_t(X_t, X_m) \times K_t(X_m, X_m) + \delta_{tm} \times K_t(x, X_t) \times K_t(X_m, X_m), \tag{15}$$

where δ_{tm} takes the value 1 if $t = m$ or 0 otherwise. X_m and X_t are the training examples in the mth and tth regions respectively. x is the test example that belongs to the tth region.

5. Experimental Results

In this paper, the performance of the proposed detector is evaluated on the CMU+MIT [45] dataset and FDDB dataset [46].

We train the Adaboost background filter, setting the minhitrate as 0.998, maxfalsealarm as 0.5, stage as 3, so Adaboost classifier can quickly remove more than 80% backgrounds. Since the local features are more robust, so we use multiple local CNNs to get the local facial features. The CNN models contain two convolutional layers (with max-pooling) to extract features, followed by the fully-connected layer and the output layer. Specific, we apply the CNN to a

Accepted manuscript

ACCEPTED MANUSCRIPT

7×7 face region. So we could obtain a 7×7 dimensional feature through the model. Since the sample is 28×28 pixels, so there are total 16 local CNNs applied to a sample. At last, we could obtain $16 \times 7 \times 7$ dimensional features from a sample. We use the features in LS-KC-SVM, so local kernels can be applied to a 7×7 region without overlap, the local kernels' number is 16 too. And the global kernel is applied to the whole facial features.

5.1. Evaluation on CMU+MIT dataset

The CMU+MIT dataset contains test sets A, B, C (test, test-low, new-test) and rotated test set. We use the three test sets (test A, B, C), without the rotated ones, containing 130 images with 511 faces. We evaluate our detector on CMU+MIT and the ROC curves are shown in fig. 3. Our performance is comparable to several popular detection algorithms on this dataset including Viola and Jones [44], Li et al. [47], Jun et al. [48], Zhou et al. [49] and Chen et al. [50]. As shown in fig. 3, our detector is more efficient than other algorithms. Especially for cases at low false positives, our detector can still achieve good results.

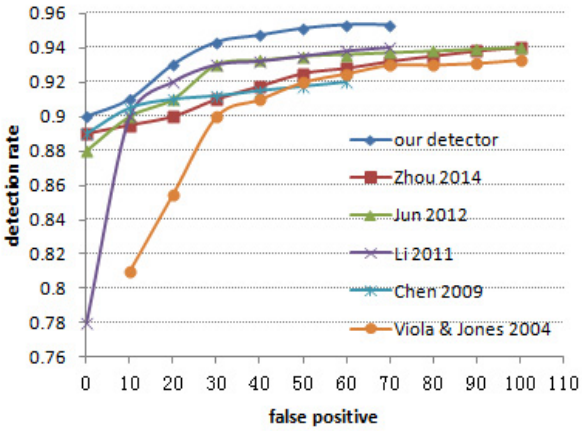


Figure 3: ROC curves of different algorithms on CMU+MIT dataset.

To illustrate the robustness to obstacles such as occlusion, illumination changes and pose changes of our detector, we conduct a set of experiments and compare the detection results with the famous Viola and Jones' detector [44]. In particular, we select several face images under above obstacles from CMU+MIT data set. As showing in fig. 4, some faces that are successfully detected by our detector but failed in Viola and Jones' detector. This is because on one hand we apply the local kernels to the local regions of the features extracted from face. So even though the face is occluded, only partial features are influenced, our method that using global and local kernels simultaneously

ACCEPTED MANUSCRIPT

can still work. On the other hand, we extract local visual features by multiple local CNNs which can represent face more effectively.

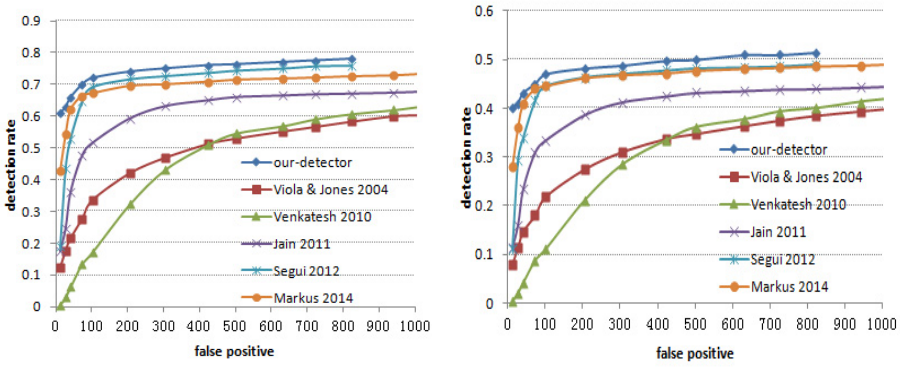
Figure 4: Detect faces under obstacles and compare results with Viola and Jones' detector. (Examples at the first and third row are detection results from Viola and Jones' detector, whereas examples at the second and fourth row are detection results from our face detector.)

5.2. Evaluation on FDDB dataset

The CMU+MIT dataset is a little out-of-date as it only contains gray, relative low-resolution images, and the size of the data set is too small to reflect nowadays data explosion status. Hence, the UMass face detection dataset and benchmark (FDDB) is introduced [46]. It contains 2845 images with a total of 5771 faces under a wide range of conditions. It is a large-scale face detection benchmark with standardized evaluation process. We follow the required evaluation procedure to report our detection performance with the toolbox provided by the authors. In this experiment, we compare to some available results on the benchmark [44, 51, 52, 53, 54] to demonstrate the effectiveness of our method, as showing in Fig. 5.

FDDB uses ellipse face annotations and defines two types of evaluations: the discontinuous score and continuous score. In the discrete setting, a detection window is considered correct if its intersection-over-union ratio with respect to an annotated face region is larger than 0.5. This criterion is commonly used in

ACCEPTED MANUSCRIPT



(a) (b)

Figure 5: (a)Continuous score ROC curves and (b) Discrete score ROC curves for different methods on FDDB dataset

object detection evaluation. In the continuous setting, the overlapping ratio is used as a weight for every detection window. This criterion is much stricter. So we can see in Fig. 5, the detection rate in continuous score ROC curve is much lower than discrete score ROC curve. But it is obvious that our detector outperforms others under both protocols.

FDDB dataset is very challenging because the faces have large appearance variations due to some real-world obstacles, such as viewpoint, extreme illuminations and expression changes. Facing the challenging dataset, our method still achieves good results. As showing in fig. 6, we show some faces under partial occlusion, pose and expression changes. For example there are some faces are partial occluded, such as (a), (b), (e) and (f), our detector can handle them well. Besides, some rotated faces are successfully detected, such as (b), (c), (g) and (h). However, if the rotation angle of face is too large, our detector cannot detect these faces, such as (c). This is because our training examples only cover $\pm 15^{\circ}$ up-down out-of-plane rotation (Pitch) and $\pm 20^{\circ}$ leftCright out-of-plane rotation, so our detector can effectively detect faces within this angle range. It will does not work if the rotation angle is too larger.

6. Conclusion

In this paper, we propose a locality-sensitive SVM using kernel combination algorithm for robust face detection. Since a number of factors such as pose changes, exaggerated expressions and extreme illuminations can lead to large intra-class variations in different face images, so we employ the locality-sensitive SVM to handle this issue. Then we introduce the combination kernel method to the locality-sensitive SVM to measure the detailed and rough similarity both, which improving the generalization ability. Besides, since local kernel is more robust than global kernel, to better exert the function of LS-KC-SVM, we adopt

Accepted manuscript

ACCEPTED MANUSCRIPT

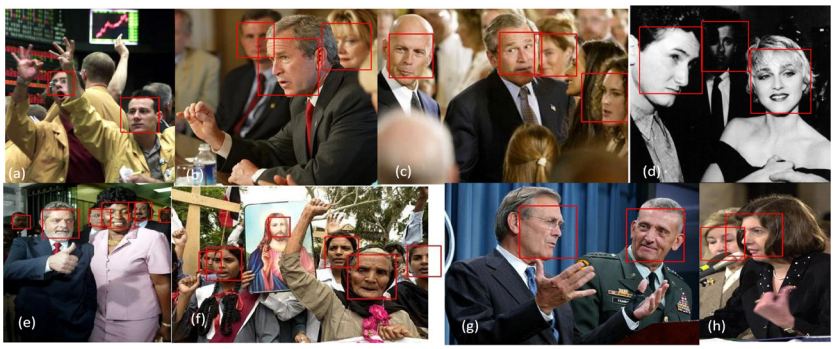


Figure 6: Some examples of face detection results on FDDB dataset.

multiple local CNNs which apply to different face region to jointly learn discriminate local features used in our LS-KC-SVM algorithm. Evaluations on the datasets demonstrate the performances of the proposed method.

7. Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 61371156, Anhui Province Science and Technology Research Programs under Grant 1401B042019. The authors would like to thank the anonymous reviewers for their helpful and constructive comments and suggestions.

References

[1] L. Lin, X. Wang, W. Yang, et al., Discriminatively trained and-or graph models for object shape detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (5) (2014) 959–972.

[2] C. Li, L. Lin, W. Zuo, et al., Sold: Sub-optimal low-rank decomposition for efficient video segmentation, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2015) 5519–5527.

[3] K. Wang, L. Lin, J. Lu, et al., Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence, IEEE Transactions on Image Processing 24 (10) (2015) 3019–3033.

[4] H. Cheng, P. Tan, R. Jin, Localized support vector machine and its efficient algorithm, Proc. SIAM Intl. Conf. Data Mining 44 (12) (2014) 2431–2442.

[5] G. Qi, Q. Tian, T. Huang, Locality-sensitive support vector machine by exploring local correlation and global regularization, IEEE conference on computer vision and pattern recognition (2011) 841–848.

Accepted manuscript

ACCEPTED MANUSCRIPT

[6] K. Hotta, Robust face recognition under partial occlusion based on support vector machine with local gaussian summation kernel, Image and Vision Computing 26 (11) (2008) 1940–1498.

[7] K. Hotta, View independent face detection based on horizontal rectangular features and accuracy improvement using combination kernel of various sizes, Pattern Recognition 42 (3) (2009) 437–444.

[8] C. Zhang, Z. Zhang, Improving multiview face detection with multi-task deep convolutional neural networks, Applications of Computer Vision (2014) 1036–1041.

[9] J. P. Papa, A. X. Falcao, V. H. C. Albuquerque, et al., Efficient supervised optimum-path forest classification for large datasets, Pattern Recognition 45 (1) (2012) 512–520.

[10] R. Y. M. Nakamura, L. M. G. Fonseca, J. A. Santos, et al., Nature-inspired framework for hyperspectral band selection., IEEE Transactions on Geoscience and Remote Sensing 52 (2014) 2126–2137.

[11] R. J. Pisani, R. Y. M. Nakamura, P. S. Riedel, et al., Towards satellite-based land cover classification through optimum-path forest, IEEE Transactions on Geoscience and Remote Sensing 52 (10) (2014) 6075–6085.

[12] A. S. Iwashita, J. P. Papa, A. N. Souza, et al., A path- and label-cost propagation approach to speedup the training of the optimum-path forest classifier, Pattern Recognition Letters 40 (2014) 121–127.

[13] N. Zhang, R. Farrell, T. Darrell, Pose pooling kernels for sub-category recognition, Computer Vision and Pattern Recognition (2012) 3665–3672.

[14] R. Farrell, O. Oza, N. Zhang, et al., Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance, International Conference on Computer Vison (2011) 161–168.

[15] Z. Liu, H. Li, W. Zhou, et al., Contextual hashing for large-scale image search, IEEE Transactions on Image Processing 23 (4) (2014) 1606–1614.

[16] W. Zhou, M. Yang, H. Li, et al., Towards codebook-free: Scalable cascaded hashing for mobile image search, IEEE Transactions on Multimedia 16 (3) (2014) 601–611.

[17] Z. Liu, H. Li, L. Zhang, et al., Cross-indexing of binary sift codes for large-scale image search, IEEE Transactions on Image Processing 23 (5) (2014) 2047–2057.

[18] Y. Pang, S. Wang, Y. Yuan, et al., Learning regularized lda by clustering, IEEE Transactions on Neural Networks and Learning Systems 25 (12) (2014) 2191–2201.

Accepted manuscript

ACCEPTED MANUSCRIPT

[19] Y. Pang, Z. Ji, P. Jing, et al., Ranking graph embedding for learning to rerank, IEEE Transactions on Neural Networks and Learning Systems 24 (8) (2013) 1292–1303.

[20] Q. Hu, D. Yu, Z. Xie, et al., Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2) (2006) 191–201.

[21] T. Liu, D. Tao, Classification with noisy labels by importance reweighting, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2015) 1–14.

[22] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2015) 1–14.

[23] C. Gong, T. Liu, D. Tao, et al., Deformed graph laplacian for semisupervised learning, IEEE Transactions on Neural Networks and Learning Systems PP (99) (2015) 1–14.

[24] C. Gong, D. Tao, K. Fu, et al., Flap: Ficks law assisted propagation for semi-supervised learning, IEEE Transactions on Neural Networks and Learning Systems 26 (9) (2015) 2148–2162.

[25] C. Xu, D. Tao, C. Xu, Large-margin multi-view information bottleneck, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8) (2014) 1559–1572.

[26] S. Ding, B. Qi, Research of granular support vector machine, Artificial Intelligence Review 38 (1) (2012) 1–7.

[27] Y. Pang, K. Zhang, Y. Yuan, et al., Distributed object detection with linear svms, IEEE Transactions on Cybernetics 44 (11) (2014) 2122–2133.

[28] S. Ding, Z. Shi, Track on intelligent computing and applications, Neurocomputing 130 (2014) 1–2.

[29] S. Ding, X. Hua, Recursive least squares projection twin support vector machines, Neurocomputing 130 (2014) 3–9.

[30] Q. Hu, L. Ding, J. He, Lp norm constraint multi-kernel learning method for semi-supervised support vector machine, Journal of Software 24 (11) (2013) 2522–2534.

[31] N. Zhang, R. Farrell, F. Iandola, et al., Deformable part descriptors for fine-grained recognition and attribute prediction, International Conference on Computer Vision (2013) 729–736.

[32] D. Huang, C. Zhu, Y. Wang, et al., Hsog: A novel local image descriptor based on histograms of the second-order gradients, IEEE Transactions on Image Processing 23 (11) (2014) 4680–4695.

Accepted manuscript

ACCEPTED MANUSCRIPT

[33] D. Huang, M. Ardabilian, Y. Wang, et al., 3-d face recognition using elbp-based facial description and local feature hybrid matching, IEEE Transactions on Image Processing 7 (5) (2012) 1551–1565.

[34] H. Li, D. Huang, J. M. Morvan, et al., Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors, International Journal of Computer Vision 113 (2) (2015) 128–142.

[35] D. Weng, Y. Wang, M. Gong, et al., Derf: Distinctive efficient robust features from the biological modeling of the p ganglion cells, IEEE Transactions on Image Processing 24 (8) (2015) 2287–2302.

[36] Q. Hu, W. Pedrycz, D. Yu, et al., Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 40 (1) (2010) 137–150.

[37] D. Huang, C. Shan, M. Ardabilian, et al., Local binary patterns and its application to facial image analysis: A survey, IEEE Transactions on Systems, Man, and Cybernetics 41 (6) (2011) 765–781.

[38] S. Ding, L. Lin, G. Wang, et al., Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015) 2993–3003.

[39] S. Ding, W. Jia, C. Su, et al., Research of neural network algorithm based on factor analysis and cluster analysis, Neural Computing and Applications 20 (2) (2011) 297–302.

[40] S. Ding, H. Jia, J. Chen, et al., Granular neural networks, Artificial Intelligence Review 41 (3) (2014) 373–384.

[41] H. Li, Z. Lin, X. Shen, et al., A convolutional neural network cascade for face detection, The IEEE Conference on Computer Vision and Pattern Recognition (2015) 5325–5334.

[42] S. Li, J. Xing, Z. Niu, et al., Shape driven kernel adaptation in convolutional neural network for robust facial trait recognition, The IEEE Conference on Computer Vision and Pattern Recognition (2015) 2522–2534.

[43] S. Ding, H. Jia, L. Zhang, et al., Research of semi-supervised spectral clustering algorithm based on pairwise constraints, Neural Computing and Applications 24 (1) (2014) 211–219.

[44] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[45] H. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transcations on Pattern Analysis and Machine Intelligences 20 (1) (1998) 23–38.

Accepted manuscript

ACCEPTED MANUSCRIPT

[46] V. Jain, E. Learned-Miller, Fddb: A benchmark for face dfetection in unconstrained settings, University of Massachusetts Amherst Technical Report (2010).

[47] J. Li, T. Wang, Y. Zhang, Face detection using surf cascade, IEEE International Conference on Computer Vision Workshops (2011) 2183–2190.

[48] B. Jun, D. Kim, Robust face detection using local gradient patterns and evidence accumulation, Pattern Recognition 45 (9) (2012) 3304–3316.

[49] S. Zhou, J. Yin, Face detection using multi-block local gradient patterns and support vector machine, Journal of Computational Information Systems 10 (4) (2014) 1767–1776.

[50] Y. Chen, C. Han, C. Wang, et al., A cnn-based face detector with a simple feature map and a coarse-to-fine classifier, IEEE Transcations on Pattern Analysis and Machine Intelligences PP (99) (2009) 1–13.

[51] B. Venkatesh, S. Marcel, Fast bounding box estimation based face detection, Proc. Workshop on Face Detection of the European Conference on Computer Vision (2010).

[52] V. Jain, E. Learned-Miller, Online domain adaptation of a pre-trained cascade of classifiers, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2011) 577–684.

[53] S. Segui, M. Drozdzal, P. Radeva, et al., An integrated approach to contextual face detection, International Conference on Pattern Recognition Applications and Methods (2012) 90–97.

[54] N. Markus, M. Frljak, I. Pandzic, A method for object detection based on pixel intensity comparisons organized in decision trees, arXiv preprint arXiv:1305.4537 (2013).

Accepted manuscript