Enhanced Pictorial Structures for Precise Eye Localization Under Uncontrolled Conditions

Xiaoyang Tan[1,2] Fengyi Song[1] Zhi-Hua Zhou[2] Songcan Chen[1]
[1]Department of Computer Science and Engineering
Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
[2]National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{x.tan, f.song, s.chen}@nuaa.edu.cn zhouzh@lamda.nju.edu.cn

Abstract

In this paper, we present an enhanced Pictorial Structure (PS) model for precise eye localization, a fundamental problem involved in many face processing tasks. PS is a computationally efficient framework for part-based object modelling. For face images taken under uncontrolled conditions, however, the traditional PS model is not flexible enough for handling the complicated appearance and structural variations. To extend PS, we 1) propose a discriminative PS model for a more accurate part localization when appearance changes seriously, 2) introduce a series of global constraints to improve the robustness against scale, rotation and translation, and 3) adopt a heuristic prediction method to address the difficulty of eye localization with partial occlusion. Experimental results on the challenging LFW (Labeled Face in the Wild) database show that our model can locate eyes accurately and efficiently under a broad range of uncontrolled variations involving poses, expressions, lightings, camera qualities, occlusions, etc.

1. Introduction

The task of detecting and localizing eye positions in a given image which contains a face is crucial for the initialization of many face processing applications such as face tracking, face recognition, face expression analysis, eye behavior analysis, *etc*. There is a subtle difference between eye detection and eye localization; that is, the latter generally requires a much more accurate prediction of the eye positions (usually only a few pixels of errors are allowed) than the former. Recent research has disclosed that an inaccurate eye localization will cause serious problems for automatic face recognition systems [19], especially for those based on techniques heavily relying on the quality of the geometrically normalized face images, such as eigenface and fisherface.

Figure 1. Illustration of the complicated appearances of eyes (*images from the LFW database [10]*).

Similar to other general object detection tasks such as face detection [22], people detection [3] and animal detection [12], the major challenge of eye localization comes from the fact that, as illustrated in Figure 1, the appearances of eyes are complicated due to various factors, from normal behaviors of eyes (*e.g.*, opening, closing, *etc*.) to environmental changes (*e.g.*, outdoor lighting, pose, scale, reflection of glasses, partial occlusion by hairs, *etc*.). To address the problem, various approaches have been proposed. These approaches can be roughly classified into three categories [23, 26], *i.e.*, template-based approaches, appearance-based approaches and feature-based approaches.

Most of the early methods, such as the deformable templates methods [24], belong to the template-based category, where a generic eye model is designed based on eye shapes and then used to search eyes in the images. These methods usually have good accuracy, however, generally they are computationally expensive and require good image contrast which is not always available in practice.

Appearance-based approaches aim to localize eyes based on their photometric appearance using various statistical classification techniques, such as principal component analysis (*e.g.*, eigeneyes [16]), support vector machines [9], neural networks [5], Boosting [14], *etc*. Everingham *et al*. [2] compared several kinds of appearance-based approaches and found that the simple Bayesian model outperforms the other methods including a regression-based and a Boosting-based method. In general, appearance-based eye localiza-

tion methods are more robust against complicated appearance changes than template-based ones due to the capability of learning from examples, yet the problem of reliably separating true small eye regions from other regions with a low false positive rate remains unsolved.

Feature-based approaches attempt to exploit special characteristics of the eyes such as the dark pupil and white sclera to distinguish the eyes from other objects [25]. Such eye-specific features can be regarded as the context in which the true eye lies. However, when the input image is with low contrast or with closed or occluded eyes, the eye-specific features are difficult to be detected.

Overall, most of the existing methods are only feasible under rather constrained conditions, and few work studies the problem of precise eye localization under uncontrolled conditions such as situations with extreme lighting changes, large expression variations and partial occlusions, although addressing these issues is important for real applications, *e.g.*, uncontrolled face recognition [6, 7].

In this paper, we present a new approach for precise and robust eye localization. Previous studies [1, 3] disclosed that the accuracy as well as the false positive rate of feature localization can be improved effectively by exploiting the context information of the object of interest. Here, by context information of eyes, we mean any facial features that are helpful for identifying the eye positions, such as the places of nose, mouth, *etc*. Compared with mouth, nose can be detected more reliably due to the fact that its appearance is less sensitive to expression changes and occlusions (say, by beard), hence being used in this work. Actually, the nose position may also be useful in face normalization algorithms.

The Pictorial Structure (PS) model [3, 4] is well suited for our purpose. It is a computationally efficient framework for part-based modelling and recognition of objects, and has been successfully applied to face identification [3], people finding [3] and other object recognition and detection tasks [12]. The essence of the PS model is to consider the components (or parts) of an object in the context of its overall interior configurations, aiming to finding out the location of each component as well as its spatial configuration through encoding them into a global object function. In contrast to pure appearance-based approaches where each object is handled by a single appearance model, the PS method provides a powerful framework for modelling an object in terms of its components and the geometrical relationship between components.

The main contribution of this paper is to enhance the traditional PS model such that it can be used to handle the complicated appearance and structural changes of eyes under uncontrolled conditions. Extensive experiments on the challenging LFW (Labeled Face in the Wild) database show that the proposed model can localize eyes accurately and efficiently under uncontrolled conditions.

The rest of this paper is organized as follows. Section 2 briefly introduces the Pictorial Structure model. Section 3 proposes our enhanced PS model. Section 4 presents methods for fitting the learned PS model to test images and handling partial occlusion. Section 5 reports on our experiments. Finally, Section 6 concludes.

2. Background

In this section, we briefly introduce the statistical framework of Pictorial Structure following the denotations in [3]. In PS an object is first decomposed into parts and then the best part candidates are searched subject to some spatial constraints such that the likelihood of generating the concerned image is maximized. Hence a PS model can also be viewed as a specific Markov Random Field (MRF) with parts as its sites.

The PS model [3, 4] can be expressed naturally in terms of an undirected graph $G = (V, E)$, where the vertices $V = \{v_1, \cdots, v_{N_p}\}$ correspond to N_p parts, and the edge set $E = \{(v_i, v_j), i \neq j\}$ characterizes the local pairwise spatial relationship between different parts. An instance of an object is given by a configuration $L = (l_1, \cdots, l_i)$, where each $l_i = (x_i, y_i)$ specifies the location of the component v_i on the image plane. For example, a face can be represented by four parts (*i.e.*, two eyes, one nose and one mouth) and the spatial relationship between these four parts. Hence the appearance and structural information are combined into a unified framework.

Specifically, given an image I containing a face, the likelihood of generating this image by the facial parts at some locations is $p(I|L, \theta)$, where θ is the model parameter. Assume that we are working on the region output by a face detector, and therefore we need not model the background. To infer the locations of the facial parts from this model, we can look for the maximum *a posterior* $p(L|I, \theta)$, *i.e.*, the probability that a face configuration is L given the model θ and an image I. According to Bayes rule, the posterior can be written as

$$p(L|I, \theta) \propto p(I|L, \theta)p(L|\theta), \tag{1}$$

where $p(I|L, \theta)$ is the generative model of appearance and $p(L|\theta)$ measures the prior probability that a face appears at the location L. Here the model parameter is denoted by $\theta = (u, c)$, where $u = (u_1, \cdots, u_{N_p})$ expresses the appearance while $c = \{c_{ij}|(v_i, v_j) \in E\}$ expresses structural constraints on edges. Felzenszwalb and Huttenlocher [3] also included the set of edges E as model parameter in order to simplify the underlying graphical structure (*e.g.*, letting it be a tree), but this is not necessary for our purpose of eye localization.

Assuming that the parts are statistically independent, we

have

$$p(I|L,\theta) = p(I|L,u) \propto \prod_{i=1}^{N_p} p(I|l_i,u_i), \tag{2}$$

where the appearance of each component can be modelled by unimodal Gaussian distribution $p(I|l_i,u_i) \propto \mathcal{N}(\alpha(l_i),\mu_i,\Sigma_i)$.

With a similar independent assumption on the spatial relationship between pairs of parts, we have the structural model

$$p(L|\theta) = p(L|c) = \prod_{(v_i,v_j)\in E} p(l_i,l_j|c_{ij}), \tag{3}$$

where the edge constrains between the components can also be modelled by Gaussian distribution

$$\begin{aligned} &p(l_i,l_j|c_{ij}) = p(x_{l_i},x_{l_j}|c_{ij})p(y_{l_i},y_{l_j}|c_{ij}) \\ &= \mathcal{N}(x_{l_i}-x_{l_j},s_{ij},\Sigma_{ij})\mathcal{N}(y_{l_i}-y_{l_j}),s'_{ij},\Sigma'_{ij}). \end{aligned} \tag{4}$$

Plugging (3) and (2) into (1), we get the global objective function

$$p(L|I,\theta) \propto \left\{ \prod_{i=1}^{N_p} p(I|l_i,u_i) \prod_{(v_i,v_j)\in E} p(l_i,l_j|c_{ij}) \right\}. \tag{5}$$

The first term at the right hand of (5) is the appearance model, while the second term expresses the spatial constraints. Taking the negative logarithm of (5), we get the energy function of the PS model. For tree-structured models, Felzenszwalb and Huttenlocher [3] have developed computationally efficient algorithms for learning from training data and fitting test images.

3. Enhanced Pictorial Structure

Possible criticisms to the PS model include that the unimodal generative appearance model may not be capable to provide a good approximation to multimodal distributions of eye appearance under uncontrolled conditions, and the local pairwise prior may impose overly strong constraints on the spatial relationship between the parts. We will address these issues in this section.

3.1. Discriminative Pictorial Structure

One important observation from our work is that in a complicated setting, the distribution of eye patterns are multimodal in nature, which could not be approximated well with a unimodal Gaussian model. [1] More importantly, our goal is to localize the eyes precisely rather than describe the whole face object using facial features. So, a discriminative model which focuses more on the points that count is more appropriate.

For this purpose, we introduce a class label $z \in \{1,\cdots,N_p\}$ for each part, denoting one of N_p possible semantic labels (*e.g.*, *Right-eye*, *Left-eye*, *Nose*, *etc*.) for that part. The appearance model of (2) can then be rewritten as

$$\begin{aligned} p(I|L,\theta) &= p(I|L,u) \propto \prod_{i=1}^{N_p} p(I|l_i,u_i) \\ &= \prod_{i=1}^{N_p} \sum_z p(I,z|l_i,u_i) = \prod_{i=1}^{N_p} \sum_z p(I|z)p(z|l_i,u_i). \end{aligned} \tag{6}$$

Note that $p(I|z)$ gives the probability of generating an image given its part labels. This can be useful if we detect eyes directly in general background (*i.e.*, without face context). In our setting, however, the localizer takes the image region output by a face detector as input, where the part labels of interest are assumed to be known. Hence we need not model any preference over the labels in image I; this implies that $p(I|z)$ is a constant and we can omit it for simplicity. This simply reduces (6) to

$$p(I|L,\theta) \propto \prod_{i=1}^{N_p} \sum_z p(I|z)p(z|l_i,u_i) = \prod_{i=1}^{N_p} \sum_z p(z|l_i,u_i). \tag{7}$$

Furthermore, we restrict the region for searching each kind of part (*e.g.*, right eye, left eye, nose, *etc*.) by collecting statistics about their true positions with respect to the corresponding output window of our face detector. This allows us to model only the parts from certain predefined region with known labels, thus not only reducing the number of candidate locations for each part significantly, but also simplifying our model significantly to

$$p(I|L,\theta) \propto \prod_{i=1}^{N_p} \sum_z p(z|l_i,u_i) = \prod_{i=1}^{N_p} p(z_i|l_i,u_i), \tag{8}$$

where the posterior distribution, $p(z_i|l_i,u_i)$, characterizes the probability that the label of the part v_i is a certain label value z_i given its appearance u_i and position l_i. To this end we derive a discriminative model completely within the PS framework. There were many work [1, 12] which use a discriminative appearance model, yet none has been designed for eye localization.

There are many ways to approximate $p(z_i|l_i,u_i)$ [8], among which the energy-based methods (*e.g.*, conditional random fields and logistic regression) are very popular due to the convenience of incorporating arbitrary functions of training examples and the nonparametric nature in the sense of no need to assume any particular distribution. Here, for simplicity we choose to model the decision boundary with

[1] One option is to replace the simple unimodal Gaussian generative model with a more complex one like Gaussian mixture model (GMM), at the cost of higher model complexity and hence lower detection efficiency.

a hyperplane in some feature space, and then fit it with a logistic sigmoid function to give an approximation probability of interest [18].

In particular, we use a support vector machine (SVM) to calculate the optimal separating hyperplane in the feature space, which corresponds to a nonlinear boundary in the complicated input space. Other options such as Relevance Vector Machine (RVM [21]) and Adaboost can also be considered. The SVM solver outputs an optimal hyperplane in the general form of $\hat{f}(u) = \hat{\beta} + \sum_{i=1}^{N} \hat{\alpha}_i K(u, u_i)$, where N is the number of training examples and K is a predefined kernel function (Gaussian kernel is used in this paper). Then, the estimate of posterior probability is fitted by $p(z = 1|u) = 1/\{1 + e^{-A\hat{f}(u)-B}\}$ using the binomial log-likelihood as loss function [18]. [2]

In practice, the learned support vectors are usually not sufficiently sparse to meet the requirement of real-time detection. One way to address this issue is to use a *reduced set* method to reduce the number of support vectors and hence the computational complexity. In this work we adopt the method in [15] for this purpose and find that typically 10 to 20 support vectors are enough for each part.

To speed up the detection further, in implementation we train two hierarchical SVM classifiers. The first-level SVM works on the simple gray-intensity feature of each part in order to reject quickly a large number of negatives. The second-level SVM uses the Gabor features as input, which are known to be robust against scale, illumination and other appearance changes. The cascaded SVMs are trained in a way similar to the Viola-Jones face detector [22], where a threshold is learned for each level of SVM on the training set according to the given performance criterion (*i.e.*, true positive rate & false positive rate). Through this process, we reduce the number of candidate positions for each parts from about 400 to 20 efficiently and effectively.

3.2. Global and Local Constraints

As introduced before, the traditional PS models the spatial configuration between a pair of parts with a separate Gaussian distribution, yet we find that such local pairwise constraints may be overly strong and only work well under normal conditions (*i.e.*, the spatial relationship does not change too much), while in practice the configuration between facial parts may be deformed largely due to variation such as scale, rotation and expression changes.

To illustrate such limitation, in Figure 2 we give an example of the fitted Gaussian distribution of relative location of the left eye with respect to the location of the right eye (*c.f*. Eq.(3)) along the vertical axis. This can be understood as some confidence score indicating to what extent a given

[2]Here we abuse the notation of z to indicate whether or not a particular facial feature (*e.g.*, left eye) with appearance parameter u is detected.

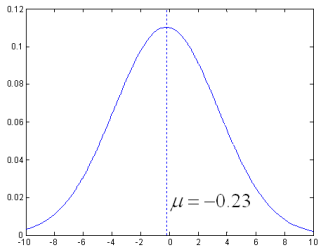


Figure 2. The fitted Gaussian distribution of relative location of the left eye with respect to the location of the right eye (*c.f*. Eq.(3)) along the vertical axis.

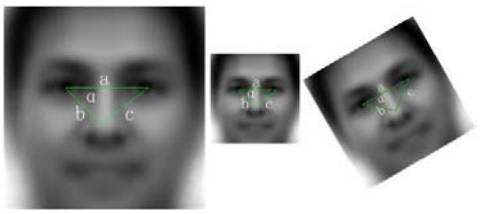


Figure 3. Illustration of some commonly encountered structural changes between eyes. From left to right: the original, scaled and rotated images.

configuration follows a prior spatial regularity. The confidence score reaches its peak at the position of mean value and decreases otherwise. In the example of Figure 2, the mean value of the distribution is very close to zero, which implies that the best vertical positions of two eyes should be on the same horizon line, thus imposing a very strong constraint on the spatial relationship between two eyes, easily being violated in the case of rotation, *e.g.*, the rightmost image in Figure 3. Similarly, the constraint on relative locations in the horizontal axis is problematic since face images may be taken under different scales, *e.g.*, the middle image in Figure 3.

To address the above-mentioned deficiency, we introduce an improved structural description method which is more robust to rotation, scale and translation. The idea is to incorporate global structural constraints with local ones. By global structure we mean the spatial relationship among more than two parts. In our case, this is simply the triangle formed by the two eyes and the nose (*c.f*. the leftmost image in Figure 3); if more facial features need to be detected, more triangles like this can be added. Moreover, instead of modelling the *relative location* between parts in previous studies, we adopt *relative distances* to encode the local constraints.

In particular, we model three pieces of structural information based on the facial feature triangle, that is, 1) the length of each edge, 2) the length ratio between a pair of edges, and 3) the inner angle between any two edges. The first one is local constraint and the latter two are global constraints. They have nice affine-invariant properties. First,

the edge lengths are invariant to 2D rotation and translation; Second, both the length ratio and inner angle between two edges are invariant to scale, 2D rotation and 2D translation. As the consequence, these three structural measurements make our model more robust against structural changes caused by expression and pose variations which are commonly encountered in real applications.

In implementation, the length of edge L_{ij} is defined by the Euclidean distance between parts v_i and v_j on the image plane, *i.e.*,

$$L_{ij} = \sqrt{(x_{l_i} - x_{l_j})^2 + (y_{l_i} - y_{l_j})^2} \quad \forall i, j \in \{1, 2, 3\}, i \neq j, \tag{9}$$

and the length ratio r_{ij} and the cosine angle $cos(\alpha_{ij})$ between edges are defined respectively as

$$r_{ij} = L_{ik}/L_{jk} \tag{10}$$

and

$$cos(\alpha_{ij}) = \frac{L_{ij}^2 + L_{jk}^2 - L_{jk}^2}{2L_{ij}L_{ik}}. \tag{11}$$

All of the above definitions can be expressed as a function of edge length e. Here, we define $e_1 \triangleq L_{12}$, $e_2 \triangleq L_{13}$ and $e_3 \triangleq L_{23}$, and thus the energy function $E(e_1, e_2, e_3)$ of our structural model is

$$E(e_1, e_2, e_3) = -\sum_{i=1}^{3} \varphi_1(e_i) - \sum_{i,j=1, i\neq j}^{3} \varphi_2(e_i, e_j) - \sum_{i,j,k=1, i\neq j\neq k}^{3} \varphi_3(e_i, e_j, e_k), \tag{12}$$

where φ_1, φ_2 and φ_3 are the potentials corresponding to the aforementioned three structural constraints, respectively, modelled by Gaussians. Considering that different subjects have different configurational biases, we weight these Gaussian potentials with their variances respectively before combining them; the weights can be considered as a prior for different constraints. The values of the above structural parameters can be learned from independent training data using the maximum likelihood method.

4. Matching Algorithm

4.1. Fitting the Model

The best fit of the enhanced PS model in an unseen test image is

$$L^* = \underset{L}{\operatorname{argmin}} \left(\sum_{i=1}^{N_p} \left(-\log p\left(z_i | l_i, u_i\right) \right) + E\left(e_1, e_2, e_3\right) \right). \tag{13}$$

An exact inference for the non-tree-structured model is difficult and so we adopt an approximation method. We first run the appearance model to filter the noise candidates out, then find the best configuration which minimizes the structural model. The major computational cost comes from the first stage where $(N_{leye} + N_{reye} + N_{nose})$ positions need to be examined. The cascaded-SVM classifier described in Section 3.1 effectively reduces the cost involved in this stage. The second stage is proven to be very efficient due to that it works on the 2D image plane and that the first stage helps reduce considerable number of candidate positions. Our current implementation has not been optimized, but it takes about only 0.1 seconds to fit a 100×100 image on a 2.8GHz P4 machine; we believe it is acceptable for many real applications.

Figure 4. Illustration of three typical eye occlusion conditions. From left to right: one or two eyes are weakly occluded; only one eye is occluded and almost undetectable; both eyes are completely occluded.

4.2. Predicting with Partial Occlusion

In many real-world images such as those collected from the web, the eyes may be partially occluded by hand, hair, glasses, *etc*. This will impose difficulty for our appearance model. Here, we adopt a heuristic method to handle the following typical partial occlusions, as illustrated in Figure 4. 1) One or two eyes are weakly occluded but can still be detected by our discriminative model (*i.e.*, its posterior exceeds some threshold). This situation does not need any special treatment and we simply use the learned model to fit it as if they were not occluded. 2) Only one eye is occluded and could not be detected reliably but the other eye and the nose have good response values. In this situation, we can use the positions of the two reliable features to predict that of the difficult one, *i.e.*, only the structural model is used to predict the position of the occluded eye. A similar method has been used by Leung *et al*. [13] in finding faces in cluttered scenes. 3) Both eyes are occluded and undetectable. This is the worst case, however, if the nose can be detected, we can still use it as the starting point to trigger the structural model fitting; otherwise we have to include more context facial features or rely on the prior positions of eyes output by the face detector.

5. Experiments

We evaluate the proposed method on databases including LFW (Labeled Faces in the Wild) [10], FERET [17], *etc*., and achieve encouraging results on all databases. Due to the

page limit, considering that LFW is the most challenging database, we only present the results on LFW in this section.

5.1. Data

LFW [10] is a large WWW database in which all faces were collected from real-life featuring variations on pose, lighting, expression, background, camera quality, occlusion and image noise. The appearance of eye region in an image is largely changed by these variations, hence posing great challenge to eye localization techniques. From the total 13,233 target face images, we randomly select 2,000 images and split them into two data sets with 1,000 images each, using one for training and the other for testing. Various transformations such as rotation, blurring, contrast modification and addition of Gaussian white noise are then applied to the initial set of training images, yielding about 17,000 new images in total. The final training set contains 3,000 images, among which 500 images are randomly selected from the initial set and 2,500 from the generated images. Some example test images are shown in Figure 1. The LFW database does not provide the ground-truth eye positions, so we invite two human volunteers to manually label the eye positions and take the average as the ground-truth. We define the ground-truth as the pupil of eyes by default, and when the eyes are completely occluded (say, by sunglasses), we define it as the center of eyeball. The extensive labor work is the reason why we have used a subset of 2,000 images instead of the total 13,233 images in LFW.

5.2. Settings

All images undergo the same preprocessing pipeline prior to analysis, as illustrated in Figure 5. For face detection, we use the publicly available implementation of the Viola-Jones face detector [22] from the OpenCV library; it outputs a bounding box indicating the predicated center of the face and its scale. After verifying with a SVM classifier, we scale the detected face images to a standard size of 100×100 pixels (*i.e.*, the maximum likelihood estimation of the scale of face detector windows), which helps reduce the amount of translation and scale variation in the image. The geometrically normalized images then undergo illumination normalization which compensates for low-frequency lighting variations and suppresses noise with a Difference of Gaussians filter. This technique has recently been shown to lead to state-of-the-art performance on face recognition [20] and we find it is also useful in eye localization. Finally, using statistics about the eye/nose positions relative to face detector window, the search regions for two eyes and the nose are separately estimated [2].

The training patches consist of both positive and negative samples. They are collected according to the ground-truth, while the negative set undergoes an additional bootstrap procedure to filter out samples with low utility value.

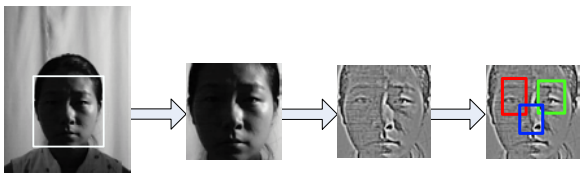


Figure 5. Illustration of the overall preprocessing pipeline (the bounding box in the rightmost image is the search region).

The size of each patch is determined by cross-validation on the training set.

To evaluate the precision of eye localization, we adopt the measure proposed by Jesorskey *et al.* [11]. The localization criterion is defined in terms of the eye center positions according to

$$d_{eye} = \frac{max(d_l, d_r)}{\|C_l - C_r\|}\,, \tag{14}$$

where C_l and C_r are the ground-truth positions and d_l and d_r are the Euclidean distances between the detected eye centers and the ground-truths. For eye detection, usually $d_{eye} < 0.25$ is required [25], but for eye localization, $d_{eye} < 0.05$ or $d_{eye} < 0.1$ is more desired.

We compare our method with the traditional PS method [3] and the Bayesian method [2]. The simple Bayesian method has recently been shown to perform better on eye localization than several classical appearance-based approaches such as regression method, Boosting-based method and SVM-based method [2].

5.3. Results

Figure 6 plots the cumulative error distribution curves of the compared methods, where the horizontal axis is normalized Euclidean distance (*i.e.*, d_{eye}) between the predicted eye position and ground-truth position, while the vertical axis is the cumulative localization score, showing the percentage of images that have been successfully processed corresponding to a certain localization error. As expected, the PS method outperforms the Bayesian method, partially due to the fact that the PS method has stronger capability of structural validation while many images in the database exhibit different extent of rotation either in the plane or out of plane. By replacing the generative appearance model with a discriminative model, and by incorporating the global shape constraints, our enhanced PS method performs the best among the compared methods. To make this clear, we tabulate the percentage of successful localization subject to $d_{eye} < 0.05$ and $d_{eye} < 0.1$ in Table 1. It can be found that our method promotes the performance of the traditional PS method from 53.4% to 80.2% when $d_{eye} < 0.05$, and achieves the best correct localization rate of 98.4% at $d_{eye} < 0.1$, about 5% higher than the other two methods. Figure 5.3 presents some example images

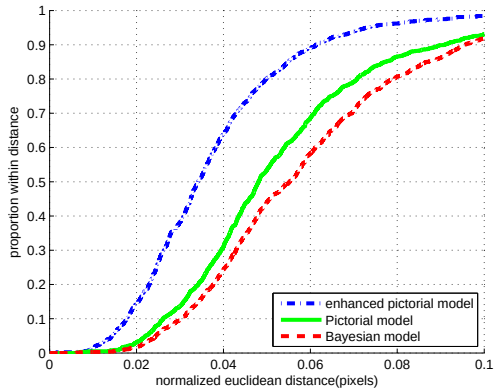


Figure 6. Comparing the cumulative error distribution curves.

in which our method correctly localizes the eyes while the other two methods failed; most of the incorrect localizations by the other two methods are caused by the complicated appearance or structural changes due to lighting, pose, expression and partial occlusions.

We also examine the relative localization performance of each eye. Breifly, here we only report the results on the right eye (results on the left eye are similar). Table 2 gives an overall estimation of the performances of the compared methods. In particular, the table discloses that in 90% of images the eye is located within 1.90 pixels by our method, much better than the other two methods (2.96 pixels by the Bayesian method, 2.72 pixels by the traditional PS method). Although it is difficult to make a quantitative comparison with other methods in literatures due to the lack of common evaluation data set, we notice that the best reported result in [2] is 2.04 pixels and 2.74 pixels with the same error measures on the FERET data set and a WWW data set, respectively.

To study the localization behavior further on the right eye in x and y axes, we report the displacements by the compared methods in Table 3. Note that although the mean value of the traditional PS method is slightly better than that of the Bayesian method in x-axis, its standard deviation is much larger than that of the Bayesian method, [3] which may due to the use of overly strong local structural constraints. It is clear that our method outperforms both the traditional PS method and the Bayesian method significantly.

We also study the scalability of the proposed method by removing the local search region for each part. In this case, the appearance model (8) is modified as: $p(I|L,\theta) \propto \prod_{i=1}^{N_p} \sum_z p(z|l_i,u_i) \approx \prod_{i=1}^{N_p} max_z p(z|l_i,u_i)$. The results are shown in Table 4. As expected, the performance of all the compared methods degenerates when the search region increases. However, the efficiency of model inference of our method is still improved since it filters out a large number of noise candidates at first.

[3] This should not be confused with the results in Figure 6 where a different evaluation criterion is used.

Table 1. Percentages of successful localizations subject to $d_{eye} < 0.05$ and $d_{eye} < 0.1$, respectively.

Method	$d_{eye} < 0.05$	$d_{eye} < 0.1$
Bayesian method	44.0%	91.9%
Traditional PS	53.4%	93.1%
Our method	80.2%	98.4%

Table 2. Errors of right eye localization, measured by Euclidean distance in pixels in normalized images. In 90% images the eyes are located within 1.90 pixels by our method.

Method	50% images	90% images
Bayesian method	1.51 pixels	2.96 pixels
Traditional PS	1.39 pixels	2.72 pixels
Our method	0.97 pixels	1.90 pixels

Table 3. Pixel coordinate error (mean $\pm$ std.) in original images when searching in reduced space.

Method	x-coordinates	y-coordinates
Bayesian method	1.38 ± 5.17	1.22 ± 3.88
Traditional PS	1.33 ± 7.18	1.38 ± 8.19
Our method	0.93 ± 4.36	0.71 ± 1.84

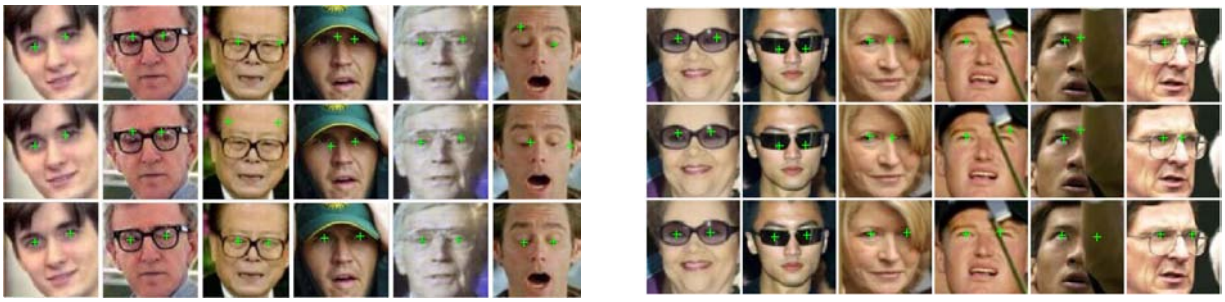
Table 4. Pixel coordinate error (mean $\pm$ std.) in original images when searching in whole image.

Method	x-coordinates	y-coordinates
Bayesian method	4.76 ± 122.62	2.00 ± 32.05
Traditional PS	1.93 ± 18.80	2.68 ± 49.78
Our method	1.06 ± 5.47	0.97 ± 11.52

Finally, we examine the effectiveness of our method for handling partial occlusions, and some typical results are shown in Figure 5.3. It can be seen that our method is more robust against partial occlusions caused by pose, hair, sunglasses, and other things like baseball pole. It is interesting to note that sometimes we human beings may be cheated by our own eyes, that is, the eyes may not be occluded by sunglasses as much as we may have thought and can still be precisely located by our method (*c.f*. Figure 5.3).

6. Conclusion

In this paper, we present a new method for eye localization under uncontrolled conditions. We enhance the Pictorial Structures (PS) model [3] by replacing the generative model with a discriminative model, incorporating global geometrical constraints on facial features, and adopting an effective heuristic method to deal with occlusion. Experiments on the challenging LFW database [10] show encouraging results on a broad range of appearance variations and imaging conditions. The proposed method is possible to be



(a) On images without occlusions (b) On images with occlusions

Figure 7. Examples results of eye localization on LFW images *(from top to bottom: Bayesian method, traditional PS method, our method)*.

extended to localize other facial features.

Acknowledgments

This research was supported by the National Science Foundation of China (60773060, 60635030, 60721002), the Jiangsu Science Foundation (BK2006187, BK2008018), the National High Technology Research and Development Program of China (2007AA01Z169), the Jiangsu 333 High-Level Talent Cultivation Program and the Project sponsored by SRF for ROCS, SEM.

References

[1] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, pages 628–641, 1998.

[2] M. Everingham and A. Zisserman. Regression and classification approaches to eye localization in face images. In *FGR*, pages 441–448, 2006.

[3] F. Felzenszwalb and P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):1573–1405, 2005.

[4] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1):67–92, 1973.

[5] C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE TPAMI*, 26(11):1408–1423, 2004.

[6] X. Geng, Z.-H. Zhou, and H. Dai. Uncontrolled face recognition by individual stable neural network. In *PRICAI*, pages 553–562, 2006.

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Individual stable space: An approach to face recognition under uncontrolled conditions. *TNN*, 19(8):1354–1368, 2008.

[8] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2002.

[9] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *CVPR*, pages 657–662, 2001.

[10] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[11] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *AVBPA*, pages 90–95, 2001.

[12] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Extending pictorial structures for object recognition. In *BMVC*, pages 789–798, 2004.

[13] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *ICCV*, pages 637–644, 1995.

[14] Y. Ma, X. Ding, Z. Wang, and N. Wang. Robust precise eye location under probabilistic framework. In *FGR*, pages 339–344, 2004.

[15] D. Nguyen and T. Ho. An efficient method for simplifying support vector machines. In *ICML*, pages 617–624, Bonn, Germany, 2005.

[16] A. P. Pentland, B. Moghaddam, and T. E. Starner. View-based and modular eigenspaces for face recognition. In *CVPR*, pages 84–91, 1994.

[17] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE TPAMI*, 22(10):1090–1104, 2000.

[18] J. C. Platt. Probabilistic outputs for support vector machines and comparisions to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[19] S. Shan, Y. Chang, and W. Gao. Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution. In *FGR*, pages 314–320, 2004.

[20] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *AMFG*, pages 168–182, 2007.

[21] M. E. Tipping. The relevance vector machine. In *NIPS 12*, pages 652–658. 2000.

[22] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[23] M. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE TPAMI*, 24(1):34–58, 2002.

[24] A. Yuille, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *IJCV*, 8(2):99–111, 1992.

[25] Z.-H. Zhou and X. Geng. Projection functions for eye detection. *Pattern Recogn.*, 37(5):1049–1056, 2004.

[26] Z. Zhu and Q. Ji. Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. *CVIU*, 98(1):124–154, 2005.