An Eye Detection and Localization System for Natural Human and Robot Interaction without Face Detection

Xinguo Yu[1], Weicheng Han[2], Liyuan Li[1], Ji Yu Shi[1], and Gang Wang[1]

[1] Institute for Infocomm Research, Connexis, Singapore 138632
{xinguo,lyli,jyshi,gswang}@i2r.a-star.edu.sg
[2] Department of Electrical and Computer Engineering,
National University of Singapore, Singapore 117543
calayanrail@gmail.com

Abstract. There were many eye localization algorithms *depending on face detection* in the literature. Differently this paper presents a novel eye detection and localization system *not depending on face detection* for natural human and robot interaction using both stereo and visual cameras. To build a robust system we use stereo and visual cameras in synergy. The stereo camera is used to localize the head of the person to replace face detection. Then our eye identification algorithm detects and localizes two eyes inside head box. In eye detection step, our algorithm uses a HOG-moment (H*istogram* O*f* G*radient*) feature to detect two eyes inside the head box. In eye localization step, we employ an iterative procedure to search the best location for eye pair. The experimental results show that the proposed eye detection and localization algorithm, not depending on face detection, has a similar robustness as the existing eye localization algorithms.

Keywords: Eye Detection and Localization, HOG, Face Detection, Disparity Image, Human and Robot Interaction.

1 Introduction

Eye detection and localization has been an important research problem in the past decades due to it is a key step in a variety of applications. In natural human and robot interaction eye identification is the key technology of eye contact between human and robot. The result of eye identification can facilitate to find out the direction of gaze. In the facial feature extraction two eyes can be located first thanks to their salience. And with the helpful innate geometrical constraint the accurate eye centers can facilitate the extraction of other facial features by providing good location estimation of other features [7-8]. Facial feature extraction is the essential step in multiple applications such as face tracking, face recognition, facial expression recognition, and human computer interface [22-23]. The facial feature extraction is to localize the facial components of interest including eyes, nose, and mouth and to estimate their scale. The accuracy of facial feature localization has big impact on the performance of the applications such as face recognition and facial expression recognition using locations of facial features as their input [22].

R. Groß et al. (Eds.): TAROS 2011, LNAI 6856, pp. 54–65, 2011.
© Springer-Verlag Berlin Heidelberg 2011

The algorithms of eye localization in the literature can be divided into active and passive two approaches. The active approach localize eyes from the images taken by near infrared (NIR) cameras [1-2], whereas the passive approach localizes eye from the images taken by visual cameras [3-22]. The principle of the former approach is red-eye effect in flash photographs, utilizing special IR illuminators and IR-sensitive CCD for imaging. In the indoor and relatively controlled conditions the spectral properties of the pupil under NIR illumination provide a very clean signal. Hence the algorithms in this approach are relatively simple and fast. And they can achieve high accuracy in eye localization when the required special conditions are met. The most significant conditions in this category are a relatively stable lighting condition, a camera set close to the subject, and open eyes [1-2].

Localizing two eyes from the visual images is more challenging than from the NIR images. In the literature there were more research efforts in this approach due to its wide variety of applications and its challenges and. These algorithms used various methods such as template matching [14], rule-based [17, 21], model-based [4, 16], feature based [5, 8, 12, 17], and hybrid [10]. In the template algorithms, multiple templates are created for searching eye pair [14]. Templates are used under certain order and rules. A number of templates are required to cope with the variety of the conditions of taking images and the variety of eye pairs. In the rule-based algorithms, several functions or transforms are used to compute several features and then rules are applied to these features to obtain the eye locations of two eyes. For instance, authors in [17, 21] used the projection functions to obtain the features, whereas authors in [3] used the radial symmetry transform to obtain the features. Authors in [8] proposed an algorithm based on the assumption that eye center is the center of isohphote curvature. However, the rules based on the prior knowledge are not easily decided. This method also has difficulty in finding features that can cope with the different conditions. In the model-based algorithms, single or multiple models are created to capture the eye characters. A model comprises of a set of equations including the eye location and some other variables of describing the considering region. The goal is to find the optimal solution of the model [4, 16]. In this method, the main challenge is that it is difficulty to define single or multiple models capable of capturing the variety of conditions. Finally, feature based algorithms used a trained learning machine to classify each region whether is an eye or an eye pair based on low-level feature. Some of examples of features include wavelets, Gabor feature, Haar-like feature, and HOG feature [8, 12]. HOG (H*istogram* O*f* G*radient*) is one of effective methods in object detection and recognition. Monzo *et al* [12] used a HOG to model two eyes and a face together. The algorithm in [12] first detects the face region and then it employs Haar feature to obtain the eye candidates. The last step is to use HOG model to evaluate each pair of eyes and pick the best location of the pair of eyes.

Though there were many eye localization algorithms in the literature, they need some further work for building the system for real applications. The existing algorithms still have two issues. First, they take the face box as their input and assume that the face box is perfectly performed. The fact is that face detection still can not achieve the perfect performance, especially for occluded, makeup, camouflage faces. Second, they are not robust for the uncontrolled open environment. For example, they are not robust to locate the eyes under the scenario of natural human and robot interaction. In this paper, we develop a robust and fully automatic eye localization

system based on the synergy of stereo and visual cameras. This system uses a procedure to replace *face detectio*n in the existing eye localization algorithms. Compared with face detection, this procedure has multiple merits. First, it can robustly locate the head for a wide range because the stereo can have a wide view. Second, it is much faster than face detection due to it is a shape analysis procedure. It is another critical merit for building the real system. Third, it still can locate the head when faces are of camouflage, makeup, or partial occluded. Fourth, it can know which head is closer to the robot.

The rest of the paper is organized as follows. Section 2 gives the overview of the proposed eye localization system. Section 3 and 4 describe the procedures of eye detection and eye localization respectively. Section 5 presents the evaluation of the algorithm. We conclude the paper in Section 6.

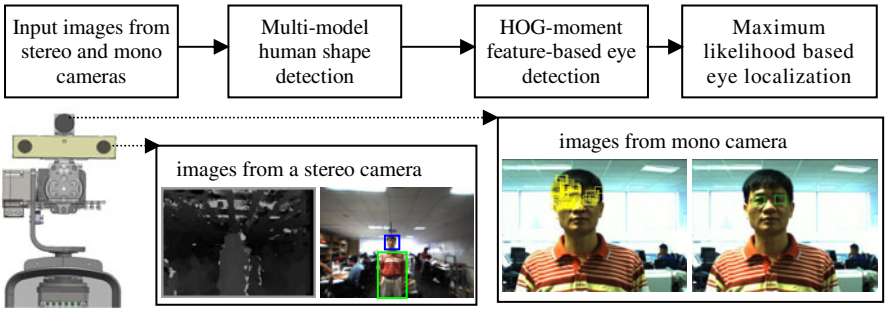

Fig. 1. Flowchart of our eye detection and localization system and the layout of stereo and mono cameras on our robot

2 Overview of the Eye Detection and Localization System

Here we give the overview of our eye detection and localization system, which is a robust and fully automatic eye localization system of using both stereo and mono cameras in synergy. As depicted in Fig 1, the system comprises of three main components: head localization, eye detection and eye localization. The component of head localization works on the images by the stereo camera. It first finds the person close to robot. Then it localizes person's head and converts the head box in the stereo-camera image into the one in mono-camera image. This component replaces the face detection in the existing eye localization algorithms in the literature [3-22]. The other two components actually are an eye detection and localization algorithm from the visual images with the known head box. Eye detection component is to obtain the eye candidates, which are the regions in certain size that probably contain an eye. Our eye candidate detection employs a scan procedure that scan the full image for all the regions. For each region, we use a SVM (Support Vector Machine) on the HOG-moment feature to evaluate whether it is an eye candidate. HOG-moment vector is the concatenation of the HOG vector and moment vector of this region. HOG (Histogram of Gradient) is a robust object detection technique and it especially has good performance in human detection. However, HOG can be fooled by some objects that have the similar edge distributions to eye. We complement HOG with moment vector.

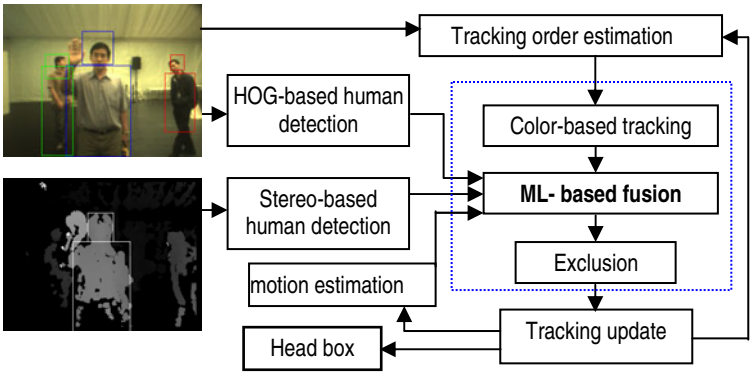

Fig. 2. The block diagram of multi-model human body and head detection and tracking

The succeeding eye localization component is to identify the eye pair from the eye candidates and find the accurate eye centers of two eyes. We first form a formula that calculates the likelihood that two points are left and right eye centers. Thus, our goal is to find two points that have a largest likelihood value. We design an iterative procedure to achieve this goal. The initial positions of two points are crucial for obtaining the best eye locations. We define a Bayesian formula to select the best pair from eye candidates. The measure used in this search is produced by integrating the appearance measure and the distance of two eyes.

3 Head Localization by Stereo Camera

3.1 Head Localization by Stereo Camera

The block diagram of our head localization component is shown in Fig 2. The input to this component is both disparity and color images from the stereo camera on the robot head. For each frame, the blocks in the diagram are executed as follows. First, the humans in the view are detected from the disparity and color images. Meanwhile, the positions of occluded humans are predicted based on their motion. To handle possible complex occlusions, multi-person tracking is performed sequentially from the closest one to the farthest (including the fully occluded ones) to approximate a globally optimal tracking process. The order is determined by the predicted 3D positions of all the tracked humans. Then, the blocks within the dotted box of the diagram in Fig 2 are executed to track humans one by one. For each human, a mean-shift tracking is first performed, and then the new position is located in the image by the ML-based fusion (*ML is the acronym of Maximum Likelihood*). Finally, the exclusion step is performed to suppress the visual features of the tracked human in both color and disparity images. This operation is to avoid other humans being trapped in the positions of those tracked humans. When the sequential multi-person tracking is completed, the system updates the 3D positions and appearance models of the tracked humans, as well as the initializations and terminations of the tracks. The results of

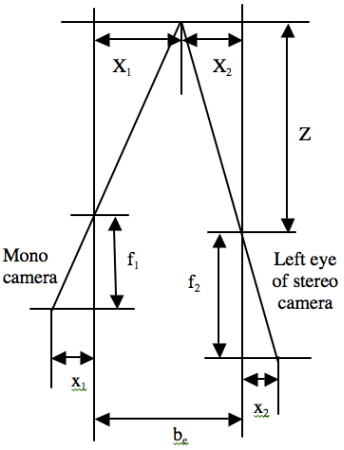

Fig. 3. The illustration of the relative relation of the variables of stereo and mono cameras

human tracking are the bounding boxes of human body and head. The following steps in this system just use the head box.

3.2 Head Location Conversion

To convert the head location in the stereo camera into the location in the mono camera we establish a mapping from mono camera to the left eye of the stereo camera. The variables $x_1, x_2, X_1, X_2, f_1, f_2, b_e, and\ Z$ are defined in the fig 3 and b_s is the baseline of the stereo camera. Then they have the following relation.

$$\begin{cases} \dfrac{x_1}{X_1} = \dfrac{f_1}{Z}, \quad \dfrac{x_2}{X_2} = \dfrac{f_2}{Z}, \quad \dfrac{x_2}{X_2} = \dfrac{f_2}{Z}, \\ x_2 = \dfrac{f_2}{Z} X_2 = \dfrac{f_2}{Z}(b_e - X_1) = -\dfrac{f_2}{f_1} x_1 + \dfrac{f_2 b_e}{Z}. \end{cases} \tag{1}$$

In stereo camera $Z = fbd^{-1}$, where d is the disparity value. Hence,

$$x_2 = -\frac{f_2}{f_1} x_1 + \frac{f_2 b_e}{f_2 b_s} d = -\frac{f_2}{f_1} x_1 + \frac{b_e}{b_s} d = k_1 x_1 + k_d d \tag{2}$$

where x_1 is the position from mono camera and x_2 is the estimated position in the color image from the left eye of the stereo camera. The coordinates should be computed with respect to the image centers, hence $x_1 = X_1 - X_{c1}$ and $x_2 = X_2 - X_{c2}$, where X_1 and X_2 are the positions in the image respect to the upper left center of the corresponding images, and X_{c1} and X_{c2} are the image centers of the corresponding images. By substituting the variables the model becomes

$$X_2 = k_1 X_1 + k_d d + C \tag{3}$$

where $C = X_{c2} - k_1 X_{c1}$ is a constant. The parameters for the model are k_1, k_2, and C.

4 HOG-Moment Based Eye Detection

Eye detection is a procedure that scans all possible eye regions and evaluates each region whether it is an eye candidate using a SVM on HOG-moment feature. Given a grey image, it first estimates the eye dimension by the head size and forms five sizes of eye regions around the estimated eye dimension. For each size of region we scan the whole image to acquire the eye candidates.

4.1 HOG Feature

Histogram of oriented gradient (HOG) is an adaptation of Lowe's Scale Invariant Feature Transformation (SIFT) approach. A HOG feature is created by first computing the gradient magnitude and orientation at each image sample point in a window (*or region*) around an anchor point. The window is divided into a W x H cells. An orientation histogram for each cell is then formed by accumulating samples within the cell, weighted by gradient magnitude. Concatenating the histograms from all the cells forms the final HOG feature vector. A W x H cells window was used to scan the image with step length at S_w pixels in horizontal and S_h pixels in vertical. The ratio between W and H is decided by object shape, which can be acquired by doing statistics on the annotated samples. And the values of W and H are chosen by considering accuracy and computation time. Each cell block in the window covers M x M pixels. Different values of M are used to detect the same target in different scales. In this way, HOG can handle the scale variance of object in images. In a cell block, the orientation of gradient of each pixel is classified into K bins. K bins evenly divide from 90 degree to -90 degree, i.e. each bin spans 180/K degrees. The value linking to a bin is in the interval [0.0, 1.0], which represents the ratio of pixels belongs to the bin. Thus, W x H x K bins are used for an anchor position. In this paper, W=4, H=3, K=9, S_w =W/2, S_h =H/2. The values of M are 16 + j*2 for j = 0 to 4 in this paper.

4.2 Moment Feature

Eye region has a distinctive gray value distribution pattern, *i.e.* the dark center iris, surrounding white sclera, then upper and lower eyelids, face skin, and eyebrow. This pattern is reflected by the spatial intensity pattern in gray eye images. This paper proposes a method to extract the moment feature that targets to capture this pattern. Let $B = (W, H)$ be the window of a normalized intensity image of an eye I(x) centered at $x_c = (x_c, y_c)$. Let us denote g_{iris} as the brightness value of the iris, which is selected as the minimum intensity value from the core center of 5 *x* 5 window centered at x_c. Then, the moments of up to the third order which characterize the spatial intensity variations related to the iris can be defined as

$$m_{ij} = \frac{1}{W^{i+1} H^{j+1}} \sum (x - x_c)^i (y - y_c)^j (I(x, y) - g_{iris}). \quad (4)$$

We obtain ten items when we limit to $0 \le i, j \le 3$ and $i + j \le 3$. Besides the moment feature the spatial intensity distribution is also employed to characterize the

eye graph pattern. We divide the window B into 3x3 grids, where the center block contains the iris and the other blocks cover the surrounding regions. The average of the intensities related to the iris brightness for each block is computed as

$$n_{kl} = \frac{9}{WH} \sum_{I(x,y) \in b_{kl}} (I(x,y) - g_{iris}) \tag{5}$$

Where $k, l = 1, 2, 3$ and b_{kl} is the block at the kth row and lth column of the grid. These values characterize how the brightness spatially changes related to the iris. The average of each block can be computed from the integral image, so that the computational cost is very low when scanning the detection window over an image.

By combining (4) and (5) a vector of 19 dimensions can be obtained v = (m_{00}, m_{10}, m_{01}, m_{20}, m_{02}, m_{11}, m_{03}, m_{30}, m_{12}, m_{21}, n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, , n_{33}). We still call it the moment feature, though it is not a pure moment feature in this paper.

5 Maximum Likelihood Based Eye Localization

5.1 Formulation

Given the eye candidates detected in eye detection step, the eye localization is to find the ideal positions of the left and right eyes (*i.e.*, $\mathbf{x}_l$ and $\mathbf{x}_r$ where $\mathbf{x} = (x, y)$) starting with these eye candidates. Let us denote the N detections as B= $\{B_i : i=1,2,\ldots,N\}$ where B_i is the box of the ith detection centered at x_i. Assuming that the detections are independent each other, using Bayes' theorem, the likelihood of that the left eye is located at x_l can be expressed as

$$L(\mathbf{x}_l \mid \mathbf{B}) = \frac{\sum_{i=1}^{N} \pi_i^l P(\mathbf{x}_l \mid B_i) P(B_i)}{P(\mathbf{x}_l)} \propto \sum_{i=1}^{N} \pi_i^l P(\mathbf{x}_l \mid B_i) P(B_i) \tag{6}$$

where π_i^l denotes the association of the ith detection with the left eye, $P(\mathbf{x}_l \mid B_i)$ is the conditional probability of $\mathbf{x}_l$ being the ideal position of the left eye given the detection B_i and $P(B_i)$ is the confidence of the ith detection. Similarly, one can express the likelihood of the right eye position as

$$L(\mathbf{x}_r \mid \mathbf{B}) = \frac{\sum_{i=1}^{N} \pi_i^l P(\mathbf{x}_r \mid B_i) P(B_i)}{P(\mathbf{x}_r)} \propto \sum_{\mathbf{i=1}}^{\mathbf{N}} \pi_i^r P(\mathbf{x}_r \mid B_i) P(B_i) \tag{7}$$

When the probability measures follow simple Gibbs distributions, the conditional probabilities $P(\mathbf{x}_l \mid B_i)$ and $P(\mathbf{x}_r \mid B_i)$ can be written as

$$P(\mathbf{x}_l \mid B_i) = e^{-\frac{|\mathbf{x}_l - \mathbf{x}_i|^2}{\sigma_d^2}} \text{ and } P(\mathbf{x}_r \mid B_i) = e^{-\frac{|\mathbf{x}_r - \mathbf{x}_i|^2}{\sigma_d^2}} \tag{8}$$

where σ_d is the spatial variance of the eye centers which is determined according to the statistics of human faces.

If $\mathbf{x}_l$ and $\mathbf{x}_r$ are the ideal positions of the two eyes, they should be separated at an interocular distance. Hence, the probability of that $\mathbf{x}_l$ and $\mathbf{x}_r$ represent a pair of eyes can be expressed as

$$P_{eye-pair}(\mathbf{x}_l, \mathbf{x}_r) = e^{-\frac{|(\mathbf{x}_r - \mathbf{x}_l) - \mathbf{d}_e|^2}{\sigma_e^2}} \tag{9}$$

where $\mathbf{d}_e = (d_x, d_y)$ is the distance vector between the two eyes, and σ_e denotes the variances of inter-ocular distances. They can be chosen according to the statistics of human faces. For an upright face, d_x equals the inter-ocular distance because $d_y \approx 0$. The vector $\mathbf{d}_e$ also indicates the pose of a front face.

Combining (6) to (9), the likelihood probability of the positions of the two eyes can be defined as

$$P_{eye}(\mathbf{x}_l, \mathbf{x}_r) = L(\mathbf{x}_l \mid \mathbf{B}) L(\mathbf{x}_r \mid \mathbf{B}) P_{eye-pair}(\mathbf{x}_l, \mathbf{x}_r) \propto$$
$$\left(\sum_{i=1}^{N} \pi_i^l e^{-\frac{|\mathbf{x}_l - \mathbf{x}_i|^2}{\sigma_d^2}} P(B_i) \right) \left(\sum_{i=1}^{N} \pi_i^r e^{-\frac{|\mathbf{x}_r - \mathbf{x}_i|^2}{\sigma_d^2}} P(B_i) \right) e^{-\frac{|(\mathbf{x}_r - \mathbf{x}_l) - \mathbf{d}_e|^2}{\sigma_e^2}} \tag{10}$$

The eye localization is to find the positions $\mathbf{x}_l$ and $\mathbf{x}_r$ that maximize the likelihood probability. Ideally, if $\mathbf{x}_l$ and $\mathbf{x}_r$ are the true positions of the left and right eyes in the face image, the likelihood reaches its maximum at the positions with $\frac{\partial P_{eye}(\mathbf{x}_l, \mathbf{x}_r)}{\partial \mathbf{x}_l} = 0$ and $\frac{\partial P_{eye}(\mathbf{x}_l, \mathbf{x}_r)}{\partial \mathbf{x}_r} = 0$. From (10), we can obtain

$$\begin{cases} \mathbf{x}_l = \dfrac{\sigma_e^2 \sum_{i=1}^{N} \mathbf{x}_i \pi_i^l P(\mathbf{x}_l \mid B_i) P(B_i)}{(\sigma_d^2 + \sigma_e^2)\, L(\mathbf{x}_l \mid \mathbf{B})} + \dfrac{\sigma_d^2 (\mathbf{x}_r - \mathbf{d}_e)}{\sigma_d^2 + \sigma_e^2} \\ \mathbf{x}_r = \dfrac{\sigma_e^2 \sum_{i=1}^{N} \mathbf{x}_i \pi_i^l P(\mathbf{x}_r \mid B_i) P(B_i)}{(\sigma_d^2 + \sigma_e^2)\, L(\mathbf{x}_r \mid \mathbf{B})} + \dfrac{\sigma_d^2 (\mathbf{x}_l - \mathbf{d}_e)}{\sigma_d^2 + \sigma_e^2} \end{cases} \tag{11}$$

Unfortunately, the system (11) does not form a close solution for maximizing the likelihood $P_{eye}(\mathbf{x}_l, \mathbf{x}_r)$ since both sides have $\mathbf{x}_l$ and $\mathbf{x}_r$. In this paper, we propose an iterative solution to this problem. It contains two steps. In the first step, a pair of good initialization positions are selected from the detections, and in the second step, an iterative algorithm is applied to refine the positions and scales of the two eyes.

5.2 Initialization

For iterative algorithms, good initialization is very important for success to real world problems. In this work, the initial eye positions are selected from the detections. Again, for each pair of two detections B_i and B_j, the probability that they are good candidates for the left and right eyes can be defined as

$$P_{eye}(i, j) = P(B_i) P(B_j) P_{eye-pair}(B_i, B_j) \tag{12}$$

where $P(B_i)$ is the confidence of the ith detection, and $P_{eye-pair}(B_i, B_j)$ is computed using (10) with d_x, the average of human interoccular distance and $d_y = 0$. The pair of maximum probability value is selected as the initial positions and scales of the two eyes.

5.3 Iterative Estimation

Let B_i and B_j be the selected initial positions of the left and right eyes. The initial center points of the left and right eyes can be denoted as $x_l = x_i$ and $x_r = x_j$. The initial scales of the two eyes can be represented as $B_l = B_i$ and $B_r = B_j$, where $B_i = (W_i, H_i)$ represents the width and height of the detection window. The distance vector between the two eyes characterizes the face pose. The initial face pose is assumed as upright face, hence, d_x is set as the average human inter-ocular distance and d_y is set as 0. From (11) and (12), an iterative algorithm for eye localization from detections can be defined. The updates in each step of the iterative algorithm can be expressed as

$$\begin{cases} \mathbf{x}_l^{t+1} = \dfrac{\sigma_e^2 \sum_{i=1}^{N} \mathbf{x}_i \pi_i^{l(t)} P(\mathbf{x}_l^t \mid B_i) P(B_i)}{(\sigma_d^2 + \sigma_e^2)\, L(\mathbf{x}_l^t \mid \mathbf{B})} + \dfrac{\sigma_d^2 (\mathbf{x}_r^t - \mathbf{d}_e^t)}{\sigma_d^2 + \sigma_e^2} \\ \mathbf{x}_r^{t+1} = \dfrac{\sigma_e^2 \sum_{i=1}^{N} \mathbf{x}_i \pi_i^{r(t)} P(\mathbf{x}_r^t \mid B_i) P(B_i)}{(\sigma_d^2 + \sigma_e^2)\, L(\mathbf{x}_r^t \mid \mathbf{B})} + \dfrac{\sigma_d^2 (\mathbf{x}_l^t - \mathbf{d}_e^t)}{\sigma_d^2 + \sigma_e^2} \\ \mathbf{d}_e^{t+1} = \mathbf{x}_r^t - \mathbf{x}_l^t \\ B_l^{t+1} = \dfrac{\sum_{i=1}^{N} B_i \pi_i^{l(t)} P(\mathbf{x}_l^t \mid B_i)\, P(B_i)}{L(\mathbf{x}_l^t \mid \mathbf{B})} \\ B_r^{t+1} = \dfrac{\sum_{i=1}^{N} B_i \pi_i^{l(t)} P(\mathbf{x}_r^t \mid B_i)\, P(B_i)}{L(\mathbf{x}_r^t \mid \mathbf{B})} \end{cases} \tag{13}$$

The update of the interocular distance vector $\mathbf{d}_e$ is derived from $\dfrac{\partial P_{eye}(\mathbf{x}_l, \mathbf{x}_r)}{\partial \mathbf{d}_e} = 0$, and the scales of the eye boxes are updated as the weighted average of the associated detection windows. Here, the association parameters π_i^l and π_i^r are computed based on the overlapping of the boxes. Let B_l^t be the estimated box of the left eye at the t iteration step, which is centered at $\mathbf{x}_l^t$. The association of B_l^t with the ith detection B_i is computed as $\pi_i^{l(t)} = \dfrac{|B_l^t \cap B_i|}{|B_l^t \cup B_i|}$, *i.e.,* the numerator is the area of the intersection and the denominator is the area of the union of the two boxes. The iterative algorithm stops when both $|\mathbf{x}_l^{t+1} - \mathbf{x}_l^t| < \varepsilon$ and $|\mathbf{x}_r^{t+1} - \mathbf{x}_r^t| < \varepsilon$ hold, where ε is a small value.

With a good initialization, the iterative algorithm converges very quickly. Since the inter-ocular distance vector is involved in the updating in the iterations, the algorithm is able to adapt to a quite large variations of head poses. The updates of the eye scales can improve the estimations of the associations in the iterative steps.

6 Experimental Results

We conduct the experiments to compare the performance of the proposed algorithm with some existing algorithms on databases BioData, CVL, and Olivia. We also conduct the experiments on the effect of the particular methods of our algorithm.

6.1 Data and Evaluation Criterion

BioIData [24] is a frequently-used database for eye localization [3-5, 8-9, 11, 17-18, 21]. The dataset consists of 1521 gray images with a resolution of 384x286. Each one shows the frontal view of a face of one out of 23 different test persons.

The CVL [25] database contains 797 color images of 114 persons. Each person has 7 images of size 640x480 pixels: far left side view, 45° angle side view, serious expression frontal view, 135° angle side view, far right side view, and smile frontal view. The algorithm in [17] used the 335 frontal view face images from this database. We carry out the experiments on the same dataset 335 images.

The Olivia is the database that contains the 3000 images in the resolution 320x240 of ten persons. These images are recorded by our robot during the interaction between the robot and persons.

The aim of eye localization is to find the center of eye and the eye center of an eye is the center point of its pupil for open eyes and the middle point of two corners for closed eyes. In our evaluation, we adopt the measure proposed by Jesorskey *et al.* [11] including the variables. The error measure, defined as localization criterion,

$$\mathbf{e} = \frac{\max(\| \mathbf{C}_l - \tilde{\mathbf{C}}_l \|, \quad \| \mathbf{C}_r - \tilde{\mathbf{C}}_r \|)}{\| \mathbf{C}_l - \mathbf{C}_r \|} \tag{14}$$

where C_l and C_r are the groundtruth positions and $\tilde{C}_l$ and $\tilde{C}_r$ are the detected eye centers of left and right eyes respectively in pixel. $\| \bullet \|$ is the Euclidean distance.

6.2 Experimental Results

To use HOG-moment feature to detect the eye candidates we need to train a SVM to judge whether a window containing an eye. The images used in our training comprise the ORL face database (AT & T) and 10% of BioID and CVL databases. For each image, we produce 2 positive samples for left eye and right eye respectively and 5 negative samples which are randomly selected windows from each image and are far enough from the correct eye region. We test our algorithm on BioID and CVL databases and compare our result with the result reported in [3-5, 8-9, 11, 17-18, 21]. For BioID database, our result is 98.55% with e<0.25. This result is not the best but in the best performance class. Our result is 90.14 with e<0.1. This result is better than all the results except the results in [5] and [11]. It is worth noticing that our algorithm does not depend on the face detection, whereas all the algorithms in [3-5, 8-9, 11, 17-18, 21] depend on the results of face detection and all assume that face detection achieved the perfect result, which is very difficulty to get. For 355 images selected from CVL, our algorithm achieve correct rate of 96.17% for e<0.25 and 90.56% for e<0.1 respectively. For e<0.25, our result 96.17% is worse than 99.7% reported in [17], whereas for e<0.1 our result 90.56% is much better than 80.9% of the result reported in [17].

Besides the experiments of comparison with the existing algorithms we also conduct some experiments to explore the performance of our algorithm under the different configurations. We compare the performance between with and without the face box and without HOG-moment features and the results are presented in Table 2. When we do not use the information of face region, our algorithm searches a larger area. When there are some objects in images that are very similar to eye in moment and appearance the algorithm would increase the rate of false alarm. However, such objects are very few in the databases used to do experiments in this paper.

Table 1. Comparison on the eye localization performance between our algorithm and the 8 algorithms in the literature

	#Ref	[3]	[4]	[5]	[8]	[9]	[11]	[17]	[18]	ours
BioID	0.25	96.00	96.1	98.00	93.00	91.8	99.9	99.46	98.49	98.55
	0.1	64.00	85.2	96.0	77.0	79.0	97.9	73.68	90.85	90.14
CVL	0.25							99.7		96.17
	0.1							80.9		90.56
Olivia	0.25									99.97
	0.1									91.65

Table 2. The performances of our algorithm when it is full one and no moment or no face box

Name	BioID	BioID	CVL	CVL	Olivia	Olivia
Threshold	0.25	0.1	0.25	0.1	0.25	0.1
Full Alg	98.55	90.14	96.17	90.56	99.97	91.65
No moment	94.61	80.34	92.63	89.38	98.33	61.58
No face box	84.42	77.71	96.17	90.56	99.94	90.10

7 Conclusions

We have presented a novel eye detection and localization system of using both stereo and mono cameras. This paper has multiple contributions in the technique development. First, it is an eye detection and localization system tested on the robot, not just an algorithm working on images. Second, it uses the stereo camera to obtain the head location in the image from mono camera. This procedure replaces the face detection in many other eye localization algorithms. Third, our algorithm is robust to the uncontrolled environment. This robustness is because of the reliable head box localization and the good distinguish ability of HOG-moment feature.

In the near future, we want to develop more robust eye localization algorithm by integrating more methods. In addition, we want to develop different eye locators and then use them in a scheme to achieve better performance of eye localization.

References

[1] Amir, A., Zimet, L., Sangiovanni-Vincentelli, A., Kao, S.: An embedded system for an eye-detection sensor. Comput. Vis. Image Underst. 98(1), 104–123 (2005)

[2] Haro, A., Flickner, M., Essa, I.: Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. In: CVPR 2000, pp. 163–168 (2000)

[3] Bai, L., Shen, L., Wang, Y.: A novel eye location algorithm based on radial symmetry transform. In: ICPR 2006, vol. 3, pp. 511–514 (2006)

[4] Campadelli, P., Lanzarotti, R., Lipori, G.: Precise eye localization through a general-to-specific model definition. In: BMVC 2006, pp. 187–196 (2006)
[5] Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: BMVC 2004, pp. 277–286 (2004)
[6] Fasel, I., Fortenberry, B., Movellan, J.: A generative framework for real time object detection and classification. Computer Vision and Image Understanding 98, 182–210 (2005)
[7] Gernoth, T., Kricke, R., Grigat, R.-R.: Mouth localization for appearance-based lip motion analysis. WSEAS Transactions on Signal Processing 3(3), 275–281 (2007)
[8] Hamouz, M., Kittler, J., Kamarainen, J.-K., Paalanen, P., Kalviainen, H., Matas, J.: Feature-based affine-invariant localization of faces. In: PAMI 2005, vol. 27(9), pp. 1490–1495 (2005)
[9] Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the Hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
[10] Jin, L., Yuan, X., Satoh, S., Li, J., Xia, L.: A hybrid classifier for precise and robust eye detection. In: ICPR 2006, Hong Kong, vol. 4, pp. 731–735 (2006)
[11] Kroon, B., Maas, S., Boughorbel, S., Hanjalic, A.: Eye localization in low and standard definition content with application to face matching. Computer Vision and Image Understanding 113(8), 921–933 (2009)
[12] Monzo, D., Albiol, A., Sastre, J., Albiol, A.: Precise eye localization using HOG features, Machine Vision and Applications (May 2010), 10.1007/s00138-010-0273-0
[13] Niu, Z., Shan, S., Yan, S., Chen, X., Gao, W.: 2D cascaded AdaBoost for eye localization. In: ICPR 2006, vol. 2, pp. 1216–1219 (2006)
[14] Rurainsky, J., Eisert, P.: Eye center localization using adaptive templates. In: CVPR Workshops 2004, pp. 67–74 (2004)
[15] Song, J., Chia, Z., Liu, J.: A robust eye detection method using combined binary edge and intensity information. Pattern Recognition 39, 1110–1125 (2006)
[16] Tan, X., Song, F., Zhou, Z.-H., Chen, S.: Enhanced pictorial structures for precise eye localization under uncontrolled conditions. In: CVPR 2009, pp. 1621–1628 (2009)
[17] Türkan, M., Pardàs, M., Çetin, A.E.: Human eye localization using edge projections. In: Proceedings of 2nd International Conference on Computer Vision Theory and Applications (VISAPP 2007), Barcelona, Spain, vol. 1 (2007)
[18] Valenti, R., Gevers, T.: Accurate eye center location and tracking using isophote curvature. In: CVPR 2008, pp. 1–8 (2008)
[19] Wang, P., Green, M.B., Ji, Q., Wayman, J.: Automatic eye detection and its validation. In: CVPR Workshops, June 20-26, pp. 164–171 (2005)
[20] Wang, P., Ji, Q.: Multi-view face and eye detection using discriminant features. CVIU 105, 99–111 (2007)
[21] Zhou, Z.H., Geng, X.: Projection functions for eye detection. Pattern Recognition 37, 1049–1056 (2004)
[22] Shan, S., Chang, Y., Gao, W., Cao, B.: Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In: Int'l Conf. Automatic Face and Gesture Recognition 2004, pp. 314–320 (2004)
[23] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, vol. 1, pp. 511–518 (2001)
[24] BioID, http://www.humanscan.de/support/downloads/facedb.php
[25] CVL, http://www.lrv.fri.uni-lj.si/
[26] http://www.cl.cam.ac.uk/research/DTG/attarchive:pub/data/att_faces.tar.Z