2009 11th IEEE International Symposium on Multimedia

Pose Robust Human Detection in Depth Image Using Four Directional 2D Elliptical Filters

Taewan Kim, Sangho Cho, Jongmin Yoon, and Daijin Kim
Department of Computer Science and Engineering
Pohang University of Science and Technology
San 31, Hyoja-Dong, Nam-Gu, Pohang, 790-784, Korea
{*taey16,scho, albedo039, dkim*}*@postech.ac.kr*

Abstract

This paper proposes a pose robust human detection method for sequences of stereo images using four directional 2D elliptical filters (4D2DEFs), which can detect humans regardless of scale and pose. Four 2D elliptical filters with specific orientations are applied to a 2D spatial-depth histogram, and threshold values are used to detect human candidates. These candidates are verified by either detecting the face or matching head-shoulder shapes. Experimental results show that human detection using the proposed method outperforms that of using the existing Object Oriented Scale Adaptive Filter (OOSAF) by 15∼20%, especially in the case of posed humans.

1. Introduction

Human detection is an essential task for Human-Robot-Interaction (HRI). Because intelligent robots should coexist with humans in a human-friendly environment, they must be aware of humans in their proximity, and identify them.

There are a variety of human detection methods using a single camera. Tuzel *et al.*[7] proposed a covariance descriptor-based human detection method, where humans can be detected when their whole bodies appear in an image. However, in mobile robot applications, we can not employ this algorithm because it is common to have their faces and torsos only. Qiang Zhu *et al.*[8] integrated the cascade-of-rejectors approach with histograms of oriented gradients (HoG) features to achieve a fast and accurate human detection system. However, it only worked for detecting whole bodies. Mikolajczyk *et al.* [6] proposed the part based human detection method, which was robust to detect partially occluded humans. However, this method needed to train each of human body parts independently, and to search them at several scales. This required to a huge amount of computation time, which was not suitable for using their method in mobile robot environment. Andriluka *et al.* [1] proposed the human detection and tracking method using a hierarchical articulation model of human body, which was based on a hierarchical Gaussian process latent variable model (hG-PLVM). Their approach used prior knowledge on possible articulations and temporal coherency of a walking people for a probabilistic gait modeling. However, it may fail to detect standing people.

There are a variety of human detection methods using a stereo camera. BGavrila and Munder [2] proposed the stereo-based pedestrian detection from a moving vehicle, which detected ROI of humans using shape template from the sparse disparity map, and then verified the detected result using texture-based neural network classifier. This method also used tree-based hierarchical shape-templates for each pose, which provided the robustness of rotation off-plane (ROP). However, this method was proposed for detecting the people far away from an autonomous navigating vehicle. Li *et al.* [5, 4] designed the Object-Oriented Scale-Adaptive Filter (OOSAF) and segmented the human candidates by applying the OOSAF whose filter parameter was changed in accordance with the distance between the camera and the human. They verified human candidates using template matching of the human head-shoulder. Their approach showed a good human detection rate and was suitable for a mobile robot platform because it did not use background subtraction and allowed to detect even upper-body of a human. However, it had a poor human detection rate when the humans were not facing the front.

We propose a pose robust human detection method of a sequence of stereo images in a cluttered environment in which the camera and the human are moving and the illumination conditions change. Fig. 1 outlines the proposed human detection system. It consists of consecutive two modules: human detection and human verification.

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.106

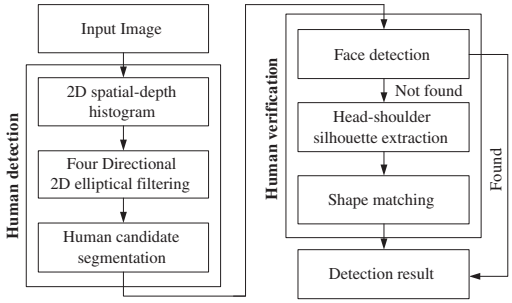


Figure 1. The proposed human detection system.

This paper is organized as follows. Section 2 describes the proposed 4D2DEFs for human detection, and verification using face detection and the head-shoulder template matching. Section 3 shows the experimental results of the human detection performances in terms of the accuracy of the estimated human rotation, human detection rate. Finally, section 4 presents our conclusions.

2 Human Detection and Verification

The OOSAF[5, 4] extracted humans by using a scale-adaptive filter whose adequate scale varied according to the distance between the human and the camera. The OOSAF performs the convolution of the 2D spatial-disparity histogram $H(x_d, d)$ with the scale adaptive filter along the x_d axis (i.e., horizontal direction of $H(x_d, d)$), where the histogram $H(x_d, d)$ was obtained by projecting the disparity map $D(x, y)$ along the y axis (i.e., vertical direction of the disparity map).

This convolution causes several problems as follows. First, two coordinates x_d and d have different natures because they represent the position and disparity, respectively. Second, the size parameter of the scale adaptive filters should be changed according to the distance between the human and the camera. Third, the disparity is inversely proportional to the distance between the human and the camera. These problems can be avoided by using the proposed 2D spatial-depth histogram $H(x_z, z)$ that can be obtained by an appropriate transformation as follows.

Let x_d and d be the horizontal spatial coordinate and the disparity coordinate in the 2D spatial-disparity histogram $H(x_d, d)$, respectively, and x_z and z be the horizontal spatial coordinate and the depth coordinate in the 2D spatial-depth histogram $H(x_z, z)$, respectively. Then, the 2D spatial-depth histogram is obtained from the 2D spatial-disparity histogram by a scale change and translation between two coordinates as

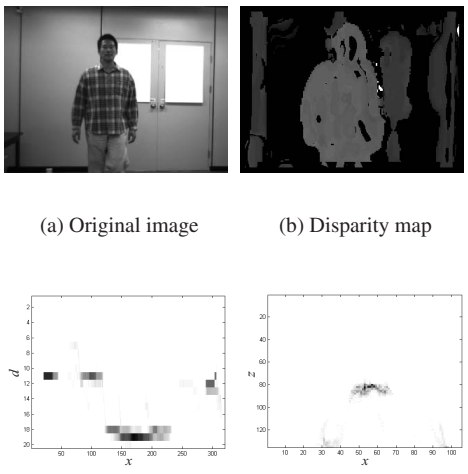
$$H(x_z, z) = H(x_d, d), \quad (1)$$

$$d = \frac{C_B C_F}{pz}, \quad (2)$$

$$x_d = \frac{C_F}{pz} x_z + c_{x_d}, \quad (3)$$

$$\therefore H(x_z, z) = H\left(\frac{C_F}{pz} x_z + c_{x_d}, \frac{C_B C_F}{pz}\right), \quad (4)$$

where c_{x_d} is the center of spatial axis x_d of the 2D spatial-disparity histogram, C_F is the focal length of the stereo camera, C_B is a baseline of the stereo camera, and p is a pixel width. In our case, we set C_F, C_B, and p to 3.8 mm, 120 mm, and 4.65 μm, respectively. Fig. 2 shows several different images: (a) the original color image obtained from a camera, (b) the disparity map obtained from a stereo camera, (c) the 2D spatial-disparity histogram and (d) the 2D spatial-depth histogram.



(a) Original image (b) Disparity map

(c) Spatial-disparity histogram (d) Spatial-depth histogram

Figure 2. Some examples of different images.

Another problem of the existing OOSAF is that it does not work well when the human is rotated from the frontal direction. Thus, a new pose-robust 4D2DEF is proposed which executes the convolution over the 2D spatial-depth histogram. It has the following properties: 1) the shapes are 2D ellipses that mimic the contour of the human body in the 2D spatial-depth space, 2) their weights are decreased smoothly along the normal direction of the body contour, and 3) they are oriented at $0°$, $45°$, $90°$, and $135°$. The filter whose orientation matches the human pose the best is selected to segment the human candidates. The estimated rotation of the human pose is also used as a cue for the human verification. A detailed explanation of how to design the proposed 4D2DEF is given below.

2.1 Four Directional 2D Elliptical Filter

The proposed 4D2DEF has an elliptical shape that resembles the body contour of a human:

$$\frac{x^2}{\left(\frac{W}{2}\right)^2} + \frac{z^2}{\left(\frac{T}{2}\right)^2} = V, \quad (5)$$

where W and T are the average width and thickness of human bodies.

The proposed 4D2DEF has a 2D kernel function $F(x, z)$:

$$F(x,z) = \begin{cases} V & \text{if } 0 \le V \le 1, \\ 2 - V & \text{if } 1 < V \le 2, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The 2D kernel function $F(x, z)$ represents the weight values of the 4D2DEFs, which is maximal (=1) on the contour of the ellipse and it decreases as the (x, z) moves away from the contour.

The proposed 4D2DEF contains an oriented 2D elliptical filter to cope with the rotation of the humans. The shape of the filter rotated by an angle θ can be represented by

$$\frac{(x\cos\theta + z\sin\theta)^2}{\left(\frac{W}{2}\right)^2} + \frac{(-x\sin\theta + z\cos\theta)^2}{\left(\frac{T}{2}\right)^2} = V_\theta. \quad (7)$$

Similarly, the 2D kernel function $F_\theta(x, z)$ of the oriented 2D elliptical filter can be represented by

$$F_\theta(x,z) = \begin{cases} V_\theta & \text{if } 0 \le V_\theta \le 1, \\ 2 - V_\theta & \text{if } 1 < V_\theta \le 2, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Then, the designed 2D kernel function is normalized such that the sum of the 2D kernel function is 1 in order to keep the sum of the filtering results as the same. Fig. 3 illustrates the 2D kernel functions of the four directional 2D elliptical filters rotated by $0°$, $45°$, $90°$, and $135°$. As shown in Fig. 3, only half of the 2D kernel functions are shown because the camera can only see the front of the person.

Fig. 4 shows the filtered results that are obtained by the convolution of the spatial-depth histogram (Fig. 2-(d)) with the four oriented 2D elliptical filters (Fig. 3) as

$$\Psi_\theta(x,z) = H(x,z) * F_\theta(x,z), \quad (\theta = 0, 45, 90, 135), \quad (9)$$

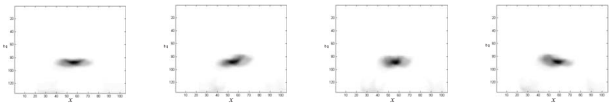
where $*$ is a convolution operator, and $F_\theta(\cdot,\cdot)$ is the 2D kernel function.

2.2 Human Candidate Segmentation

After obtaining the filtered spatial-depth histograms $\Psi_\theta(x, z)$, the human candidate segmentation is performed as follows.

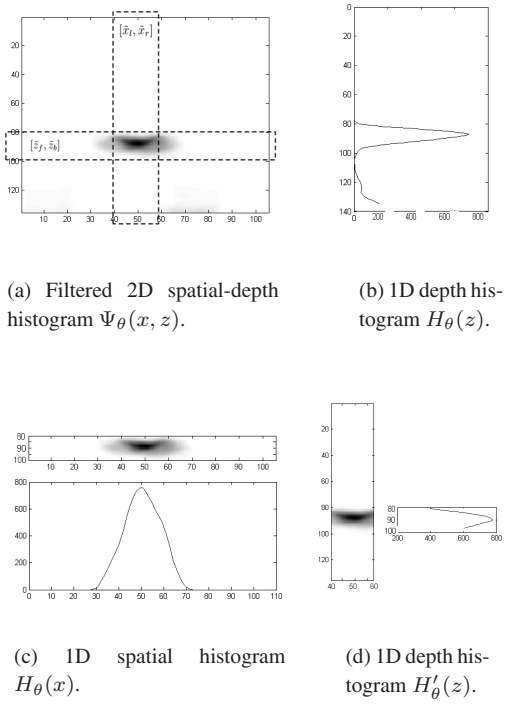
(a) F_0. (b) F_{45}. (c) F_{90}. (d) F_{135}.

Figure 3. Four directional 2D elliptical filters with four type of poses.



(a) $\Psi_0(x, z)$. (b) $\Psi_{45}(x, z)$. (c) $\Psi_{90}(x, z)$. (d) $\Psi_{135}(x, z)$.

Figure 4. Example of filtered results using four different 2D kernel functions.



(a) Filtered 2D spatial-depth histogram $\Psi_\theta(x, z)$. (b) 1D depth histogram $H_\theta(z)$.

(c) 1D spatial histogram $H_\theta(x)$. (d) 1D depth histogram $H'_\theta(z)$.

Figure 5. Several histograms used for the human candidate segmentation.

First, a 1D depth histogram $H_\theta(z)$ is obtained by projecting $\Psi_\theta(x, z)$ along the spatial axis. (Fig. 5-(b)) The value on the 1D depth histogram corresponds to the area of the human body's X-Y cross section.

Second, the value of $H_\theta(z)$ is scanned along the depth direction and the front-end of the human body is found by using the inequality

$$H_\theta(z) > \tau_\theta, \tag{10}$$

where τ_θ is a threshold that is determined appropriately as $\tau_\theta = \frac{1}{5}A_\theta$. Here, we use the subscript θ to clarify that the threshold is a function of the orientation of the filter and A_θ is the average cross section area of human bodies of a specific rotation angle θ. Preliminary experiments showed that a value of $\frac{1}{5}$ gave the best results. The depth that satisfies the inequality first is denoted as the front-end depth $\tilde{z}_f$. The back-end of the human body $\tilde{z}_b$ can be obtained as

$$\tilde{z}_b = \tilde{z}_f + T,$$

where T is the average thickness of a human body.

Third, the 1D histogram $H_\theta(x)$ is obtained by projecting $\Psi_\theta(x, z)$ along the depth axis within $[\tilde{z}_f, \tilde{z}_b]$, as shown Fig. 5-(c). The position whose $H_\theta(x)$ value is maximal is considered to be the center x_c of the body. We estimate the left-end $\tilde{x}_l$ and the right-end $\tilde{x}_r$ of a body as

$$\tilde{x}_l = x_c - \xi \times W, \tag{11}$$

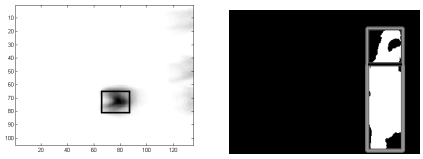
$$\tilde{x}_r = x_c + \xi \times W, \tag{12}$$

where W is the average width of the human body and ξ is a width factor. We set it to 0.4 rather than 0.5 in order not to merge two adjacent people into one in the disparity map. The front-end, back-end, left-end and right-end make a bounding box where the human body is plausibly located.

Fourth, the 1D depth histogram $H'_\theta(z)$ is computed by projecting $\Psi_\theta(x, z)$ along the spatial axis within $[\tilde{x}_l, \tilde{x}_r]$ (Fig. 5-(d)), and the true front-end z_f and the true back-end z_b are obtained by finding the sign changing point of the slope of $H'_\theta(z)$ within the interval of $[\tilde{z}_f, \tilde{z}_b]$ along the depth direction, starting from the peak value of $H'_\theta(z)$. Similarly, the true left-end x_l and the true right-end x_r are obtained by finding the sign changing point of the slope of $H'_\theta(x)$ within the interval of $[\tilde{x}_l, \tilde{x}_r]$ along the spatial direction, starting from the peak value of $H'_\theta(x)$. The true front-end, back-end, left-end, and right-end make a true bounding box where the human body is really located.

Finally, the current segmented human body is extracted from the filtered spatial-depth histogram $\Psi_\theta(x, z)$ by setting all values within the current bounding box to zero. This process will be repeated until all humans are found.

The above section has explained how to segment the human candidates using one 2D elliptical filter. All of the above procedures are repeated for the other differently oriented filters. This is advantageous because while one specific filter may fail to segment a human the other filters may succeed. Fig. 6 shows a typical example of the human candidate segmentation result, where the left and the right sub-figures are the segmented human body within the true bounding box in the $\Psi_\theta(x, z)$ and the corresponding human body in the disparity map, respectively.



(a) The segmented human body in the $\Psi_\theta(x, z)$.

(b) The corresponding human body in the disparity map.

Figure 6. A typical example of the human candidate segmentation method.

2.3 Human Verification

2.3.1 Human Verification using Face Detection

Before the face detection is performed, the head candidate segmentation is performed using the method of Li *et al.* [5] within the bounding boxes of the segmented human candidates. Then, face detection is performed using the Adaboost-based face detector [3] in the region of the head candidates. If a face is found, the segmented human candidate is a real human. However, the face detector cannot find the face when it is not frontally posed. This problem is solved by head-shoulder shape matching.

2.3.2 Human Verification using Head-Shoulder Shape Matching

We perform human verification using Active Shape Model (ASM) based scale invariant matching. To surmount rotation off-plane pose variation, we trained 4 head-shoulder shape models correspond to the specific orientations ($0^\cdot$, $45^\cdot$, $90^\cdot$, and $135^\cdot$). We use Mahalanobis distance for efficient feature matching.

3 Experimental Results and Discussion

Our experiments used four different types of image sequences. TYPE1 image sequences consisted of four image sequences where each image sequence consisted of one

human moving back and forth with arbitrary human poses. TYPE2 image sequence consisted of two image sequences where each image sequence consisted of three humans who were moving back and forth with arbitrary human poses and were allowed to overlap with each other. TYPE3 image sequences consisted of four image sequences where each image sequence consists of one human moving back and forth with four different human poses. TYPE4 image sequences consisted of two image sequences where each image sequence consisted of one human moving back and forth under the clutter background environment.

The proposed human detection system was implemented on a PC platform with 3.4GHz Pentium-4 CPU and 2GB RAM.

3.1 Human Detection

We measured the human detection method performance using four different types of images sequences: TYPE1, TYPE2, TYPE3, and TYPE4. We defined the detection accuracy f_d by the ratio of the number of detected humans N_d over the total number of humans N_t as $f_d = \frac{N_d}{N_t}$, where N_d and N_t was counted manually by human's visual inspection.

Table 1 compares the human detection performances of the existing OOSAF and the proposed 4D2DEF using the TYPE1 image sequence. It shows that (1) the human detection rate of the proposed 4D2DEF is higher than that of the existing OOSAF by almost 20%, and (2) the face detector is not effective at improving the human detection rate in the case of the TYPE2 image sequence because the image sequence does not contain frontally posed face images.

Table 1. Comparison of human detection rates (%) between the OOSAF and the 4D2DEF using the TYPE1 image sequence.

	OOSAF		4D2DEFs	
	Without face detection	With face detection	Without face detection	With face detection
Sequence 1	73.25	75.58	**96.05**	**96.05**
Sequence 2	75.27	75.27	**90.63**	**90.63**
Sequence 3	75.00	75.50	96.50	**97.00**
Sequence 4	77.57	79.00	91.41	**93.08**
Average	75.27	76.34	93.65	**94.19**

TYPE2 image sequences show three human's walking. The detection of multiple humans has an additive problem. When the humans are overlapped while moving, the extracted shape of the human candidate can not be obtained exactly. In this case, the face detection method shows its ability to verify humans.

Table 2 compares the human detection performances of the existing OOSAF and the proposed 4D2DEF using the TYPE3 image sequence. The human detection rates of the 4D2DEF are higher than those of the OOSAF. However, the human detection rates using the TYPE2 image sequence are less than those of using the TYPE1 image sequence, because there are some detection failures when two humans overlap each other. The human detection rate for the image sequence 2 is poorer than that for the image sequence 1, because image sequence 2 has a higher frequency of overlapped humans. The human detection accuracy with the face detector increases a little rather than that of without face detector in the case of the TYPE2 image sequence, because the face detector is more effective than the shape matching method for human verification when the humans overlap.

Table 2. Comparison of human detection rates (%) between the OOSAF and the 4D2DEF using the TYPE2 image sequence.

	OOSAF		4D2DEFs	
	Without face detection	With face detection	Without face detection	With face detection
Sequence 1	70.70	73.18	92.09	**94.19**
Sequence 2	65.16	66.04	88.28	**89.38**
Average	67.93	69.61	90.19	**91.79**

Figure 7. Examples of the distant human detection results.

If the proposed human detection method is applied for the human robot interaction (HRI), it should be work even when the camera (i.e., robot) is far away from people. Fig. 7 shows that the proposed algorithm can detect humans although they are far from the camera, up to 4 meters, where the figures below the detected humans are the estimated distance between the camera and the people. The distance (Z) was obtained by following equation $Z = \frac{C_B C_F}{pd}$, where C_B is a baseline and C_F is a focal length of the stereo camera, d is a disparity value positioned at the center of the detected human, and p is a pixel width. As the human is far from the camera (i.e., robot), the size of human decreases. This

makes the human detection difficult, because the small human size decreases the amount of detailed information for stereo matching. Moreover, it makes the human face detection and the head-shoulder shape matching for human verification difficult. If the human is too close to the camera (i.e., robot), the size of the human increases. This makes the human detection slow, because the large size of human increases the stereo matching time. Therefore, there is an appropriate distance range that the proposed human detection method works, i.e., from 1 to 4 meters.

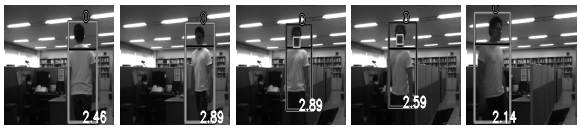


Figure 8. Examples of the human detection results in spot-like illuminations and cluttered background conditions.

Basically, the proposed human detection method does not use any background information at all. This makes our algorithm insensitive to cluttered background environments. We carried out the human detection task under spot-like illuminations and clutter background conditions using the TYPE4 image sequence. Fig. 8 shows the human detection results, which proves that the proposed human detection method still works well in a the cluttered background conditions.

3.2 Execution time

From the viewpoint of practical applications, execution time is very important. Table 3 shows execution times for TYPE1 through TYPE3 sequences. The execution time was measured by the following metric, $\text{FPS} = \frac{\text{Total execution time(sec.)}}{\text{Total number of frames}}$. We note that TYPE1 and TYPE3 sequences show almost the same FPS, but TYPE2 sequences show roughly half of them, because there is only one human in the TYPE1 and TYPE2 sequences and there are three humans in the TYPE3 sequences. As the number of detected humans increases, we inevitably need more computational time.

4 Conclusion

We proposed a pose robust human detection method from a sequence of stereo images using four directional 2D elliptical filters (4D2DEFs), which detects and identifies humans regardless of scale and pose.

The experimental results show that the human detection rate using the 4D2DEF is better than that of using the OOSAF by approximately 20% in most types of image sequences, and the execution time is more than 6 FPS.

Table 3. Execution time (FPS) on TYPE1 through TYPE3 sequences.

	TYPE1	TYPE2	TYPE3
Sequence 1	12.87	7.07	13.12
Sequence 2	12.81	6.91	12.97
Sequence 3	12.82	N/A	13.04
Sequence 4	12.86	N/A	13.82

5 Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University, and the Ministry of Education of Korea for its financial support toward the Division of Mechanical and Industrial Engineering, and the Division of Electrical and Computer Engineering at POSTECH through BK21 program.

References

[1] M. Andriluka, S. Rothk, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *IEEE International Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[2] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007.

[3] B. Jun and D. Kim. Robust real-time face detection using face certainty map. *The 2nd International Conference on Biometrics*, pages 29–36, 2007.

[4] L. Li, S. S. Ge, T. Sim, Y. T. Koh, and X. Hunag. Object-oriented scale-adaptive filtering for human detection from stereo images. *Proceedings of the 2004 IEEE Conference on, Cybernetics and Intelligent Systems*, pages 135–140, 2004.

[5] L. Li, Y. T. Koh, S. S. Ge, and W. Huang. Stereo-based human detection for mobile service robots. *Proceedings of 8th International Conference on Control, Automation, Robotics, and Vision*, pages 74–79, 2004.

[6] K. Mikolajczyk, C. Schmid, and A. Zisserman. Detection and tracking of humans by probabilistic body part assembly. *European Conference on Computer Vision*, pages 69–82, 2004.

[7] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

[8] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1491–1498, 2006.