The human detection in images using the depth map

Dmitriy Tatarenkov, Dmitry Podolsky
Scientific educational center "Mediacentr"
SUT
Saint-Petersburg, Russia
duferob@gmail.com

In today world the necessity for the autonomous mobile robots and vehicles is increasing. The safety autonomous moving demands the reliable and fast detection algorithms. The Histogram of Oriented Gradients (HOG) descriptors show significantly outperforms the existing feature sets for a human detection. Though the given method has a lot of type I errors. The amount of these errors can be decreased by using the object distance information. This paper presents a robust human detection method using pairs of color frame and depth map. During the experiment, we used color images and maps of depth received from the Kinect v2 visual sensor. During the first step in our detection experiment we processed the whole frame with the HOG descriptor and received regions of interest. Then on the second step we determined the approximate distance to this region and compare its value to the range of possible human height and width values on that distance. The experimental results show that the new proposed method of HOG and distance restriction combining provides lower false positive and increase the precision in comparison to the HOG method without using the depth map. It gives opportunities to train more sensitive classifiers, which can provide the higher recall values. Consequently, we can increase the safety moving of the autonomous mobile robots and vehicles.

Keywords—computer vision; depth map; human detection; histogram of oriended gradients

I. INTRODUCTION

The human detection in images is the most important challenge for computer vision. It is a complicated problem considering that the human appearance wide variously, there are few poses which man is able to assume. Besides, the development of a reliable system for a human detection in a complex scene is rather difficult due to the variation of illumination, background environment, location, overriding objects. At the present time researchers try to resolve two main problems, i.e. the feature extraction and developing of a classifier training algorithm. Among the vast variety, it is worth highlighting the histogram of oriented gradients (HOG) features [1], suggested by Navneet Dalal and Bill Triggs for pedestrian detection in 2005. That method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in being computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization to improve the accuracy. The support vector machine (SVM) by Corinna Cortes and Vladimir Vapnik has been used mostly for the classifier training [2].

978-1-5386-1786-1/17/$31.00 ©2017 IEEE

II. HISTOGRAM OF ORIENTED GRADIENTS (HOG) DESCRIPTOR

HOG classifier is a robust way of describing local object appearances and shapes by their distribution of intensity gradients or edge directions. This method has been used successfully as a low level feature in various object recognition tasks. The gradient vector is formed by combining the partial derivatives of the image I in the x and y directions:

$$\nabla I = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right), \tag{1}$$

The gradients in two directions can be computed using a vertical and horizontal [−1, 0, 1] filters:

$$\begin{cases} \dfrac{\partial I}{\partial x} = I * [-1 \quad 0 \quad 1] \\ \dfrac{\partial I}{\partial y} = I * [-1 \quad 0 \quad 1]^T \end{cases}, \tag{2}$$

The gradient orientations of the image are computed as:

$$\theta = \arctan\left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right), \tag{3}$$

Then the image is divided into M cells of $N \times N$ pixels. A histogram with H bins is computed and normalized by the 1 weighted gradient at each pixel, for each of the cells. The number of histogram bins used is 9, therefore each bin corresponds to a 20 degree orientation interval. The concatenation of the histograms from each cell yields a $H \times M$ length feature vector for the image. Fig. 1 shows the gradient and histogram for a sample image. The kernel is implemented for computing the gradient over the entire image, both for computing normalized histograms for each cell, and for concatenating those into the feature vector used by the SVM.

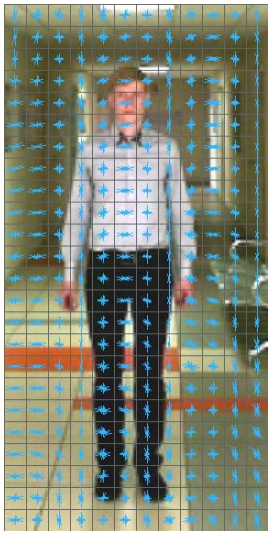

Fig. 1. Orientations of gradients on the image.

III. Support Vector Machine Classification

The SVM belongs to the maximum margin classifiers class. This algorithm searches for a decision surface that has a maximum distance to the closest points (support vectors) in the training set, which can separate two classes. The problem of learning a binary classifier can be expressed as:

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i k(x, x_i) + b, \quad (4)$$

where a set of points $x_i \in \Re^n$, $I = 1, 2, \ldots, N$, with each point x_i belongs to one of two classes identified by the label $y_i \in \{-1,1\}$ and where N is the number of training patterns, α_i and b are learned weights, and k is a kernel function. Note that $f(x)$ does not correspond to the dimensionality of the feature space. An important family of kernel functions is the polynomial kernel:

$$k(x, y) = (1 + x * y)^d, \quad (5)$$

where d is the degree of the polynomial. The weights α_i and b are selected so that the number of incorrect classifications in the training set is minimized, while the distances from this hyperplane to the support vectors are maximized.

IV. Depth Map Receiving

In our experiment we use 6 image sets for different scenes. The first scene contains a group of people who walk across the corridor randomly, deliberately opening and closing doors. The second and third scenes contain a standing and sitting man images accordingly to a distance between 1m to 13m sequentially. The fourth scene contains a group of people who enter and leave the room through the doorway. The man walks inside the room with different visual obstacles between the man and the visual sensor. The whole video information was received from a Kinect v.2 visual sensor.

The principle of operation is the measure of the time a light signal needs to travel from the camera to the object and back to the camera, or the measure of the phase shift between emitted and received signals using some high-frequency periodic waveform.

These devices have a built-in shutter in the image sensor, that opens and shuts at the same rate as the light pulses are sent out. Since the part of each returning pulse is blocked by the shutter according to its time of arrival, the amount of received light relates to the distance the pulse traveled. Such principle was invented by Antonio Medina in 1992 [3]. The distance z can be calculated using the equation:

$$z = R\,(S_2 - S_1) / 2(S_1 + S_2) + R / 2 \quad (6)$$

for an ideal camera. R is the camera range, determined by the round trip of the light pulse, S_1 the amount of the light pulse that is received, and S_2 the amount of the light pulse that is blocked. Therefore each frame of the scene is represented by the color full HD image and the corresponding depth map with the resolution of 512x424 pixels.

We manually selected the human being on each frame. The coordinates of each object had been saved to the database. Also we analyzed the second scene where the man stands in front of the camera. It allowed us to get the table of a height and width of the man in pixels corresponding to the distance.

$\alpha + \beta = \chi.$ (1) (1)

TABLE I. The correlation between a human size and his distance to a camera

Distance, m	*Height, pixel*	*Width, pixel*
1	840	245
2	473	137
3	319	146
4	240	92
5	188	70
6	160	55
7	138	40
8	119	35
9	105	30
10	98	28
11	89	26
12	81	24
13	72	21

V. Human Detection

The HOG descriptor was implemented on the OpenCV library. This detector uses SVM with the linear kernel. The OpenCV library contains linear SVM models trained for solving pedestrian detection problems. The package includes a

classifier, trained on the basis of INRIA [4], with the size of the detection window of 64x128. Through using the classifier, the parameters which can be varied include the horizontal and vertical detection window step, and the multiplicative scale change step.

During the experiment, the classifier trained on the basis of INRIA for each scene was successively applied, changing the value of the horizontal and vertical window stride in the range [2, 4, 6, 8, 10, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 56, 64, 72, 88, 96, 112, 128, 160, 192, 256] for all possible combinations with a scaling step of 1.05. As a result, we obtained the number of false positive and false positives, as well as the number of missed objects. On the basis of these data, the true positive rate (TPR) is calculated by the equation:

$$TPR=TP / (TP+FN) \quad (7)$$

where TP is the number of correct detector operations, and FN is the number of undetected objects. Also we calculate the accuracy (PPV):

$$PPV=TP / (TP+FP) \quad (8)$$

where FP is the number of false positive detections [5].

An analysis of the obtained data shows that as the step of the sliding window increases, the performance of the detector and the value of the accuracy parameter increase, though the completeness parameter decreases. As a result of the processing, the precision/recall (PR) graphs were obtained for each scene under study. The concatenation is given on Fig. 2. The detector shows more correct operations if the TPR value is higher (the highest point of the PR chart), but on the other hand it has a large number of false positives. The more to the right is the point of the graph, the lesser false responses the detector has, but also the fewer true responses occur. Therefore, to solve the problem it is necessary to determine the optimal detection parameters, where PPV and TPR take the highest values. The dependency graph allowed choosing the optimal values of the minimum height and width of the classifier window for each of the scenes, as well as for all scenes in the aggregate.

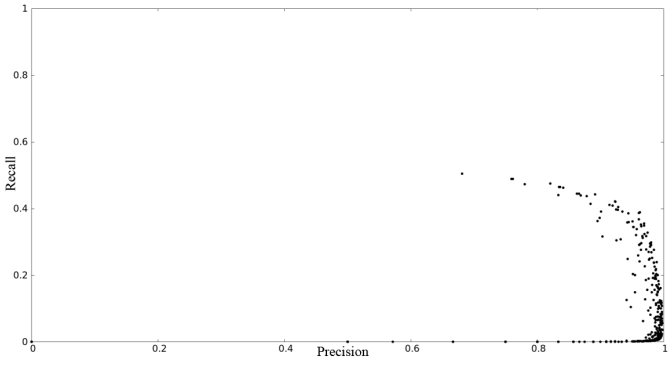

Fig. 2. Graph of dependence accuracy / completeness when changing the step of the sliding window for the classifier INRIA.

The second stage of the research was to identify the level of scale factor affected the results of human detection with the image. The step of the scanning window was selected according to the results of the previous part of the experiment; it took the values successively i.e. 4x4, 8x8 pixels. The zoom factor allowed changing the number of iterations of the scaling of the input image. The smaller this step was, the greater the probability of detecting a person on the image, but on the other hand the number of false positives increases together with the processing time of the processor. As the scale step increases, the number of undetected objects increases as well. This is due to an inaccurate matching of the sliding window with the part of the input frame containing the image of the person. Fig. 3 shows the skip detection on the middle frame of three consecutive frames when the scale factor is increased by 0.01. During the study, the classifier was sequentially used with the scaling factor step values [1.01, 1.02, 1.03, 1.04, 1.05, 1.06, 1.07, 1.08, 1.09, 1.1, 1.11, 1.12, 1.13, 1.14, 1.15, 1.2, 1.3] for each frame of the tested scenes.

$$\alpha + \beta = \chi. \quad (1) \quad (1)$$

Fig. 3. The dynamics of human detection in the image when a value of the scale factor is changed.

The analysis of data and graph (Fig. 4) showed that the optimal parameters for detecting people on the images in real conditions with the visual sensor are following: the scaling step factor is 1.03 and the sliding window step is 8x8 pixels. For these values, the value of TPR equals to 0.32, and the value of PPV equals to 0.92.

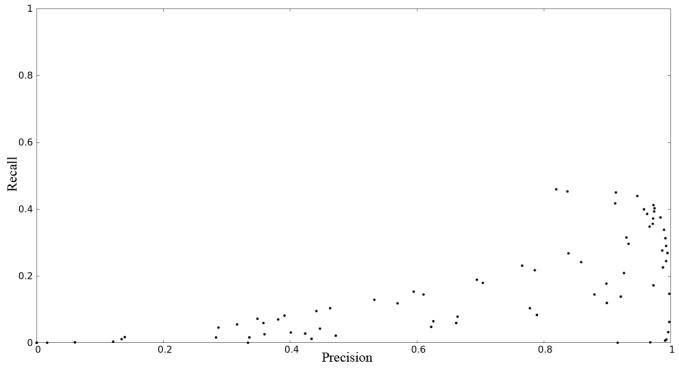

Fig. 4. The graph of precision / recall on the value of the scale factor and the value of the step of the scanning window for the classifier INRIA.

At the same scale factor step value and the minimum value of the sliding window step of 4x4 pixels the TPR equals to

0.48 and the precision is 0.57. Obviously, with an improved value of recall, a large number of false positives occur. During the experiment, a method of false positives detection based on anthropometric features of a person, i.e. his growth, was developed. The use of such data implies the presence of accurate information about the distance from the visual sensor to the intended person located in the candidate area. To obtain the distance to the intended person, a partial sampling of the values from a corresponding, pre-calibrated depth map was performed. Within the framework of the work, an assumption was made about the range of human growth from 0.4m to 2.5m. Also in the course of the experiment, the fact that the classifier was trained in images of people standing in a standing position was taken into account. In each image of the training sample, the center of the human head was at a distance from the upper edge of the image, on average, at a distance of 32 pixels (Fig. 5). It is obvious that the horizontal line at the corresponding level along the vertical line must correspond to the condition of the equation:

$$\sum_{x=0}^{(w-1)/3} d(x) > \sum_{x=(w-1)/3}^{2(w-1)/3} d(x) < \sum_{x=2(w-1)/3}^{w-1} d(x) \qquad (9)$$

where w is the width of the candidate area, d is the depth of the pixel. This formulation of the condition significantly reduces processing time to reliability of the HOG detector operation, and also makes it possible to use a depth map (or point cloud) obtained from any LIDAR type sensor.

Fig. 5. The test line on the depth map.

During the experiment, each region of the depth map corresponding to the region of the color frame was classified by the detector, as containing the person region, which was tested for compliance with the two above conditions. In case the image depth data did not satisfy at least one of these conditions, the candidate area was marked as a false positive response. The results of applying this filtration are given in Table 2, where scale factor equals to 1.03. The number of real objects detected by hand is 6106.

TABLE II. THE RESULT OF THE DETECTOR ON THE SCENES

Window stride, px	*True positives*	*False positives*	*Missed objects*	*Depth map control*
4x4	2944	2231	3162	No
8x8	1944	154	4162	No
4x4	2944	241	3162	Yes

In accordance with the above results, with a scale factor of 1.03, a sliding window stride of 8x8 pixels, the recall of test scenes equals to 0.32, which is significantly higher than the 8x8 sliding window stride without the introduced depth map classifier. The precision corresponds to the level of 0.92, which is much higher than the performance of the detector without filtering by depth.

VI. CONCLUSIONS

$$\alpha + \beta = \chi \qquad (1) \qquad (1)$$

The detector based on HOG-classifiers showed good results being applied to the human detection systems. The results of the experiments show that in the solution of the problem of detecting a person at a distance of up to 13m from the Kinect v.2 visual sensor, the most optimal is to use a classifier trained on images of 64x128 pixels with a window stride of 8x8 pixels. Such classifier parameters allow achieving an optimal ratio of the recall/precision. During the experiments we also discovered that despite of the decrease of the detection accuracy value with a decrease in the value of the sliding window stride, the recall value is much higher. During the research, a method of reducing false positives was proposed. The method uses information about the depth of the image. This approach was experimentally confirmed to allow increasing the value of the recall of detection without a significant decreasing of the precision value.

References

[1] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection.” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. – San Diego, CA, USA, pp.886-893, 2005.

[2] V.N. Vapnik, “The nature of statistical Learning theory.” Spring Press. New York, USA, 1995.

[3] Medina, Antonio. "Three Dimensional Camera and Rangefinder". January 1992. United States Patent 5081530.

[4] INRIA Person Dataset [http://pascal.inrialpes.fr/data/human].

[5] M. Enzweiler, D.M. Gavrila, “Monocular Pedestrian Detection: Survey and Experiments” Pattern Analysis and Machine Intelligence. V. 31, № 12, pp. 2179-2195, 2009.