Human detection in fish-eye images using HOG-based detectors over rotated windows

Article · September 2014

DOI: 10.1109/ICMEW.2014.6890553

An-Ti Chiang
New York University

HUMAN DETECTION IN FISH-EYE IMAGES USING HOG-BASED DETECTORS OVER ROTATED WINDOWS

An-Ti Chiang and Yao Wang

Department of Electrical and Computer Engineering, Game Innovation Lab,
Polytechnic Institute of NYU, Brooklyn, NY 11201, USA
atc327@nyu.edu, yw523@nyu.edu

ABSTRACT

Fish-eye cameras are efficient means to provide an omni-view video recording over a large area using a single camera. Although effective algorithms for human detection in images captured by conventional cameras have been developed, human detection in fish-eye images remains an open challenge. Recognizing that humans typically appear on radial lines emitted from the center in fish-eye images, we propose to apply the popular human detection algorithm based on the Histogram of Oriented Gradient (HOG) features after rotating each search window on a radial line to the vertical reference line. We extract positive and negative examples by such rotations to train the SVM classifier using HOG features. To detect humans in a given image, we rotate the image successively and detect windows containing humans along the reference line after each rotation using the trained classifier. We use multiple window sizes to detect people with different appearance sizes. We further develop an algorithm to discover multiple overlapping windows covering the same person and identify the window that encloses the person the best. The proposed method has yielded highly accurate human detection in low-resolution, low-contrast images containing multiple people with varying poses and sizes.

Index Terms— human detection, fish-eye camera, histogram of oriented gradient, support vector machine

1. INTRODUCTION

Human detection in video footage is an important task in many applications, including video surveillance in dynamic scenes, driving assistance system, content-based retrieval, etc. Effective algorithms have been developed for human detection in video captured by conventional cameras [1]. For video surveillance, fish-eye cameras are often used because they can cover a large region using a single camera. Because of the special characteristics of video frames captured by fish-eye cameras, human detection in fish-eye video remains an open challenge. Some previous works [2] [3] depend on first knowing intrinsic or extrinsic parameters of the fish-eye camera, then using these parameters to support their detection process. Satio *et al.* use the geometric relations to calculate the rough height of human in certain place [4] and use this information to guide human detection. Their algorithm is further limited to the application where people only walk through certain regions in the surveyed area. Other approaches first warp the fish-eye view to a normal view and then apply human detection algorithms for normal views. This approach suffers from the inaccuracy in camera calibration and the distortion from the warping process.

One key in human detection is finding a robust feature as the descriptor to describe human beings. Oren *et al.* use 4 non-standard Haar wavelets as the feature [5]. Lienhart *et al.* apply an extended set of Haar-like features to accelerate the detection speed [6]. Dalal *et al.* proposed a histogram of orientation gradients (HOG) feature, which is robust even in cluttered background or low luminance environment [7]. The human detection algorithm based on the HOG feature proposed in [7] is thus far the most popular [1].

In this paper, we consider how to apply the HOG – based approach for human detection in fish-eye video. Recognizing that if we rotate the bounding box covering a person on any particular radial line to the vertical line passing the center (the reference line), the resulting rotated boxes from different radial lines would have similar structures as a human in a conventional video, we propose to apply the popular HOG-based human detection algorithm after rotating each search window on a radial line to the reference line. We rotate each radially oriented bounding box containing a human body or non-human objects to the reference line and resample all resulting images to the same size, and use these as positive and negative samples for training the HOG-based classifier. To detect humans in a given video frame, we rotate the frame by different degrees and classify all possible windows along the reference line with various sizes using the trained classifier. The method has yielded highly accurate human detection in low-resolution, low-contrast images containing multiple people with varying poses and sizes.

The rest of this paper is organized as follows. In Section 2, we will describe the training and detection method of the proposed algorithm. Experimental results to demonstrate the effectiveness of the proposed algorithm are shown in Section 3. Finally, we conclude this paper in Section 4.

2. TRAINING AND DETECTION METHODS

The flowchart of the proposed algorithm is shown in figure 1. It contains two parts: training part and the detection part. We will explain different components in subsequent subsections.

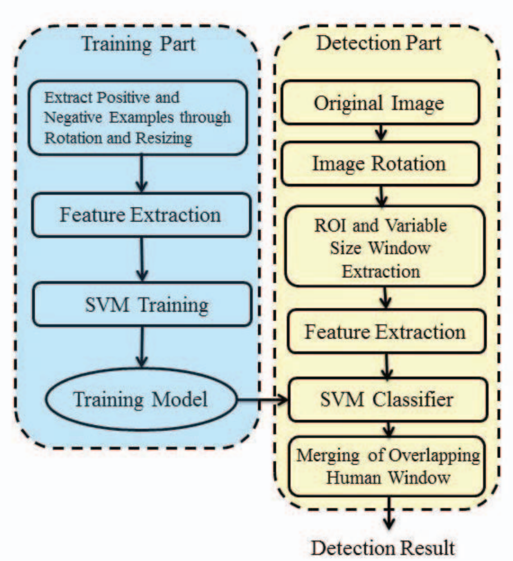


Figure 1. Proposed system for human detection in fish-eye images.

2.1. Image Rotation

Figure 2 shows a comparison between a typical normal camera image and a typical fish-eye camera image. As we can see, in the normal camera image, humans always stand perpendicular to the ground and have relatively common appearance in terms of their overall contours which can be enclosed by a bounding box of a fixed aspect ratio in the conventional video frame. However, in a fish-eye video, humans typically stand on radial lines emitted from the center (where the camera is located), and the aspect ratio of the bounding box for each person depends on his/her radial distance to the center.

Because of this special characteristic of fish-eye camera, the traditional method of using a sliding window to search in a rectangular region of interest in the original image will not work. To deal with this problem, we apply an image rotation process. We rotate the image in angle increments of a chosen step size to a reference line, which is the vertical line through the center of the image. As shown figure 3(a), the blue line represents the radial line that passes through a human and the green line is the reference line. The image rotation process rotates the image so that the radial line fits the reference line.

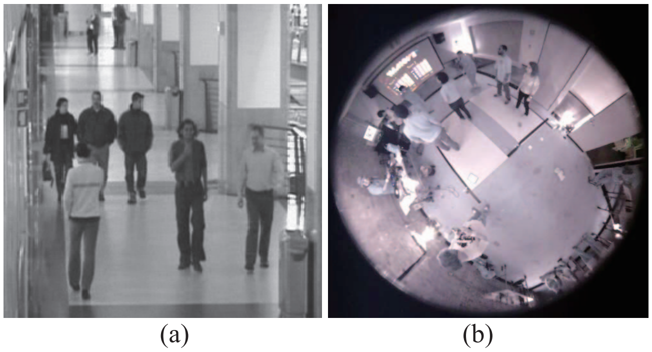


Figure 2. Typical human appearances in (a) conventional camera view, and (b) fish-eye camera view.

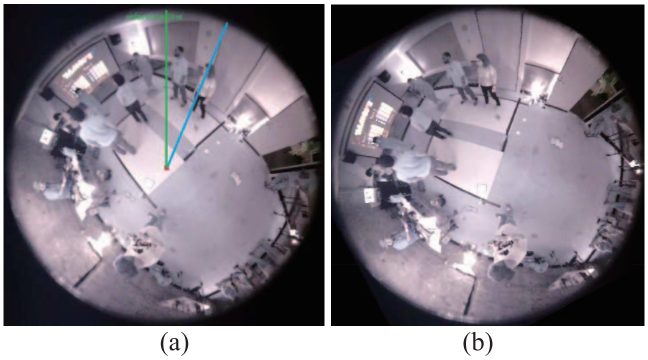


Figure 3 (a) Original image: Vertical line (green) represents the reference line and the slanted one (blue) represents the radial line intersecting with a human; (b) After image rotation.

2.2. Multi-Scale Search Along the Reference Line After Each Rotation

A Region of Interests (ROI) is a region that may contain human. As described earlier, in the fish-eye camera image, humans stand on the radial lines, so the ROIs are also on all possible radial lines. Therefore, starting with the original image, we rotate the image with respect to the camera center by a small angle (to be called angle stepsize, which is 4 degree in our experiment) each time. After each image rotation, we define a ROI along the reference line, which is indicated by the yellow region in figure 4. Within each ROI, we examine all possible detection windows of a chosen size. We choose to set the ROI width to be the same as the detection window width, so that we only search vertically all possible detection windows. The height of the ROI depends on the radius of the fish eye image, which is half of the image height. In our experiment, the resolution of the fish-eye image is 384 by 384, and hence the ROI height is 192.

The appearance size of a human in the fish-eye view depends on the radial distance of the human to the center in addition to the actual human size. To detect humans with different appearance sizes, we use multiple detection window sizes, including 32x64, 40x80, 54x102, and 60x120.

For each candidate window size, we go through all possible image rotations. After each rotation, we determine the ROI along the reference line with the same width as the detection window size. We examine all windows within this ROI with vertical inter-distance of 5 pixels using the HOG+SVM approach and label each window as either human or non-human. A window labeled as human is rotated back to its original radial line to indicate the location of a detected person.

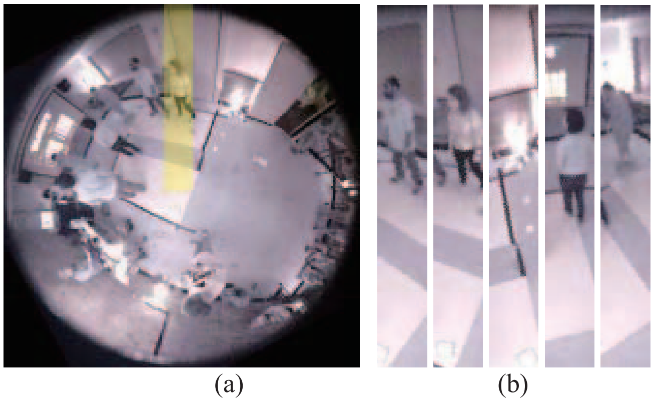


Figure 4. (a) Yellow region represents the region of interest (ROI) around the reference line. (b) Some examples of extracted ROIs after different rotations.

2.3. HOG+SVM for Each Detection Window

For each detection window, we extract the HOG features and further apply a pre-trained SVM classifier on the extracted feature vector. The HOG+SVM method for human detection, pioneered by Dalal [7], uses the Histogram of Oriented Gradients (HOG) as features for each detection window [7]. HOG features describe the distribution of intensity gradient orientations and are computed on a dense grid of uniformly spaced cells within the window. Through the use of overlapping local contrast normalization, HOG features can well describe the object appearance and yet are not sensitive to the luminance and contrast variations. The feature vector derived for each detection window is fed into a trained support vector machine (SVM) with a Gaussian kernel, for classification as either human or non-human. The SVM classifier is adopted because of its superior performance in the binary classification problem. The SVM classifier actually outputs a score, indicating the likelihood of the underlying window being a human. A window with a score higher than a preset threshold (denoted by T_SVM) is labeled as human. By varying this threshold, one can achieve different tradeoffs between the true positive detection rate and the false positive detection rate.

2.4 Merging of Overlapping Detected Windows

Because the detection windows are overlapping, the presence of a human typically lead to multiple adjacent windows being declared as containing a human after using the HOG plus SVM detector on each window. We have developed a window-merging algorithm that identifies each group of overlapping detected windows covering the same person and keeps only one window that has the highest likelihood to contain a human.

Our merging algorithm consists of two steps. The first step examines overlapping windows that are close in their angular positions and tries to merge windows that have substantial overlap along the radial line. Specifically, for each detected window containing a human, we check whether it has any neighboring detected window that is within an angle difference of +/- 4 degree, and has an overlap ratio (=overlap area/area of the smaller window) larger than a threshold (=85%). If yes, we keep the window that has a larger SVM score and remove the other window. We do this for all pairs of overlapping windows. The second step examines remaining pairs of windows that have relatively large angle differences (with angle difference up to +/- 24 degree) and yet have overlap ratio larger than a looser threshold (=62.5%). Again for each pair, we keep only the window with the higher SVM score.

2.5 Collection of Training Samples

The human appearances in fish-eye images are quite different from the human regions extracted from conventional camera views, even after the proposed rotation. Therefore, the normal human data set (e.g. INRIA [8]) is not appropriate to train our classifier. Instead we manually extract bounding boxes containing humans along all the radial lines and rotate them to the reference line. The original extracted bounding boxes with this procedure are of different size and aspect ratios, chosen to cover actual entire bodies with a desired margin. These boxes are then resized to a standard size of 32x64. Figure 5 shows some examples of the collected training data. As can be seen, the extracted samples have different appearances, differing in clothing, illumination and background. We also extract regions not containing humans along the radial lines, to collect negative samples. Figure 6 shows some examples of the negative samples.

3. EXPERIMENTAL RESULTS

3.1. Training and Testing Data Sets

Frames from a video sequence captured during an in-door dance game with a lot of on-lookers are used to train and test the proposed algorithm. The video is 2.5 hour long with a frame rate of 30 fps. The frame size is 384x384.

We created a user interface that facilitates manual specification of a rectangular bounding box containing a human along a radial line. Each extracted box is then rotated to the reference line and resized to the standard size of 32x64 pixels. Using this interface on 59 video frames containing humans, we manually created 583 positive (human) samples. These samples are each flipped horizontally to generate a total of 1166 positive samples. The initial negative (non-human) images are blocks extracted from human-free video frames through angular uniform sampling with an angle spacing of 3 degree. The block sizes vary in both the width and aspect ratio (possible widths are 32, 40, 48 and 56, and possible heights are 64, 80 and 112). All these blocks are also rotated to the reference line and resized to 32x64. The total number of negative examples is 1513. Using these positive and negative samples as the initial training set and additional negative samples extracted from other human-free frames as test images, we determine those samples that are falsely detected as humans. We add these 'false positive examples' as negative samples into the original training set to generate the final training set. The total number of negative samples in the final training set is 3363 from 6 human-free frames.

We apply the same procedure on frames not used for training sample extraction to obtain additional positive and negative examples to form the testing set. Positive examples are extracted from 7 video frames and the negative examples are extracted from 2 human-free video frames. The numbers of positive and negative examples are 120 and 594, respectively.

It was reported in Dalal's paper [7] that adding some empty space around the human in the detection window will significantly improve the detection accuracy. Our positive training samples are generated by having the bounding box containing a horizontal margin that is about a quarter of the box width, and a vertical margin that is about one eighth of the box height. For example, in a 64x128 bounding box, we will have 16 pixels in the margin on all four sides.

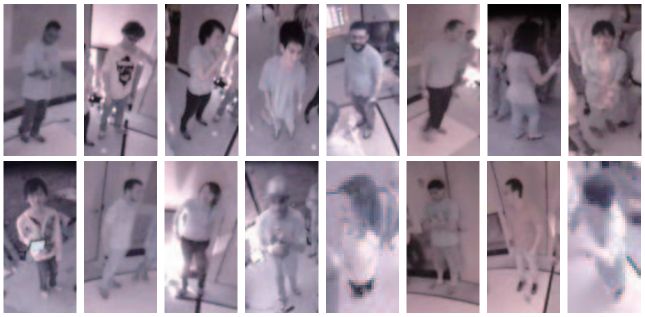
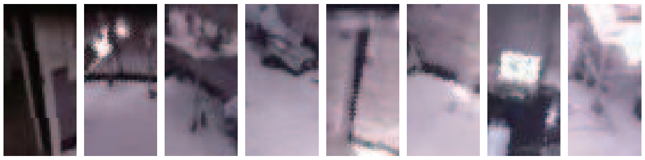


Figure 5. Positive images from our data set.



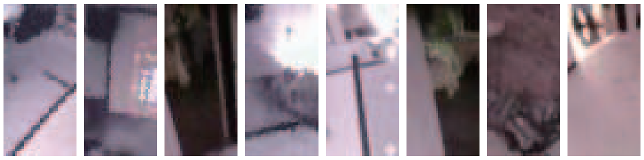


Figure 6. Negative images from our data set.

3.2 Classification Results for Training Samples and Testing Samples

We first report the detection results for pre-collected training samples and testing samples, using the trained detector.

For each sample corresponding to a resized 32x68 window, we use the method in [7] to extract HOG features, using the software provided in [9]. The HOG parameter settings are as following: the number of orientation bins is 9, cell size is 4x4, and block size is 2x2. We did not invoke local contrast normalization. We then train the SVM classifier (using the software provided in [10]) using the training set. The margin for the SVM was set to 0.01 as reported in [7]. We choose the threshold T_SVM to be -0.3 to achieve a good tradeoff between the true positive rate and the false positive rate. Table 1 shows the classification accuracy for the training set and the testing set, respectively.

Testing set	ground true = positive	ground true = negative
Classifier = Positive	118 (98.3%)	10 (1.7%)
Classifier = negative	2 (1.7%)	584 (98.3%)

(a)

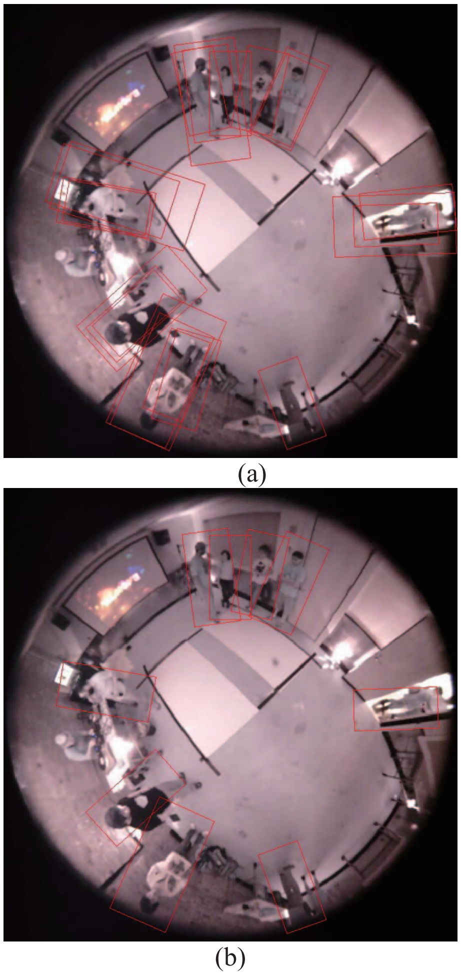
Training set	ground true = positive	ground true = negative
Classifier = Positive	1148 (98.5%)	4 (0.1%)
Classifier = negative	18 (1.5%)	3359 (99.9%)

(b)

Table 1 Classification results for (a) the testing set and (b) the training set. The overall accuracy (true positive plus true negative) is 98.3% for the testing set and 99.5% for the training set.

3.2. Detection Results on Fish-Eye Views Directly

As described earlier, we use an angle step of 4 degree to search for humans along different radial line. For each radial line, we rotate the image so that the radial line aligns with the vertical reference line, and apply multiple detection window sizes of 32x64, 36x72, 40x80, 54x102 and 60x120, along the reference line with a vertical spacing of 5 pixels. Each detection window is labeled as either human or non-human based on its HOG features, using the trained SVM classifier. A window labeled as human is rotated back to its position along the original radial line. Finally we apply the merging algorithm of Sec. 2.4 to merge multiple windows covering the same person.

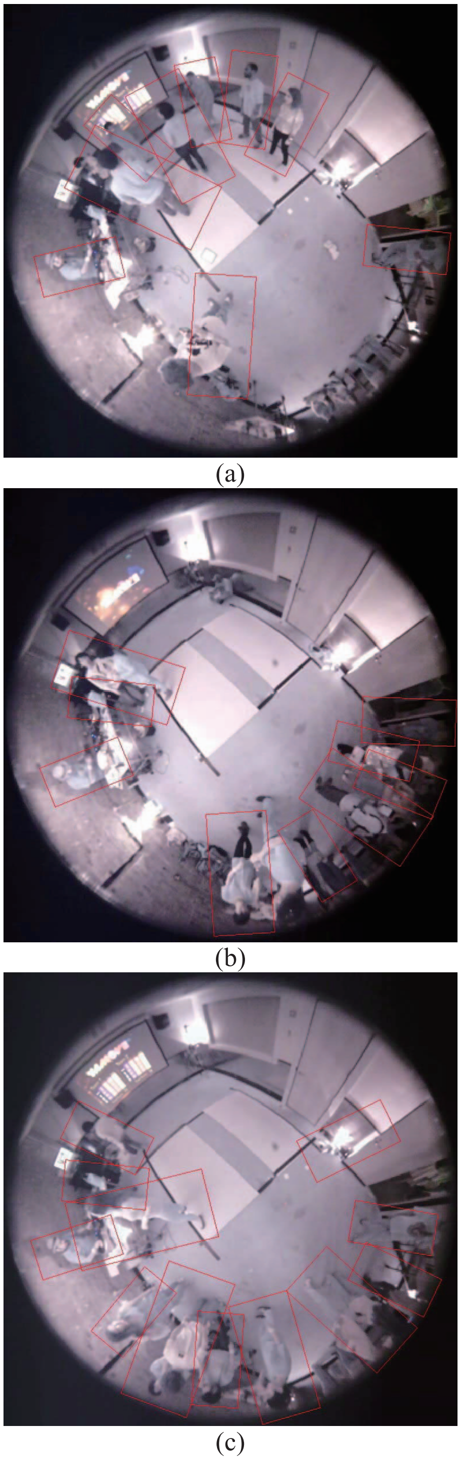

(a)

(b)

Figure 7. (a) Initial detection results in multiple detected windows for the same person; (b) detection result after the proposed window-merging process.

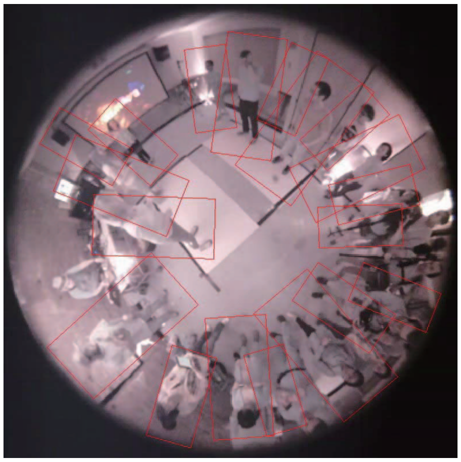
Figure 7(a) shows the initial detection result for a sample frame. Figure 7(b) shows the result after merging. We can see that the merging algorithm is very effective in removing redundant detected windows.

Figure 8 shows the final detection results for several video frames not included in the training set. We can see that most humans are detected accurately, with a few false positives and false negatives. The false negative (miss) cases usually happen in the crowded scenario. For example, in Figure 8(d), there are two undetected humans in the right of the image. One missed person is largely blocked by other detected people; the other missed person carries a big bag in front of the body and has some camera equipment on his shoulder, making his appearance very different than other humans in this video. In figure 8(c) there is a false negative detection box next to the DJ. We believe this is because the object structure in this box is somewhat similar to a human body. It is very encouraging that the detection is quite accurate in spite of the fact that the video frame has low resolution and low contrast, and humans are often partially occluded and appear in different poses.


(a)

(b)

(c)



(d)

Figure 8. Detection results for selected frames.

4. COCLUSION

In this paper, we show that the popular HOG+SVM approach for human detection in conventional camera views can be effectively applied in fish-eye images after we rotate a search window along a radial line to a vertical reference line. In spite of the fact that people located at different radial distances appear somewhat different, a single SVM classifier was able to handle people at different distances (but not too close to the center). Experimental results reveal that the proposed method can effectively detect human with different poses and appearance sizes in the omni-view images even in a scene with crowded people and cluttered background, and under relatively low contrast. The proposed window-merging algorithm was very effective in eliminating redundant detected windows corresponding to the same person.

5. REFERENCES

[1] P. Dollar, C. Wojek, B. Schiele, P. Perona,"Pedestrain detection: An evaluation of the state of the art" IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp.743-761, Aug. 2011.

[2] H. Yu, W. Liu, S. Zhang, H, Yuan, H. Zhao, "Moving object detection using an in-vehicle fish-eye camera," IEEE Int. Conf. Wireless Communications Networking and Mobile Computing, Chengdu, China, pp. 1-6, Sept. 2010.

[3] Y. Kubo, T. Kitaguchi, J. Yamaguchi, "Human tracking using fisheye image," The Society of Instrument and Control Engineers (SICE) annual Conf., Takamatsu, Japan, pp. 2013-2017, Sept 2007

[4] M. Saito, K. Kitaguchi, G. Kimura, M. Hashimoto, "People detection and tracking from fish-eye image based on probabilistic appearance model," The Society of Instrument and Control Engineers (SICE) annual Conf., Tokyo, Japan, pp. 435-440, Sept. 2011

[5] M. Oren, C. Papageorgion, P. Sinba, E. Osuna, T. Poggio, "Pedestrian detection using wavelet templates," IEEE Proc. Computer Vision Pattern Recogn., San Juan, Puerto Rico, pp. 193-199. June 1997.

[6] R. Lienhart, J. Maydt, "An extended set of Haar-like features for rapid object detection," IEEE Int. Conf. Image Process., New York, USA, vol. 1, pp. 900-903, 2002.

[7] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," IEEE Proc. Conf. Computer Vision Pattern Recogn., San Diego, USA, vol. 1, pp. 886-893, June 2005.

[8] http://pascal.inrialpes.fr/data/human/

[9] http://thebrainiac1.blogspot.com/2012/07/v-behaviorurldefaultvmlo.html

[10] T. Joachims. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning. The MIT Press, Cambridge, MA, USA, 1999.