Body parts detection for people tracking using trees of Histogram of Oriented Gradient descriptors

Etienne Corvee and Francois Bremond
INRIA Sophia Antipolis, Pulsar group
2004 Route des Lucioles 06902 Sophia Antipolis
http://www-sop.inria.fr/pulsar

Abstract

Vision algorithms face many challenging issues when it comes to analyze human activities in video surveillance applications. For instance, occlusions makes the detection and tracking of people a hard task to perform. Hence advanced and adapted solutions are required to analyze the content of video sequences. We here present a people detection algorithm based on a hierarchical tree of Histogram of Oriented Gradients referred to as HOG. The detection is coupled with independently trained body part detectors to enhance the detection performance and to reach state of the art performances. We adopt a person tracking scheme which calculates HOG dissimilarities between detected persons throughout a sequence. The algorithms are tested in videos with challenging situations such as occlusions. False alarms are further reduced by using 2D and 3D information of moving objects segmented from a background reference frame.

1. Introduction

A large variety of video applications require objects of interest to be detected, recognized and tracked in a particular scene in order to extract semantic information about scene activity. In particular, most video surveillance applications rely on the detection of human activities captured by static cameras. In this domain, although cameras remain mostly fixed, many issues occur. For example, outdoor scenes can display varying lighting conditions (e.g. sunny/cloudy illumination, shadows), public spaces can be often crowded (e.g. subways, malls) and images can be obtained with a low resolution and can be highly compressed. Hence, detecting and tracking objects in such complex environment remains a delicate task to perform. Although the techniques presented in the state of the art of this domain show great results, their success is relative to the environment, the camera location as well as the evaluation context in which the techniques are tested.

One of the major difficulty is encountered when detecting and tracking humans in occlusion scenarios since their bodies overlap onto the image plane and their foreground pixels cannot be separated when they are simply thresholded from a background reference frame. Therefore vision algorithms need to extend their analysis using information held by the underlying pixels such as shape and colors.

In this paper, attention is focused on using information provided by the human silhouette. Our approach is designed to track people using human silhouette features as well as 2D and 3D information provided by a standard object detection algorithm. Human silhouette features are trained in section 3. The people detection algorithm is presented in section 4 with the combination of body parts and a geometrical moving object information. The HOG based tracking algorithm is then described in section 5. Experimental results are given in section 6 and a conclusion on these works is given in section 8.

2. State of the art

Object detection has been studied for many decades with different approaches depending on the object of interest to be detected and the application purposes. Haar features have been studied intensely for the detection of objects, in particular for face detection [20]. One major feature used for object detection is provided by Histograms of Oriented Gradients i.e. HOG as evaluated in [5]. Pedestrians, faces and bicycles are successfully detected when represented by HOG [4, 1]. A boosting technique is often used to model and rapidly detect objects [9] such as humans [22]. SVM coupled with HOG is often used [4] for this task. Although Covariance features can be computationally expensive to estimate, their have strong discriminative powers. Tuzel and al. [19] use a Logiboost algorithm on Riemannian manifolds. Covariance features in a Riemannian geometry are trained allowing the classification of pedestrians.

Many recent papers use body parts to enhance people detection performance. There are many ways to combine body parts; for instance Mohan et al. [13] studied different voting combination of body parts classifiers. In [12], Mikolajczyk et al. use 7 body part detectors independently trained to better detect humans. Hussein and Porikli [6] introduce the notion of deformable features in a Logiboost algorithm to allow body parts to have non fixed locations in a people image template. Detected faces in profile and frontal views can also help tracking humans [anonymous].

The use of hierarchical trees has shown great interests to classify multiple object classes into clusters. Mikolajczyk et al. [11] classifies object edge based features using a hierarchical classification approach. They are capable of detecting simultaneously several object classes in a single, scale and rotation invariant model. Probability distributions of model parameters are estimated using a Bayesian rule on the object appearance cluster in the hierarchical tree as well as the geometric distribution. A 128 dimensional SIFT [10] descriptor is used on the dominant edge orientation of feature regions. They have tested their techniques for recognizing pedestrians, cars, motorcycles, bicycles and rocket propelled grenade launcher shooters in a large image database. Wu et al. [21] use a PCA trained based body part segmentation and the detected body parts are then combined to detect humans. Occlusion hypothesis using 3D location of people in a scene are analyzed to better track people during occlusion.

In terms of tracking in video surveillance applications, performance is best when the tracking scheme is well adapted and used robustly detected objects. Nevertheless, people can never be all successfully detected in crowded scene viewed by a single camera. Moreover, objects interaction in a scene can be complex and rules need to be understood by a tracker to handle difficult cases such as occlusions. For instance a person is allowed to disappear when entering his/her own car or a person has to re-enter a scene after a certain time after entering a cloakroom. A generic tracking algorithm needs to provide consistent people trajectories before robust semantic information can be extracted from a scene. Trackers often use motion prediction model such as Kalman filtering [anonymous] or scene context information [3] to provide consistent people trajectory. Singh et al. [17] first detect high confidence partial segments of trajectories called tracklets. They first detect 4 body parts using a Bayesian combination of edgelet part detector [21] which makes the people detection robust. Using a delay, tracklets of newly detected people can be merged with previously fragmented tracklets. They have evaluated with success their algorithm in occlusion scenarios present in the Caviar [2] datasets. The tracker uses a multi hypothesis theme: data association is performed using a combination of probabilities obtained from color model, motion model and a 3D human height model.

Unlike most tracking algorithms which track detected objects, features can be instead tracked independently without having to classify objects, as performed with HOG features in [1]. The detection of objects can be constrained with pixel motion information [7] or pedestrian optical flow [4].

According to the state of the art in people detection and tracking in video surveillance applications, our people tracker needs to combine body parts to handle occlusion cases. Motion of objects in the 2D image plane and a 3D calibrated needs also to be used to filter out false alarms. We propose a person classifier using a tree of HOG features. An tree example is shown in figure 1 where each image is the average of the training samples of people corresponding with similar features. Finer level of granularity of human posture is reached as we go down the tree of features. The classified HOG features allow us to elaborate our first approach to track people based on a temporal analysis of the detected people using geometrical and HOG information.

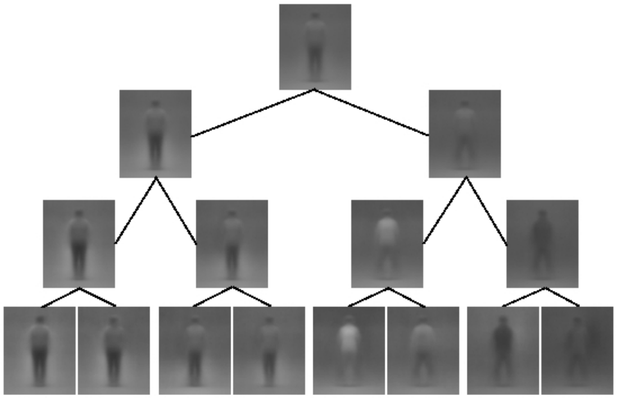

Figure 1. Tree of human appearance

3. Training

Having a set of positive samples (the people image database), we seek to find the most dominant cells i.e. the areas best describing human features. HOG features are extracted from the sampling of pixel edge orientation into a $N_b = 8$ bins histogram in a given area of the image referred to as a HOG cell. Edge magnitude and orientation are estimated using a first order Sobel kernel. HOG histograms are constructed from the edge magnitude response as formulated in equation 1 for 1 bin b. Histograms are then normalized for all N_b bins. In equation 1, $\|\nabla I(\mathbf{x})\|$ is the magnitude and $\Theta(\mathbf{x})$ is the orientation of the edge response operator for all pixel locations $\mathbf{x}$ within the cell c. The term T denotes the sampling rate which here corresponds to $360^o/N_b$ i.e. 45^o.

$$h(c,b) = \sum_{\mathbf{x}\in c} \begin{cases} \|\nabla I(\mathbf{x})\| & \text{if } \frac{\Theta(\mathbf{x})}{T} = b \\ 0 \text{ else} \end{cases} \quad (1)$$

$$H(c) = argmax_b\,(h(c,b)) \quad (2)$$

A dimension reduction of the HOG feature is performed in equation 2 by extracting the most dominant orientation $H(c)$ from the histogram. $H(c)$ is simply the bin giving the maximum histogram response which is thresholded to discard low contrasted cells. Applying a dimensionality reduction means a loss of information and cells should not be too large. In the presented works, a cell dimension of size 8×8 pixels is employed. HOG cell orientation is extracted at every 4 pixels across image samples.

Once the dominant orientation $H(c)$ is found for all cells in a image sample, we need to find the cells best describing the global human appearance in the entire database composed of N_p positive images in the database. Each cell is associated with N_p dominant orientations for which a histogram is calculated according to equation 3 where b, c and i represent the bin index, the cell index and the image index in the database respectively. The trained most dominant cell edge orientation $M(c)$ is given by the histogram maximum probability of occurrence as defined in equation 4. Cell strength is estimated by a weight $w(c)$ directly given by $M(c)$ as expressed in equation 5.

$$m(c,b) = \sum_{i} \begin{cases} 1 \text{ if } H_i(c) = b \\ 0 \text{ else} \end{cases} \quad (3)$$

$$M(c) = argmax_b(m(c,b)) \quad (4)$$

$$w(c) = \frac{max_b(m(c,b))}{\sum_b(m(c,b))} = \frac{m(c,M(c))}{\sum_b(m(c,b))} \quad (5)$$

The first node (top of the tree) constitutes of the whole training dataset and of the cell giving the maximum $M(c)$ among all possible cells. Two sub nodes are initiated by splitting the database in two according to the distribution of cell orientation error $\Delta\Theta_i(c)$ as follows:

$$\Delta\Theta_i(c) = \frac{\|H_i(c) - M(c)\|}{N_b/2} \quad (6)$$

where $H_i(c)$ is the cell HOG feature (defined in equation 2) of the i^{th} sample. Errors are normalized in the range 0 and 1. Sub-nodes are in turn initiated from each node using the maximum cell's $M(c)$ unused in the parent nodes. An image sample has a path in the tree and is then associated with the sum of weighted errors for each level as follows:

$$e_i = \frac{\sum_n w(c,n)\Delta\Theta_i(c,n)}{N_{lv}} \quad (7)$$

where n is the node index along the tree of N_{lv} levels. An example of a 4 level tree is shown in figure 1. The number of features to describe objects increases with the number of levels, however a N_{lv} levels tree requires to allocate $2^{N_{lv}-1}$. To limit this high memory requirement, a tree is divided into smaller successive smaller trees: an image sample undergoes iteratively several tree classification instead of one single tree. For N_{it} iterations, the overall sample error becomes the average of the error e_i. The overall training takes about 10 minutes for 5500 training samples.

body part	Height %		Width %		colour
	top	bottom	left	right	
torso	20	55	20	80	green
head-shoulder	10	50	25	75	magenta
left arm	20	55	20	40	blue
right arm	20	55	60	80	cyan
legs	50	95	20	80	yellow

Table 1. Body parts location in the person image template

4. Person detection

A scanning window of the size of the training samples i.e. 48x96 pixels scans each image of a video to find object candidates. The scan is performed with a sampling rate of 4 pixels vertically and horizontally across the image. This searching operation is repeated over multiple lower resolutions (15%) of the original image in order to find people with various sizes in the image. The Y color channel is extracted from the input image.

The integral image technique [20, 16] is then used for fast computation of the HOG features at the most discriminative cells along the tree. The HOG error defined in section 3 is computed for each scanning window and compared to a threshold evaluated in section 6 for people detection.

4.1. Combination with Body parts detection

Five body part detectors are trained using the same training scheme described above to refine people detection results. These detectors are trained on smaller and restrained locations of the person image template. The body part locations are selected manually in percentage of the template height and width from the top left image origin as shown in table 1.

Figure 2 shows an example of detected body parts combined with the detected person in two TrecVid [18] sequences. The bottom row figures display the HOG cells used for the detection and the top row figures display these cells bounding boxes. The thick white box represents the person template size and the red box represents the bounding box of the detected full body. Body parts are drawn with color referenced in table 1. False person alarms can be filtered out if it is not associated with enough body parts as evaluated in section 6. More false alarms can be filtered out using motion information as explained in the next section.

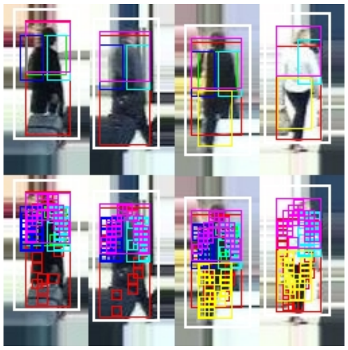

Figure 2. Examples of detected body parts and the corresponding HOG cells

4.2. Motion filtering using 2D and 3D information

False person candidates detected in regions where there is no object detected from a simple background reference frame subtraction are filtered out. An integral image is calculated from the binary foreground image map in order to quickly estimate the percentage of foreground pixels within the bounding box of a person cells and its body parts cells. If the percentage of motion of a person is below a threshold of 50% then this person is considered to be a false alarm. In the other case, body parts are eliminated if their motion percentage is below a threshold of 75%. A false alarm is detected if the number of body parts is below a threshold which is evaluated in section 6.

Given a 3D calibrated environment, 3D position and dimension of people candidates are calculated. The Tsai calibration technique is used to calibrate a scene. If the 3D dimensions do not fit a pre defined 3D human model, the person is no longer a valid person candidate.

5. HOG based people tracking

For a tracking algorithm to be performant, it needs to be provided with reliable detected object with as much as information as possible i.e. its position in the scene under view and a visual signature as unique as possible. However, occlusion occurs under various forms: it can be static or dynamic. In both cases, moving objects goes from partially to totally occluded by static object(s) or by other dynamic object(s). When the degree of occlusion is too high, detection algorithms can no longer detect these objects using a single frame. Therefore a tracking algorithm needs to retrieve the lost information in later frames in order to pursue the tracking of objects coming out of occlusions or with low degrees of occlusion.

In these works HOG descriptors integrated in the tracking scheme. The proposed tracker uses a simple 2D dissimilarity measure to link newly detected persons with the previously detected persons. A link between two persons is created if their bounding box distance is less than half of the maximum bounding box height between the two objects. However, when an object is subject to an occlusion many links or no links can be associated with a person. An error term based on HOG dissimilarity is then added to make the tracking algorithm more robust.

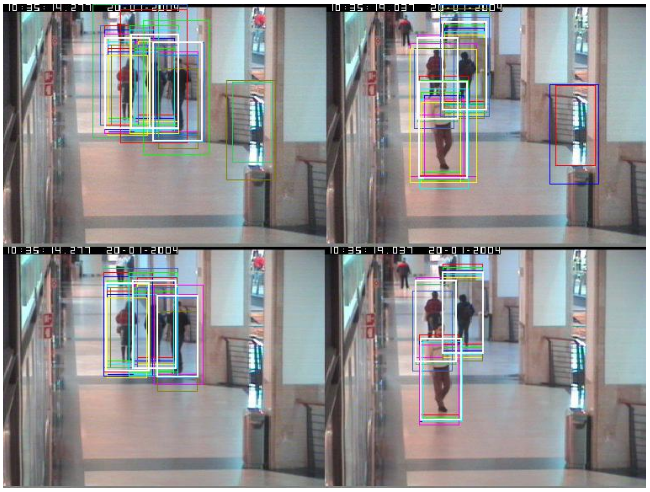

Figure 3. Examples of filtered detected person candidates in a Caviar sequence

A link between two objects detected in two different frames is associated with three error terms:

- The link geometric error e_g calculated by the dice coefficient of the bounding box areas of the two objects.
- The link HOG error e_h which is the weighted average of the cells dissimilarity between the objects. Cell weights are provided by equation 5 and two cells dissimilarity is given by the product of these two error terms:
  - the geometric error calculated by the normalised cell 2D displacement between the two closest cells with respect to the object size, and
  - the orientation error calculated by the angle difference between the cell orientation (based on equation 6 without any trained values).
- The link combined error $e_{gh} = e_g e_h$.

Link errors are also calculated for the detected body parts and averaged with the full body person link error. Every non updated persons i.e. who do not matches any previously detected persons or whose link is weaker than another link using the same matched persons, lead to the creation of new trajectory.

algorithm	TD%	FD%
OpenCv HOG	25.01	0
proposed HOG (Th1)	29.76	0
proposed HOG (Th2)	36.04	0.04

Table 2. Evaluation using Nicta testing database: TD = true detection rate and FD = false detection rate

6. Experimental results

The people detection algorithm is evaluated in section 6.1. The combination of body parts detection is evaluated in section 6.2 and the people tracking algorithm is then evaluated in section 6.3.

6.1. People detector evaluation

The detector is trained with a positive database composed of 500 images of persons provided by the MIT university [15] and negative database composed of 5000 images of various background scene provided by the NICTA project [8]. The database shows high variability in its content: people have many various poses with various degrees of occlusion and often wearing clothes similar to the background. Five iterations and a five level tree are necessary for the system to be performant while remaining relatively fast. Figure 4 shows the performance of the full body detector for 10 possible normalised thresholds ($[0, 0.1...1]$) when evaluated against a testing database composed of 424 positive and 5000 negative samples (different than training database). The results in table 2 show that our technique performs better than the HOG based detector provided by OpenCv [14] for two optimal thresholds chosen from figure 4.

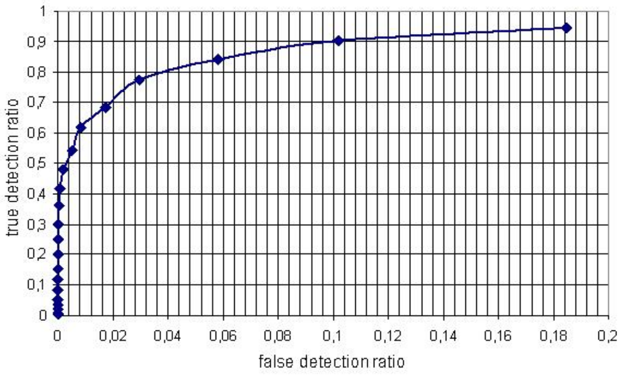

Figure 4. People detector performance for various thresholds

6.2. Body part detector evaluation

The detection of people combined with body parts detection (described in section 4.1) is here evaluated using two Caviar video sequences [2] entitled 'ceecp1' and 'cwbs1' provided with annotated bounding boxes. A comparison is performed with a system using the geometric constraints provided by motion filtering (described in section 4.2).

The system is evaluated using two criteria: the false alarm rate FA which defines the average number of false alarms per frame and the missed detection rate MD which is the average number of mis-detected ground truth object. A detected person and a ground truth object are assumed associated when their bounding box areas overlap with a minimum of 60% intersection (this threshold is subject to the large size difference of ground truth bounding boxes and the training database template). The FA and MD results are shown in figure 5 for various number of body parts detected (including the person detection) in the sequence 'ceecp1'. When no body parts are used, the results show that motion filtering the rate of false alarms considerably decreases than without filtering, but with slightly higher missed detection rates. We can also see that results looks similar when at least 2 body parts are used: hence body parts combination has similar effect than filtering. In this sequence, we perform better than the OpenCv HOG detector for a minimum of 2 body parts with or without motion filtering.

For the other sequences, similar behaviors are observed. The performance figures are displayed in table 3 for a minimum of 3 body parts where FA is the average false alarm rate and MD the average missed detection rate over the 3 sequences. The results show that our approach performs better than OpenCv and filtering induces lower false alarms but with higher missed detections than without filtering. In terms of tracking, false alarms induces false alarm tracks associated with small spatio-temporal trajectories easily filtered out by thresholding operations. Missed detected persons induce broken trajectories on a frame to frame basis and which can be avoided by merging fragmented tracks.

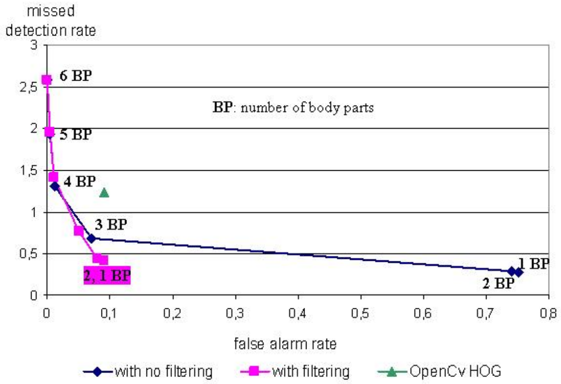

Figure 5. Evaluation of the body part detector

algorithm	FA	MD
OpenCv HOG	0.68	1.42
Our HOG - no filtering	0.22	1.57
Our HOG - with filtering	0.19	1.61

Table 3. Detection evaluation: FA = number of false alarms per frame and MD = number of missed detected ground truth per frame

6.3. People tracking evaluation

The people provided by the HOG based people detector are tracked by the HOG based people tracker described in section 5. This tracking approach is here evaluated. In simple scenarios where people are not occluding each other, the results showed that people are 100% successfully tracked whether HOG information is combined or not with geometric dissimilarities. We have hence chosen sequences where occlusion also occurs. Table 4 displays the figures obtained when tracking people in 5 TrecVid sequences captured at 5 fps and containing between 46 and 140 frames. A total of 25 persons were manually annotated throughout these sequences. The evaluation is performed in terms of:

- MF - mean fragmentation rate over the sequence: the mean number of detected tracks (i.e. IDs) per ground truth track.
- MLT - mean longest track life over the sequence: the mean of the longest fragment for each ground truth track.
- MTT - mean total track life over the sequence: the mean of all the fragment lifetimes for each ground truth track.

The results show relatively similar tracking behaviors: in average, 73% of the ground truth persons are well tracked with 3 different trajectories and the longest correct tracks covers more than 50% in average of each ground truth trajectory. The results also show that the tracker is more performant in terms of MF and MLT when combining geometric and HOG cues. A slightly higher MTT is obtained for the tracker using solely HOG information followed by the combined cues tracker. We reach a frame rate of 1.94 frames per second for half PAL size images for the entire vision system using a 2.40 GHz 32 bit processor with 1GB memory.

Figure 6 shows examples of tracked persons in occlusion scenarios in the Caviar and TrecVid sequence.

7. Acknowledgement

We would like to thank the ANR project 'VideoId' who partially founded this work.

algorithm	MF	MLT%	MTT%
Tracker Geo	3.33	52.2	72.1
Tracker Hog	3.27	56.7	73.5
Tracker Comb	2.88	57.3	73.4
Rank	C,B,A	C,B,A	B,C,A

Table 4. Tracking evaluation in occlusion scenarios in TrecVid camera 1: MF = Fragmentation rate, MLT = mean longest track life, MTT = mean total track life, Tracker Geo: tracker using solely geometric cues, Tracker Hog: tracker using solely HOG cues, Tracker Comb: tracker using all cues

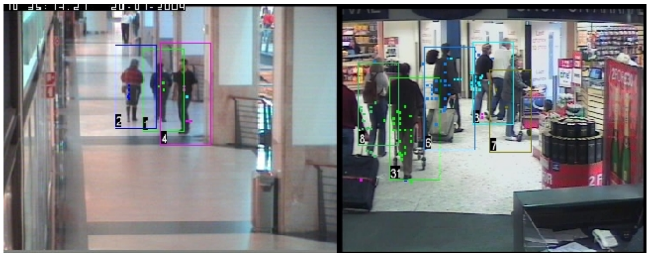

Figure 6. Examples of detected HOG descriptors and tracked persons

8. Conclusion and future works

A novel tree based approach is proposed to classify the various people appearances according to HOG descriptors characterized by most dominant edge orientation. This system is quickly trainable and hence easily adaptable while allowing people to be reliably detected. The combination of body parts with motion filtering showed performance enhancement of the proposed detection algorithm and to performs better than the HOG detector provided by the OpenCv library. A novel tracking scheme is adopted in order to investigate HOG dissimilarities of persons and their body parts in a video with challenging occlusion cases. Better tracking performance is obtained when tracking combines geometric and HOG dissimilarities. Additional cues shall be added to enrich visual signatures of people and help merge fragments of trajectories.

The proposed approach reaches a frame rate of about 2 fps which can be increased when using 3D constrains of a calibrated scene. The performance of people detectors varies considerably with the nature of the database used for training and the nature of the video sequences. Hierarchical classifying trees could provide useful information about training databases for them to be better used. Body parts provide information which deserves better attention and analysis to cope with occlusion and understand people dynamics in complex environments.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragment-based tracking using integral histogram. In *Computer Vision and Pattern Recognition - CVPR*, 2006. 1, 2

[2] CAVIAR. Context Aware Vision using Image based Active Recognition. In *http://www.homepage.inf.ed.ac.uk/rbf/CAVIAR/*. 2, 5

[3] D. P. Chau, F. Bremond, E. Corvee, and M. Thonnat. Repairing people trajectories based on point clustering. In *In the International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009. 2

[4] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition - CVPR*, 2005. 1, 2

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: a benchmark. In *CVPR*, 2009. 1

[6] M. Hussein, F. Porikli, and L. Davis. Object detection via boosted deformable features. *IEEE International Conference on Image Processing (ICIP)*, 2009. 2

[7] M. Jones, P. Viola, P. Viola, M. J. Jones, D. Snow, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *In ICCV*, pages 734–741, 2003. 2

[8] S. Kong, M. Bhuyan, C. Sanderson, and B. Lovell. Tracking of persons for video srveillance of unattended environments. In *19th International Conference on Pattern Recognition (ICPR2008)*, 2008. 5

[9] I. Laptev. Improvements of object detection using boosted histograms. In *Proceedings of the British Machine Vision Conference*, 2006. 1

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2:91–110, 2004. 2

[11] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. *CVPR*, 2006. 2

[12] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004. 2

[13] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:349–361, 2001. 2

[14] OpenCv. Intel Open Source Computer Vision Library. In *http://sourceforge.net/projects/opencvlibrary/*. 5

[15] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *International Conference on Image Processing: ICIP99*, 1999. 5

[16] F. Porikli. Integral histogram: a fast way to extract histogram in cartesian space. In *CVPR*, 2005. 3

[17] V. K. Singh, B. Wu, and R. Nevatia. Pedestrian tracking by associating trackets using detection residuals. *IEEE Workshop on Motion and video Computing (WMVC)*, pages 1–8, 2008. 2

[18] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid,. In *MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006. 3

[19] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *PAMI*, 30(10), 2008. 1

[20] P. Viola and M. Jones. Robust real-time face detection. In *International Journal of Computer Vision*, 2004. 1, 3

[21] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *CVPR*, pages 951–958, 2006. 2

[22] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition - CVPR*, 2006. 1