A fast and precise HOG-Adaboost based based visual support system capable to recognize Pedestrian and estimate their distance.

Takahisa Kishino[1], Sun Zhe[1],Ruggero Micheletto[1]

[1] Yokohama City University, Graduate School of Nanobioscience, 22-2 Seto Kanazawa-ku, 236-0027 Yokohama, Japan

Abstract

In this paper,we present a visual support system the visually impaired. Our detection algorithm is based on the well known Histograms of Oriented Gradients (HOG) method, due to its high detection rate and versatility[4]. However, the accuracy of object recognition rate is reduced because of high false detection rate. In order to solve that, multiple parts model and triple phase detection have been implemented. These additional filtering stages were conducted by separate action on different area of the sample, considering deformations and translations. We demonstrated that this approach has raised the accuracy and speed of calculation. Through an evaluation experiment based on a large dataset, we found that false detection has been improved by 18.9% in respect to standard HOG detectors. Experimental tests have also shown the system ability to estimate the distance of the pedestrian by the use of a simple perspective model. The system has been tested on several photographic datasets and have shown excellent performances also in ambiguous cases.

Keywords: Pedestrian detection, HOG methods, distance evaluation, single-camera, Adaboost

1 Introduction

In this study, we present steps toward the realization of an automatic pedestrian detection system for the visually impaired. Today, the visually impaired has many obstacles to walk outside and despite the advancement of technology in many fields, we often see blind people accompanied by friends or by a dog. Widespread devices to help these persons are not common yet. However, we think that audio support devices that are based on a GPS route navigation systems[7] are promising and may be successful. These can point the location and indicate directions to the user goal. On the other hand, they also present dangers because of possible collision by other pedestrian. To reduce this risk, there is high demand for system that notifies of approaching pedestrian indicating also the distance range. Therefore here we aim to develop a simple and low cost active vision system which warns the visually impaired based on camera images with a single-camera portable device.

Many of the pedestrian detecting systems are to be used with a fixed camera. These systems typically apply standard computer vision techniques, such as background subtraction or background modelling[9]. Since the camera is given to the visually impaired, the images are shot directly from the user viewpoint. Therefore background is changing without predictable rules and a highly variable image is unavoidable, making unusable the most simple subtraction elaboration techniques. As an alternative for background subtraction, it is possible to use the approach of detecting multiple and deformable parts. This approach will be used in this study and its results shown. We will demonstrate that it is possible to reduce the effect of a variable and dynamic background[5] and obtain high quality detection and robust results. Of course, estimating the distance to the pedestrian is an essential function for notifying a collision risk. To achieve this function, usually stereo or multiple camera systems are used. These systems are bigger and more costly, moreover since they need two or more images to estimate the distance at for every single frame, they requires more memory than our proposed single camera system. Commercially available portable devices of this kind, stereo or fitted with multiple cameras are very few in number. They are specialized, high cost and difficult to realize. Therefore here we focus in the effort to develop a single camera system, simple and portable able to recognize the pedestrians as possible collision target, estimate approaching distance and giving out a proper warning signal that can be audio or can be relayed to another device for different operations.

Our system consists of an algorithm that acts in two phases: pedestrian detection and estimation of distance. In the detection phase, it scans the image for the pedestrian shape standing in front of the camera. We improved a standard HOG method introducing a multiple part model to it. Multiple part detection algorithms are usually considered slow in processing time, because the elaboration is repeated for each candidate object. To surpass this problem and reduce processing time, we used a cascade method for the multiple part model. This approach results in the ability to scan images more efficiently, in an optimized manner. The multiple part algorithm will still require more time than a simple HOG model, but the amount of increased processing is much less than the standard non-optimized multiple part HOG model. Our algorithm is able to simultaneously estimate the pedestrian distance. In this elaboration phase, the distance is calculated based on the result of the previous detection phase. By using a perspective projection, the location on the image of specific parts of pedestrian gives us the distance from the camera. In other words, we used real-world pictures of people located a different distances, and determined an algorithm that associate position of head, feet of target pedestrian to the average location in pixel on the image. Once the first detection phase is finished, the second algorithm use the location information obtained to determine the estimated distance, and increase or decrease in distance of a pedestrian within a defined frame range, results in the ability to calculate the collision risk.

2 System model

This section describes the pedestrian detection system and the estimating distance to them.

2.1 Histograms of Oriented Gradients:

The underlying building blocks of our method are the Histograms of Oriented Gradients (HOG)[3]. HOG representation captures the gradient structure that is characteristic of the human shape. A magnitude m and orientation θ of gradients at each pixel are given by the equation:

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \tag{1}$$

$$\theta(x,y) = \tan^{-1}\frac{f_x(x,y)}{f_y(x,y)} \tag{2}$$

where $f_x(x,y) = L(x+1,y) - L(x-1,y)$, $f_y(x,y) = L(x,y+1) - L(x,y-1)$ and $L(x,y)$ is proportional to the brightness of a pixel. θ is discretized into one of nine orientation bins. Each pixel is assigned the orientation of its gradient, with a strength that depends on m. The image is divided into $n * n$ not overlapping pixel regions that are called *cells* and each group of cells is integrated into a *block*. The blocks can overlap with each other. These gradient features are represented as one-dimensional histogram, as shown in Figure 1.

2.2 Multiple parts model:

The pedestrian have diverse postures (e.g. looking down, checking the phone). From the point of view of the camera held by the subject, the relative changes in the parts (e.g. head, arm) position are especially important. So our system is required to deal with various poses without making mistakes. One of the methods to detect diverse human poses, consists of divide the target object into several parts [5] and consider the whole as the composition of them. Thus our method treats the person model as a cluster consisting of three parts, that define the complete body, the head and the legs, see Figure. 2. The model of the pedestrian is composed by a base filter and two secondary models (P_1,P_2). The base filter F_0 covers the entire human and defines the rough pedestrian location. The secondary models represent the head and feet parts. In general each secondary model is given by the relation:

$$P_i = (F_i, v_i, s_i)$$

Here i is the part number, F_i is the part filter for the specific ith element, v_i is a two dimensional vector locating the center of a box enclosing the part and finally s_i gives the size of the box. There are three sizes for each box to reduce error.

We have considered that using higher resolution is essential to reduce false detections. Therefore each part filter f_i has a higher HOG resolution than its base. The position of each filter can move at each scene, remaining confined within its base filter range. The filter can act externally to the base filter provided it is overlapping with it at least 50% of its area. In this way the variability is enough to recognize the region of the body in diverse position without incurring in errors.

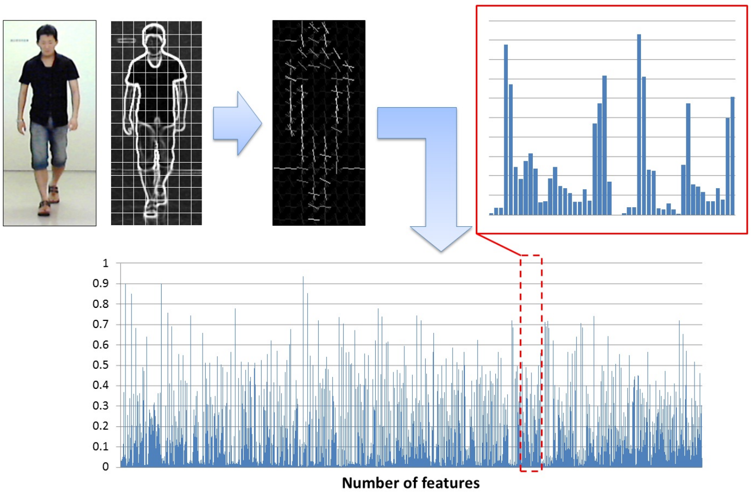


Figure 1: An example of the HOG process. The HOG analysis acts on *cells* that contain the images edges of the image captured by the camera. The edges are evaluated for orientation θ and an *orientation magnitude* m value is calculated for each cell and normalized . The magnitudes m are normalized by $blocks$ (see text for details). These normalized magnitudes are stored as an one-dimensional histogram and represent the main feature of the image that will be compared with a database of positive or negative samples for the human recognition process.

2.3 Classifier construction:

To optimize performances we used a cascade AdaBoost classifier[11, 12] trained by the main HOG features[10]. When the processing is complete, the final classifier $H(x)$ is a linear combination of several weak classifiers $h_t(x)$. The number of weak classifiers is T and it is equal in number as the learning samples.

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where x is the input feature data, t is number of learning round. α_t is the weight of the tth learning data, this it is given by the equation:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - e_t}{e_t}$$

where e_t is the weight summation of each learning sample at tth learning round.

In other words, the output value of the final classifier represents the level of resemblance to human profile of a particular area on the image. When the classifier is applied to the image area where a human is present, it outputs a high value. By iteration of raster scans, a map that associates the human likelihood to the image is generated. As shown in Figure 3, our classifier likelihood is mainly distributed over the human profile in the example.

The features of HOG are compared with positive samples (human images) and negative ones (background object and other non-human images). The features that exhibit greater differences between positive and negative samples are considered for an efficient classifier, about 200 classifiers were selected in this study.

We also compared the AdaBoost classifier algorithm with the SVM classifier and we realized that a weak classifier AdaBoost with high-speed detection is more suitable for pedestrian recognition[6].

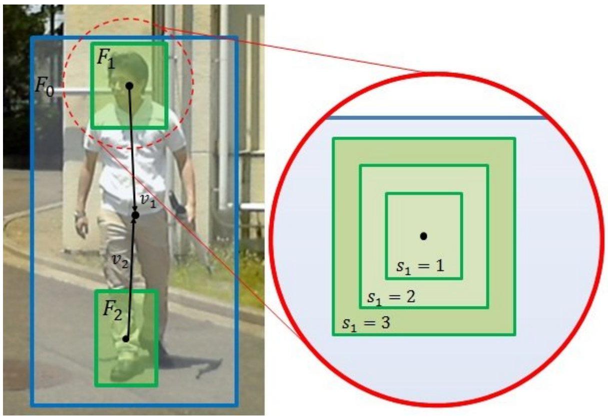


Figure 2: A sketch representing how the multiple part analysis is done. We apply the HOG algorithm firstly on the whole screen and by the use of a mean-shift clustering procedure the position of the person body is located. To realize a more robust analysis we apply the HOG algorithm again twice in order to locate the head and the feet parts of the image. This further analysis is made using the three different sizes s_i as mentioned in the text to reduce error.

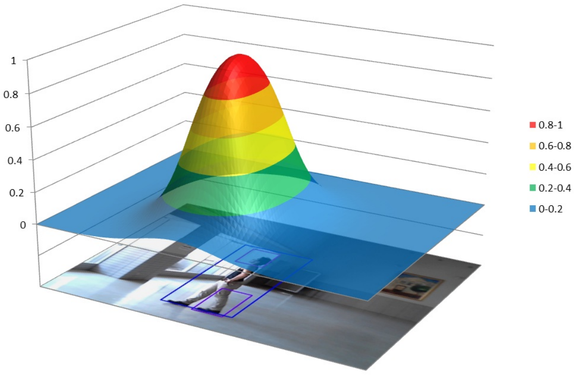


Figure 3: An example of the likelihood distribution of a image. For each image frame captured by the camera, a likelihood map is generated in real time. This map is obtained from output values of the classifiers that scans the image in an iterative process. When a region of the image is found to have similar features to the human profile accordingly to a database of learning data, that area is assigned a higher likelihood.

2.4 Distance range estimation:

To notify the user of collision risks, our system estimate the distance to pedestrian that are located ahead the camera. As the image is shot from a single camera the distance should be estimated by perspective projection. There are two clues to estimate distance by perspective: the size of pedestrian in the image and its position. Since the size of pedestrian depends on the person body height, our system derives the approximate range using the position of the pedestrian. We assume a planar road surface and that a camera optical axis is parallel to the road surface, even though the system is robust enough to accept small perturbations around these conditions (see 4 with a diagram of the imaging geometry).

The camera is held by the user at height h. The distance to the pedestrian from the camera is d. The point on the road at the distance d is projected onto the image plane at the position y. This is the image coordinates given by the equation:

$$y = \frac{fh}{d} + \frac{H}{2}$$

where f is the camera focal length expressed in pixel and H is the full screen size.

When a first pedestrian is at a distance d_1 and a distant pedestrian is at d_2, the points of contact are projected onto the image at

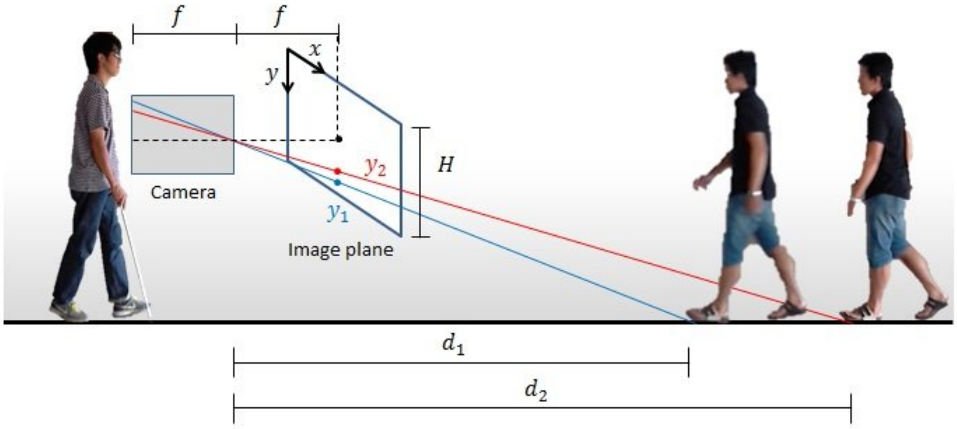


Figure 4: Ideal image geometry in the perspective model for distance estimation. See text for details.

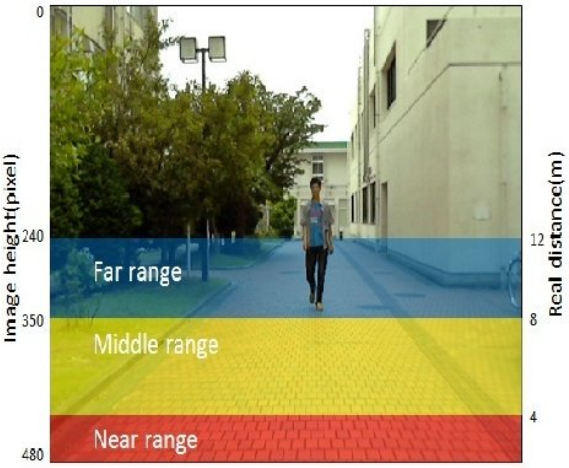


Figure 5: The relation with distance shown on the camera image.

y_1 and y_2. As shown in the example of Figure. 4, y_1 results to be smaller than y_2. To calculate the distance to a pedestrian the system detects the person foot and its range is estimated from the following formula:

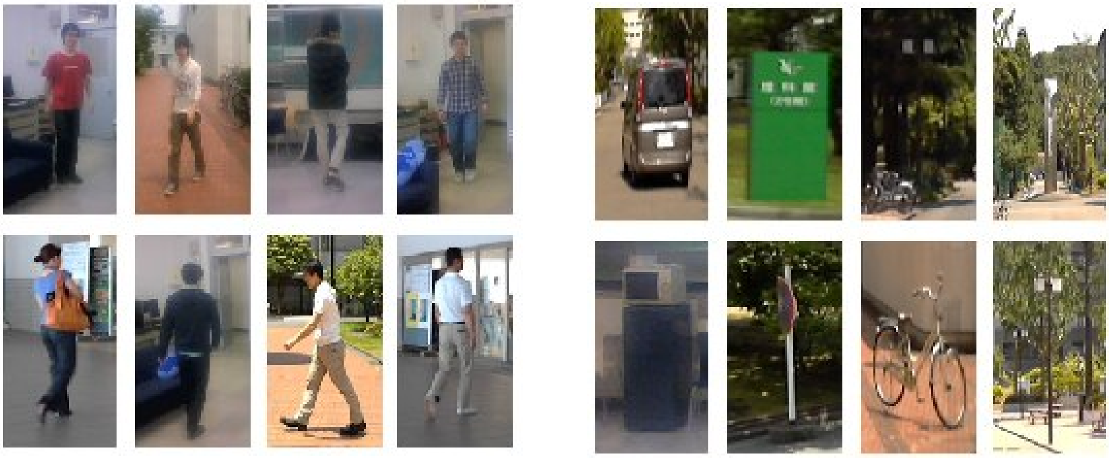
$$d = \frac{1}{fh}(y - \frac{H}{2})$$

In the current algorithm the camera is set at $h = 1.4$m and as an indication to notify of the collision risk, pedestrian are grouped by their distance into three range levels: near-range (4m or less), medium-range (between 4m and 8m), and far-range (8m or more). Figure.5 shows this distant relationship superimposed to the camera image.

3 Experiment and results

3.1 Experiment:

To evaluate detection accuracy, we prepared a dataset. Figure. 6 shows some examples of this dataset images. These images are clipped out the frame image shot with a camera held by an experimenter. The complete dataset consists of 1000 images of a pedestrian over a background and 1000 images of backgrounds without pedestrian.



(a) Positive images

(b) Negative images

Figure 6: Some examples of positive and negative test images.

Using this dataset we compare two methods of detection, conventional HOG and our multi parts detector. Table **??** shows the accuracy result.

Table 1: Accuracy rate for HOG and our multiple model method.

	precision	miss	false
HOG	91.90	9.10	25.21
Multiple Model	88.70	11.30	6.29

Even if the conventional HOG detector has a better detection rate than our method, the difference in accuracy is just 3.2%. On the contrary, compared to the conventional HOG method, our method have a 18.92% better false detection performance. Since our method have a flexible positioning of each part the decrease in detection rate is low. Instead the false detection result, the conventional HOG method is more likely to falsely detect objects with a complex texture. Overall, our method makes a slight sacrifice in detection rate, to obtain an improved lower false detection rate.

3.2 Pedestrian detection:

The process of detecting the pedestrian from an image consists of a detection window that scans the image over and over. The scale and position of this window are changed. In this way it is possible to detect humans whose size is diverse. These images are taken in multiple locations, with a resolution of 640*480 pixels. They include images in which people and cars are passing through streets, tree leaves are flickering, and conditions are varying.

In Figure.7 we show a comparative result of pedestrian detection examples. The conventional method shows false detections due to complex background. Because our method operates with three phase detection, it works better and is more robust than the conventional HOG method. Our method is able to detect pedestrian accurately in these non-ideal environments. Figure. 8 shows the result of estimating the distance of each person. We see that the position of each pedestrian can be estimated with good approximation. Based on this result, our system can compute the collision risk and notify the user of approaching pedestrian.

3.3 Comparison of HOG, SIFT and PCA-SIFT

We realized that HOG method is more suitable than other methods because of its geometrical and optical transfer invariance, and because it shows low computational complexity and high velocity. Instead, the SIFT method it is heavier and implies a larger amount of calculation. On the other hand, PCA-SIFT have better speed than SIFT, but multi-dimensional data are filtered out and this is not suitable for the recognition of objects with minor differences. [8]

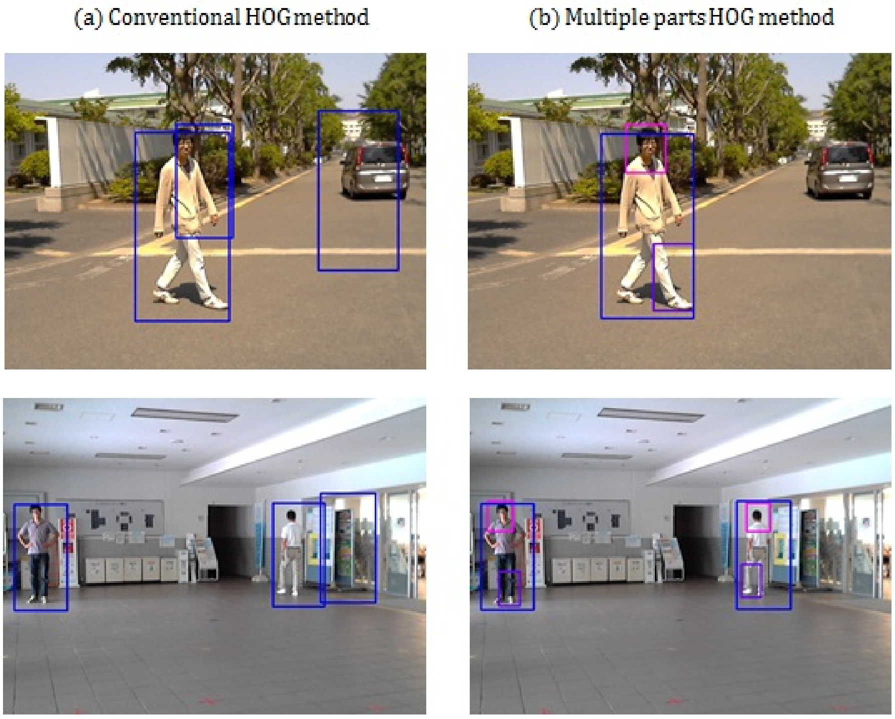


Figure 7: Examples of detection images over complex and moving backgrounds. On the left are shown the results with a conventional HOG approach: moving car and other objects are falsely detected. On the right our multiple part HOG approach solves this problem when tested on the same video frames.

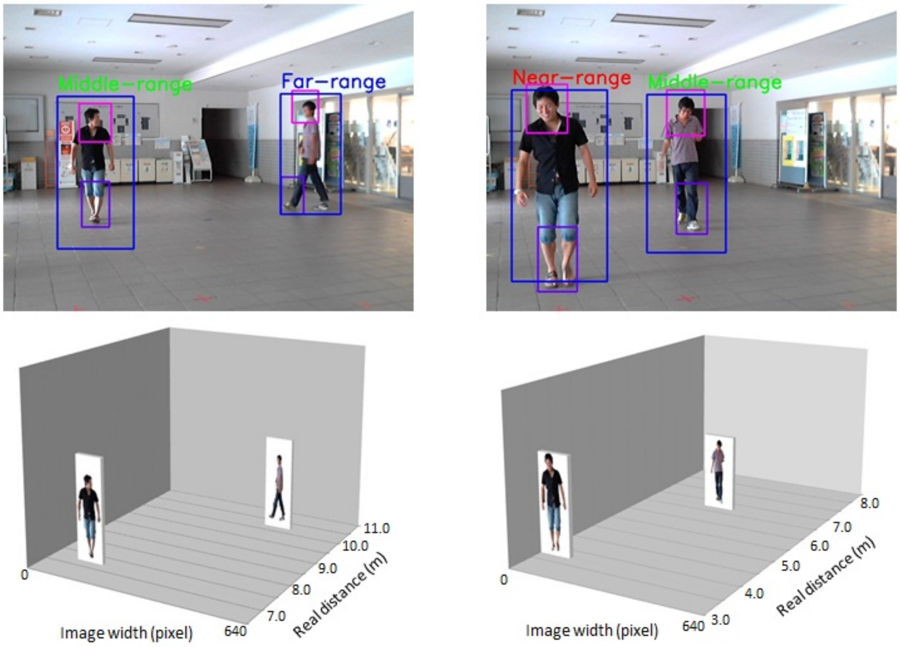


Figure 8: Examples of detection images and the corresponding distance estimation range.

4 Conclusions and perspective

In this study, we proposed a pedestrian detection system which can estimate the distance to targets with a single camera by the use of a multiple-parts HOG model. The result of test experiments shows that our detection method has a 18.92% improvement

Table 2: A rough comparison between HOG, SIFT and PCA-SIFT methods for pedestrian recognition.

	Speed	Scale	Rotation
HOG	best	best	common
SIFT	common	good	good
PCA-SIFT	good	common	best

in false detection rate against the conventional HOG method. The system was also able to detect pedestrian in complicated backgrounds moving environments. The system was able to estimate the distance of pedestrian using the single camera. It is possible to derive collision risk from the estimated distance. Our system can run with only a camera and a computer, it does not require ultrasonic sensors and multi camera systems. Therefore it can be implemented on simple and convenient devices (e.g. smartphones and tablet computers). We are planning to improve the system developed up to now to work as a support application in real environments. Especially there we want to focus on two parameters, real time processing and pedestrian tracking. Our system takes more CPU time than the simple HOG method. We have to optimize processing time and simultaneously realize the tracking of the pedestrian to support the user to decide the direction of avoidance. We plan to use time-series filtering, Kalman filter [2] or Particle filter [1]. By applying a time-series filter algorithm, faster process time is expected because of the determinate scan area. The new filtering will also enable to improve the accuracy because of the tracking information relative to the target person.

References

[1] P. Brasnett, L. Mihaylova, D. Bull, and N. Canagarajah. Sequential monte carlo tracking by fusing multiple cues in video sequences. *Image vision Computing*, 25(8):1217–1227, 2007.

[2] E. Cuevas, D. Zaldiver, and R. Rojas. Kalman filter for vision tracking. *Fachbereich Mathematik und Informatic, Technical Report B, Freie Universitat Berlin*, 2005.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, pages 886–893, 2005.

[4] A. Bensrhair F. Suard, A. Rakotomamonjy and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. *Proc. IEEE Conf. Intell. Vehicles*, pages 206–212, 2006.

[5] P. Felzenszwalb, R. Girschick, and D. McAllester. Cascade object detection with deformable part models. *CVPR*, pages 1–8, 2010.

[6] Ryan Cekander Harpreet S. Sawhney Feng Han, Ying Shan and Rakesh Kumar. A two-stage approach to people and vehicle detection with hog-based svm. *Performance Metrics for Intelligent Systems Workshop in conjunction with the IEEE Safety, Security, and Rescue Robotics Conference*, pages 134–136, 2006.

[7] A. Helal, S. Moore, and B. Ramachandran. Drishti: An integrated navigation system for the visually impaired and disabled. *Fifth International Symposium on Wearable Computers (ISWC01)*, pages 149–156, 2001.

[8] Luo Juan and Oubong Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing*, 3:143–152, 2009.

[9] Sebastian Montabone and Alvaro Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *IMAGE AND VISION COMPUTING*, 28(3):391–402, MAR 2010.

[10] Mei-Chen Yeh Qiang Zhu, Shai Avidan and Kwang-Ting Cheng. Fast human detection using a cascade of histograms of oriented gradients. *MITSUBISHI ELECTRIC RESEARCH LABORATORIES*, pages 1491–1498, 2006.

[11] Xunshi Yan and Yupin Luo. Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier. *NEUROCOMPUTING*, 87:51–61, JUN 15 2012.

[12] Tianzhu Zhang, Si Liu, Changsheng Xu, and Hanqing Lu. Boosted multi-class semi-supervised learning for human action recognition. *PATTERN RECOGNITION*, 44(10-11, SI):2334–2342, OCT-NOV 2011.