A COMPARISON OF PEOPLE COUNTING TECHNIQUES VIA VIDEO SCENE ANALYSIS

Poo Kuan Hoong, Ian K.T. Tan, and Chai Kai Weng
Faculty of Computing and Informatics, Multimedia University, Malaysia
E-Mail: khpoo@mmu.edu.my

ABSTRACT

Real-time human detection and tracking from video surveillance footages is one of the most active research areas in computer vision and pattern recognition. This is due to the widespread application from being able to do it well. One such application is the counting of people, or density estimation, where the two key components are human detection and tracking. Traditional methods such as the usage of sensors are not suitable as they are not easily integrated with current video surveillance systems. As video surveillance systems are currently prevalent in most places, using vision based people counting techniques will be the logical approach. In this paper, we compared the two commonly used techniques which are Cascade Classifier and Histograms of Gradients (HOG) for human detection. We evaluated and compared these two techniques with three different video datasets with three different setting characteristics. From our experiment results, both Cascade Classifier and HOG techniques can be used for people counting to achieve moderate accuracy results.

Keywords: human tracking, people counting, Histograms of Oriented Gradients (HOG), Cascade Classifier, OpenCV.

INTRODUCTION

The automatic objects recognition in images is one of the most difficult problems in computer vision. Correspondingly, real-time human detection and tracking video is one of the most active areas in computer vision due to its widespread applications. There are many applications that are able to detect and track humans that are useful for applications such as video surveillance, smart vehicle, human computer interaction, and content based indexing. At the same time, being able to detect human beings accurately is crucial to applications such as the visual surveillance system where it can be applied to diverse application areas; including gait classification in human, people and gender identification, and abnormal event detection such as fall detection for the elderly.

Generally, in order to detect and track human as an object, there are several crucial steps. Firstly, the ability to detect the objects of interest by finding the image regions that will correspond to the objects. Secondly, the ability to track the identified objects across different video frames is required. Additionally, it is also important to be able to continuously track the same object across time. Other challenges faced are the variations in appearance, illumination, and shadows. Due to the complexity of the human movement and the non-rigidity of the body, it is difficult to detect, recognize and track humans in videos. Various methods have been proposed in human detection [1]. Most of them relied on detection by changes caused in subsequent image frames due to human motion.

In summary, people detection can be divided into three classes:

1) **Background modeling:** The need to learn background model over time and then subtract the background model from the video frame to obtain the target foreground object [2],
2) **Human body part detection:** Human body part such as head, upper body, arms and legs are detected in video frames [3], and
3) **Human body shape detection:** Approximation of body parts into some special shapes such as rectangles and ellipses are detected in video frames [4].

People counting and density estimation have been an important component in video surveillance systems. It has wide applications such as estimation of people flow in public areas such as train stations, airports, and even commercial centers such as shopping malls.

Traditional methods such as the usage of manual counting, for example using a Tally Counter as illustrated in Figure-1 or infrared sensor (as in Figure-2) counting device have drawbacks. Manually counting suffers from errors by human operators, while infrared sensor methods can only capture a very restricted area and hence are inaccurate. Accuracy for infrared sensors can be improved through deploying multiple devices but will then suffer from duplicate counting.

Figure-1. Manual Tally Counter.

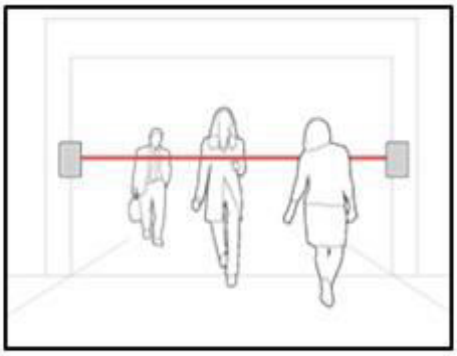

Figure-2. Infrared sensor beam (source: http://www.easpartners.com/useruploads/images/peoplecounters.jpg).

With video surveillance systems that are prevalent everywhere, one of the approaches is to directly detect and count people from an image or videos. There has been previous works done in the area of human detection. Techniques that generally perform well in scenes that are less crowded have been proposed by several groups [5-10]. Among all the techniques used, the two commonly used techniques are 1) Cascade Classifier [11] and 2) Histograms of Oriented Gradients (HOG) [12]. In this paper, we evaluate these two techniques by comparing the people detection results vis-a-vis three video datasets which have three different scene characteristics.

The rest of this paper is organized as follows: Section II introduces some related works; Section III provides an outline of our approach in this paper; Section IV presents the results and discussion, and finally we conclude our findings in Section V.

RELATED WORKS

There are many people counting techniques available. Traditional counting systems are generally based on installing physical hardware such as infrared sensors or pressure sensors that are placed near to the entrance or exit. Fang Zhu *et al.* [6] proposed a new method for infrared people counting based on the characteristics of the time continuous data collected by infrared sensors. These solutions are generally low cost but not easy to integrate with video surveillance system.

Other popular methods that can easily integrate with video surveillance systems are commonly vision-based. Huang *et al.* [13] proposed a stereo-based head detection method for human detection in a crowded scene. They applied 3D scale-filtering to extract the likelihood evidence of heads from the stereo image. They managed to obtain good results of detecting human heads in crowds have been obtained from the experiments on real scene. Other researchers such Zhao *et al.* [14] proposed people counting method based on face tracking combining a new scale invariant Kalman filter with kernel based tracking algorithm. Their proposed method demonstrated a good performance where the people counting accuracy was reported to be approximately 93%.

In this paper, we applied the two commonly used techniques for people detection in video: 1) Cascade Classifier and 2) Histograms of Gradients (HOG).

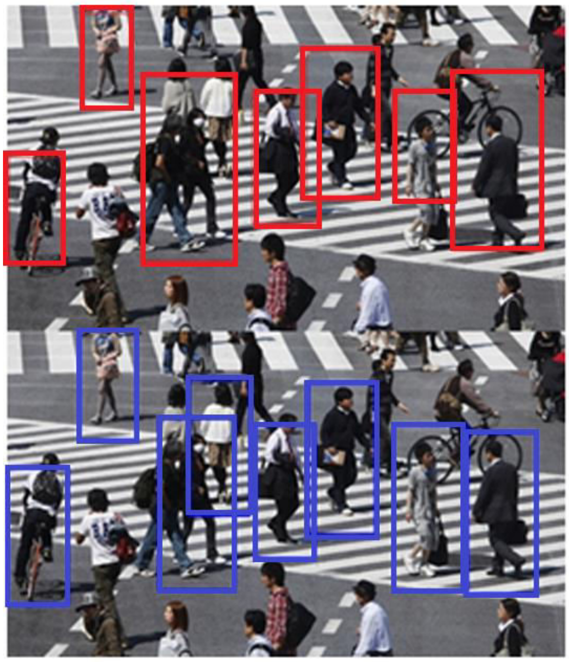

Figure-3. People Detected by HOG (Top) and Cascade Classifier (Bottom).

Cascade Classifier

Object Detection using Haar feature-based cascade classifiers is an effective object detection method proposed by Paul Viola and Michael Jones [11]. In a nutshell, cascade classifier is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. Information collected from the output from a given classifier will be used as additional information for the classifier in the cascade which is then used to detect objects in other images.

Histograms of Gradients (HOG)

The HOG descriptor proposed by Dalal *et al.* [12] is one of the most successful human detectors in static images. Their algorithm uses a dense grid of HOG to represent a detection window. It captures the edge direction or the distribution of local intensity gradients of objects. With the development of the hardware, the HOG descriptor has been able to achieve the real-time effect on the human detection. HOG method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid.

Figure-3 provides a visual illustration of people detected by Cascade Classifiers and HOG techniques.

MATERIALS AND METHODS

In this paper, we compare the accuracy of people counting using Cascade Classifier and HOG techniques. Our experiments consist of two main components: 1) people detection: people in the image or video frame will be detected and 2) people counting: people counting will be carried out in the image or video frame, by counting the boxes that are drawn by people detection process. Figure-4 gives an overview of our experiment process flow.

Our experiments are conducted in an emulated environment, where a virtual machine is setup to run the algorithms and simulate the result. The followings are the experiment setup:-

Hardware

- 512 MegaBytes of memory
- Dual-Core processor
- 15 GB of HDD space (used)

Software

- Linux (Lubuntu 32-bit)
- OpenCV 2.4.9
- GNU C/C++ Compiler

Open Source Computer Vision (OpenCV)

OpenCV is an open source computer vision library programming functions that is written in C and C++. It is originally developed by Intel research center in Nizhny Novgorod, Russia and it contains over 500 functions associated with vision, and also contains a general-purpose machine learning library. In this paper, OpenCV was chosen as it is open-source and free for academic and commercial use. Besides that, it is widely used and has a well and proper documentation online [15].

1) **Cascade Classifier:** Cascade Classifier is used to detect human in a video frame by detecting human faces where the face aspect ratio remains approximately fixed. During the initial stage, the algorithm needs large number of positive images (images filled with faces) and negative images (images without any faces) in order to train the classifier. OpenCV comes with an utility for creating training samples from one image, which can automatically apply alteration in background, lightning and rotation for the output images. This approach is suitable for cases where the object does not vary in appearance, for example the training on stop signs, cars, and logos. Figure-5 shows the process flow for Cascade Classifier.

The documentation for using the createsamples utility could be found in the OpenCV install directory under the OpenCV/apps/HaarTraining/doc. It uses the CascadeClassifier class to detect objects in a video stream. As for face detection, OpenCV already contains many pre-trained classifiers for face, eyes, smile and other features. Those XML files are stored in opencv/data/haarcascades/ folder. [16].

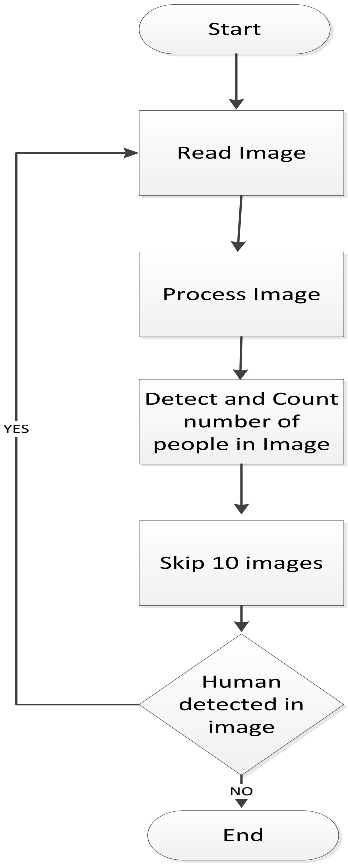

Figure-4. A flow chart illustrating the process flow for this study.

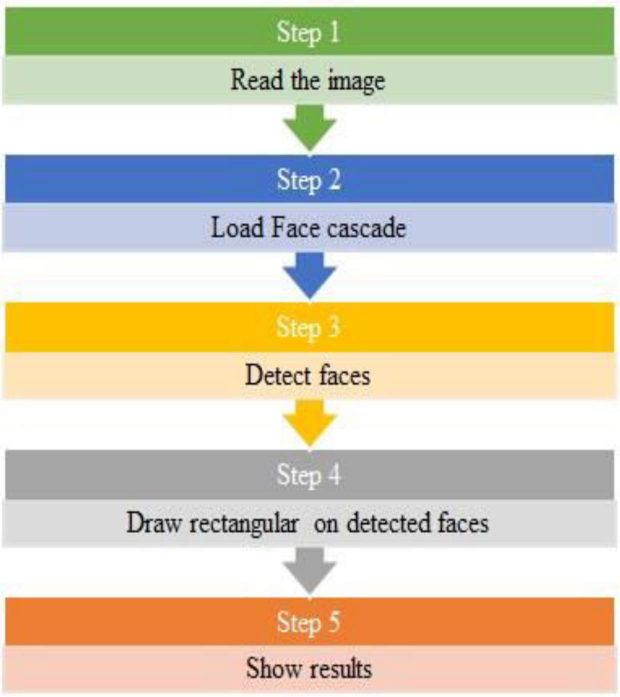

Figure-5. The process flow for Cascade Classifier.

2) **HOG:** HOG is a type of "feature descriptor" where the intent of a feature descriptor is to generalize the object in such a way that the same object (in this case a person) produces as close as possible to the same feature descriptor when viewed under different conditions. Based on these descriptors, an object can be classified as people in an image or video frame. In OpenCV, it has peopledetect.cpp that come with Dalal (Dalal & Triggs, 2005) default person detector. With the HOGDescriptor::compute in OpenCV, one can compute the HOG of the given image. It is required to train using the Support Vector Machine (SVM) machine learning algorithm with Train/negative and Train/Positive image set. After that, test the SVM model with the Test/negative. The trained SVM model is a file containing support vectors. With the support vectors, one can use them to predict people/non-people classification. It is noted that OpenCV can only use one vector to detect people [17]. Figure-6 shows the process flow for people detection using HOG.

Datasets

For this study, there are three different datasets used (as shown in Figure-7):

1. Town Center Dataset,
2. Sunny Day Dataset, and
3. Mall Dataset.

1) Town Center Dataset:

In Town Center Dataset [18], the camera was position at the top with 125 degrees facing downwards. Recording was done outdoor with normal lighting conditions. In the recording, there were a lot of people can be seen entering and exiting from frame to frame.

2) Sunny Day Dataset:

In Sunny Day Dataset [19], the camera was positioned at 90 degrees. Recording was done outdoor with a bright sunny day lighting conditions. In the recording, it can be observed that many people passing and entering the area from frame to frame. This dataset will be used to evaluate on the two techniques where people movement is high in bright lighting condition.

3) Mall Dataset:

The Mall Dataset [20] was collected from a publicly accessible video surveillance camera. The crowd density or the number of people in this dataset is quite high. The camera is positioned on top, facing 125 degrees downwards from top. The lightning of the environment is indoor condition and it can be seen that there are a lot people going in and out from frame to frame.

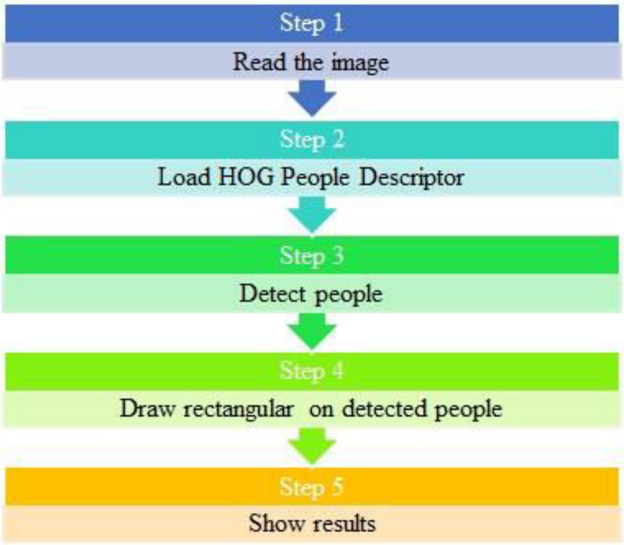

Figure-6. The process flow for people detection using HOG.

Experiments Setup

In our experiments, the video frames from three datasets are converted to static images where the images in the datasets are named sequentially with ascending order so that the algorithm can read image to image for human detection. After reading the images, the algorithm will process the images using Cascade Classifier and HOG techniques. Subsequently, both algorithms will record down the people counted using the Cascade Classifier and HOG techniques into a text file. The name of the text file will be named according to the time the file is created.

Figure-7. Screenshots from the three datasets: Sunny Day, Town Center and Mall datasets.

As shown in Figure-4, our method will check whether there are any remaining images in the sequence. If there are, the program will loop itself and process the next image. It is noted that OpenCV has built-in an image sequence reading that will automatically set the next image to be read. At the end of the program if there are no more images, the program will exit.

The followings are the assumptions for our experiments:

Ground truth will be based on people counted visually (manually) for all frames.

A person that is detected partially or more than half visible will be counted as one, as the person exits or enters the frame.

When two or more people are standing together (in very close proximity), they will be grouped in a box, hence it is counted as one person visually.

For larger image datasets that is created by high frames per second (fps) videos, one in every 10 images will be used. The rationale is that there will not be significant changes between the video captured frames.

The first 30 images from the resultant datasets will be used in this study.

The experiment will start from the second frame in order to give time for the algorithm to stabilize.

Mean Squared Error (MSE)

Mean Squared Error (MSE) is used to measure the average squared errors between the actual number of people ($\hat{\theta}$) as compared to number of people detected (θ) by Cascade Classifier and HOG techniques.

$$\mathrm{MSE}(\hat{\theta}) = \frac{\sum[(\hat{\theta} - \theta)^2]}{\sum(\mathrm{data})}$$

Accuracy Measurements

In term of accuracy measurements, we measure the accuracy of people counted for each frame image for both techniques by calculating the Percentage of Error in comparison to the people detected visually by human.

% of Error = Actual - (Detected - False Detection)

The accuracy of people counting can be calculated using the following formula:

Accuracy = 1-Error

RESULTS AND DISCUSSIONS

Town Center Dataset

As shown in Figure-8, both of the techniques were evaluated using the Town Center Dataset. The evaluation was carried out for 30 frame images with 10 frame images skipped after one frame image was processed. Results obtained shown that both techniques achieved similar performance in terms of the percentage of errors. In comparison with the actual number of people in frame image, the highest number of person seen in the frame image is 26 people at frame image 92. Both Cascade Classifier and HOG counted 15 and 18 person respectively.

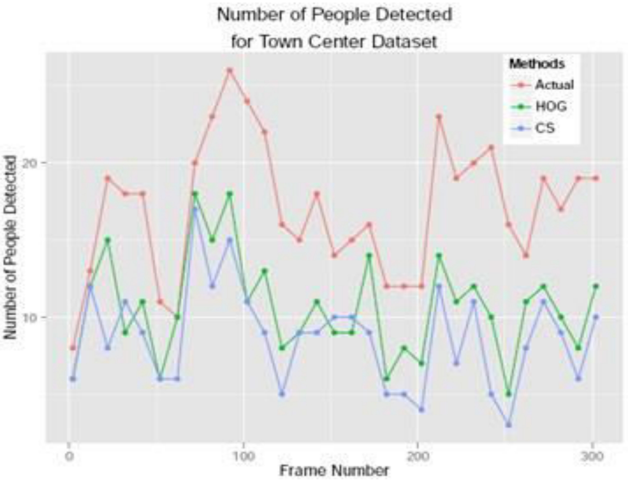

Figure-8. Number People Detected for Town Center Dataset.

On the other hand, the lowest number of people seen in one frame image is 8 people at frame image 2. Both Cascade Classifier and HOG only managed to count 6 people. Results from both techniques for the Town Center Dataset show that both of techniques detected lower number of people in comparison to actual number of people in the frame images. Table-1 shows that the average percentage of error for both Cascade Classifier and HOG techniques, which are 52.94% and 47.28%.

Mall Dataset

Figure-9 shows the evaluation results obtained using the two techniques for the Mall Dataset. The results show that the number of people detected for both techniques are lower than the actual counted number of people detected manually using human visuals. The highest actual number of people detected visually was 35 people at frame image 12. Both Cascade Classifier and HOG detected 17 and 11 number of people respectively. Meanwhile, for frame image with the lowest number of people detected visually was 10 people at frame image 302. Both Cascade Classifier and HOG only managed to detect 5 and 4 people respectively. Table-1 shows the average percentage of error for Cascade Classifier and HOG are 49.17% and 49.59%, where the average percentage errors are about the same, around 50%.

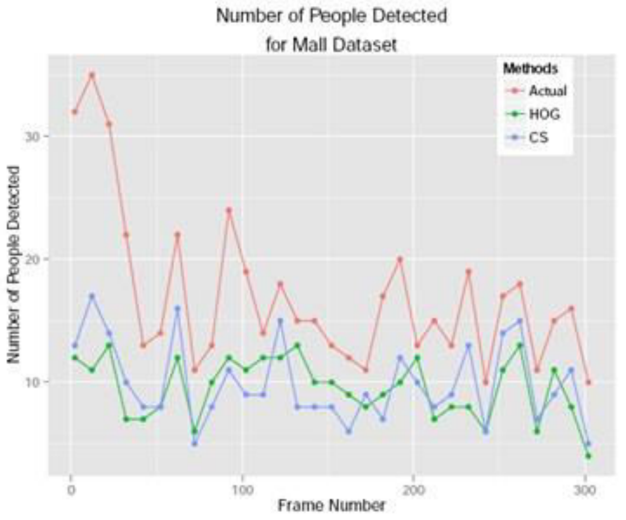

Figure-9. Number People Detected for Mall Dataset

Sunny Day Dataset

Lastly for Sunny Day Dataset, results are as shown in Figure-10 where the highest number of people detected visually is at frame image 282 with 10 people. On the other hand, the lowest number of people detected visually is at frame image 262 with 3 people. Both Cascade Classifier and HOG managed to detect 7 and 6 people respectively. Meanwhile, for frame image number 262, both Cascade Classifier and HOG managed to detect 5 and 4 number of people respectively. Table-1 shows the average percentage error for Cascade Classifier and HOG are about the same, 32.29% and 31.87% respectively.

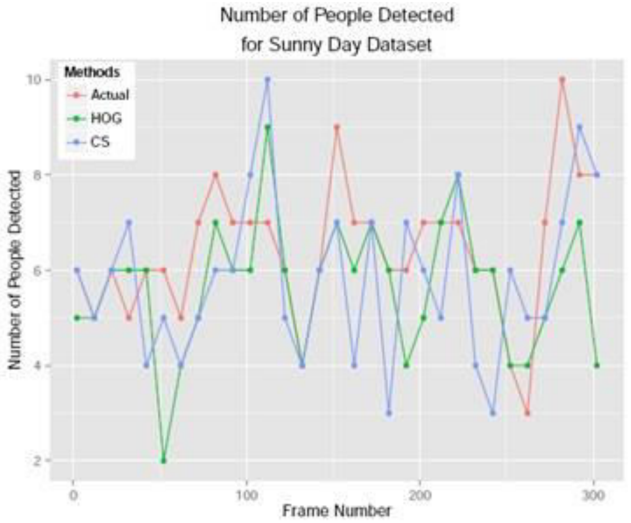

Figure-10. Number People Detected for Sunny Dataset.

Both Cascade Classifier and HOG techniques perform reasonably well for Mall and Town Center Datasets in comparison to Sunny Day Dataset. This is mainly due to both techniques are able to detect heads and faces that are clearly visible in Mall and Town Center Datasets. While in the Sunny Day Dataset, people are moving away from the camera with their backs facing the camera which lead to both techniques unable to detect people in the video frames hence lower people count.

Table-1.MSE & Average Percentage Error for both Cascade Classifier and HOG.

Dataset	Cascade Classifier		HOG	
	MSE	% of error	MSE	% of error
Mall Dataset	69.42	49.17%	84.77	49.59%
Town Center Dataset	82.61	52.94%	50.87	47.28%
Sunny Day Dataset	3.03	31.87%	2.65	31.87%

CONCLUSIONS

Based on our experiment evaluation results for three video datasets, we can conclude that for both of Cascade Classifier and HOG techniques for people detection and people counting, they achieved similar results in terms of accuracy. The average accuracy for three datasets for Cascade Classifier is 56.87% and 59.04% for HOG. It can be noted that there is a small differences of 3% for both techniques. The average percentages of error for both Cascade Classifier and HOG for the three datasets are 46.46% and 44.29% respectively. In summary, based on our experiment results, both Cascade Classifier and HOG techniques can be used for people counting to achieve moderate accuracy results. However, in order to achieve higher accuracy result, further improvements are needed.

REFERENCES

[1] P. Dollar, C. Wojek, B. Schiele and P. Perona. 2012. Pedestrian Detection: An Evaluation of the State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 34, No.4, pp.743–761.

[2] B. Wu, R. Nevatia and Y. Li. 2008. Segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008). pp. 1–8.

[3] M. Andriluka, S. Roth and B. Schiele. 2008. People-tracking-by-detection and people-detection-by-tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008). pp. 1-8.

[4] T. Zhao, R. Nevatia and B. Wu. 2008. Segmentation and Tracking of Multiple Humans in Crowded Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 30, No.7, pp. 1198–1211.

[5] D. Conte, P. Foggia, G. Percannella and M. Vento. 2010. A Method Based on the Indirect Approach for Counting People in Crowded Scenes. In: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp.111-118.

[6] F. Zhu, X. Yang, J. Gu and R. Yang. 2009. A New Method for People-Counting Based on Support Vector Machine. In: Second International Conference on Intelligent Networks and Intelligent Systems. pp. 342–345.

[7] H. Yang, H. Su, S. Zheng, S. Wei and Y. Fan. 2011. The large-scale crowd density estimation based on sparse spatiotemporal local binary pattern. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6.

[8] R. Ma, L. Li, W. Huang and Q. Tian. 2004. On pixel count based crowd density estimation for visual surveillance. In: IEEE Conference on Cybernetics and Intelligent Systems. Vol. 1, pp.170–173.

[9] S.-F. Lin, J.-Y. Chen and H.-X. Chao. 2001. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans. Vol. 31, No.6, pp. 645–654.

[10] Z. Cai, Z. L. Yu, H. Liu and K. Zhang. 2014. Counting people in crowded scenes by video analyzing. In: IEEE 9th Conference on in Industrial Electronics and Applications (ICIEA). pp. 1841–1845.

[11] P. Viola and M. Jones. 2001. Robust Real-time Object Detection. International Journal of Computer Vision 4. pp 51-52.

[12] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 886–893.

[13] X. Huang, L. Li and T. Sim. 2004. Stereo-based human head detection from crowd scenes. International Conference on Image Processing (ICIP '04). Vol. 2, pp. 1353–1356.

[14] X. Zhao, E. Delleandrea and L. Chen. 2009. A People Counting System Based on Face Detection and Tracking in a Video. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09). pp. 67–72.

[15] OpenCV [Wiki] Available: http://code.opencv.org

[16] OpenCV Cascade Classifier [Documentation] Available: http://docs.opencv.org/modules/objdetect/doc/cascade_classification.html

[17] OpenCV SVM HOG [Documentation] Available: http://docs.opencv.org/modules/gpu/doc/object_detection.html

[18] Town Center Dataset [Online] Available: http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html

[19] Sunny Day Dataset [Online] Available: https://data.vision.ee.ethz.ch/cvl/aess/dataset/

[20] Mall Dataset [Online] Available: http://www.eecs.qmul.ac.uk/ccloy/downloads_mall_dataset.html