15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

A PTZ Camera Based People-Occupancy Estimation System (PCBPOES)

Arun Kumar Chandran, Aravind Subramaniam, Wai Choong Wong
Interactive and Digital Media Institute, National University of Singapore 119613
idmakc@nus.edu.sg, arasubu@gmail.com, elewwcl@nus.edu.sg

Junjing Yang, Karn Ashokkumar Chaturvedi
Department of Building, National University of Singapore 117566
bdgyj@nus.edu.sg, bdgkac@nus.edu.sg

Abstract

A single PTZ Camera Based People-Occupancy Estimation System (PCBPOES) is proposed to estimate the number of people occupying a region of interest with acceptable accuracy. The PTZ camera aids this objective by efficiently monitoring a wide area by dividing it into zones, capturing high resolution zone images aided by the optical zoom for detecting human head patterns. A seminar room is used as a test location for evaluating this system. Cascade trained support vector machine models are used to detect head patterns in the images. Features such as HAAR, HOG were explored for modeling the head patterns. We adopt a zone-based detection and ensemble approach to combine the zone detection results. A temporal probabilistic selection approach improves the head detection accuracy. Experiments prove that the system has high detection accuracy achieving average precision scores of 0.92 and recall scores of 0.82. A bio-inspired retina model is adopted to detect heads in low illumination. The system is currently used to estimate people-occupancy in the test location.

1 Introduction and Related works

Vision based techniques are non-obtrusive methods to sense objects of interest, such as humans. Researchers have used such techniques to build numerous applications which are not limited to monitoring applications, operation planning, crowd counting and so on. Traditionally, human observers were used to sense objects of interest. They do a post-scene analysis of the record to infer outcomes. This is a time consuming process and prone to human errors. With vision based techniques, such application and decision making can be performed with less human effort with better accuracy in a very short time. Such intelligent applications could detect abnormal activities [1], pedestrian bottlenecks [2] , predict pedestrian motion [3], pedestrian foot count [4] and many more [5]. Retail planners are interested in the popularity of shops and ways to better position shops to increase the foot count [6]. Motion based techniques are generally used to sense objects of interest in these applications as they tend to concentrate on the moving pedestrians in locations of transit and passages. Very few works are focused on detecting static people. A lecture room is a good example where people are static. The major environmental challenges faced in such locations are different from the other locations. Low illumination, a very wide viewing area and lack of prominent human motion are the major challenges in these locations.

In this paper, we propose a system to address these challenges and estimate the number of people occupying a region of interest. The system uses a cascade detector [7] of support vector machines (SVM) [8] to detect human heads. A single PTZ camera is used to cover the wide area of the test location. It is a cost effective solution compared to a multiple camera solution. The PTZ camera is automated to capture six different images representing six different zones of the test location. A probabilistic temporal selection approach is adopted to improve the detection accuracy followed by region of interest based merging of the zone detection results.

2 PTZ Camera Based People-Occupancy Estimation System

The stages in estimating the people-occupancy of a region are explained using figure 1. The individual stages are explained below.

PTZ Automation

We selected a seminar room as the test location. The location, shown in figure 2, has a capacity of 220 seats, consisting of 11 rows and having two walking passages. A PTZ camera is used to cover the entire test location instead of multiple static cameras. Figure 2 shows snapshots of the six zones captured by six different pan, tilt and zoom settings of the PTZ camera. The camera PTZ routine is described in the below pseudocode.

PSEUDOCODE: Camera PTZ Routine

1) $Checktime()$
2) $If\, time == 09.00, capture()$
$Else If\, time == 19.00, kill\, capture()$
$Else\, Idle()$
3) $Capture()$
$For(i = A : F)$ where i = A to F
3.1) Set $camera(P_{zone_i}, T_{zone_i}, Z_{zone_i})$
$For(burst = 1 : 3)$ Capture 3 images one second apart
3.2) $Image\, capture()$
3.3) $Delay(100ms)$ A delay for the camera to stabilize in the new P, T and Z values.

Detector

Captured images are scanned for head patterns by passing them through a cascade head detector of support vector machines (SVM). In training, Histogram of Oriented Gradients (HOG) [9] feature space is learned in the cascade detector. Human head patterns form the positive samples, while the other image regions form the negative samples. The training process involves an

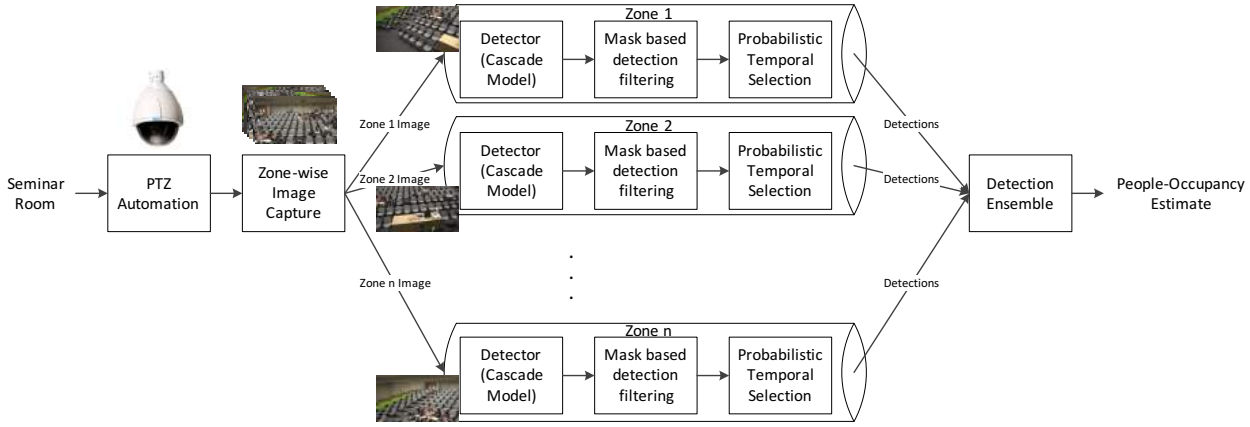

Figure 1. PTZ Camera Based People-Occupancy Estimation System flow.

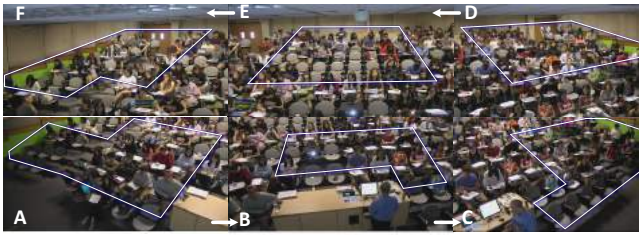

Figure 2. Different zones captured with the PTZ automation shown with the sequence of image capture (white arrows) and the masks (white polygons) used for the individual zones.

Figure 3. Samples from test location.

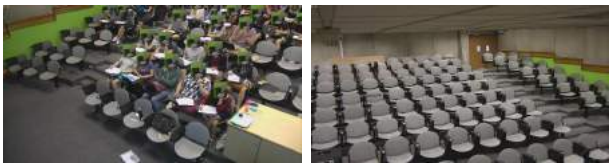

Figure 4. Negative sample collection. Blank images and images with head regions covered are used.

adaptive boosting routine (AdaBoost) [14],[15], where a false alarm termination threshold is used to terminate every learning stage. A semi-supervised learning framework is adopted from [16], providing a generalized approach for object detection. It speeds up the samples collection process and also improves the SVM model quality over time.

The Istituto Italiano di Tecnologia (IIT) [17] and coffebreak head samples data set (HOCoffee) [18] are used to build the SVM model. These data sets contain samples with frontal and side views but are recorded in a controlled lab environment without much occlusions. In order to consider real-world conditions, which is an uncontrolled environment, a data set of head samples is created with diverse poses and perspectives - NUS lecture data set. Such a diverse sample collection is required to detect heads from different zones which have different camera perspectives (especially in zone A and zone C, where the angle of observation is quite steep when compared to the rest). Samples are collected from recorded video clips of the test location. Figure 3 shows samples extracted from the test location. Negative samples are recorded from image frames that have no humans present. Image frames with people are used after covering the head regions with other environment objects like chairs. Figure 4 shows some of the image frames used to generate the negative samples.

Automatic retina model activation

Head detection is affected at low illuminations because of insufficient head features. Low illumination occurs due to the various activities in the region of interest. For example, lecturers switch off the lights in the front side during a lecture for better visibility of the projected slide. We adopt a bio-inspired method to improve the local luminance of the image frames in such situations - retina model [10], [11]. The model mimics a human eyes parvo and magno cell characteristics to analyze the image frames. The parvo cell channel has the ability to adjust the sensitivity with respect to the luminance of its neighborhood cells [12]. Figures 5 and 6 shows the improvement in luminance achieved using the parvo cell channel. This image processing is activated when a low illumination condition is automatically sensed by measuring the average image pixel intensity and comparing it against the average pixel intensity of a ''lights on' situation.

Mask based detection filtering

The detected head regions are filtered using this method. Pre-defined masks are created for each zone to remove the effect of overlapping regions between zones. This filter eliminates false positives which do not hap-

Figure 5. Enhancement in head samples with retina pravo channel luminance adjustment.

Figure 6. Enhancement in detection results with retina pravo channel luminance adjustment.

pen in the seating area. Figure 2 shows the created masks, represented by the blue polygons. Detections from the overlapping regions in the zones are filtered out using a blob filtering technique [13] based on the location of the lower edge of the bounding box in the mask region.

Probabilistic temporal selection

This method is employed to reduce the number of sporadic false positives. Here, the detection results are checked over a temporal window of image frames. A detection d_{i,f_j} in frame f_j is selected according to equation 1.

$$d_{i,f_j} \epsilon d_s \Leftarrow P_o(d_i) > 0.6 \quad (1)$$

where, $P_o(d_i)$ is the probability of occurrence of d_i measured across a temporal window of image frames, to select d_i to a set d_s. We select a temporal window of 4 consecutive image frames, sampled at 1 second intervals. The detection d_{i,f_j} is associated with a detection $d_{i,f_{j+1}}$ in a subsequent frame f_{j+1}, if their centroids are not separated by more than D_{thresh} distance. The distance between their centroids are calculated using equation 2.

$$\sqrt{(x_{i,f_j} - x_{i,f_{j+1}})^2 + (y_{i,f_j} - y_{i,f_{j+1}})^2} < D_{thresh} \quad (2)$$

Detection Ensemble

Detection counts generated from all the zones are combined together to generate a single people-occupancy estimate for the test location. The count estimate is stored in a database for generating a history of people-occupancy estimates. An sqlite [19] database is used to store the head count estimate.

3 Evaluation

This section explains the evaluation performed on the developed system. Precision, recall and F scores [20] are recorded to show the performance of the PCBPOES system. Detection accuracy in terms of the number of heads detected in the data sets are also presented.

Detection accuracy on data sets

The IIT [17] dataset is used in this validation to decide which feature space to use for the proposed system. HOG [9] and HAAR [21] features are tested. From table 1, it is clear that the HOG model outperforms the HAAR model. Hence, we select the HOG feature space.

The system's performance with a model built using HOG is tabulated. From tables 2 and 3, it is clear that the system has high detection accuracies.

Precision and Recall scores

Table 1. Detection accuracy with Histogram of Orient Gradient (HOG) and HAAR features.

IIT Head Pose Data set	HOG	HAAR
Total Positive samples	6000	6000
Number of detections	5610	4391
Total Negative samples	6000	6000
False positives	1090	720
Precision	0.8373	0.8590
Recall	0.9350	0.7310

Table 2. Pose-wise detection accuracy on IIT Headpose and HOCCoffee pose-wise data sets.

	No. of Images	Detected	Rejected
IIT Front Left	2000	1782	218
IIT Front	2000	1931	69
IIT Front Right	2000	1741	259
IIT Left	2000	1816	184
IIT Right	2000	1788	212
HOC Front Left	847	825	22
HOC Front	382	379	3
HOC Front Right	687	671	16
HOC Left	1627	1538	89
HOC Right	1974	1907	67

Precision determines the capability of the technique in rejecting pedestrian-like confusers (false positives). While, recall determines the sensitivity of the technique in detecting the pedestrian heads (true positives). F-score is a combination of the two. Precision, Recall and F- scores range from 0 to 1. Scores close to 1 indicate high performance of the technique under investigation.

Evaluation is performed on video clips (30 mins each) recorded from the test location. From figure 7, we can see that the system consistently has high precision and recall values of above 0.8 and 0.9, respectively. The ground truth for the video clip analyzed in figure is 190. The F score value observed is on average 0.88. The system misses people whose faces are not visible when they are bent down. Enriching the sample data set by including head samples with side poses and poses where people are bent down can improve the performance.

The performance improvement achieved with the probabilistic temporal selection method is highlighted in figure 8. This analysis was done in moderate crowd conditions with ground truth of 126. We could see a significant increase in precision scores with the probabilistic method. The method does not affect the recall scores significantly.

Field Testing

The results over a seminar session are visualized to understand the performance of the proposed system.

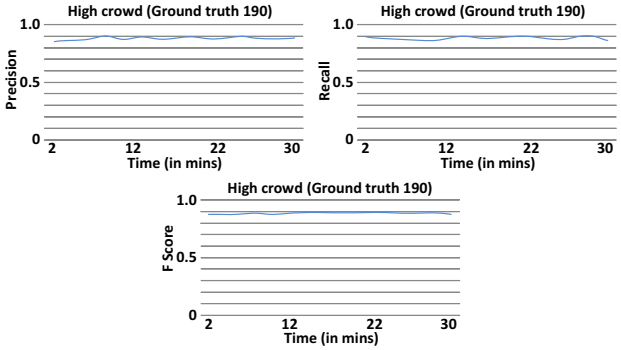

Figure 7. Precision, Recall and F score for a high crowd video clip.

Table 3. System performance on standard data sets.

	IIT Headpose	HOCoffee
Total Positive samples	10000	5517
Number of detections	9058	5320
Total Negative samples	10000	5500
False positives	2130	1080
Precision	0.8110	0.8312
Recall	0.9058	0.9642

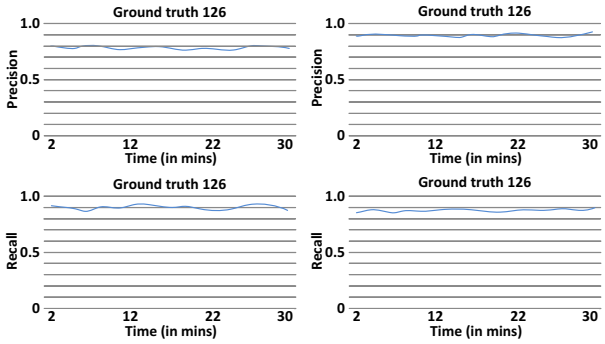

Figure 8. Performance plots for non-probabilistic method (left); probabilistic (right).

From figure 9, we can observe that the PCBPOES system follows the ground truth of people in the test location. It does miss out on detecting people whose face is not visible when they are bent down. Retina model is adopted to better detect head patterns at low light conditions. Results clearly show that the pedestrians who were not detected previously under low light conditions (when the light were switched off in the front rows) are detected after pre-processing the input image frames.

4 Conclusion and Future Works

The proposed PCBPOES system estimates the number of people in the test location, which is improved by the use of the parvo channel of the retina model. Experimental evaluations prove the superior performance of the proposed system on research data sets and in video streams from the test location. The qualitative analysis also highlights the proposed system's capability in handling environmental illumination variations.

Automatic mask recognition based on segmentation of the region of interest is one of the future works. We are also exploring on improving the detection performance by extending its purview beyond one single detection approach. A hybrid detection approach is currently explored to improve the performance.

Acknowledgment This research was carried out at the NUS-ZJU Sensor-Enhanced Social Media (SeSaMe) Centre. It is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative and by National University of Singapore under CiBEST (BEE Hub).

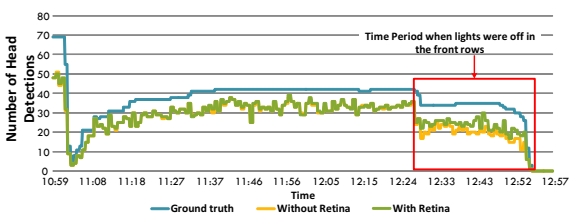

Figure 9. People count values over a seminar session.

References

[1] F. B. Lung et al.: "Spatio-temporal descriptor for abnormal human activity detection", IAPR International Conference on Machine Vision Applications (MVA), pp.471-474, 2015.

[2] B. Solmaz et al.: "Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems", IEEE Trans. Pattern Anal. Mach. Intell. , vol.34, pp.2064-2070, 2012.

[3] I. Ardiyanto et al.: "Human motion prediction considering environmental context", IAPR International Conference on Machine Vision Applications (MVA), pp.390-393, 2015.

[4] D. Winnie et al.: "Identifying Sources and Sinks and Detecting Dominant Motion Patterns in Crowds",The Conference on Pedestrian and Evacuation Dynamics, pp.22-24, vol.2, pp.195-200, 2014.

[5] Julio Cezar Silveira Jacques Junior et al.: "Crowd Analysis Using Computer Vision Techniques [A survey]", Signal Processing Magazine, IEEE, 2010.

[6] Aloys Borgers et al.: "A Model of Pedestrian Route Choice and Demand for Retail Facilities within InnerCity Shopping Areas", Geographical Analysis, 1986.

[7] P. Viola et al.: "Rapid Object Detection using a Boosted Cascade of Simple Features", proceedings IEEE Conf. on Computer Vision and Pattern Recognition(CVPR 2001), 2001.

[8] C.Cortes et al.: "Support-vector networks", Machine Learning, vol.3, pp.273, 1995.

[9] N.Dalal et al.: "Histograms of oriented gradients for human detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol.1, pp.886-893, 2005.

[10] J. Herault et al.: "Modeling visual perception for image processing", Proc. Computational and Ambient Intelligence, pp.662-675, 2007.

[11] W. Beaudot et al.: "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", Computer science, INPG, Grenoble 51.

[12] A. Benoit et al.: "Using human visual system modeling for bio-inspired low level image processing", Computer vision and Image understanding, pp.758-773, 2010.

[13] Arun Kumar Chandran et al.: "Real-time Identification of Pedestrian Meeting Events from Surveillance Videos using Motion Similarity", Journal of Real-Time Image Processing, 2016.

[14] F.Yoav et al.: "A decision-theoretic generalization of online learning and an application to boosting", Journal of Computer and System Sciences, vol.55, 1997.

[15] Robert Schapire et al.: "Explaining Adaboost", IEEE International Conference on Image Processing, vol.1, 2002.

[16] Arun Kumar Chandran et al.: "Pedestrian Crowd Level Estimation by Head Detection using Bio-inspired Retina Model", IEEE Conference on Technologies for Smart Nation (TENCON), 2016.

[17] IIT Head pose data set: https://sites.google.com/site/diegotosato/ARCO/iit.

[18] HOCCoffee data set: https://sites.google.com/site/diegotosato/ARCO/coffeebreak.

[19] Grant Allen et al.: "The Definitive Guide to Sqlite (2nd ed.)", Apress, Berkely, 2010.

[20] Powers et al.: "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation", Journal of Machine Learning Technologies, vol.2, 2011.

[21] R.Lienhart et al.: "An extended set of Haar-like features for rapid object detection", ICIP02, pp.900-903, 2002.