Improved PAM-based Traffic Behavior Recognition Using Trajectory-Wise Features

Thien Huynh-The *, Dinh-Mao Bui *, Sungyoung Lee *, Yongik Yoon †
* Department of Computer Engineering
Kyung Hee University, Gyeonggi-do, 446-701, Korea
Email: thienht,mao,sylee@oslab.khu.ac.kr
† Department of Multimedia Science
Sookmyung Women's University, Seoul, 140-172, Korea
Email: yiyoon@sookmyung.ac.kr

Abstract—**Recently CCTV-based behavior recognition have gained considerable attention in the transportation surveillance systems to identify normalities, such as traffic jams, accidents, and dangerous driving. An improved method is presented in this paper for the traffic behavior surveillance system by discovering more highly specific features based on the trajectory information. The multiple sparse feature comprising the object location, moving direction, speed, and appearance time length obtained from the moving object detection and tracking stage is modeled by the Pachinko Allocation Model. This hierarchical probabilistic model captures the correlation among the traffic activities and behaviors through the sparse features as the visual words. In the classification phase, the Support Vector Machine constructed from Decision Tree Architecture is utilized. Compared with existing methods, the proposed method outperforms 3-8% approximately in overall classification accuracy.**

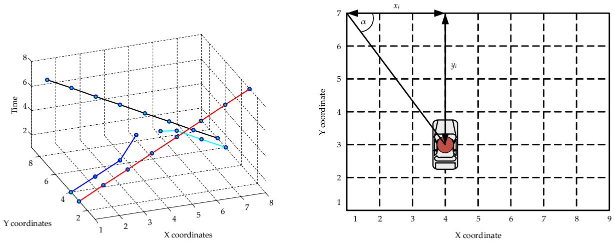
I. Introduction

Human Behavior Analysis (HBA) which is integrated into many video surveillance systems as an important component is a research area that has recently attracting attention from the computer vision and artificial intelligence communities. The principle purpose of visual surveillance includes detecting, recognizing, and tracking the moving objects from input videos captured by CCTV cameras, and to further understand and describe behaviors. Visual surveillance has been considered in practical applications such as security guard services in smart buildings, traffic surveillance in urban areas, and access control in specific places. In these applications involving people or vehicles, the behaviors can be analyzed based on the human postures, the object trajectories and the tracking information. This information can be combined to recognize more complex contexts such as vehicle interactions [1], human interactions [2], and human-vehicle interactions [3]. Given the large amount of surveillance video data available from closed-circuit television (CCTV) systems and the real-time nature of surveillance applications, it is desirable to provide an automatic operating system that may reduce human intervention.

One of the most important applications of surveillance systems, automatic road surveillance has received increasing interest in recent years. In this domain, the learning of the traffic behavior appears to be the most complex task, especially in dynamic environments. Dynamic Bayesian Network (DBN) was used for Behavior Recognition in Road Detection system (BRRD) [4] through Vehicle Sensor Networks (VSNs) to infer road events. Moreover, group detection using collaborative filtering provides an improvement in detection performance. In the method proposed by Zhang et al. [5] describes an extension of Stochastic Context-Free Grammar (SCFG) to model complex relations between atomic activities in the temporal dimension. Another approach of Sanroma et al. [6] to unify simple and complex activity recognition, the tracking information and activity zones are modeled by Hidden Markov Model (HMM) with the Stochastic Grammar. In testing, to parse a given primitive moving tracking, a Multi-Thread Parsing (MTP) algorithm is executed to further recognize the interesting complex events that include parallel temporal relations. However, the drawback of HMM-based approaches is the requirement of large amounts of training data because sometimes they do not scale as well for complex behavior scenarios. A multi-class supervised joint topic model is proposed in the research of Hospedales et al. [7] to address the issue of rare and subtle behavior learning. Besides no leveraging of typical activities as components to explain rare activities, Hospedales's model is further nonadaptive.

The use of topic models for context learning has recently been introduced. The Delta-Dual Hierarchical Dirichlet Process (dDHDP), which is an extension of HDP, was designed by Haines et al. [8] for jointly learning both normal and abnormal behavior using weakly supervised training examples. A new topic model is introduced by Hospedales et al. [9] to overcome drawbacks on sensitivity, robustness and efficiency of object behavior mining. The topic model, namely Markov Clustering Topic Model (MCTM), builds on existing dynamic Bayesian network models and Bayesian topic models. This model was demonstrated to succeed on the unsupervised mining of behaviors in complex and crowded public scenes. An efficient method developed on Pachinko Allocation Model (PAM) was presented in [10] to model the activity and behavior from with fully flexible correlation.

In this paper, the authors continuously improves the method in [10] by considering more specific features to enhance classification accuracy. Firstly, the feature-book comprises the object location, moving direction, speed, and appearance time length, which is constructed from object trajectory information in the temporal-spatial dimension. Traffic activities and behaviors are then generated from the identified trajectories with a flexible topic model, namely the Pachinko Allocation Model (PAM). PAM provides a full correlation between features-activity and



(a) (b)

Fig. 1. The object trajectory: (a) in the temporal-spatial dimension and (b) the direction of motion path.

activities-behavior based on an arbitrary Directed Acyclic Graph (DAG) structure. Finally, a multi-class Support Vector Machine (SVM) technique is employed to classify the activity and behavior according to the outputs of the PAM model.

II. Methodology

A. Feature Extraction

As a preprocess for improving the quality of input video sequences, an efficient histogram equalization [11] is used to enhance the overall contrast. The object trajectories in the input video captured from the CCTV system are then extracted using a combined background subtraction and tracking technique. The Adaptive-K Gaussian Mixture Model (AK-GMM) [12] is used to establish the model for background estimation on account of its robustness under changing environments The moving objects are distinguished from the foreground using a background subtraction technique [13]. The Kalman filter is used for tracking objects and it enables prediction of an objects future location, reduction of noise introduced by inaccurate detections, and facilitation of the association of multiple objects to their tracks.

In the proposed method, each moving object is described by four features comprising the location, moving direction, speed, and appearance time length. The object trajectories are represented in the temporal-spatial dimension. Example object trajectories illustrated in the temporal-spatial domain are shown in Fig. 1(a). To determine the orientation of the object trajectory, the absolute angle α of the current location is calculated through the following equation:

$$\alpha = \arcsin\left(\frac{y_t}{\sqrt{x_t^2 + y_t^2}}\right) \tag{1}$$

where (x_t, y_t) are the coordinates of the object at the t^{th} frame. A direction computation example is shown in Fig. 1(b). During a specific time period of the input video from t_a to t_b frame, the trajectory of the k^{th} object is formed as:

$$\mathcal{O}_k^{t_{a-b}} = \left[\left(X_k^{t_a}, Y_k^{t_a}\right), \left(X_k^{t_a+1}, Y_k^{t_a+1}\right), \ldots, \left(X_k^{t_b}, Y_k^{t_b}\right)\right] \tag{2}$$

where $X_k^t = (x_k^t, y_k^t)$ is the coordinate of an object. $Y_k^t = (\alpha_k^t, \upsilon_k^t, \tau_k^t)$ contains information of moving direction α_k^t and speed υ_k^t; and appearance time length τ_k^t measures the number of frames in which an object is considered from t_a to t. The moving speed is calculated as the Euclidean distance of an object in two adjacent frames:

$$\upsilon_k^t = \left\|X_k^{t-1}, X_k^t\right\| \tag{3}$$

Assume that each input video has n frames, the trajectory is defined as follows:

$$\mathcal{O}_k^n = \left[\left(X_k^1, Y_k^1\right), \left(X_k^2, Y_k^2\right), \ldots, \left(X_k^n, Y_k^n\right)\right] \tag{4}$$

The features extracted from the video can be expressed as the feature-book $\mathcal{C}$:

$$\mathcal{C} = \begin{bmatrix} \left(X_1^1, Y_1^1\right), \left(X_1^2, Y_1^2\right), \ldots, \left(X_1^n, Y_1^n\right) \\ \left(X_2^1, yY_2^1\right), \left(X_2^2, Y_2^2\right), \ldots, \left(X_2^n, Y_2^n\right) \\ \vdots \\ \left(X_K^1, Y_K^1\right), \left(X_K^2, Y_K^2\right), \ldots, \left(X_K^n, Y_K^n\right) \end{bmatrix} = \begin{bmatrix} \mathcal{O}_1^n \\ \mathcal{O}_2^n \\ \vdots \\ \mathcal{O}_K^n \end{bmatrix} \tag{5}$$

where K is the number of detected objects.

B. Codebook Construction

For the codebook construction, the authors utilize the k-means clustering algorithm based on the Euclidean distance metric to cluster the extracted feature dataset. Concretely, each element (X_k^t, Y_k^t) in the feature-book $\mathcal{C}$ is considered as a codeword. In the k-means clustering, the center of each cluster is regarded to be a codeword. The parameter K, the number of clusters and also the size of the codebook (the number of vocabulary words) is set in advance. Therefore a frame sequence describing more complex traffic activities can be represented by the histogram of codewords for multi-object.

C. Topic Modeling

The Pachinko Allocation Model concretely described in our work [10] is continuously applied to model codewords into activities and behaviors. Compared to Latent Dirichlet Allocation (LDA), a topic model, PAM provides more flexibility and greater expressive power than LDA model because it captures not only correlations among the words (as in LDA), but also correlation among topics. In the following subsection, the details of the proposed model based on PAM are introduced with the algorithm for the estimation of the parameters. Although PAM employs arbitrary DAGs to model the topic correlations, this work proposes a four-levels hierarchy structure as a special case of PAM [14]. This structure consists of one root topic, u super topics at the second level $\mathcal{P} = \{p_1, p_2, \ldots, p_u\}$, v subtopics at the third level $\mathcal{Q} = \{q_1, q_2, \ldots, q_v\}$ and the words at the bottom. Words refer here to the object features comprising the location and direction information, which were organized in the previous stage. The super topic and subtopic correspond to the traffic behavior and activity, respectively. The root is associated with behaviors, the behaviors are fully associated with activities, and the activities are fully connected to the features, as shown in Fig. 2(a). The multinomials of the root and behaviors are sampled for each frame based on a single Dirichlet distribution $g_r(\delta_r)$ and $g_j(\delta_j)|_{j=1}^{u}$, respectively. The activities are modeled with multinomial distributions $\phi_{q_j}\Big|_{j=1}^{v}$ sampled from Dirichlet distribution $g(\beta)$ which is used for sampling the location and direction features. Fig. 2(b) depicts a graphic model for the four-levels PAM. After modeling, the probability distribution presenting the implicit activity - behavior - frame correlation, is generated. For classification, the authors used the Binary Tree of SVM [15] to solve the N-class pattern recognition problem.

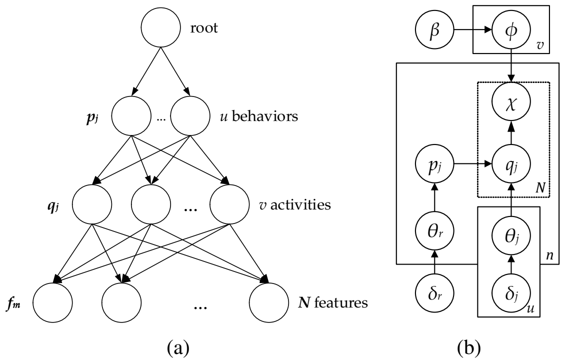


Fig. 2. Pachinko Allocation Model: (a) Hierarchical topic model (b) Graphic model.

TABLE I. ACTIVITY DESCRIPTIONS OF TWO MAIN BEHAVIORS

		VERTICAL TRAFFIC	
Activity	Color	Fig. 3	Description
V1	Orange	(a), (b), (c)	Bottom to top flow
V2	Blue	(c)	Bottom to top and turn left at the intersection
V3	Pink	(c)	Bottom to top and turn right at the intersection
V4	Yellow	(a), (b), (c)	Top to bottom flow
V5	Green	(b), (c)	Top to bottom and turn left at the intersection
V6	Cyan	(c)	Top to bottom and turn right at the intersection
		HORIZONTAL TRAFFIC	
Activity	Color	Fig. 3	Description
H1	Black	(d)	Vertical flow for pedestrian on the left side
H2	White	(f)	Vertical flow for pedestrian on the right side
H3	Pink	(d), (g)	Left to right flow
H4	Yellow	(d), (g)	Left to right and turn right at the intersection
H5	Blue	(g)	Left to right and turn left at the intersection
H6	Cyan	(e), (f), (h)	Right to left flow
H7	Green	(e), (f), (h)	Right to left and turn right at the intersection
H8	Orange	(h)	Top to bottom and stop at the intersection

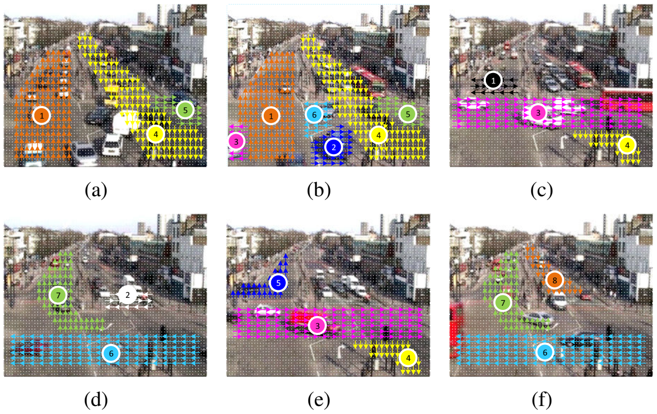


Fig. 3. Traffic activities discovered by PAM. (a)-(b): the vertical traffic activities V1-V6. (c)-(f): the horizontal traffic activities H1-H8.

III. EXPERIMENTAL RESULTS

The experiments are performed on the QMUL dataset [16] with a long time video recorded at 30fps in frame rate and 360×288 in frame resolution. The video sequence is divided into non-overlapping 4-second clips. There are totally 750 clips comprising 320 vertical traffic flow clips and 430 horizontal traffic flow clips. In this work, $u = 2$ for vertical and horizontal traffic behaviors; and $v = 14$ for traffic activities involving six vertical and eight horizontal activities. The description of the discovered activities outlined in Fig. 3 is referenced in Table I. In the PAM modeling, the Dirichlet distribution over behaviors and activities was produced with the parameter 0.01; the Gibbs sampling was processed with 1,000 burn-in iterations. For each vertical and horizontal traffic dataset, the proposed method is evaluated using the 10-fold cross-validation. All of the experiments are performed with MATLAB R2013a on the desktop PC operating Windows 7 with a 2.67 GHz Intel Core i5 CPU and 4GB RAM.

In the vertical and horizontal traffic datasets, the numbers of clips presenting particular activities discovered by PAM were not equivalent. For example, the occurrence of activity V1 and V4 in the vertical dataset and activities H3 and H6 in the horizontal dataset consumed more than 60% of the full video length. Therefore, they can be regarded as the primary activities corresponding to each dataset. For the vertical traffic, activities were discovered by PAM as shown in Fig. 3(a)-(b), while the horizontal traffic activities were presented in Fig. 3(c)-(f). The improved method using four feature types was evaluated and compared with LDA and PAM approaches [10] using only location and moving direction features. The confusion matrices of classification are reported in Fig. 4 for the traffic activities and in Fig. 5 for the horizontal activities. Due to using two additional specific features, i.e. moving speed and appearance time length, the improved version provided the better results in activity classification for vertical and horizontal traffic. Especially, some rare activities, such as V3, V6, and H8 were detected and recognized with highest accuracy. The improved approach outperformed with enhancement in overall accuracy 9.07% for LDA and 2.50% for PAM corresponding to the vertical dataset; and 6.98% for LDA and 2.56% for PAM corresponding to the horizontal dataset. In the merging of all clips to classify the behavior, the improved approach still showed the highest accuracy rate. This results can be explained that some activities described by the same features in location and moving direction are distinctly distinguished through the moving speed and time length of appearance.

IV. CONCLUSION

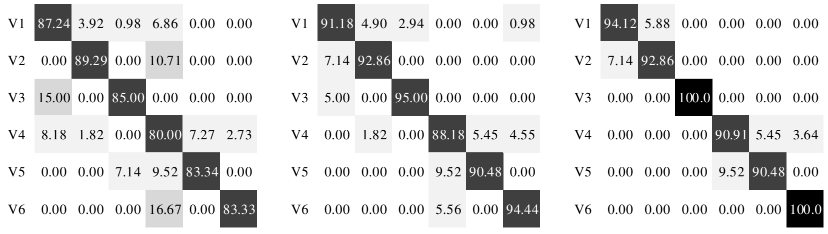
In this paper, we improved the classification performance in accuracy of traffic behavior recognition [10] by using highly specific features comprising object moving speed and appearance time length. In the our work, the PAM algorithm is utilized for automatic activity and behavior modeling from the sparse features. The generated probabilistic model is then provided for the SVM classifier. By exploiting high specific features, some activities covered by others with the same feature are distinguished more evident. Therefore, when compared with the previous method using location and direction information, it provides the better classification accuracy.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0101-15-1282-00010002, Suspicious pedestrian tracking using multiple fixed cameras).

REFERENCES

[1] L. Zhao, L. Shang, Y. Gao, Y. Yang, and X. Jia, "Video behavior analysis using topic models and rough sets [applications notes]," *Computational Intelligence Magazine, IEEE*, vol. 8, no. 1, pp. 56–67, Feb 2013.



(a)

V1	87.24	3.92	0.98	6.86	0.00	0.00
V2	0.00	89.29	0.00	10.71	0.00	0.00
V3	15.00	0.00	85.00	0.00	0.00	0.00
V4	8.18	1.82	0.00	80.00	7.27	2.73
V5	0.00	0.00	7.14	9.52	83.34	0.00
V6	0.00	0.00	0.00	16.67	0.00	83.33

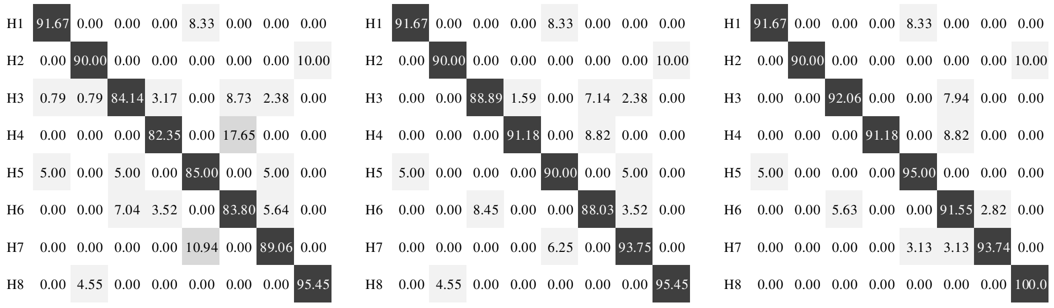
(b)

V1	91.18	4.90	2.94	0.00	0.00	0.98
V2	7.14	92.86	0.00	0.00	0.00	0.00
V3	5.00	0.00	95.00	0.00	0.00	0.00
V4	0.00	1.82	0.00	88.18	5.45	4.55
V5	0.00	0.00	0.00	9.52	90.48	0.00
V6	0.00	0.00	0.00	5.56	0.00	94.44

(c)

V1	94.12	5.88	0.00	0.00	0.00	0.00
V2	7.14	92.86	0.00	0.00	0.00	0.00
V3	0.00	0.00	100.0	0.00	0.00	0.00
V4	0.00	0.00	0.00	90.91	5.45	3.64
V5	0.00	0.00	0.00	9.52	90.48	0.00
V6	0.00	0.00	0.00	0.00	0.00	100.0

Fig. 4. Confusion matrices of classification for the vertical activity: (a) Two features + LDA, (b) Two features + PAM, (c) Four features + PAM. The overall classification accuracy rates of these approaches are 84.06%, 90.63%, and 93.13%, respectively.



(a)

H1	91.67	0.00	0.00	0.00	8.33	0.00	0.00	0.00
H2	0.00	90.00	0.00	0.00	0.00	0.00	0.00	10.00
H3	0.79	0.79	84.14	3.17	0.00	8.73	2.38	0.00
H4	0.00	0.00	0.00	82.35	0.00	17.65	0.00	0.00
H5	5.00	0.00	5.00	0.00	85.00	0.00	5.00	0.00
H6	0.00	0.00	7.04	3.52	0.00	83.80	5.64	0.00
H7	0.00	0.00	0.00	0.00	10.94	0.00	89.06	0.00
H8	0.00	4.55	0.00	0.00	0.00	0.00	0.00	95.45

(b)

H1	91.67	0.00	0.00	0.00	8.33	0.00	0.00	0.00
H2	0.00	90.00	0.00	0.00	0.00	0.00	0.00	10.00
H3	0.00	0.00	88.89	1.59	0.00	7.14	2.38	0.00
H4	0.00	0.00	0.00	91.18	0.00	8.82	0.00	0.00
H5	5.00	0.00	0.00	0.00	90.00	0.00	5.00	0.00
H6	0.00	0.00	8.45	0.00	0.00	88.03	3.52	0.00
H7	0.00	0.00	0.00	0.00	6.25	0.00	93.75	0.00
H8	0.00	4.55	0.00	0.00	0.00	0.00	0.00	95.45

(c)

H1	91.67	0.00	0.00	0.00	8.33	0.00	0.00	0.00
H2	0.00	90.00	0.00	0.00	0.00	0.00	0.00	10.00
H3	0.00	0.00	92.06	0.00	0.00	7.94	0.00	0.00
H4	0.00	0.00	0.00	91.18	0.00	8.82	0.00	0.00
H5	5.00	0.00	0.00	0.00	95.00	0.00	0.00	0.00
H6	0.00	0.00	5.63	0.00	0.00	91.55	2.82	0.00
H7	0.00	0.00	0.00	0.00	3.13	3.13	93.74	0.00
H8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.0

Fig. 5. Confusion matrices of classification for the horizontal activity: (a) Two features + LDA, (b) Two features + PAM, (c) Four features + PAM. The overall classification accuracy rates of these approaches are 85.58%, 90.00%, and 92.56%, respectively.

TABLE II. BEHAVIOR CLASSIFICATION COMPARISON BETWEEN PAM AND LDA

	LDA			**PAM**			**Improved method**		
Behavior	Vertical	Horizontal	Recall (%)	Vertical	Horizontal	Recall (%)	Vertical	Horizontal	Recall (%)
Vertical	259	61	80.94	286	34	89.38	290	30	90.63
Horizontal	66	364	84.65	59	371	86.28	45	385	89.53
Precision (%)	79.69	85.65	-	82.90	91.60	-	86.56	92.77	
Accuracy (%)		**83.07**			**87.60**			**90.00**	

[2] T. Wang and H. Snoussi, “Detection of abnormal events via optical flow feature analysis,” *Sensors*, vol. 15, no. 4, pp. 7156–7171, 2015.

[3] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, no. 1, pp. 206–224, March 2010.

[4] Y. Zeng, K. Xiang, and D. Li, “Applying behavior recognition in road detection using vehicle sensor networks,” in *Computing, Networking and Communications (ICNC), 2012 International Conference on*, Jan 2012, pp. 751–755.

[5] Z. Zhang, T. Tan, and K. Huang, “An extended grammar system for learning and recognizing complex visual events,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 240–255, Feb 2011.

[6] G. Sanroma, L. Patino, G. Burghouts, K. Schutte, and J. Ferryman, “A unified approach to the recognition of complex actions from sequences of zone-crossings,” *Image and Vision Computing*, vol. 32, no. 5, pp. 363–378, 2014.

[7] T. Hospedales, J. Li, S. Gong, and T. Xiang, “Identifying rare and subtle behaviors: A weakly supervised joint topic model,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 12, pp. 2451–2464, Dec 2011.

[8] T. Haines and T. Xiang, “Delta-dual hierarchical dirichlet processes: A pragmatic abnormal behaviour detector,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 2198–2205.

[9] T. Hospedales, S. Gong, and T. Xiang, “Video behaviour mining using a dynamic topic model,” *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303–323, 2012.

[10] T. Huynh-The, O. Banos, B.-V. Le, D.-M. Bui, Y. Yoon, and S. Lee, “Traffic behavior recognition using the pachinko allocation model,” *Sensors*, vol. 15, no. 7, pp. 16 040–16 059, 2015.

[11] T. Huynh-The, B.-V. Le, S. Lee, T. Le-Tien, and Y. Yoon, “Using weighted dynamic range for histogram equalization to improve the image contrast,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, 2014.

[12] H. Zhou, X. Zhang, Y. Gao, and P. Yu, “Video background subtracion using improved adaptive-k gaussian mixture model,” in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol. 5, Aug 2010, pp. V5–363–V5–366.

[13] O. Barnich and M. Van Droogenbroeck, “Vibe: A universal background subtraction algorithm for video sequences,” *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[14] W. Li and A. McCallum, “Pachinko allocation: Dag-structured mixture models of topic correlations,” in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 577–584.

[15] B. Fei and J. Liu, “Binary tree of svm: a new fast multiclass training and classification algorithm,” *Neural Networks, IEEE Transactions on*, vol. 17, no. 3, pp. 696–704, May 2006.

[16] QMUL Junction Dataset. [Online]. Available: http://www.eecs.qmul.ac.uk/ccloy/downloads_qmul_junction.html