TRAFFIC CONGESTION ANALYSIS: A NEW PERSPECTIVE

*Jia Wan, Yuan Yuan, Qi Wang**

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China

ABSTRACT

In this paper, a new perspective of congestion is presented to promote the development of traffic video analysis. Our main contributions are threefold: a) An unified and quantifiable definition of congestion is proposed to describe the traffic state in video. b) Based on the definition, a congestion dataset which contains multiple traffic scenes is constructed as a platform for the research community. At the same time, a precise labeling method is introduced to get the ground truth of congestion level accurately. c) An algorithm based on Inverse Perspective Mapping (IPM) and pairwise regression is proposed to analyze traffic videos and serves as a baseline. We further compare the proposed method with two deep learning methods. Intensive experiments justify the effectiveness of the proposed method.

***Index Terms*—** Traffic video, image understanding, signal processing, congestion detection

1. INTRODUCTION

Traffic congestion analysis has attracted increasing attention because of the congested traffic status [1]. As the development of society, traffic congestion has become a common phenomenon around the world. Many accidents occur and time is wasted on the road under this circumstance. If we can get an understanding of traffic congestion level automatically, it will be easier to manage the traffic. Since the camera is mounted almost every corner of the cities, it is possible to achieve it by analyzing the videos in these cameras.

However, given some video data, how to analyze congestion level is a challenging task. The illumination variations, the stopped vehicles, the changes of heights and angles of cameras, and the different road conditions make the congestion traffic hard to analyze.

Many algorithms have been proposed to remedy the congestion detection problem but most of these works can only solve the congestion detection in a specific scene. Besides, many algorithms rely on background subtraction which is insensitive to stopped vehicles and the vehicles far from camera. Moreover, the literatures seldom consider the perspective transformation which can seriously affect the stability of congestion level. All these problems limit the usage of congestion detection in real applications.

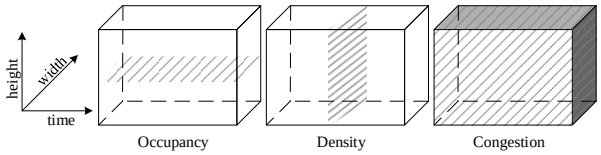

Fig. 1. The proposed definition is based on the time-space congestion.

The fundamental reason of these problems is the definition. How to define congestion in a unified, quantifiable way is the key to solve congestion detection. In this paper, we first propose an appropriate definition of congestion. Then, a dataset containing different scenes is constructed to serve as the platform for the research community. Based on the proposed definition, we precisely label the congestion level of the frames in the dataset. At last, we propose a congestion level detection algorithm with respect to the proposed definition and dataset.

2. RELATED WORK

We briefly review some congestion detection methods in this section. Existing algorithms can be divided into two classes. One is based on the detection or segmentation of moving objects (mainly vehicles). Another is based on the feature extraction of frames or videos.

The motivation of the first class is simple and straightforward: more vehicles indicate more congested traffic. For example, [2, 3] propose two algorithms which classify congested traffic videos based on the segmentation of moving vehicles on lanes. Both of them utilize the pixel number of vehicles and the speed of pixels as features. The calculation of pixel number relies on foreground detection [4, 5, 6] or other techniques to detect or segment moving objects. The speed of vehicle is calculated by tracking methods, such as optical flow

*Corresponding author.

This work is supported by the National Natural Science Foundation of China under Grant 61379094 and Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264.

[7, 8] and Kanade-Lucas-Tomasi (KLT) [9]. Firstly, many vehicles can not be detected when they stop or move slowly, and the vehicles far from the camera are small and hard to detect. These problems affect the detection of vehicles and then result in poor performance. Secondly, velocity of vehicle will be affected by perspective transformation, but these methods seldom consider that.

Another class of methods are based on congestion related features extraction and classification. Motivated by visual dynamics, [10] proposes a system including Spatiotemporal Orientation Analysis as features. To encode the motion information, [11] proposes a motion vector statistical feature to detect traffic congestion. Symbolic representation is another feature proposed in [12] which combines appearance and motion clues together. These methods don't rely on object detection as preprocessing, but the classification results depends on the accurate labeling since most algorithms divide congestion videos into 2-5 levels which is inaccurate for real applications.

3. THE PROPOSED DEFINITION AND DATASET

To remedy the problems mentioned above, we first define what is congestion and how to calculate it. Then, we construct a dataset for the task.

3.1. Definition

Generally, congested traffic condition can be measured by two aspects: spatial congestion and temporal congestion. The spatial congestion (i.e. occupancy) is the area between vehicles and road at a particular time. The temporal congestion (i.e. density) is often detected by loop detector [13]. It is calculated by the percent of time a point on the road is occupied by vehicles. The occupancy can only represent the congestion level at a point of time. The density can only represent the congestion level at a point of space. Both of them are one-sided. In this paper, we define congestion in the domain of time-space. As shown in Figure 1, the occupancy, density and the proposed congestion can be seen as the ratio of vehicles in the area with shadow.

Formally, given an video clip, we define

$$f(x,y,t) = \begin{cases} 1, & \text{occupied} \\ 0, & \text{not occupied} \end{cases} \tag{1}$$

where x, y refer to the position in one frame, and t is the time of that frame in video. The $f(w,h,t)$ can indicates that whether a point at the frame is occupied by a vehicle.

Then, the congestion can be formally expressed as:

$$congestion = \frac{\sum\limits_{x,y,t} f(x,y,t)}{width \times height \times time} \tag{2}$$

where $congestion \in (0,1)$ indicates congestion level.

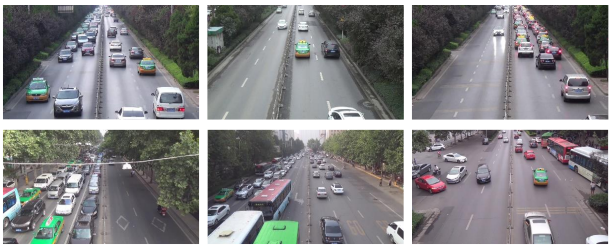

Fig. 2. Typical images in the proposed congested scene dataset.

3.2. Dataset and Labeling

Since there exists no available data for multiple scenes congestion detection, we construct a dataset which contains multiple conditions. We first collect 6 videos on different roads which contain 2-4 lanes. Typical images of these scenes can be seen in Figure 2. The resolution of these videos are 1080×720. The average length of these videos is 30 minutes.

Since the labeling is not precise in the previous works, we propose a more accurate and quantifiable labeling method based on the proposed definition. To accurately label the congestion level, we have to segment every vehicle on the road, which is very time-consuming. To simplify the labeling, we suppose that the length of a vehicle is equal to the length of lanes. This is reasonable since most of the vehicles are moving along the lane. Based on this assumption, Equation 2 can be reduced to:

$$congestion = \frac{\sum\limits_{y,t} f(y,t)}{height \times time}. \tag{3}$$

With this simplified definition, we only have to draw lines on vehicles to represent the length of them. Then, the congestion can be automatically calculated. Note that we skip every 10 frames when labeling, and others are calculated with linear interpolation.

4. OUR METHODS

With the definition and dataset, we propose a system to detect congestion level. We first consider the perspective transformation which is essential when multiple scenes are considered. Based on this, texture is utilized as low level features since it is more stable across different scenes through the experiment. At last, pairwise linear regression is included for the final congestion detection. The pipeline of the proposed method is shown in Figure 3.

4.1. Inverse Perspective Mapping

How to detect congestion level in videos which have different angles and lane numbers is the main problem. To remedy this,

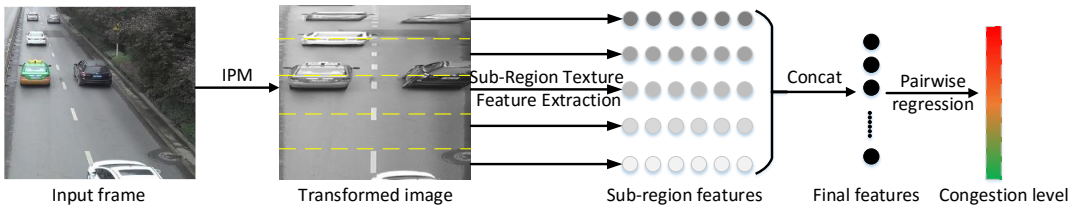

Fig. 3. The pipeline of the proposed method.

an Inverse Perspective Mapping (IPM) method [14] is utilized to remedy the effect of perspective transformation. Note that the region of interest (i.e. road) is predefined in the study. Typical results can be seen in Figure 4.

4.2. Texture Feature Extraction

Based on the IPM, we split the transformed image into several sub-regions. Then, we extract texture features in all regions. At last, all these features in sub-regions are concatenated as the final representation.

Most previous work utilized key point or the segmentation of moving objects as low-level features. However, these features will be highly affected by the change of camera angles and lane numbers. The texture on roads is more stable across different scenes through the experiment, thus we utilize texture as low-level features.

As shown in Figure 4, the top and bottom of the transformed images have different textures. That is the motivation of why we split the transformed image into sub-regions. Then, the texture feature is extracted in each sub-region. At last, these features are concatenated to form the final representation. Note that, the LBP [15] is utilized to extract texture feature.

4.3. Pairwise Regression

After feature extraction, the congestion level can be calculated by regression. Since the congestion level won't change much between adjacent frames, a pairwise regression is proposed to model that relationship.

Given a feature vector extracted in an image, the congestion level can be calculated by linear regression which is formulated as: $\hat{y} = wx$, where $\hat{y}$ is a real value which indicates the congestion level, x refers to final representation, and w is the parameter.

Given n pairs of feature representations of two adjacent frames $X = \{x_1, x_2, x_3, ..., x_{2n}\}$ and the corresponding labels (i.e. congestion level) $Y = \{y_1, y_2, y_3, ..., y_{2n}\}$, the parameters can be learned by minimize the loss function below:

$$\begin{aligned} loss &= \|Xw - Y\|^2 + \lambda\|X_1 w - X_2 w\|^2 \\ &= \|Xw - Y\|^2 + \lambda\|w\|^2 \end{aligned} \tag{4}$$

where $X_1 = \{x_1, x_2, x_3, ..., x_n\}$ and $X_2 = \{x_{n+1}, x_{n+2}, x_{n+3}, ..., x_{2n}\}$. Note that, x_i and x_{n+i} are representations of two adjacent frames. This is a ridge regression problem which can be solved by analyzing the ridge trace [16].

5. EXPERIMENTS

To confirm the effectiveness of the proposed method, extensive experiments are conducted. Since there exists no dataset for multiple scenes congestion detection, the experiments are performed only on the proposed dataset. The effect of IPM is evaluated at first. Then, the effectiveness of texture feature is confirmed. At last, the proposed method is compared with two deep learning methods.

5.1. Experimental Settings and Evaluation Protocol

The whole dataset has 6 different scenes. We extract 5000 frames for training and 1500 images for testing in each scene. In experiments, we split the transformed images into 32 sub-regions when we extract texture features. After that, the LBP feature [17] is extracted in each sub-region. The λ is set as 10 through the analysis of ridge trace.

Most previous works solve it as a classification problem [18] and the performance is evaluated by accuracy. Since our labeling is more precise (a real value instead of a class label), we solve it as a regression problem. Thus, the Mean Squared Error (MSE) is included as the evaluation protocol which can be calculated as:

$$mse = \frac{1}{2n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{5}$$

5.2. The Effect of IPM

Since different scenes contain different camera angles and lane numbers, how to eliminate the variations among different scenes is the key problem of the traffic congestion detection. The IPM is employed to remedy this problem.

To confirm the effectiveness of IPM, we compare the proposed method to the algorithm without IPM. As the results shown in Figure 5, the error bar with IPM is lower than the error bar without IPM which means the performance of the

Fig. 4. Typical results of IPM.

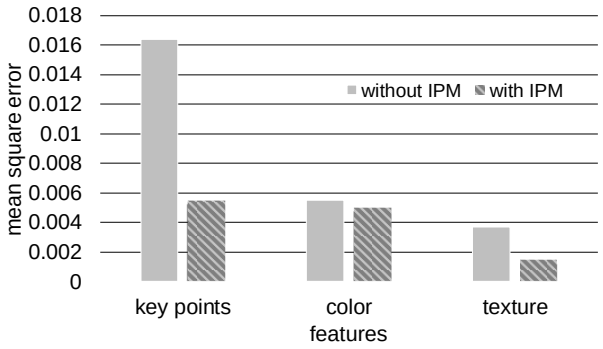

Fig. 5. The experimental results of different features with or without IPM.

Methods	CNN	CNN-LSTM	ours
MSE	0.002	0.0025	0.0015

Table 1. The comparison with deep learning methods.

method with IPM is superior to the other algorithm. Since we utilized MSE as evaluation protocol, the lower MSE indicates better performance.

After IPM, the difference among scenes has been reduced which eliminates the variation between different scenes. Furthermore, the effect of background has been reduced as well which makes IPM effective.

5.3. The Effectiveness of Texture Features

After IPM, the transformed image is split into sub-regions and then the texture feature is extracted. The congested traffics in different scenes have similar texture. Thus, the texture is included as low-level features since it is more stable among different scenes.

We compare the texture feature to the key point number and the color features. The key points in images are first detected by Harris Corner detector [19]. Then, the number of key points is treated as the feature representation. Besides, the histogram of color is used as color features.

The experimental result is shown in Figure 5. The performance of texture feature is superior to the key point number and the color features, no matter with or without IPM. The key point is hard to detect in low-light condition. Thus, the point number will be affected by the light condition. Although the color of vehicles is different to the color of road, the color of vehicles is arbitrary. Thus, it is hard to distinguish different congestion level via color features.

5.4. Comparison with Deep Learning

Deep learning has show its potential in many tasks including image classification, object detection, etc. We compare the proposed method with two end-to-end deep learning methods. The first deep network is a minor variation of AlexNet [20]. The second network is proposed in [21], namely CNN-LSTM. The LSTM is included in this network to encode temporal information. Note that we change the last classification layer to regression layer in both networks.

The experimental result is shown in Table 1. We can see that the performances of deep learning methods without careful design are worse than conventional method. Furthermore, the CNN-LSTM with temporal information encoded is worse than the simple CNN. The reason is that the label of congestion detection is too weak. Under this circumstance, the label is hard to guide the networks to learn reasonable information, and a more complex network makes the learning more difficult. That is why the performance of the CNN-LSTM is worse than the simple CNN.

6. CONCLUSION AND FUTURE WORKS

Most algorithms treat congested video analysis as a classification problem and only one scene is considered in literature. However, a real application should be a regression problem since the congestion is continuously changed. Furthermore, multiple scenes should be considered since it is hard to train multiple regression models for different cameras. To remedy the multiple congested video analysis as a regression problem, a quantifiable and unified definition of congestion is first introduced. Based on the definition, a multiple scenes traffic congestion dataset is constructed to serve as a platform for the community. Then, an IPM based algorithm is proposed as a baseline to congestion analysis.

Since the deep learning method do not work well, a stronger label including the position of vehicles and a carefully designed deep network shall be exploited in the future.

7. REFERENCES

[1] Juan José Vinagre-Díaz, Ana Belén Rodríguez-González, and Mark Richard Wilby, "Bluetooth traffic monitoring systems for travel time estimation on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 123–132, 2016.

[2] Andrews Sobral, Luciano Oliveira, Leizer Schnitman, and Felippe De Souza, "Highway traffic congestion classification using holistic properties," in *International Conference on Signal Processing, Pattern Recognition and Applications*, 2013.

[3] Shan Hu, Jiansheng Wu, and Ling Xu, "Real-time traffic congestion detection based on video analysis," *Journal of Information and Computational Science*, vol. 9, no. 10, pp. 2907–2914, 2012.

[4] Huichi Zeng and Shanghong Lai, "Adaptive foreground object extraction for real-time video surveillance with lighting variations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 1201–1204.

[5] Kedar A. Patwardhan, Guillermo Sapiro, and Vassilios Morellas, "Robust foreground detection in video using pixel layers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 746–751, 2008.

[6] Jing-Ming Guo, Yun-Fu Liu, Chih-Hsien Hsia, Min-Hsiung Shih, and Chih-Sheng Hsu, "Hierarchical method for foreground detection using codebook model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 6, pp. 804–815, 2011.

[7] Hanlin Tan, Yongping Zhai, Yu Liu, and Maojun Zhang, "Fast anomaly detection in traffic surveillance video based on robust sparse optical flow," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 1976–1980.

[8] Nelson Monzón, Agustín Salgado, and Javier Sánchez Pérez, "Regularization strategies for discontinuity-preserving optical flow methods," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1580–1591, 2016.

[9] Jianbo Shi and Carlo Tomasi, "Good features to track," in *Conference on Computer Vision and Pattern Recognition*, 1994.

[10] Konstantinos G. Derpanis and Richard P. Wildes, "Classification of traffic video based on a spatiotemporal orientation analysis," in *IEEE Workshop on Applications of Computer Vision*, 2011, pp. 606–613.

[11] Amina Riaz and Shoab A Khan, "Traffic congestion classification using motion vector statistical features," in *International Conference on Machine Vision*, 2013, pp. 90671A–90671A.

[12] Mahsa Ghasembaglou and Abolfazl Toroghihaghighat, "Symbolic classification of traffic video shots," *Advances in Intelligent Systems and Computing*, vol. 225, no. 1, pp. 207–219, 2013.

[13] Fred L Hall, "Traffic stream characteristics," *Traffic Flow Theory. US Federal Highway Administration*, 1996.

[14] Zhenqiang Ying and Ge Li, "Robust lane marking detection using boundary-based inverse perspective mapping," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 1921–1925.

[15] Fuxiang Lu and Jun Huang, "An improved local binary pattern operator for texture classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 1308–1311.

[16] Arthur E. Hoerl and Robert W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.

[17] Matti Pietikäinen, "Local binary patterns," *Scholarpedia*, vol. 5, no. 3, pp. 9775, 2010.

[18] Yuan Yuan, Jia Wan, and Qi Wang, "Congested scene classification via efficient unsupervised feature learning and density estimation," *Pattern Recognition*, vol. 56, pp. 159–169, 2016.

[19] Chris Harris and Mike Stephens, "A combined corner and edge detector," in *Proceedings of the Alvey Vision Conference*, 1988, pp. 1–6.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[21] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.