Intelligent Recognition of Traffic Video Based on Mixture LDA Model

Xiaowei Tang[1(✉)], Xin-Lin Huang[1], Si-Yue Sun[2], Hang Dong[1], Xin Zhang[1], Yu Gao[1], and Nan Liu[1]

[1] Department of Electronics and Information Engineering, Tongji University, Shanghai 201804, People's Republic of China
{xwtang, xlhuang, dh, mic_zhangxin, gaoyul631643}@tongji.edu.cn

[2] Shanghai Engineering Center for Micro-satellites, Shanghai, People's Republic of China
sunmissmoon@163.com

Abstract. In this paper, an efficient unsupervised model is proposed to recognize simple actions and complex activities in traffic scenes which is named mixture LDA model. Under this framework, we use hierarchical Bayesian models are to describe three important components in traffic video: basic visual features, simple actions, and complex activities. This model adopts an unsupervised way to learn how to recognize traffic video. Moving pixels can be divided into different simple actions and short video clips can be divided into different complex activities in a long traffic video sequence, then we can achieve the purpose of recognizing different activities in the surveillance video.

Keywords: Bayesian model · Mixture LDA model · Traffic video identification

1 Introduction

With the development of economy, video analysis and recognition technology has been applied widely in different urban public facilities and maintaining public security. There is growing demand in security control, dynamic condition records and active alarming system. Video analysis and recognition technology is playing an important role in the development of the whole society. Intelligent Transportation System (ITS) [1] is put forward by Japan and the United States at first. They called this research "intelligent vehicle system". It was used in road monitoring and intelligent vehicle research. However, along with the deepening of the research work, the system functions are extended too. The rapid development of intelligent transportation monitoring system will have a crucial impact on both transportation system and people's lifestyle.

Because of the complexity of the urban road traffic, the mature vehicle behavior recognition and detection algorithm is mainly used in expressway at present. However, most of the traffic accidents usually occur in the urban traffic route and its traffic environment are usually more complicated than expressway. Abnormal behavior recognition and detection belongs to the traffic incident detection. It refers to the

X.-L. Huang (Ed.): MLICOM 2016, LNICST 183, pp. 356–363, 2017.
DOI: 10.1007/978-3-319-52730-7_36

unexpected events on the road including emergency brake, illegal parking, and illegal turning left or right, and illegal lane change and running through red light. Frequent abnormal vehicle events bring huge property loss and casualties to the society and usually cause traffic jams. Hence, applying vehicles abnormal behavior detection to urban traffic road can help to save lives and prevent property loss.

Present study shows that the recognition methods of video are mainly divided into two categories including supervised methods and unsupervised methods, from the perspective whether using the training data set [2].

Video recognition based on supervised methods usually first set up learning model of normal behavior, and then judge the test data whether match the model. If they two do not match, the test data will be regarded as abnormal behavior. J. Snoek used the HMM model to identify the fall event in stairs [3]. First of all, they filtered out the noise and shadows in the background by using adaptive background reduction method, then obtained the low-level features of moving objects in video with optical flow, then divided the video into several segments based on time according to the theory of condition random field, then set up a reasonable threshold referring to the normal events and determined whether the event under test belongs to abnormal event by combining the HMM model. Yin Qingbo completed the anomaly detection under supervised methods [4]. The method brought in clusters in the process of establishing the training data set. Zhu Dandong adopted the theme hidden markov model in the identification of human abnormal activities [5].

Researchers began to shift emphasis from supervised detection method to unsupervised method recently years. Unsupervised anomaly detection method can effectively make up for the disadvantages of supervised method because it doesn't need to get any prior sample sets in advance. It only needs to obtain continuous sample data sets to finish the recognition and classification of normal events, regard the events with low probability as abnormal event and finally use similarity features for abnormal judgment. Wang and others introduced HDP into the hidden dirichlet distribution to realize the behavior identification in complicated and crowned scene [6]. Hospedales used LDA model to describe the spatial correlation and used HMM model to describe time correlation [7]. He combined the two models to form a new model called Markov Clustering Topic Model (MCTM) to achieve the purpose of simplifying related sequence of time and space when describing model. D. Kuettel combined HMM model with LDA model and promoted them to the infinite dimension [8]. He proposed Dependent Dirichlet Processes HMM (DDP-HMM) relied on dirichlet process and adopted this model to detect the abnormal behaviors in video.

2 System Scheme

2.1 Recognition Model of Low-Level Visual Features

In this paper, our experiment data is a 40-second-long traffic video obtained from complicated areas. The traffic video includes many different simple actions and complex activities. There are also some hard-solving problems in the traffic video, such as changeable light, different kinds of car types and environment changes. Hence, at the

first of this article, we should decide how we can divide the traffic video sequence from complicated or crowded area into simple actions and complex activities. In this section, we regard simple actions such as car going straight, car making a U-turn, people walking across the road and so on as basic element to describe more complex activities. Simple actions can always result in coherent behavior which usually couldn't quit in the middle way. We define the complex activities as the group of a variety of simple actions which take place in the meantime. For example, a car is making a U-turn while another car next to it is turning right. However, we didn't consider those complex but short activities. For example, two pedestrians are crossing the road together then turning into different directions. This article only consider those simple actions taking place in the meantime and we adopt the mixture LDA model.

We consider the moving pixels as the elementary unit to recognize the traffic video. Adopting this method can successfully refrain from tracking problems in complex area. The reason that we didn't use the overall motion characteristics is that varieties of simple actions always take place in the meantime in crowded areas, and the purpose of this article is to divide the traffic video into several kinds of motion clusters. We obtain the basic data sets by compute the location and direction of each moving pixels. When a long traffic video is given, we can divide it into several short video clips. Moving pixels which have similar simple actions always show up in one video clip. Our data sets totally have two hierarchical structures, including long traffic video sequences divided into short clips and moving pixels divided into simple actions.

Moving pixels in each frame are first obtained by using optical flow. The pixels are compared between two continuous frames. We judge the pixel as a moving pixel when the discrepancy of a pixel is larger than the threshold we set up at first. The direction of each moving pixel can also be obtained by optical flow. According to the size of the traffic video (600×800), we divide the traffic scene into cells whose size is 10×10. Then each frame totally has 60×80 cells. We use four characters to describe the direction of each cell including left, right, up, and down. Hence, we can describe the moving pixels in each frame according to the codebook whose size is $60 \times 80 \times 4$.

2.2 Mixture LDA Model

Suppose the traffic video is divided into M short video clips and these M video clips will be classified into L groups. Each group is consisted of K themes, where K is an unknown figure at first. Each theme obeys a multinomial distribution. Each group c has a Dirichlet prior which equals to α_c. For a video clip j, the probability of group label c_j is first obtained from the discrete distribution η. π_j is the probability that the theme belongs to the group j, and it can be obtained from $Dir(\pi_j|\alpha_c)$ (Fig. 1). For each simple action or complex activities i in video clip j, z_{ij} represents the probability that action i belongs to theme j, and it can be obtained from probability π_{jk}. $\beta_{z_{ji}}$ is also a discrete distribution which represents the probability of each simple action or complex activity.

Overall, π_{jk} and z_{ij} are two hidden variables in our Mixture LDA model. If α_C,β and η is given, the function relationship between these three hidden variables c_j,π_j z_j and observed variable x_j is

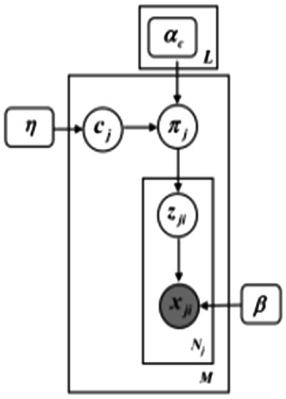

Fig. 1. Mixture LDA model [6]

$$\begin{aligned} &p(x_i, z_i, \pi_j, c_j | \{\alpha_c\}, \beta, \eta) \\ &= p(c_j|\eta) p(\pi_j|\alpha_{c_j}) \prod_{i=1}^{N} p(z_{ji}|\pi_j) p((x_{ji}|z_{ji}, \beta) \end{aligned} \tag{1}$$

The maxi-likelihood of video clip j is

$$\log p(x_j|\{\alpha_c\}, \eta, \beta) = \log \sum_{c_{j-1}}^{L} p(c_j|\eta) p(x_j|\alpha_{c_j}, \beta) \tag{2}$$

With the help of EM algorithm, $p(x_j|\alpha_{c_j}, \beta)$ can be estimated by creating a lower bound $L_1(\phi_{jc_j}, \gamma_{jc_j}; \beta, \alpha_{c_j})$, this step is called E-step in EM algorithm.

$$\begin{aligned} &\log p(x_j|\alpha_{c_j}, \beta) = \log(\int \sum_{z_j} p(x_i, z_i, \pi_j|\alpha_{c_j}, \beta) d\pi_j \\ &= \log \int \sum_{z_j} \frac{p(x_i, z_i, \pi_j|\alpha_{c_j}, \beta) q\left(z_i, \pi_j \middle| \gamma_{jc_j}, \phi_{jc_j}\right)}{m\left(z_i, \pi_j \middle| \gamma_{jc_j}, \phi_{jc_j}\right)} d\pi_j \\ &\geq \int \sum_{z_j} m\left(z_i, \pi_j \middle| \gamma_{jc_j}, \phi_{jc_j}\right) \log p(x_i, z_i, \pi_j|\alpha_{c_j}, \beta) d\pi_j \\ &- \int \sum_{z_j} m\left(z_i, \pi_j \middle| \gamma_{jc_j}, \phi_{jc_j}\right) \log m\left(z_i, \pi_j \middle| \gamma_{jc_j}, \phi_{jc_j}\right) d\pi_j \\ &= L_1\left(\gamma_{jc_j}, \phi_{jc_j}; \alpha_{c_j}, \beta\right) \end{aligned} \tag{3}$$

As soon as the lower bound is created, we can maximize the lower bound by M-step. EM algorithm is an efficient solution to estimate the hyper parameters which is also called hidden variables. After M-step is done, we can continue to use E-step to create a lower bound and then use M-step to maximize the lower bound until the hyper parameters are estimated.

$$\begin{aligned}
&\log p\left(x_j|\{\alpha_c\},\eta,\beta\right) \geq \log \sum_{c_{j-1}}^{L} p\left(c_j|\eta\right) e^{L1\left(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta\right)} \\
&= \log \sum_{c_{j-1}}^{L} m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right) \frac{p\left(c_j|\eta\right) e^{L1\left(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta\right)}}{m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right)} \\
&\geq \sum_{c_{j-1}}^{L} m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right)\left[\log p\left(c_j|\eta\right) + L1\left(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta\right)\right] \\
&- \sum_{c_{j-1}}^{L} m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right) \log m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right) \\
&= L_2\left(m\left(c_j\middle|\gamma_{jc_j},\phi_{jc_j}\right),\{\alpha_c\},\beta,\eta\right)
\end{aligned} \tag{4}$$

L_2 can reach the maximum when $m(c_j\left|\gamma_{jc_j},\phi_{jc_j}\right.)$ is chosen:

$$m(c_j\left|\gamma_{jc_j},\phi_{jc_j}\right.) = \frac{p(c_j|\eta)e^{L_1\left(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta\right)}}{\sum_{c_j} p(c_j|\eta)e^{L_1\left(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta\right)}} \tag{5}$$

3 Simulations

3.1 Simulation Steps

In the simulation, we consider a 40-second video based on complicated and crowned scenes. Its pixel is 600 × 800 and its frame rate is 15 fps. The specific simulation steps and parameter setup are as follows (Table 1):

Table 1. Steps of simulation

(1) Divide the video into consecutive frames and combine these frames into video image sequences. Divide these video image sequences into 20 short video clips. Each clip is 2 s long
(2) Use optical flow method to obtain the basic features in the traffic video and put these low-level features into WS and DS, where WS is an action set and DS is a clip set
(3) Extract semantic features in low-level features obtained in step (2) by using mixture LDA model. In this step, low-level features will be clustered into different themes
(4) Process the clustering result obtained in step (3), and range the clustering result according to the size of the probability. Set up a threshold value and discard those data sets which are lower than the threshold
(5) Take data obtained in step (4) back to the video scene, so that we can get different activities and interactions

3.2 Simulation Results

In this section, low-level features obtained by optical flow will be clustered by using mixture LDA model and EM algorithm. In mixture LDA model, M video clips will be

Table 2. Simulation setup of optical flow

Document	WO	β	α	Iterations	Correlation factor	Frame rate
20	19200	0.01	50/T	20000	3	15

grouped into L clusters, and each cluster has its own Dirichlet prior α_c which decide the distribution of themes in each video clip (Table 2).

Figure 2 shows us that the 40-second video is divided into 30 themes and we can see the probability of each theme from the figure. Figure 3 shows us the probability

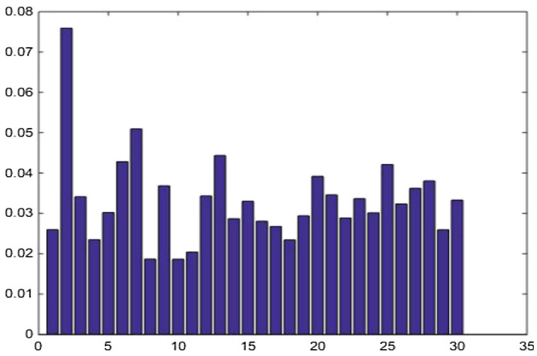

Fig. 2. Probability distribution of each topic

Fig. 3. Probability distribution of words in each topic

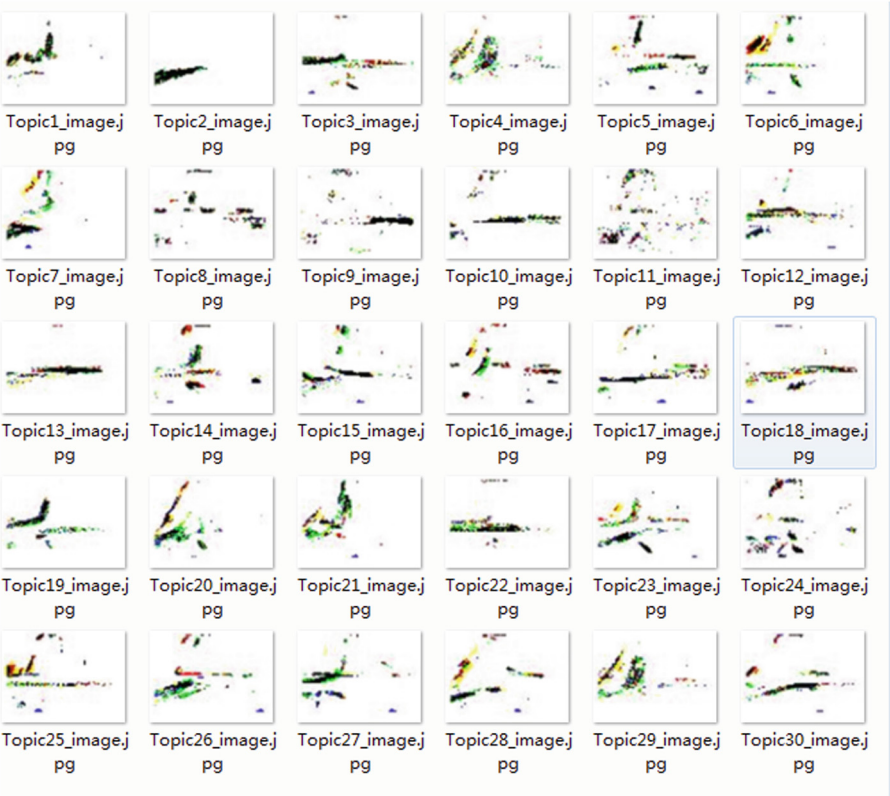

Fig. 4. Activities and interactions in each topic

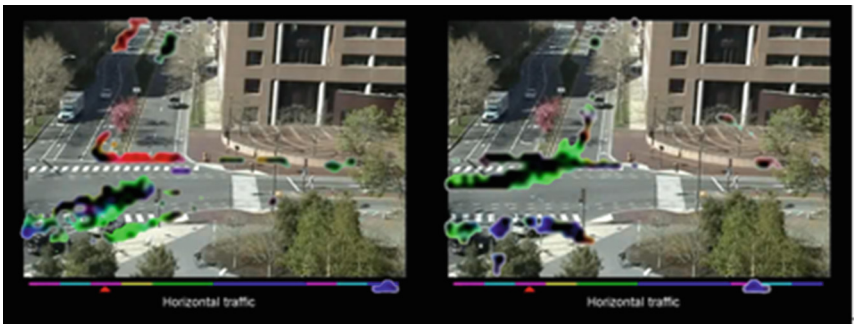

Fig. 5. Topic 26 and 27 (Color figure online)

distribution of actions or activities in each theme. We can tell that Neighboring words are usually grouped into the same cluster. After the low-level features are returned to the original video scene, we can obtain Fig. 4 which shows us the different kinds of activities and interactions in each topic.

3.3 Results Analysis

In order to make it convenient to analyze simulation results, we capture a frame from the original video as the background. Figure 5 is motion pattern in topic 26 and 27 respectively. In Compared with original traffic video, we can clearly tell that the red path represents a bicycle traveling in the motor vehicle lane and turning left into the pavement directly in topic 26. In topic 27, we can see that vehicles turning left have a conflict against pedestrian passing the pavement.

4 Conclusion

In this paper, we propose the mixture LDA model to recognize and detect different motion patterns in surveillance video. From the simulation results, we can see that the model proposed in this paper can classify different kinds of activities and interactions clearly.

References

1. Lu, H., Li, R., Zhu, Y.: Intelligent transportation system standard research. Highway Traffic Sci. Technol. **7**(21), 91–94 (2004)
2. Saligrama, V., Konrad, J., Jodoin, P.M.: Video anomaly identification. IEEE Signal Process. Mag. **27**(5), 18–33 (2010)
3. Snoek, J., Hoey, J., Stewart, L., et al.: Automated detection of unusual events on stairs. Image Vis. Comput. **27**(1), 153–166 (2009)
4. Yin, Q., Zhang, R., Li, X.: Supervised clustering anomaly detection method research based on vector quantization analysis. Comput. Res. Dev. **z2**, 414–418 (2006)
5. Zhu, X., Liu, Z.: Human abnormal behavior recognition based on hidden markov models. Comput. Sci. **39**(3), 251–255 (2012)
6. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 539–555 (2009)
7. Hospedales, T., Gong, S., Xiang, T.: A markov clustering topic model for mining behaviour in video. In: IEEE 12th International Conference on Computer Vision, pp. 1165–1172 . IEEE (2009)
8. Kuettel, D., Breitenstein, M.D., Van Gool, L., Ferrari, V.: What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1951–1958. IEEE (2010)