ATLANTIS PRESS
Advances in Intelligent Systems Research, volume 130
6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016)

A Method Based on Dense Trajectory for Violent Video Classification

Nan Wang[1, a*], Wei Song[2, b], Jianjun Hou[1, c] and Jing Yu[1, d]

[1]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

[2]School of Information Engineering, Minzu University of China, Beijing 100081, China

[a]wangnan1989@yeah.net, [b]sw_muc@126.com, [c] houjj@bjtu.edu.cn, [d] ssoohay@139.com

Keywords: Gradient; Optical flow; Dense trajectory; Extreme learning machine; Bag of words

Abstract. At present, the internet technology develops so rapidly and the video becomes the major component of the internet traffic. The content security of massive public videos is an important factor to the social stability. Among them, violent video is an important class of unsafe videos. We proposed a novel method based on dense trajectory and extreme learning machine to recognize them. The spatial-temporal characteristics were well expressed by the use of optical flow and gradient. The experiment on the benchmark dataset named Movies indicated our proposed method had a better accuracy than the state-of-the-art methods. Our proposed method is an efficient method for violent video classification.

Introduction

In the 21st century, science and technology has a fast development. Digital camera becomes the mainstream technology to capture the realistic scenes. With the advantages of large storage capacity and rewritable memory, the applications of digital camera are so wide, such as video surveillance, webcast and video playback platform. So, the video data is increasing large and the security of the video content has been gradually paid attention.

Because of the similarity between violence and fight, the algorithms on the action recognition can give us some inspirations on violence recognition. In recent years, the trajectory feature has been paid more attention, for its reflections on the movement trends of the moving object in video and representations on the dynamic characteristics of video. At first, the extraction of trajectory features depends on tracking the interest point. For example, Messing et al. [1] using KLT tracker [2] to track Harris 3D. Sun et al. [3] extracted the trajectory feature by matching SIFT interest point in consecutive frame. Jargalsaikhan et al. [4] used local features around the trajectory to describe the characteristic. Then, Wang et al. [5] put forward a method by tracking dense sampling optical flow to extract trajectory characteristics. One year later, they proposed dense trajectory method based on motion boundary [6], which further improved the recognition accuracy. In the same years, Murthy et al. [7] proposed a more efficient algorithm to describe trajectory. This algorithm had a high efficiency in motion recognition by the means of reducing redundant information and computational burden. In the past two years, algorithms based on trajectory got a wider range of applications. Kataoka et al. [8] combined trajectory with extended co-occurrence HOG to achieve better results in fine motion recognition. Yang et al.[9] also used dense trajectory for automatic recognition of workers' behavior.

Above all, the characteristic representation algorithm based on dense trajectory has obtained good results in action recognition. In this paper, we proposed a method by extracting the dense trajectory of gradient image's motion boundary, and using extreme learning machine to improve the accuracy of violent video classification.

Relative Works

DT. Dense Trajectory Feature reflects the continuous changing trend of the motion field in the video. It predicts the change of the position of the sampled pixels by calculating the optical flow field

between adjacent frames to obtain the trajectory, and then combines the local descriptors around trajectory to describe the trajectory information.

The extraction of DT feature is shown in Fig. 1. Each frame of a video is down-sampled into multi-scale images, and the optical flow field of adjacent frames is calculated at each scale. The pixel position of the next frame is predicted by the optical flow field, and so on. The changed information of different pixels on different scales is obtained. The predicted pixel position is shown in Eq. 1.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + Flow\big|_{(x_t, y_t)} \tag{1}$$

Where $P_{t+1} = (x_{t+1}, y_{t+1})$ is the predicted position of the original position (x_t, y_t) in the next frame, and $Flow$ is the optical flow field between the two frames. It can be seen that once the optical flow filed between two frames is obtained, the position of each pixel in the video frame can be tracked. The trajectory can be expressed as $(P_t, P_{t+1}, P_{t+2}, \ldots)$. In order to avoid the problem of position tracking drift, the length of the tracked trajectory is set as L. The local descriptors HOG, HOF and MBH around the trajectory are used for the representation of the extracted trajectory as shown in Fig. 1. A spatial body with length L, width N and height N is divided into $n_\sigma \times n_\sigma \times n_\tau$ small space bodies. The HOG, HOF and MBH of the small space bodies are calculated respectively to characterize a motion trajectory by cascading them to a feature descriptor.

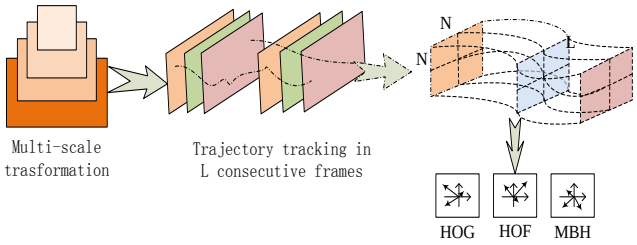

Figure 1. the scheme for DT

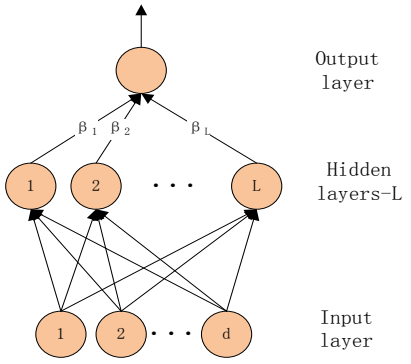

Figure 2. framework of ELM

Principle of ELM. Extreme learning machine proposed by Huang [13,14] is a single-hidden layer feedforward neural network. It can be applied to regression and classification. Its input weight and bias can be assigned randomly, and they needn't be adjusted during the training process. ELM only needs to train to solve the output weight of the hidden layer nodes. Moreover, the training and classification of ELM has sufficient theoretical support. Therefore, it is widely concerned by scholars in recent years.

The learning model of ELM is shown in Fig. 2. Assuming that $\boldsymbol{X} \in R^d$ is the input vector, the output function of the ELM is:

$$f_L(\boldsymbol{X}) = \sum_{i=1}^{L} \beta_i h_i(\boldsymbol{X}) = h(\boldsymbol{X})\boldsymbol{\beta} \tag{2}$$

Where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_L]^T$ is the output weight of the hidden layer. $h(\boldsymbol{X}) = [h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), \ldots, h_L(\boldsymbol{X})]^T$ is the output of the hidden layer and maps the input data from d dimension to L dimension.

Different from traditional learning algorithms, ELM requires minimizing the training error and output weight modulus. The corresponding optimization problem is as follows:

$$\begin{aligned} &\min\ L_{ELM} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\xi_i^2 \\ &s.t.\quad h(X_i)\boldsymbol{\beta} = t_i - \xi_i,\quad i = 1, \ldots, N \end{aligned} \tag{3}$$

With the KKT Optimization Conditions for Solving Dual Problem of Eq.3, Eq.4 is obtained:

$$\boldsymbol{\beta} = \boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{C} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T} \tag{4}$$

Where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{T}=[t_1,...,t_N]^{\mathrm{T}}$. In this case, the output function of ELM can be expressed as follows:

$$\begin{aligned} f_L(\boldsymbol{X}) &= h(\boldsymbol{X})\boldsymbol{\beta} \\ &= h(\boldsymbol{X})\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{C} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T} \end{aligned} \tag{5}$$

Proposed Method

Since the dense trajectory is the position prediction of the pixels in the interval sampling, it will contain many background trajectory traits. In order to suppress the background interference, this paper pre-processed each frame of the video by gradient. On this basis, we calculated the dense trajectory of motion boundary and achieved a good result by the means of ELM.

Image Pre-Processing. The judgement of violence recognition mainly comes from the movements of the human body. The motion of background caused by camera shake or other factors is easy to interfere in the extraction of the human motion information in the video. This paper proposed the gradient method to reduce the background interference and highlight the moving object edge information. The gradient processing mainly included the following steps:

(1) Gamma correction method for image color space normalization;

(2) Calculating the gradient amplitude of each pixel in the image.

In order to reduce the influence of local illumination, the image was compressed by Eq. 8.

$$I(x,y) = I(x,y)^{Gamma} \tag{6}$$

The gradient at position (x,y) was:

$$\begin{aligned} G_x(x,y) &= H(x+1,y) - H(x-1,y) \\ G_y(x,y) &= H(x,y+1) - H(x,y-1) \end{aligned} \tag{7}$$

In this paper, we used the gradient amplitude at any pixel to express the texture information of the image. The amplitude was calculated as follows:

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{8}$$

The outcomes of gradient processing are partly shown in Fig. 3.

Figure 3. results of gradient pre-processing

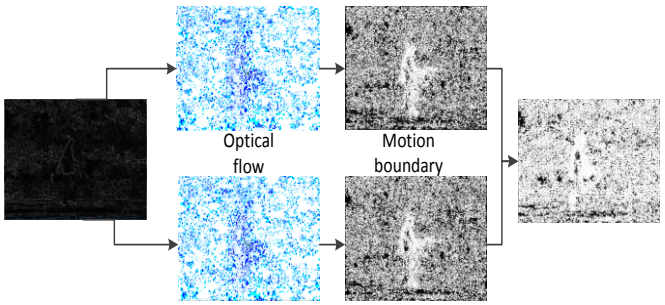

Figure 4. the calculation process of motion boundary

Dense Trajectory of Motion Boundary. In order to extract the motion information and suppress the redundant information, this paper used optical flow to process the gradient frames. The algorithm converted the preprocessed image into the optical flow field using Horn-Schunck algorithm. The

results contained horizontal component v_x and vertical component v_y. Then the gradient of each component was calculated and synthesis gradient is obtained. The gradient calculation method is the same as the preprocessing, and the optical flow field is shown in Fig. 4.

Then the dense trajectory was extracted using the method proposed by Wang et al.[5]. Finally, using Bag of Words for dimension reduction of features, each video could be represented by a compact vector.

Training for Classification. We used the feature vector of each training video as the row vector of input matrix. Assuming there are N training videos and the feature dimension is M, we obtained a matrix with dimension N×M. Using this matrix as the input of the ELM and the corresponding label as the output; we trained the output weight matrix of the hidden layer and obtained a classifier for violent video with fixed output layer parameters.

Experiment

In this paper, the feature extraction and classification model training were implemented on the Windows operating system. The feature extraction was implemented on the visual studio 2010 based on opencv2.4.9 library. The feature dimension reduction and classification were implemented on the matlab2014.

The benchmark dataset is the public dataset Movies [10]. Movies contain 200 video clips, of which 100 violent video clips were taken from action movies, and 100 non-violent video clips were collected from the action recognition public datasets. The violent and non-violent labels were labeled artificially. Partial frame taken from a dataset is shown in Fig. 5, where the first row is violent frames and the second row is non-violent frames.

Fig. 5 Movies dataset

The main parameters involved in this experiment were set as shown in Table 1.

Table 1 the main parameter setting

Parameter	Value
Gamma	1/2
Sampling interval	5[pixel]
Trajectory length	15[frame]
n_σ	2
n_τ	3
D（BoW）	250

After extracting the features from the dataset using the parameters in Table 1, each video was presented by a 250-dimention vector. Subsequently, the result in this paper compared with the conventional method is recorded in Table 2 and Table 3, respectively. Among them, Table 2 is the comparison with 5-fold cross validation experiment. Table 3 is the comparison with 10-fold cross validation experiment.

Table 2 comparison of 5-flod cross validation results

Method	Recognition accuracy[%]
HOG+SVM(HIK)[12]	49
HOF+SVM(HIK)[12]	59
MoSIFT + BoW+SVM[12]	89.5
Proposed method	99.0

Table 3 comparison of 10-flod cross validation results

Method		Recognition accuracy[%]
HOG[13]	SVM	82.5
	Adaboost	74.5
HOF[13]	SVM	84.2
	Adaboost	86.5
MoSIFT[13]	SVM	85.4
	Adaboost	98.9
Proposed method		99.0

It is easy to see from the above table that the method in this paper obtains the highest recognition accuracy on the test data set compared with the state of art methods, and the effect is far superior to the algorithm of using HOF and HOG as feature descriptors. It is because HOG only describes the texture information of the human body or object in the video, but lacks the dynamic information of them. Although HOF describes the temporal dynamic information in the video, it is relatively lack of spatial shape information and is susceptible to camera shake. From the experiment results, it can be seen that the MO SIFT feature combines the static feature points with the dynamic trajectory characteristics, and obtains higher recognition accuracy than the HOF and HOG methods. The algorithm proposed in this paper considers the spatial and temporal characteristics of video and combines the gradient feature with the optical flow characteristics. It suppresses the redundant information and highlights the temporal and spatial dynamic characteristics of video. The continuous change trajectory of the sampling point is the characteristic description unit of the video, which expresses the content of video well. ELM has an open library and doesn't require initialization. The operation is very simple and practical. In the last, the experiment achieves better classification accuracy than the current methods.

To sum up, the proposed algorithm highlights the temporal and spatial dynamic information of video, minimizes the interference of useless information and uses fast and effective ELM model to do classification. It is an effective algorithm for violent video classification.

Conclusions

This paper gets an idea from the field of action recognition on the basis of the characteristics in the violent video content. First, the video is pre-processed using gradient calculation. Then the movement of texture information is calculated. The HOG, HOF and MBH traits in the neighborhood are used to describe the trajectory, and then the BOW model is used to dimension reduction of massive features. Therefore, one compact feature is used for representing each video. Finally, the video features are classified by ELM and the higher classification rate of violent video is obtained. In this paper, we propose a new method to classify the violent video, but there are still many areas to improve. The calculation of the dense trajectory is very expensive. So how to speed up the trajectory extraction and reduce the computational burden is the focus of the next step.

Acknowledgements

This paper is supported by “the Fundamental Research Funds for the Central Universities”.

References

[1] R. Messing, C. Pal and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009, p. 104.

[2] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In IJCAI, Vol. 81 (1981) No. 1, p. 674.

[3] J. Sun, X. Wu, S.C. Yan, L. F. Cheong, T. S. Chua and J.T. Li, "Hierarchical spatio-temporal context modeling for action recognition," Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Miami, FL, 2009, p. 2004.

[4] Jargalsaikhan, S. Little, C. Direkoglu and N. E. O'Connor, "Action recognition based on sparse motion trajectories," 2013 IEEE International Conference on Image Processing, Melbourne, VIC, 2013, p. 3982.

[5] H. Wang, A. Kläser, C. Schmid and C. L. Liu, "Action recognition by dense trajectories," Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, Providence, RI, 2011, p. 3169.

[6] H. Wang, A. Kläser and C. Schmid, Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision, Vol. 103 (2013) No. 1, p. 60.

[7] V. R. Murthy and R. Goecke, "Ordered Trajectories for Large Scale Human Action Recognition," Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, Sydney, NSW, 2013, p. 412.

[8] H. Kataoka, K. Hashimoto, and K. Iwata, “Extended co-occurrence HOG with dense trajectories for fine-grained activity recognition,” Asian Conference on Computer Vision. Springer International Publishing, 2014, p. 336.

[9] J. Yang, Z. Shi, and Z. Wu, “Automatic recognition of construction worker activities using dense trajectories.” In International Symposium on Automation and Robotics in Construction and Mining, Vol. 6 (2015), p. 7.

[10] G. B. Huang, Q. Y. Zhu and C. K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, Vol. 2 (2004), p. 985.

[11] G. B. Huang, Q. Y. Zhu and C. K. Siew. "Extreme learning machine: Theory and applications." Neurocomputing, Vol. 70 (2006) No.1-3, p.489.

[12] E. B. Nievas, O. D. Suarez and G. B. Garc á, “Violence detection in video using computer vision techniques.” International Conference on Computer Analysis of Images and Patterns. Springer Berlin Heidelberg, 2011, p. 332.

[13] Deniz, I. Serrano, G. Bueno and T. K. Kim, "Fast violence detection in video," Computer Vision Theory and Applications (VISAPP), 2014 International Conference on, Lisbon, Portugal, 2014, p. 478.