Research on the Attribute Extraction Algorithm of People and Vehicles Based on Video Structure

Chaoying Zhang\textsuperscript{a,}\textsuperscript{*}, Baolin Liang\textsuperscript{b}, Sicong Ma\textsuperscript{c}, Chao Xiang\textsuperscript{d}, Zhen Bai\textsuperscript{e} and Jianxiu Wang\textsuperscript{f}

China Telecom Corporation Limited Beijing Research Institute, Beijing 102209, China.

\textsuperscript{*}Corresponding author e-mail: zhangchy50@chinatelecom.cn
\textsuperscript{b} liangbl6@chinatelecom.cn, \textsuperscript{c} masc@chinatelecom.cn, \textsuperscript{d} xiangchao@chinatelecom.cn, \textsuperscript{e} baizhen@chinatelecom.cn, \textsuperscript{f} wangjianxiu@chinatelecom.cn.

Abstract. With the development of sensor technology and computer vision, the speed of video generation is increasing exponentially. The workload of video browsing and the amount of stored data are increasing. In order to achieve fast browsing and efficient storage, video structure technology is increasingly important. This paper focuses on the video structure technology in the construction of smart cities, outlines the system process of video structure, and studies the extraction algorithm of human and vehicle target attributes based on deep learning. And combined with the actual video monitoring scene, the experiment is carried out. Compared with traditional image processing methods, the video structure target attribute extraction model based on deep learning has high detection accuracy and strong algorithm generalization ability. The intelligent analysis and processing of data by video structure technology is conducive to quickly increasing the speed of target search, reducing storage capacity, solving the problem of long-term storage, and realizing full automation of surveillance video.

Keywords: Video Structure, Deep Learning, Target Detection, Person and Vehicle Attribute Extraction.

1. Introduction

In this era of rapid development of information technology, a variety of information floods people's lives. And video data is one of them. As information medium, video has rich expressive force and is easy to be understood. It is widely used in our daily communication and entertainment. But it contains too much data and lacks structural information expression, which makes it very difficult to manage.

For a large number of chaotic video data, the video itself has the characteristics of low abstraction, high information complexity and no structure, which makes it very difficult to extract the key information. If manual labeling is used to classify and mark key information, it will cost a huge amount of labor and time. Moreover, because of the physiological characteristics of human beings, the accuracy of manual annotation will decrease sharply after repeated work or fatigue work, resulting in problems such as video labels being inconsistent with reality. Therefore, video data structure technology is
becoming a research hotspot in the field of video image analysis. And it has wide application value in daily life. At present, the technology has been studied at home and abroad.

In foreign countries, Mecocci et al. introduced a video surveillance system that can detect abnormal behaviors and events in [1]. The system can adapt to different scenarios and uses self-learning technology to learn typical behaviors of targets in specific environments. If the system detects that the target’s behavior deviates from the typical learning model, it can issue an abnormal alarm. Duque et al. proposed the observer video system [2]. The system segmented specific targets by background subtraction, and then used the appearance-based target tracking method to match and track the target. By comparing with normal behavior, the system could display the degree of abnormal behavior of targets. In addition, Hori [3] et al. proposed a multi-modal fusion video description method based on attention mechanism. This method can translate video into description sentences corresponding to video content. Its model integrates features of video, motion features and audio features, and realizes the translation from video to descriptive sentences using recurrent neural network and attention mechanism.

Domestically, many companies, institutions and universities have participated in the research of video structure, such as Hikvision, Tsinghua University and National University of Defense Technology, etc., which have greatly promoted the development of it. In the early research, Zhuang Yueting et al. [4] proposed a method to extract video catalogs. This method divides the video into shots, key frames, and scenes through traditional image processing technology, and establishes a video catalog. This method facilitates quick retrieval of video content. Huang et al. [5] proposed a video description generation method based on relation mining. This method realizes the conversion of continuous video frames to text descriptions through a double-layer LSTM, and obtains the corresponding video descriptions. Combining the image and audio information in the video, YingchunShi and others proposed a video structured model [6]. This model analyzes the content structure characteristics of news videos, and realizes the structuring of news videos through a hierarchical video catalog structuring model based on multi-semantic abstract level representation.

Judging from the research history and development status at home and abroad, video structure technology is still making continuous progress to meet the increasingly high demands of the current society. Firstly, the field of video structured application has been gradually expanded, from the first monitoring video application to medical, media and other industries. Secondly, the information extracted by video structure technology is more and more abundant. From the beginning of target recognition and action recognition, it develops to the current statement sentence description based on video key information. Thirdly, the algorithm applied is transferred from traditional image processing algorithm to artificial intelligence algorithm based on deep learning. All in all, the development of emerging media and communication technologies has led to a large accumulation of video data. Video storage and retrieval will become a major problem in the field of video analysis in the future. Based on this problem, the automatic video data processing method combining deep learning and artificial intelligence algorithm will become one of the research hotspots in the world.

This paper mainly summarizes the video structured system and commonly used target detection models, and studies the target attribute extraction algorithms for people and cars based on deep learning. Combined with the actual video surveillance scene, this paper conducts video structured human-vehicle attribute extraction experiments. Finally, the conclusion and future work are summarized.

2. Video structured system
Frame is the most basic component of video. Its essence is an image, and it is also the main content of video structure research. In the research of video structure, the most informative video frames can be extracted by key frame extraction technology, and then the main contents of these video frames can be transformed into high-level semantic information for structured information storage.

The original surveillance video is a kind of unstructured data. The target in the video can only be viewed directly, and cannot be directly read and recognized by the computer. Video structure is to label the target in the unstructured data such as video, and transform it into structured data that can be searched by certain conditions. The structure of video information refers to the expression of key information in
video data through hierarchical and easy to understand by human beings. In this way, the retrieval can be carried out according to the content level or key information in video retrieval, which can save the retrieval time. Moreover, when the amount of video data is too large and the storage conditions are insufficient, only storing structured video information can effectively reduce the space needed for storing video, and the key information of video can be retained. Video data contains rich information. In video structure technology, the information extracted from video data can be divided into low-level feature information, key video image information, and high-level semantic information, as shown in Fig. 1.

The bottom feature information generally refers to the extraction of global features, local features and structural features of an image. Global features are the basic features of image such as color, texture and shape. The feature points set of video image is extracted by local features, and the descriptors of feature points are calculated for feature matching. The structure features reflect the relationship of geometry and space-time domain between image features.

Key video image information refers to the extraction of key frames according to some underlying features and target information of the image. By fusing different underlying feature information, the information difference between frames or the information richness of video frames can be expressed, and then the most representative video frames are selected.

High-level semantic information refers to the semantic summary and description based on the goals and content contained in the video. Using deep learning, according to the appropriate number of image sets, we can train the targeted model, extract the target semantics, scene semantics, image semantics and so on. Based on the extracted semantic information, text statements are extracted from the model to summarize the events reflected in the video, which can be used for intuitive understanding. Furthermore, five key technologies are used to achieve different functions, namely key frame extraction, target detection, action recognition, scene recognition and image description. Key frame extraction technology can extract the key information in video and save storage space. Target detection can identify the type, number and location of the target in the key frame. Action detection can identify the target action. Image description can transform abstract image data into easily understood text description, which is convenient for storage and retrieval. The specific information extraction process is shown in Fig. 2.

![Fig. 1. Video structured information extraction.](image1)

![Fig. 2. Video structured process.](image2)
In terms of the content of video structured description, the main video information includes people, vehicles, scenes and objects. In the aspect of algorithm processing, when the monitoring video is transmitted to the equipment, the first stage is mainly for the detection and recognition of specific targets. The target objects in the video image are extracted by means of target detection, feature extraction and object recognition. In terms of detection and recognition, the available models are Faster RCNN [7], SSD[8], YOLOV3[9] and so on. According to the configuration of the cluster hardware and the speed of model processing and analyzing the video, the optimal model is evaluated and selected. After the model is loaded and initialized, the video stream data of the monitoring video platform is input. After data preprocessing, the trained model is called to detect and identify the target object. And the recognition result is output, as shown in Fig. 3.

![Fig. 3. Target detection process.](image)

### 3. Target detection based on deep learning

Target detection based on deep learning is the mainstream detection method at present, and it is widely used in many fields of real life. This kind of method firstly uses convolutional neural network (CNN) to extract the features of the image, then uses the prediction box to locate the target, and finally realizes the target recognition and classification. The following section briefly introduces several typical target detection models based on deep learning.

Region-CNN first combines deep learning theory with target detection technology and achieves a better detection effect. The design process of the model is mainly divided into two stages: the generation of candidate regions and the classification of candidate regions. Specifically, in the stage of candidate region generation, the selective search method is mainly used to effectively remove redundant candidate regions, thereby greatly improving the calculation speed. In the stage of candidate area classification, CNN is used to extract features from the area, and support vector machine is used to classify the candidate area. Finally, the detection task of target category and location information is completed. Although the recognition rate of R-CNN model is significantly improved compared with the traditional methods, the overall processing speed of the model is slow because the method divides the candidate region generation and target recognition into two stages, and the number of candidate regions is still large [10].

The Single Shot MultiBox Detector (SSD) is an idea of one stage, that is to input images directly to the network without going through the stage of generating candidate regions, and then output the detection results. Specifically, the entire network design of SSD draws on the idea of Region Proposal Network, using Visual Geometry Group (VGG) as the basic architecture. In down sampling, the feature map of each layer is predicted. At the same time, a priori box is used to predict the target position, and the bounding box with overlapping positions is fitted by the method of non-maximum suppression. Since SSD uses CNN for feature extraction, the convolutional layer will reduce the image dimension and resolution, and affect the detection effect of small targets. In order to solve this problem, SSD uses a
multi-scale feature detection method to predict on different scales, which improves the detection effect of small targets to a certain extent. But the detection speed of the model needs to be further optimized.

The You Only Look Once (YOLO) uses a single CNN to achieve end-to-end target detection[11]. It integrates the generation of candidate frames and target classification into a network, so that target detection is transformed into regression problem, and the network outputs detection results directly. Compared with R-CNN, YOLO is a unified framework. The design of this network structure also makes the detection speed of YOLO significantly faster than other algorithms. YOLO is divided into three steps. The first step is to convert the input image size to $448 \times 448$, and the image is divided into $7 \times 7$ grids, where each grid is a cell. The second step is to use the CNN for feature extraction. Finally, the full connection layer is responsible for the position regression of each bounding box, as well as the prediction of target confidence and category selected by the box. In the third step, it uses non maximum suppression to get the final detection frame and the category information of the target. The network framework is shown in Fig. 4.

![YOLO framework](image)

Fig. 4. YOLO framework.

The YOLO framework designs the CNN to complete the detection. Both training and prediction are end-to-end methods. The simple algorithm design makes YOLO obtain a good detection speed. In addition, the method of YOLO feature extraction is convolution of the whole image, so it can obtain a larger receptive field when detecting the target. And it is not easy to cause false judgment on the background. Secondly, the YOLO detection framework also has good generalization capabilities, and the robustness of the model during migration is better. However, when YOLO performs target detection, each cell uses only two bounding boxes to predict the same category. There is a certain degree of missed detection, and YOLO does not solve the problem of small target detection well. However, in subsequent versions of the YOLO series, methods such as multi-scale feature extraction and more a priori boxes are used to improve these shortcomings of the model, and gradually make the YOLO series detection framework the best detection algorithm in the field of target detection.

4. Human and vehicle attribute extraction experiment

For pedestrian attribute extraction in surveillance video, this paper first uses YOLOV3 to detect pedestrian in video frame, and then extracts pedestrian attribute through attribute person recognition (APR) network[12].

Among them, YOLOV3 uses Darknet-53 as its backbone network. In the network architecture, there are mainly three modules, which are connected through feature maps of different sizes to obtain receptive fields of different sizes, thereby detecting and identifying targets with different feature sizes. Since the deeper the network level, the smaller the size of the feature map. Therefore, the deeper the network, the weaker the detection ability of small targets, which will also cause the smaller the target to be more difficult to be detected in the deeper network layer. In YOLOV3, due to the connection between the FPN module and the upper layer feature map, the current layer network can obtain the information of the lower layer and deeper network. So it can increase the number of multi-size feature maps in the network and the features saved, while achieving the goals of large, medium and small Detection and identification.

After YOLOV3 analyzes the original surveillance video, pedestrians in the video can be effectively detected. And the detected pedestrian pictures are sent to the attribute extraction model in the video structured system for further identification and analysis.
The APR model is constructed from pedestrian recognition and attribute recognition, but this paper only needs the construction of attribute recognition. The APR model structure is shown in Fig. 5.

![Fig. 5. The APR model structure.](image)

The training data set uses the Market-1501 and DukeMTMC-reID data sets. The two data sets include more than 20 types of pedestrian attributes. It takes Resnet-50[13] as the benchmark network, and conducts pre training on Imagenet, and finetune the two models with new annotation attributes and currently available identity tags. After the model is trained, input a picture for testing, and some of the pedestrian attribute results obtained from the experimental test are shown in Fig. 6. By analyzing the results of attribute extraction, it can be seen that the model can accurately extract the basic characteristics of pedestrian such as gender, clothing, hair, etc., and meet the basic requirements of pedestrian attribute extraction in video structure.

![Fig. 6 Pedestrian attribute extraction results.](image)

For the extraction of vehicle attributes in surveillance videos, the models used in this paper include YOLOV3 and B-CNN [14], so as to realize vehicle detection and multi-label recognition of vehicle attributes. The identification of vehicle attributes includes vehicle color, vehicle orientation, and vehicle type. Among them, YOLOV3 is used for vehicle detection, and the detected vehicle graphics are sent to BCNN for vehicle multi-label attribute extraction. For the B-CNN model, it is mainly used to train end-to-end fine-grained classification, and the structure is shown in Fig. 7.

![Fig. 7 The B-CNN model structure.](image)
As shown in Fig. 7, two convolutional neural networks are used for feature extraction of images, and then a bilinear pooling function is used to combine the two groups of features extracted by CNN, and finally put into the softmax layer for classification. Among them, network A and network B can be two symmetric networks, or two asymmetric networks. In order to take into account the inference speed and accuracy of the program, unlike the VGG16 used in the paper, the basic network of the model in this paper uses Resnet-18. After the model is trained, input a picture for testing. The vehicle attribute results obtained from the experimental test are shown in Fig. 8. Through the analysis of the results of vehicle attribute extraction, it can be seen that the model can accurately extract three basic attribute features of vehicle type, vehicle color and vehicle orientation, which meets the basic requirements of vehicle attribute extraction in video structure.

Fig. 8 Vehicle attribute extraction results.

5. Conclusion
This paper focuses on the monitoring video structure technology in the construction of smart city, summarizes the system flow of video structure, and studies the algorithm of extracting human and vehicle target attributes based on deep learning. Combined with the actual video monitoring scene, the experimental verification is carried out. Compared with the traditional image processing methods, the video structured object attribute extraction model based on deep learning has better detection accuracy and faster running speed. With the development of video structure and the demand of more massive video information processing, there are still many deficiencies in video structure technology that need to be improved. There are different standards for different scenes, and more specific key frame extraction algorithms for different scenes need to be found.

References
[1] Mecocci A, Pannozzo M, Fumarola A. Automatic detection of anomalous behavioural events for advanced real-time video surveillance[C]. IEEE International Symposium on Computational Intelligence for Measurement Systems & Applications. IEEE, 2003.
[2] Duque D, Santos H, Cortez P. The OBSERVER: An Intelligent and Automated Video Surveillance System[C]. International Conference on Image Analysis and Recognition (ICIAR 2006) pt.1. University of Minho, Department of Information Systems, Campus de Azurem, 4800-058 Guimaraes, Portugal, 2006.
[3] Hori C, Hori T, Lee T Y, et al. Attention-Based Multimodal Fusion for Video Description[C]. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.
[4] Zhuang Yueting, Wu Yi. Video Directory—a New Method of Video Structure[J]. Pattern Recognition and Artificial Intelligence, 1999(4):408-415.
[5] Y Huang, B Bao, C Xu. Automatic generation of video description driven by relationship mining[J]. Nanjing Xinxi GongchengDaxue Xuebao, 2017, 9(6): 642-649.
[6] Shi Yingchun, Fang Pengfei, Zhou Xianzhong. News video structured model comprehensively using audiovisual features[J]. Computer Engineering and Applications, 2004, 40(32):99-101.
[7] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region
Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[9] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.

[10] Girshick R. Fast R-CNN[J]. Computer ence, 2015.

[11] Redmon J, Divvala S K, Girshick R B, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. computer vision and pattern recognition, 2016: 779-788.

[12] Lin Yutian, Zheng Liang, Zheng Zhedong, et al. Improving Person Re-identification by Attribute and Identity Learning[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017.

[13] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, 770-778

[14] Zhu X, Bain M. B-CNN: Branch Convolutional Neural Network for Hierarchical Classification[J]. 2017.