DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 008
http://www.cnki.net/kcms/detail/34.1294.TN.20170105.1512.002.html, published online January 5, 2017

Action Recognition in Surveillance Videos with Combined Deep Network Models

ZHANG Diankai[1], ZHAO Rui-Wei[2], SHEN Lin[1], CHEN Shaoxiang[2], SUN Zhenfeng[2], and JIANG Yu-Gang[2]
(1. ZTE Corporation, Nanjing 210012, China;
2. Fudan University, Shanghai 201203, China)

Abstract

Action recognition is an important topic in computer vision. Recently, deep learning technologies have been successfully used in lots of applications including video data for sloving recognition problems. However, most existing deep learning based recognition frameworks are not optimized for action in the surveillance videos. In this paper, we propose a novel method to deal with the recognition of different types of actions in outdoor surveillance videos. The proposed method first introduces motion compensation to improve the detection of human target. Then, it uses three different types of deep models with single and sequenced images as inputs for the recognition of different types of actions. Finally, predictions from different models are fused with a linear model. Experimental results show that the proposed method works well on the real surveillance videos.

Keywords

action recognition; deep network models; model fusion; surveillance video

1 Introduction

Action recognition in surveillance videos has long been an important research topic in the computer vision community. Automatic or machine aided surveillance technologies can be widely used in public areas like airport, banks, shops, etc.

The currently used action analysis methods usually contain the following steps: detection, tracking and recognition. Traditionally, some handcrafted visual features are required to be extracted from the video in order to make further computation for recognition. These traditional features include color histograms, scale-invariant feature transform (SIFT), histogram of oriented gradient (HoG), etc. The detection, tracking and recognition algorithms usually rely on these kinds of computed feature values. However, these manually designed features may suffer from their limitation in describing complicated actions in complex environment. Therefore, recognition algorithms relying on this kind of features do not always work very well in many real applications [1]–[3]. In recently years, features automatically learnt from deep neural networks are widely used in the computer vision community because of their great successes in many real applications [4], [5]. So far, there have been some work concentrating on applying deep learning methods on video analysis applications, for example, 3D convolutional neural networks (3D-CNN) [6], recurrent neural network [7] and two-streams models [8], [9]. All these methods highlight on general framework for video analysis. While for the specific problem of action recognition in the surveillance videos, there exists room for further improvement.

For example, the recognition algorithm for different types of human actions could be treated in different ways. Some actions could be comparatively easy to be recognized by single frames because of their characteristic appearance. Such actions include fighting with others or riding a bicycle. However, some other types of actions may not be easy to distinguish by merely single frames. For instance, some frames of walking and running, especially jogging, may look very similar in appearance. In order to distinguish these actions, the temporal information that describing the motion of the actions is more helpful. Therefore, the single model is not adequate to make classification for all types of human actions in surveillance videos. A composite method that combines both single image based recognition and images sequence based models could potentially be a good solution to this problem.

Motivated by the issues discussed above, we propose novel action recognition for surveillance videos in this paper. The proposed method contains improvement modifications in human detection and tracking methods, as well as a novel integrated action recognition strategy with the help of three different types of deep neural network models. The novelty and advantages of the proposed algorithm are as follows.

The detection algorithm in the proposed algorithm uses true detections from the detector and the motion points extracted from foreground motion information to compensate for miss detections. This modification could tackle with the issue of rapid appearance changes and occlusion of objects that often occur in the surveillance videos.

In the proposed network framework for action recognition, a novel fusion strategy from the spatial frames based network model and from the temporal frames based network model is

This work is supported by ZTE Industry-Academia-Research Cooperation Funds.

used to make final action recognition. Specifically, models of different structures are used to tackle with the different types of actions in surveillance videos, considering their static and dynamic properties. In the end, recognition results from both the models are fused to produce the final action classification result. Specifically, the proposed recognition algorithm relies more on a sufficiently sophiscated network model based on single frame input for some action that is distinguishable from static image frame like fighting among people or riding a bicycle. While for human actions not very distinguishable from single input image, the proposed algorithm would switch to rely more on a temporal model based on stacked image frames to provide more accurate action prediction results. For both spatial and temporal recognition models, the input image patches are expanded to the minimal square regions that completely cover the detection boxes of the objects.

Moreover, the cascade classification strategy is applied to the spatial recognition network model and temporal network model to improve the algorithm efficiency. To realize this, some thresholds are set on some predicted scores from the spatial recognition network. If the predicted scores on these layers are higher than the specified thresholds, the corresponding target object belongs to that specific action. In this way, the recognition process of the succeeding temporal network is skipped by the algorithm. The advantage is that in the prediction process, more computing resources could be saved on recognizing these simple actions and runtime efficiency could be improved.

Besides, a special network was applied to provide more accurate judgment on discerning whether detected image regions contain real foreground content or are false alarms from the detection algorithm instead. To realize this, the proposed algorithm applies a comparatively simple structured network to the detection regions in the image frame and its only aim is to check whether the image regions are real foreground or not. We intentionally set low threshold to this model so that only background regions with very high confidence are filtered out and the chances of missing any real foreground objects in the video are accordingly reduced. As this simple network model is applied only on the regions returned by the detection and tracking algorithm, the computation burden introduced by this background filtering network is in fact very limited.

2 The Approach

2.1 General Framework

In general, the proposed method first receives the captured video data as input. Then improved human detection and tracking algorithms are used to detection human regions in the original videos. Based on the detection and tracking results, proposed action recognition algorithms are applied to predict the action labels for every object obtained.

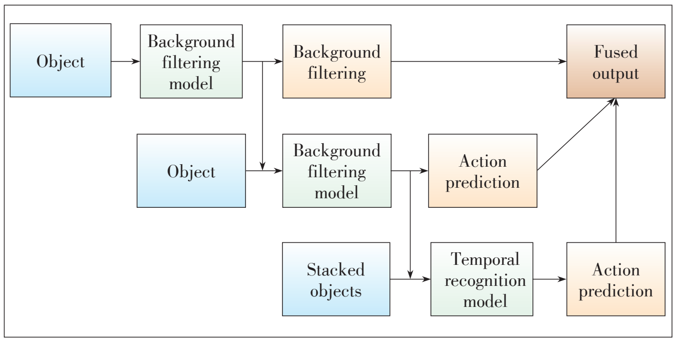
The first step is to detect and track persons in the video. We adopt the widely used tracking-by-detection method to achieve this. Specially, we calculate the motion points of the figures during the person detection and use these motion points extracted from foreground motion information to compensate for miss detections in the video.

The recognition module includes three different neural networks (**Fig. 1**). The proposed recognition algorithm is based on the neural network models to distinguish different actions in the videos. All the contextual information of the target person in the video is utilized. The proposed recognition algorithm uses both spatial and temporal information in the video sequences to achieve better recognition results.

The first network model (the upper row of Fig. 1) is used to further filter non-human patches returned by the detection and tracking algorithm. This network is comparatively simple in structure and takes non temporal data as its input. If the obtained prediction by this network shows that the input image patch is of high confidence of background image, then the algorithm will skip all the rest part of the recognition module.

If the obtained prediction by the first background filtering network shows that the input image patch is not of high confidence of background image, this image data will be sent into the second action recognition network model in the middle row in Fig. 1. This recognition model is much more complicated in structure to help discern different types of actions in each image patch received. Different from the background filtering network, this recognition network expands the input image patch to square size region on the original frame, so that it includes more contextual information to improve recognition actions on some action classes. If the output of this network shows that the input data belongs to some action with a high confidence, then the algorithm will take the corresponding action as the final prediction.

If the output of the spatial recognition network fails to give prediction to some action with high confidence, the designed algorithm will rely on the results of object tracking and use a temporal action recognition model that takes stacked sequence of image patches of some object to recognize its action, as shown in the bottom row in Fig. 1. Because this model takes the stacked sequence of image patches as input, it is better at capturing motion information of the target object. Similar to the



▲Figure 1. Network fusion for action recognition.

spatial network, this temporal network requires the input image patches to be expanded to the minimal square bounding box covering the original detection region to include more contextual information.

Finally, the results from the spatial and temporal networks are fused by a linear model to get the final predicted scores on each action class concerned.

Each component of the proposed method is described in detail in the rest of this section.

2.2 Foreground Motion Information Extraction

The first step in the designed algorithm is to detect and track human regions in a surveillance video. The purpose here is to track every person appeared in the surveillance video in real time. Our method is based on the tracking-by-detection philosophy and consists of three parts: foreground motion information extraction, object detection and object tracking. Our object detection and tracking is aided by foreground motion information.

First, we adopt a background subtraction method called ViBe [10] to filter the static background of video frames. Then edge and contour detection is performed on the resulting image. We only keep the centers of contours to represent the detected objects for simplicity.

So far, these centers represent the moving parts of our objects in a single frame. We made the centers more informative by tracking them. We adopt the Kalman Filter [11] to track these center points so that every point is assigned by an ID during its life time. These center points with ID are called "motion points" and play an important role in improving the performance of detection and tracking.

2.3 Object Detection

There are many successful person detectors proposed in previous years [12]. In this work, we adopt the aggregated channel features (ACF) [13] detector due to its advantages on both speed and accuracy. We train the detector on pedestrian dataset consisting INRIA Person and our labeled video. We also propose two methods, filtering and compensation, to improve the detector in our application.

We use filtering to deal with obvious yet common false positives, such as lamp posts, trees and traffic cones. Because our surveillance cameras are stationary, the size of a person is usually linearly related to where he appears in the region. Therefore, objects like traffic cones and trees can be filtered out easily.

In video, rapid appearance change and occlusion of objects often happen. So our person detector cannot detect a person in every frame of a video. We propose a novel method by using the true detections from the detector and the motion points extracted from foreground motion information to compensate for missed detections.

Supposing there is detection D from the detector in frame F, we use the following procedure to determine if the object contained in it is lost in frame $F+1$ or not.

1) Let N be the set of motion points contained in D, $n=|N|$, $k=0$.

2) For every motion point in N, if we can find the motion point in frame $F+1$ with the same ID and belongs to some bounding box, put it into set S, otherwise $k=k+1$.

3) If $k/n<0.5$, most of the motion points are still in some bounding box in frame $F+1$. In this case, we think the bounding box that contains the most points from S is the detection that matches D (**Fig. 2a**).

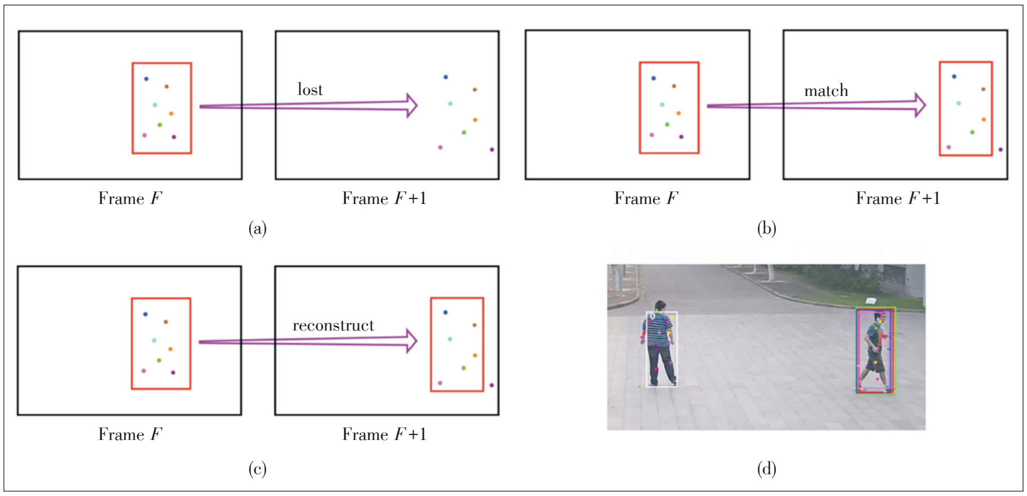
4) If $k/n>0.5$, it is very likely that detection D is lost in frame $F+1$ (**Fig. 2b**). We use set S to reconstruct it.

5) S contains the "moved" motion points of D. For every point in S, we put a bounding box that has the same relative position to it as D around it. These boxes represent the place where D should be in frame $F+1$ with respect to the movement of every motion point. We then find a bounding box that covers all these boxes, rescale it to the size of D, and make it the reconstructed detection of D in frame $F+1$ (**Fig. 2c**).

In **Fig. 2d**, the red box is the detection from last frame, green box is the covering box and blue box is the compensated detection. The idea here is similar to object tracking, however, it only works for moving objects.

2.4 Object Tracking

Our person tracker is an online tracker that integrates the detector's responses and appearance templates. The tracking is performed by assigning every target a tracker with the Kal-



▲Figure 2. Illustration of human detection with motion information.

man Filter and the target's appearance templates. At each frame, the tracker will predict the target's new position with the Kalman Filter, calculate a voting map with appearance templates, do mean-shift tracking on the voting map and update the Kalman Filter. Then detections are associated with every tracker. Every tracker updates its appearance template when assigned a new detection. We also adopt the idea of tracker hierarchy [14] to select the most effective tracking strategy.

The result of human detection and tracking is saved as a sequence of object ID and object bounding box, denoted as

$$O(i,t)=\left\{I_t^{(i)},R_t^{(i)}\right\}, \tag{1}$$

where $O(i,t)$ is the target information of object ID i at time step t in the video, and $I_t^{(i)}$ is the image content of this object detected at time step t. While $R_t^{(i)}$ is the bounding box information about this object at time step t in the video. The values in $R_t^{(i)}$ is a four dimensional vector (x,y,w,h) marking the upper left corner location, width and height of the bounding box.

2.5 Background Filtering by Simple Spatial Neural Networks

During the real time recognition process, the extracted target image region is input into the background filtering network model. As shown in **Fig. 3**, the background filtering network is based on single image input, which is the image data returned by the detection and tracking algorithm. Sample input images patches for this network (**Fig. 4**) may include both real human figures and some false alarms from the previous detection and tracking step.

After the input layer, there exist several convolution layers to extract the visual features from the image. To improve the training effect, the rectified linear units (ReLU) and max pooling layer are connected to each convolution layer in the network. Afterwards, several fully connection layers are appended to the last convolution layer to further encode the learnt visual features. In all fully connection layers except the last one, ReLU layers and dropout layers are applied to achieve non-linear transformation of feature values and generalization improvement. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs whether the input image data is foreground human figure or background image. The number of convolution and fully connection layers could be set to meet the computation capacity of the deploying hardware.

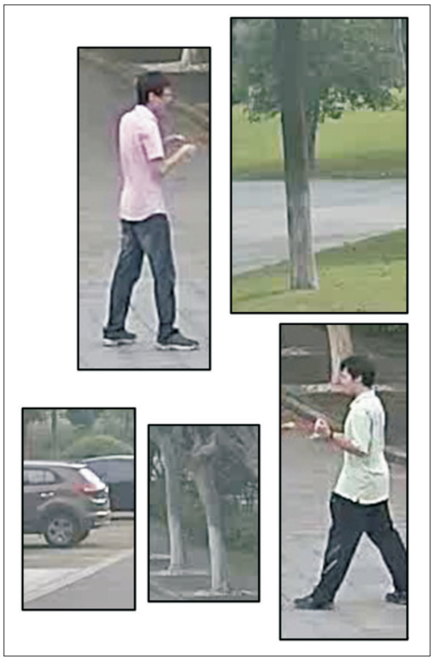
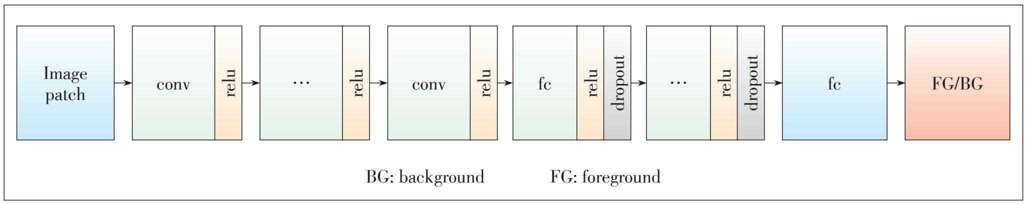


Figure 4. ▶ Sample input data to the background filtering network model.



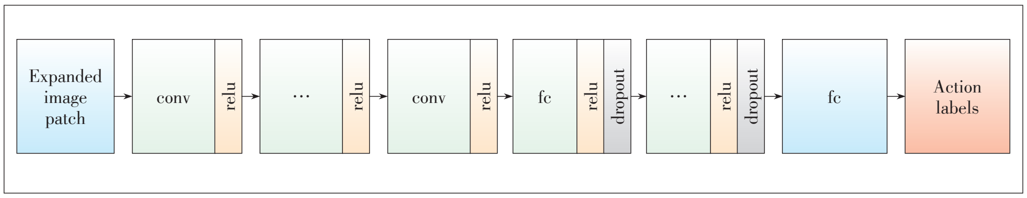
▲Figure 3. Structure of the background filtering network model.

After this step, the algorithm could filter out the non-foreground areas returned by the detection and tracking algorithm. Therefore, when used in combination, the detection and tracking algorithm could adjust the decision threshold to allow more potential foreground images to pass in order to reduce the missing rate of real human objects. Because this model is computed only on the image regions returned by the detection and tracking algorithm rather than the sliding windows on the whole image, no intensive computation consumption is introduced by this network model during running time.

2.6 Spatial Action Recognition Model Based on Single Image

After the background filtering, only foreground figure images are supposed to be obtained. In this step, the algorithm introduces a more complicated recognition network in structure to recognize the action type in the image region. In this network (**Fig. 5**), context information about the target region is also taken into consideration to improve the recognition performance.

The general network framework is very similar to the background filtering network, but with more complicated network structures like more convolution layers and fully connection layers. The input layer of this network expands the input image regions to cover more context information about the target object. Specifically, it expands the detection region to

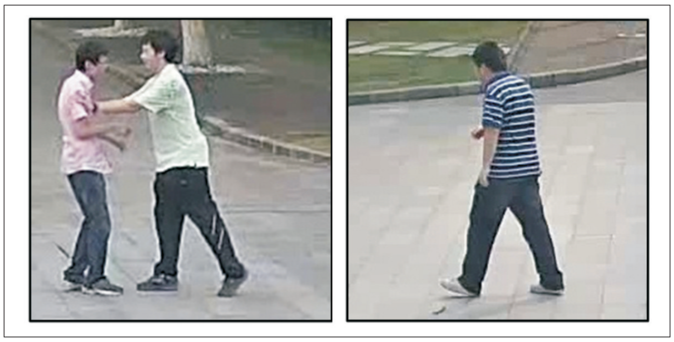


▲Figure 5. Structure of the spatial recognition network model.

the minimal square areas that cover the original object regions. Some sample input image data to the recognition network are shown in **Fig. 6**. As we can see, when expanded to square regions, the input image covers more contextual information to the target human. For example, in the upper image of Fig. 6, the person in red shirt is included as the context of the target person in green at the center of the whole image patch. This contextual information would be very helpful to recognize that the target person's action is to fight with another person. In the lower image of Fig. 6, the expanded square box image region would cover more road context with regard to the person in center, giving more indication that person is walking alone the road and not possible in fighting with someone else.

After the input, the algorithm also uses convolution layers to extract the visual features from the image, together with ReLU and the max pooling layer connected to each convolution layer in the network. Afterwards, fully connection layers are appended to the last convolution layer to further encode the learnt visual features. Also, in all fully connection layers except the last one, ReLU layers and dropout layers are applied to achieve non-linear transformation of feature values and generalization improvement. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs the probability of some specified action classes. The number of convolution and fully connection layers could be set to meet the computation capacity of the deploying hardware. In order to learn more complicated feature representation, this network uses more convolution layers and fully connection layers than the background filtering network model.

As mentioned, the output of the spatial recognition network is a vector of action class probabilities. It is then possible to set certain a decision threshold. When the output value is greater than the specified decision threshold, the algorithm will take this high confidence class label as the final predicted action class for the target person.



▲Figure 6. Sample input data to the spatial recognition network model.

2.7 Temporal Action Recognition Model Based on Stacked Images

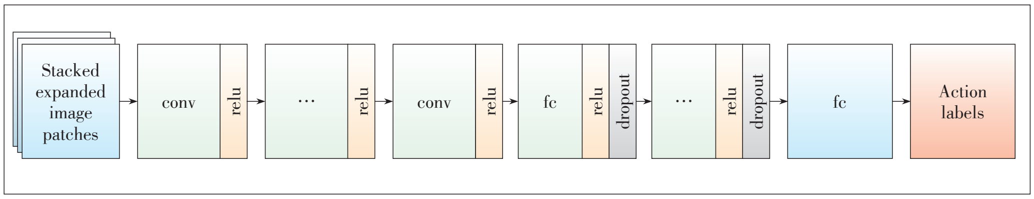
If the previous spatial action recognition model fails to output the action class prediction probabilities higher than the specified threshold, the temporal action recognition model will be introduced to help action predictions. Different from the spatial network model, this temporal model takes stacked image patches as its input data. The stacked image patches are ordered sequence of the target object in the video. Because these sequenced images contain the object information at different time, this temporal recognition model has more advantages at motion information capture. This would be advantageous in distinguishing those actions featured by human motion at a small period of time. The general framework of the proposed temporal model is depicted in **Fig. 7**.

In order to get the stack motion images of a certain object, the temporal recognition model utilizes the tracking results obtained by the tracking algorithm. The stacked image patches could be represented as

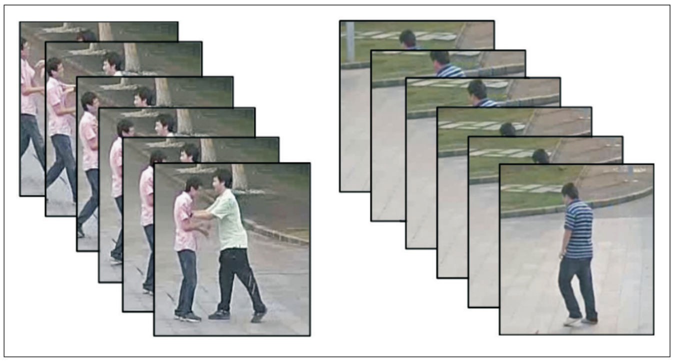
$$input(M_3) = \left[I_{t-k}^{(i)}; \ldots; I_t^{(i)} \right], \tag{2}$$

where $I_t^{(i)}$ is the image content of this object detected at time step t. This means that the stacked image patches come from the current and history object appearances in a short time interval. In the temporal network, each convolution filter size in the first convolution layer is h by w by c, k, where h, w is the height and width of the filter, c is the number of input image channels, and k is the number of stacked image patches in the input data. Some sample stacked image patches are illustrated in **Fig. 8**. These image patches are also expanded to minimum covering square areas to include contextual information.

As the image patches input into the convolution layers are stacked according to the ordered temporal sequence, content of the first few channels in the input data must contain actions happening earlier than those on the bottom channels. This makes the parameters of the first few channels in the convolution layers always to be applied to action patches happening earlier than those of the last few channels during the training phase. As the interactions of earlier behavior and later behavior are modeled by weighted addition of products between convolution parameters and image content at different time stages in sequenced order, temporal information is automatically and properly learned by this layer architecture in the process of error back-propagation algorithm during network training. Besides, the time span between the image patches also affects the



▲Figure 7. Structure of the temporal recognition network model.



▲Figure 8. Sample input data to the temporal recognition network model.

tuning of convolution parameters, which is part of the learning of temporal information.

After the input, the algorithm also uses convolution layers to extract the visual features from the image, together with ReLU and the max pooling layer connected to each convolution layer in the network. Afterwards, fully connection layers are appended to the last convolution layer to further encode the learnt visual features. Also, in all fully connection layers except the last one, ReLU layers and dropout layers are applied. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs the probability of some specified action classes.

2.8 Fusion of Recognition Models

After the spatial and temporal recognition models are trained, we apply a fusion strategy to better combine the action predictions from them. The basic idea is to use a linear function to aggregate both classification scores returned by the two models. Therefore, for each action to be recognized, the final prediction score after fusion could be written as

$$p(action) = w_S p_S(action) + w_T\, p_T(action) + b, \tag{3}$$

where $p_S(action)$ and $p_T(action)$ are the action prediction scores returned by the spatial recognition model and temporal recognition model respectively. Besides, w_S and w_T are the weights balancing these two scores, and b is the bias parameter. The values $of\, w_S$, w_T $and\, b$ could be learnt on the validation dataset by using the linear regression model.

3 The Experiment

All the data used in the experiment are collected from outdoor surveillance videos. These videos include actions of walking, running, kicking, fighting and riding.

In order to train the models for recognition, we manually labeled the human actions in a set of videos. In total, the labeled video length is more than 30 hours and we collected more than 5000 labeled trails of human figures in the videos. The tool we used to annotate the videos is the open-sourced ViPER software.

The recognition models are trained from scratch on some existing large scaled image and video datasets and fine-tuned on the surveillance data. For the spatial and temporal recognition models, we first use ImageNet image data to train the structured models from randomly initialized network parameters. Separate and stacked image frames extracted from UCF-101 dataset are then used to fine-tune the networks for the first round. Finally, the pre-trained networks are used to fine-tune our surveillance data with the action labels of interest. During the training, we adopt image patch mirroring, different scaling ratios and random cropping techniques to improve model generalization.

In our experiment, we set five convolution layers and three fully connection layers for the first background filtering network. For the second spatial recognition network model, we set 16 fully connection layers for the three in the first background filtering network. For the third temporal recognition network model, we set five fully connection layers for the three in the first background filtering network to balance recognition performance and time efficiency.

The test set contains videos of around 22 minutes. The obtained detection rate of the proposed method is 0.9751 on the test set. The average recognition precision of the recognition algorithms on the test set is 0.9251. **Table 1** compares recognition precisions on the tested action classes and the performance of the baseline algorithm with ordinary single image based CNN model. As we can see in the table, our proposed method outperforms the baseline model in predicting all the evaluated action classes. For those actions with comparatively distinguishing static appearances such as fighting and riding,

▼Table 1. Comparision of recognition precisions

Action	CNN	Our method
Walk	0.7961	0.8500
Run	0.8332	0.8859
Kick	0.9246	0.9535
Fight	0.9383	0.9596
Ride	0.9558	0.9764
Average	0.8896	0.9251

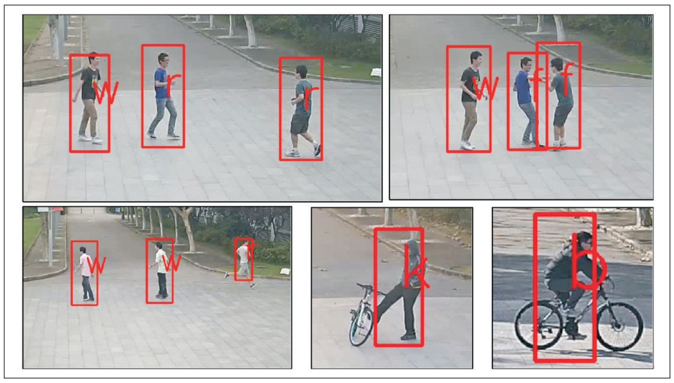
CNN: convolutional neural networks

the performance of the baseline CNN method is comparatively close to ours. While for some ambiguous actions more dependent on temporal information to distinguish, such as walking and running, our method outperforms the baseline method by a larger margin. Therefore, fusion with temporal information with the stacked images based model is able to improve the classification performance for different kinds of actions.

Fig. 9 shows some recognition results in the test. The red bounding boxes in the figure marks the detection area on the human figures. Different characters stand for different action label predicted. Here "w" stands for walking, "r" stands for running, "k" stands for kicking, "f" stands for fighting, and "b" stands for riding.

4 Conclusions

In the paper, a novel action recognition algorithm is proposed to deal with the automatic video surveillance problems. The proposed algorithm mainly features in motion compensation during detection, background filtering before recognizing, different fusion settings of spatial and temporal network models for different types of action to recognize. The experiment shows that the proposed method produces good results on the surveillance video data.



▲**Figure 9. Examples of the recognition results.**

References

[1] O. Russakovsky, J. Deng, H. Su, et al. (2014, Sept. 02). *ImageNet large scale visual recognition challenge* [Online]. Available: https://arxiv.org/abs/1409.0575

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012. doi: 10.1109/TASL.2011.2134090.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. (2013, Oct. 17). *Distributed representations of words and phrases and their compositionality* [Online]. Available: https://arxiv.org/abs/1310.4546

[4] A. Karpathy, G. Toderici, S. Shetty, et al., "Large-scale video classification with convolutional neural networks," presented at IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[5] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. (2015, Apr. 07). *Modeling spatial-temporal clues in a hybrid deep learning framework for video classification* [Online]. Available: https://arxiv.org/abs/1504.01561

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013. doi: 10.1109/TPAMI.2012.59.

[7] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. (2015, Jun. 05). *Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video* [Online]. Available: https://arxiv.org/abs/1506.01911

[8] K. Simonyan and A. Zisserman. (2014, Jun. 09). *Two-stream convolutional networks for action recognition in videos* [Online]. Available: https://arxiv.org/abs/1406.2199

[9] H. Ye, Z. Wu, R.-W. Zhao, et al., "Evaluating two-stream cnn for video classification," presented at 5th ACM on International Conference on Multimedia Retrieval, New York, USA, 2015.

[10] O. Barnich and M. Van Droogenbroeck, "ViBe: a universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011. doi: 10.1109/TIP.2010.2101613.

[11] E. V. Cuevas, D. Zaldivar, and R. Rojas. (2005). *Kalman filter for vision tracking* [Online]. Available: http://www.diss.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000000473/2005_12.pdf

[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012. doi: 10.1109/TPAMI.2011.155.

[13] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532– 1545, Aug. 2014. doi: 10.1109/TPAMI.2014.2300479.

[14] J. Zhang, L. Lo Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," in *IEEE Ninth International Conference on Advanced Video and Signal - Based Surveillance (AVSS)*, Beijing, China, pp. 379– 385, 2012. doi: 10.1109/AVSS.2012.51.

Manuscript received: 2015-12-08

Biographies

ZHANG Diankai (zhang.diankai@zte.com.cn) received his BE degree in electronic information engineering and MS degree in signal and information processing from Nanjing University of Posts and Telecommunications (NUPT), China in 2006 and 2009. He is a senior video and image algorithm engineer of ZTE Corporation. His research interests include video and image processing, pattern recognition, and computer vision.

ZHAO Rui-Wei (rw.du.zhao@gmail.com) received BS degree in 2005 and MS degree in 2009 from Tongji University, China. He is currently a PhD candidate in the School of Computer Science at Fudan University, China. His research interests include deep learning methods for image and video recognition.

SHEN Lin (shen.lin2@zte.com.cn) received her BE degree in communication engineering and MS degree in computer application from Nanjing University of Science and Technology (NUST), China in 2007 and 2009. She is a senior video and image algorithm engineer of ZTE Corporation. Her research interests include video and image processing, pattern recognition, and computer vision.

CHEN Shaoxiang (forwchen@gmail.com) is currently a research student in the School of Computer Science at Fudan University. His research interests include object detection and tracking in video data.

SUN Zhenfeng (zf_sun@foxmail.com) is currently studying at School of Computer Science, Fudan University towards the degree of Doctor of Engineering. His research interests include multimedia software systems and computer vision.

JIANG Yu-Gang (yugang.jiang@gmail.com) received the PhD degree in computer science from the City University of Hong Kong, China in 2009. During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, USA. He is currently a full professor of computer science at Fudan University. His research interests include multimedia retrieval and computer vision. He is one of the organizers of the annual THUMOS Challenge on Large Scale Action Recognition, and served as a program chair of ACM ICMR 2015. He is the recipient of many awards, including the prestigious ACM China Rising Star Award (2014).