Movement Human Actions Recognition Based on Machine Learning

https://doi.org/10.3991/ijoe.v14i04.8513

Honghua Xu, Li Li(✉), Ming Fang
Changchun University of Science and Technology, Changchun, Jilin, China
dilihan0128@163.com

Fengrong Zhang
Northeast Normal University, Changchun, Jilin, China

Abstract—In this paper, the main technologies of foreground detection, feature description and extraction, movement behavior classification and recognition were introduced. Based on optical flow for movement objects detection, optical flow energy image was put forward for movement feature expression and region convolutional neural networks was adopt to choose features and decrease dimension. Then support vector machine classifier was trained and used to classify and recognize actions. After training and testing on public human actions database, the experiment result showed that the method could effectively distinguish human actions and significantly improved the recognition accuracy of human actions. And for the different situations of camera lens drawing near, pulling away or slight movement of camera, the solution had recognition effect as well. At last, this scheme was applied to intelligent video surveillance system, which was used to identify abnormal behavior and alarm. The abnormal behaviors of faint, smashing car, robbery and fighting were defined in the system. In running of the system, it obtained satisfactory recognition results.

Keywords—optical flow, support vector machines, convolutional neural networks

1 Introduction

The recognition technique of moving human behavior detects and traces moving human objects coming from video sequences by computer visual technique. Then we can further understand the human behaviors. This process is a complicated task, and it integrates some research fields including image process, machine learning, pattern recognition etc. The recognition technique of moving human behaviors is widely applied in intelligent video surveillance, human-computer interaction, action analysis, and video retrieval etc.

In the field of intelligent video surveillance, the traditional surveillance systems can't fully play a role in real-time and active supervision because they only record the video data output from cameras. When something happens, security guards will

search for the recorded data to find the truth. Sometimes it is too late. Intelligent video surveillance system [1-2] will totally change the mode of security guards' surveillance and analysis. It can solve the problem of human resources waste as well as realizes the real-time alarm.

In the field of human-computer interaction, intelligent human-computer interaction is to get rid of the limitations of the traditional input devices such as keyboard, mouse etc. It aims to have natural interaction with human through voice, actions etc. But because the voice subjects to the distance and the surrounding noise, whereas human action is without limitations, computers must understand human actions in the intelligent environment.

In the field of movement analysis, human actions analysis and recognition can be applied in many aspects for example, in sports and dancing training, the visual method adopted to set up human geometric model through tracing and analyzing joint movement and rectifying the trainee's action. In the field of medicine, human normal gait can be modeled based on visual movement analysis [3-4] in order to judge the condition of the leg injury or the extent of cleft malformation. Besides, gait analysis can also be considered as the feature used in distant identity recognition.

In the field of video retrieval, for effectively manage and visit massive video database, videos need to be retrieved based on the content. Video streams include many meaningful incidents. Through analyzing and recognizing the movement information of these incidents, video content can be correctly labeled.

2 Research background

The key technique of human action recognition based on computer vision is to describe and understand human behaviors by means of computer vision. [5] From the aspect of technology, after inspecting the wide research of overseas researchers and some relative businessmen, we may find that the recognition process of moving objects includes foreground target detection, feature extraction and description, behavior analysis and recognition. Among them, moving target detection is referred to as picking out the interested moving targets from video sequence. This process is the basic link. The correctness of moving target detection exerts direct influence on the analysis and recognition of human behavior. Feature extraction and description is to select representative features to describe moving targets. Selecting proper features and description methods not only can reduce dimension to decrease calculation but also can distinguish between different types of target behaviors to secure the accuracy of target classification and recognition. Human behavior recognition refers to classifying the targets according to the extracted features and then further analyzes the targets in terms of predefined behavior mode.

These three basic links are interrelated. The processing effect of the former link will have an important impact on the following processing. The research content of each link crosses each other but at the same time these three links have their own research values and focuses.

2.1 Mmoving target detection

Moving target detection [6-7] is to separate video change areas from background images, i.e. to correctly split out moving target areas or contours. It is critical for the following processing to effective segmentation as the pixels of the corresponding moving areas are only considered in the following processing. Currently used methods are background difference method, adjacent frames difference method and optical flow method and so on.

2.2 Feature description and extraction

Feature description and extraction [8]is to display the features of input image or video by means of some form such as a value, a vector or a function etc. As for the feature extraction of human movement object, it is to display the features of human movement in video sequence by means of some form such as a value or a vector etc. Generally speaking, larger feature quantity can lead to larger data volume, high dimension data, and increased computation amount and computation complexity. Therefore, when choosing features, we must consider extracting essential and focused ones of the prospect target according to various application scenes. The chosen features should sport better classification ability and could better distinguish different targets. As for the description of features, large amount of high dimension data should be avoided in order to obtain simpler and faster computation. The common methods of feature extraction and description are: target area, centroid, velocity, angle, track etc.

2.3 Classification and recognition for human actions

Human behavior recognition [9-11] is to recognize the behaviors of individuals between humans or between humans and scenes by means of the methods of continuously observing of moving objects, collecting and classifying data of object behaviors, template matching etc. Human behavior recognition belongs to the advanced processing link of moving object behavior. There are several common classification and recognition methods: HMM, DBN and CRF etc. In this paper, the author adopts deep convolutional neural networks of soft computing technique to train experimental database so as to realize the classification and recognition of target behavior in video.

3 Algorithm of moving targets detection

Using the pixel changes in image sequence and the correlation between adjacent frames, the moving information can be worked out. Optical flow is the instant velocity of phase motion on the observing image plane of moving objects. [12] In general, optical flow is obtained from the movement of the prospect target in the scene, the movement of camera or the movement of both.

3.1 The principle of optical flow

According to visual perception principle, the object movement in space embodies certain continuity. The flat image of retina projected by the object movement process actually is another continuous changing process. Hence, we may suppose the grey-scale value of pixel *(x,y)* in the image at moment is *I(x,y,t)*. Then we may use *u(x,y)* and *v(x,y)* to demonstrate the horizontal and vertical velocity components of pixel *(x,y)*. Then during a period of time t, the pixel moves from *(x,y)* to *(x+dx, y+dy)* and then the greyscale value is *I(x+dx, y+dy, t+dt)*. Suppose the intensity is constant, i.e. as for certain movement target in a group of continuous two-dimensional image sequence, the corresponding pixels of frames along the motion trajectory curve have the same greyscale values. We may consider that when *dt* is quite small, the corresponding point greyscale on image is constant, i.e.:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

Unfold the right part of the above equation by means of Taylor and get equation 2.

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x} + \frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} + O(\partial^2) \tag{2}$$

Omit the high order item (∂^2), and then we may obtain the famous optical flow constraint equation 3.

$$I_x u + I_y v + I_t = 0 \tag{3}$$

Change equation 3 into the vector form and we can get equation 4.

$$\nabla I \cdot U + I_t = 0 \tag{4}$$

I_x, I_y, I_t *a*re point *(x, y)* partial derivatives of greyscale value *I(x ,y ,t)* along three directions *x, y, t.* $\nabla I = (I_x, I_y)^T$ is the spatial gradient of the image greyscale. $U = (u, v)^T$ is optical flow. Optical flow constraint equation is also called optical flow basic equation that demonstrates the constraint relation between image greyscale gradient and optical flow. However, optical flow vector $U = (u, v)^T$ in the basic equation contains two variables, which means that the equation can't be solved. So it is unsuitable to solve the optical flow by means of basic equation. Therefore, we need to introduce more restrictions which can bring more changes to the calculation methods of optical flow. Optimum estimation method of optical flow based on gradient is a classic calculation method of optical flow.

3.2 Optimum estimation of optical flow based on gradient

Horn and Schunck introduced another restriction based on the assumption that intensity is constant, i.e. overall smooth constraint hypothesis. This hypothesis is that the motion vector of the object is locally smooth or changes slowly. Especially in rigid body movement, the neighboring pixels have the same velocity of movement, i.e. velocity is smooth. The gradient value of optical flow vector squared is the small-

est. Horn combined the basic equation with the overall smooth constraint of optical flow to solve optical flow. Since the neighboring pixels have the same motion velocity, spatial rate of change of local zone velocity is zero and the gradient of the optical flow vector should tend to be zero. Then we use laplacian operator squared of component *x* and component *y* to show the smoothness of optical flow. Smooth constraint asks for the minimized smooth constraint item *Es*.

$$E_s = \iint \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right) dxdy \tag{5}$$

Whereas basic equation asks for the minimized *Ec*.

$$E_c = \iint (I_x u + I_y v + I_t)^2 dxdy \tag{6}$$

The solution of optical flow may be concluded into the solution of variation problem.

$$E = \iint \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right) dxdy + \alpha \iint (I_x u + I_y v + I_t)^2 dxdy \tag{7}$$

Among them, *a* is a parameter of credibility to image data and constraints. *a* is the weights between error *Es* and error *Ec* to adjust optical flow. When image data is more accurate, the error that deviate the basic equation should have greater weight. At that time, parameter *a* takes a smaller value. On the contrary, when the image data constrains louder noise, the credibility of the original image data is lower and then depends heavily on smoothness constraints. Hence, *a* should takes a larger value. Seek guidance of *u* and *v* respectively from equation 7.

$$E_x^2 u + E_x E_y v = -\alpha^2 \nabla u - E_x E_t \tag{8}$$

$$E_y^2 u + E_x E_y u = -\alpha^2 \nabla v - E_y E_t \tag{9}$$

Make $\bar{u}$ and $\bar{v}$ represent the values of field *u* and field *v* and make $\nabla u = u - \bar{u}$ and $\nabla v = v - \bar{v}$. Then equation 8 and equation 9 can be changed respectively.

$$(E_x^2 + \alpha^2) u + E_x E_y v = -\alpha^2 \bar{u} - E_x E_t \tag{10}$$

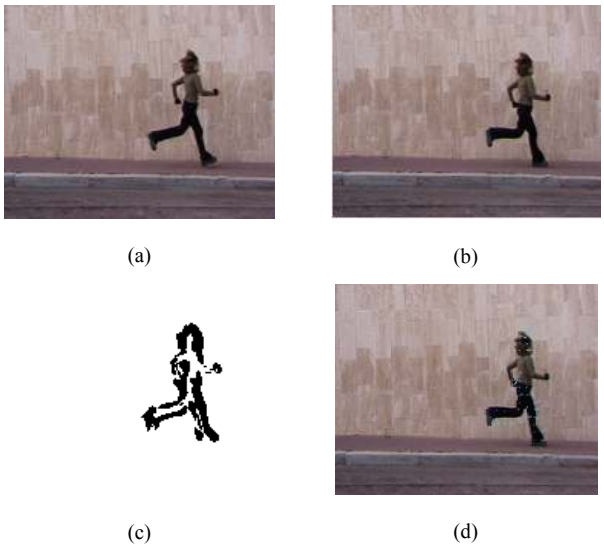
$$\left(E_y^2 + \alpha^2\right) v + E_x E_y u = -\alpha^2 \bar{v} - E_y E_t \tag{11}$$

We may obtain solution from the above equations.

$$u^{(n+1)} = \bar{u}^n - \frac{E_x [E_x \bar{u}^{(n)} + E_y \bar{v}^{(n)} + E_t]}{\alpha^2 + E_x^2 + E_y^2} \tag{12}$$

$$v^{(n+1)} = \bar{v}^n - \frac{E_y [E_x \bar{u}^{(n)} + E_y \bar{v}^{(n)} + E_t]}{\alpha^2 + E_x^2 + E_y^2} \tag{13}$$

Interaction initial values can be *u(0) =0, v(0) =0*. This is the optical flow of Horn and Schunck based on gradient overall optimization. figure 1 shows the result of optical flow detection.



(a) (b)

(c) (d)

Fig. 1. (a) 40th frame of video, (b) 41th frame of video, (c) body contours based on optical flow detection, (d) optical flow of movement body

3.3 Optical flow energy image

Motion energy image and motion history image [13] are effective ways to express human actions. Among them, MEI reflects human actions posture occurring in the region and intensity, and in a certain extent, MHI reflects the body movement posture with time changing. This is a representation based on the appearance. This method does not need any human body structure information and three-dimensional reconstruction, which is helpful to reduce the computational complexity, improve the efficiency of the algorithm and ensure the real-time requirement.

Considering differences of individual, the impact of the environment and the speed of the actions, will lead to the same action to produce a different sequence of images, so that the extracted features will be some differences.

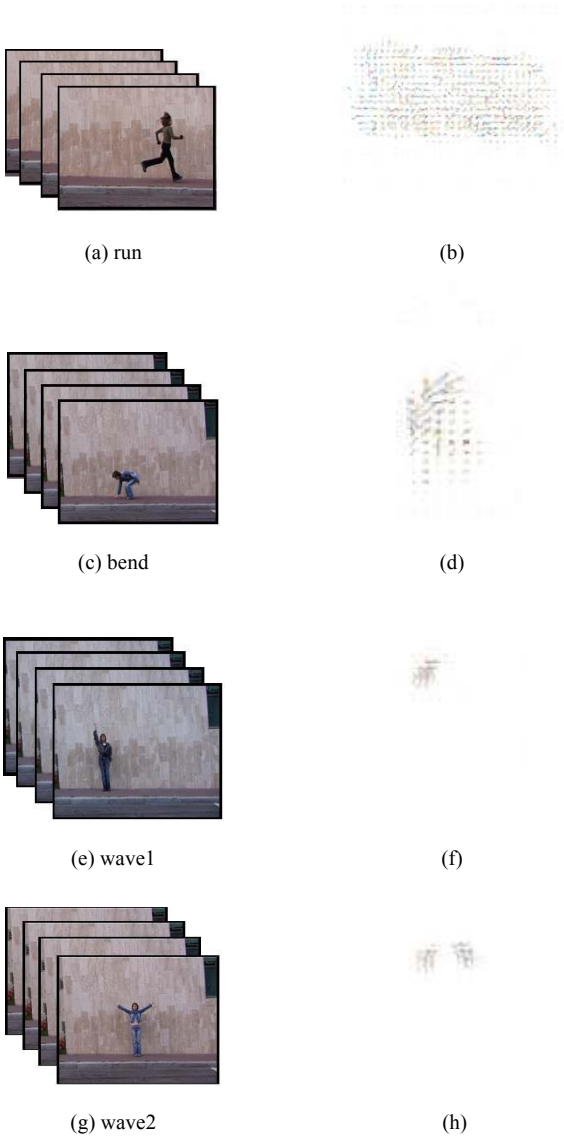
In order to deal with these differences, our method is to accumulate the motion information into one image of the video sequence frames. The motion information of a certain period of time is recorded on one frame image, and the motion cumulative energy image is constructed. We extract the features of the motion cumulative energy image as the features of human actions.

At the point of *(x, y)*, the cumulative energy image of optical flow at time t is given by Equation 14, the optical flow energy image (OFEI).

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} O(x, y, t - i) \tag{14}$$

Where $O(x, y, t)$ is the optical flow sequence representing the binary image of the human actions generation region, and the parameter τ indicates the duration of human actions.

Therefore, OFEI describes the whole human action posture occurring in the region and features. The optical flow energy images of different actions are shown in figure 2. As seeing from the diagram OFEIs, the various actions can be distinguished.



(a) run (b)

(c) bend (d)

(e) wave1 (f)

(g) wave2 (h)

Fig. 2. (a), (c), (e), (g), video segments corresponding to run, bend, wave1, wave2. (b), (d), (f), (h), corresponding optical flow energy image.

4 Convolutional neural networks as feature extrater

Deep learning [14] uses neural networks to learn useful representations of features directly from data. Neural networks combine multiple nonlinear processing layers, using simple elements operating in parallel and inspired by biological nervous systems. Deep learning models can achieve state-of-the-art accuracy in object classification, sometimes exceeding human-level performance.

Convolutional neural networks (CNN) [15-16] are a quite common tool for deep learning. They are specifically suitable for images as inputs, although they are also used for other applications such as text, signals, and other continuous responses. They differ from other types of neural networks in a few ways.

CNN is inspired by the biological structure of a visual cortex, which contains arrangements of simple and complex cells. These cells are found to activate based on the sub regions of a visual field. These sub regions are called receptive fields. Inspired by the findings of this study, the neurons in a convolutional layer connect to the sub regions of the layers before that layer instead of being fully-connected with other types of neural networks. The neurons are unresponsive to the areas outside of these sub regions in the image.

These sub regions might overlap, hence the neurons of a CNN produce spatially-correlated outcomes, whereas in other types of neural networks, the neurons do not share any connections and produce independent outcomes.

In addition, in a neural network with fully-connected neurons, the number of parameters (weights) can increase quickly as the size of the input increases. A convolutional neural network reduces the number of parameters with the reduced number of connections, shared weights, and down sampling.

A CNN consists of multiple layers, such as convolutional layers, max-pooling or average-pooling layers, and fully-connected layers.

The architecture of the CNN model is shown in figure 3. The model contains eight parameter layers. The first five are convolutions and the latter three are fully connected.

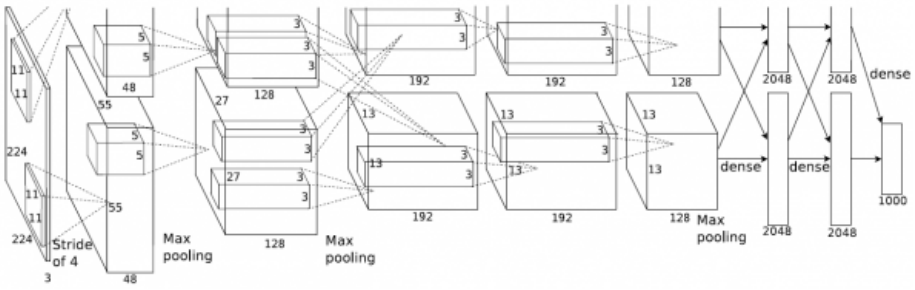


Fig. 3. The architecture of the CNN

The architecture visually demonstrates the tasks that two GPUs are responsible for. The convolution layers of the upper and lower parts in the figure are each mounted on

a GPU. The network of the two GPUs only interacts at some level. The architecture has the following characteristics.

1. Use rectified linear units as non-linearized neurons to improve the training speed.
2. Use multiple GPUs for parallel job training.
3. Local response normalization is used to prevent saturation.
4. Use overlapping pooling technology to prevent over fitting.

5 Support vector machine as classifier

A support vector machine (SVM) [17-18] is a supervised learning algorithm that can be used for binary classification or regression. Support vector machines are popular in applications such as natural language processing, speech and image recognition, and computer vision.

A support vector machine constructs an optimal hyper plane as a decision surface so that the margin of separation between the two classes in the data is maximized. Support vectors refer to a small subset of the training observations that are used as support for the optimal location of the decision surface.

Support vector machines fall under a class of machine learning algorithms called kernel methods and are also referred to as kernel machines.

Training for a support vector machine has two phases:

1. Transform predictors to a high-dimensional feature space. It is sufficient to just specify the kernel for this step and the data is never explicitly transformed to the feature space. This process is commonly known as the kernel trick.
2. Solve a quadratic optimization problem to fit an optimal hyper plane to classify the transformed features into two classes. The number of transformed features is determined by the number of support vectors.

Only the support vectors chosen from the training data are required to construct the decision surface. Once trained, the rest of the training data are irrelevant.

5.1 Separable data

We can use a support vector machine (SVM) when your data has exactly two classes. SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other. The best hyper p lane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyper plane that has no interior data points.

The support vectors are the data points that are closest to the separating hyper plane; these points are on the boundary of the slab. The figure 4 illustrates these definitions, with + indicating data points of type 1 and - indicating data points of type -1.

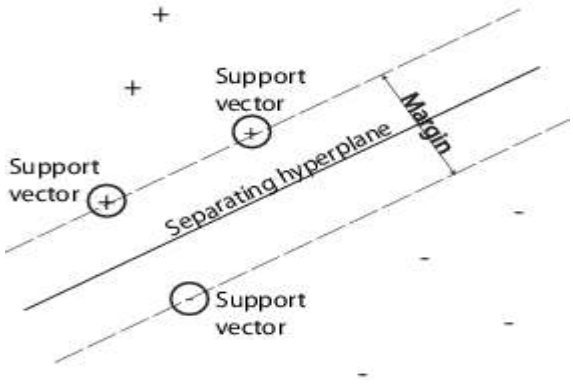


Fig. 4. The support vectors

Our data might not allow for a separating hyper plane. In that case, SVM can use a soft margin, meaning a hyper p lane that separates many, but not all data points. There are two standard equations of soft margins. Both involve adding slack variables and a penalty parameter.

5.2 Nonlinear transformation with kernels

Some binary classification problems do not have a simple hyper plane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyper plane.

This approach uses these results from the theory of reproducing kernels. Popular kernels used with SVMs include Gaussian or radial basis function, linear, polynomial and sigmoid.

The mathematical approach using kernels relies on the computational method of hyper lanes. All the calculations for hyper plane classification use nothing more than dot products. Therefore, nonlinear kernels can use identical calculations and solution algorithms, and obtain classifiers that are nonlinear. The resulting classifiers are hyper surfaces in some space S, but the space S does not have to be identified or examined.

6 The experiment of human actions recognition

The experiment is to serialize the video fragments, extract the optical flow of the moving object in each frame firstly, and accumulate and superimpose the optical flow to get the optical flow energy image of the moving target. Then, the optical energy image is used as the input of the convolution neural network, and the feature selection and dimensionality reduction are carried out. At last, the SVM classifier is trained by using the features extracted by convolution neural network to realize the classification of human actions. The specific experiment process is shown in figure 5.

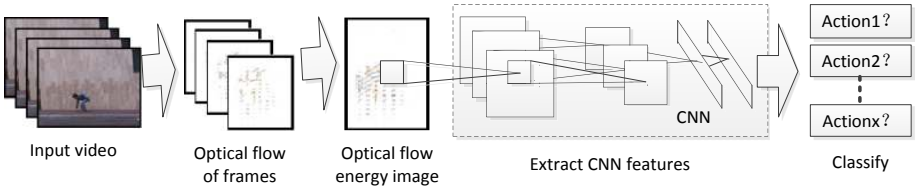


Fig. 5. The experiment process of human actions recognition

6.1 Database

This experiment was conducted using the human actions database of Weizmann and KTH. The database of Weizmann includes a total of 90 videos, and each person has 10 different actions (bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2). Video background, perspective and camera are stationary. The database of KTH includes six kinds of actions (walking, jogging, running, boxing, hand waving, hand clapping). It is performed by 25 different people, respectively, in four scenes, a total of 599 videos. The background is relatively static, the lens is closer, far away or the camera has a slight movement.

6.2 Expression of movement human actions

Because the optical flow can detect the movement of the object without knowing any information of the scene in advance, and can accurately calculate the speed of the moving object. It is still effective when the camera is in motion. In the experiment, we used optical flow method in the third part for movement foreground detection. Human action expression uses the optical flow energy image. In the experiment on the Weizmann database, the optical flow energy image is a good way to distinguish different actions. Likewise, the actions in the database of KTH where the lens is pulled closer, far away or the camera has a slight movement also achieve a better distinction. Part OFEIs of video segments are shown in figure 6.

6.3 Feature extraction based on CNN

A Convolutional Neural Network is a powerful machine learning technique from the field of deep learning. CNN is trained using large collections of diverse images. From these large collections, CNN can learn rich feature representations for a wide range of images. These feature representations often outperform hand-crafted features such as HOG, LBP, or SURF [19-20]. An easy way to leverage the power of CNN, without investing time and effort into training, is to use a pre-trained CNN as a feature extractor.

The intermediate layers make up the bulk of the CNN. These are a series of convolutional layers, interspersed with ReLU and max-pooling layers. Following these layers are 3 fully-connected layers. The final layer is the classification layer and its properties depend on the classification task.

Fig. 6. Part OFEI of video segments

Each layer of a CNN produces a response, or activation, to an input image. However, there are only a few layers within a CNN that are suitable for image feature extraction. The layers at the beginning of the network capture basic image features, such as edges and blobs. To see this, visualize the network filter weights from the first convolutional layer. This can help build up an intuition as to why the features extracted from CNN work so well for image recognition tasks.

The first layer of the network has learned filters for capturing blob and edge features. These "primitive" features are then processed by deeper network layers, which combine the early features to form higher level image features. These higher level features are better suited for recognition tasks because they combine all the primitive features into a richer image representation.

In order to improve the recognition accuracy and speed, one is to use GPU parallel computing, the other is the introduction of region features, that is, a significant reduction calculation by the features of target area. The output of the first convolution layer of the CNN is shown in figure 7. The first 48 and the last 48 operations can be performed on both GPUs.

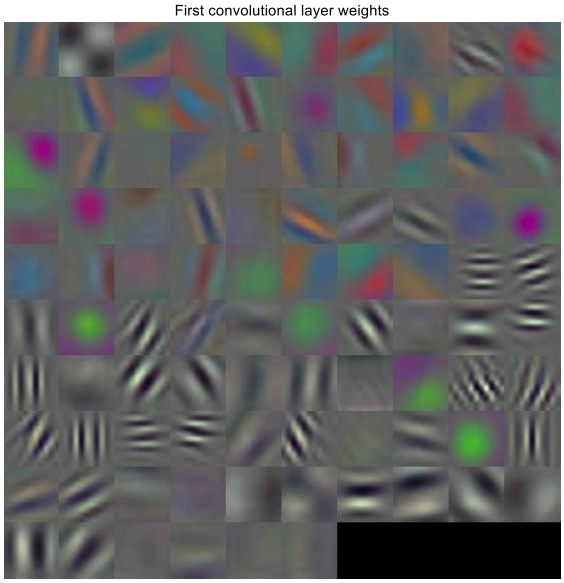


Fig. 7. The output of the first convolution layer

6.4 Action recognition using support vector machines

As with any supervised learning model, we first train a support vector machine, and then cross validate the classifier. Use the trained machine to classify (predict) new data. In addition, to obtain satisfactory predictive accuracy, you can use various SVM kernel functions, and we tune the parameters of the kernel functions. The classifier workflow is as follows.

1. Training an SVM Classifier
2. Classifying New Data with an SVM Classifier
3. Tuning an SVM Classifier

The experimental process uses cross validation. We split the sets into training and validation data. Pick 30% of images from each set for the training data and the remainder 70%, for the validation data. Randomize the split to avoid biasing the results. The overall recognition accuracy of the experiment on the Weizmann database was 93.5%. The confusion matrix is shown in table 1.

Table 1. Confusion Matrix of Actions recognition ratio

	bend	**jack**	**jump**	**pjmp**	**run**	**side**	**walk**	**wave1**	**wave2**
bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jack	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jump	0.00	0.00	0.92	0.00	0.00	0.08	0.00	0.00	0.00
pjump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.08	0.00	0.00	0.76	0.05	0.11	0.00	0.00
side	0.00	0.00	0.07	0.00	0.00	0.93	0.00	0.00	0.00
walk	0.00	0.08	0.00	0.00	0.12	0.00	0.80	0.00	0.00
wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

From the recognition results, it could be seen that error recognition mainly occurred between actions of jump and side, run and walk. The recognition accuracy on the KTH database by the same solution was 89.9%. The main factors that affected the recognition accuracy were effects which the lens was pulled closer, far away or the camera was moved slightly.

In addition, in append experiments, we found that increasing the depth of convolution neural network, the feature extraction effect could be enhanced, and further improved the recognition accuracy. However, the cost of training time was increased significantly.

7 Abnormal behaviors recognition in intelligent video surveillance system

The technique of human behavior recognition was widely applied in the intelligent video surveillance system. Under different scenes, abnormal behavior had different definitions. According to the surveillance and management demands of parking lots, we designed and implemented a set of intelligent surveillance systems. In the system, we defined abnormal behaviors as faint, smashing cars, robbery and fighting. This system aims to recognize abnormal behaviors in videos and alert about them.

7.1 Structure and function

The structure of intelligent video surveillance system can be divided into three layers. They are data-driven layer, behaviors recognition layer and user interaction layer respectively. Data-driven layer is responsible of sampling and managing video data. Behaviors recognition layer is to recognize abnormal behaviors. This layer includes foreground detector, feature extractor and classifier and it is the central layer of the intelligent video surveillance system. User interaction layer provides us with interactive operating interface and it is to display recognize result, manage information and input user operation. There exists data interaction among the three layers. They hand down users' operation demands from top to bottom, and the video data are transmitted from bottom to top.

The intelligent video surveillance system includes the following two sub-systems which are model training system and behavior recognition system. The training and recognition process is shown in figure 5. In the training sub-system, videos with artificial labels are input. In the sub-system of behavior recognition, videos to be tested are input and the results of classification and recognition are output. Before recognizing behaviors, model training sub-system should be trained firstly. During this stage, training samples are as input and the following steps are executed: get moving object by foreground detection, extract CNN features by feature extractor and recognize behavior by SVM classifier. Behavior recognition sub-system is to recognize abnormal behaviors. If abnormal behaviors are found, abnormal information will be displayed on user interaction layer and then an alarm is given out. The information such as the happing time and the location etc. will be recorded in the system.

7.2 Abnormal behaviors recognition

Based on the previous research work, we adopt CASIA [21] database to do the experiment of abnormal behavior recognition. The moving features of several abnormal behaviors are shown in figure 8.

For the abnormal behaviors that were defined, we got the confusion matrix with100% recognition accuracy. In parking environment, we deployed the intelligent video surveillance for abnormal actions recognition and alarm, and it worked well.

8 Conclusion

The recognition of human actions is an important research direction in the field of computer vision and pattern recognition. The technology study is very important for theoretical research significance and it is wide application prospect. In this paper, we studied the expression method of actions sequences. In order to solve the situation that many actions expressions were pretreated for extracting better features, we propose OFEI for feature expression based on optical flow. We used CNN for feature selection and dimensionality reduction, and then used the features extracted by CNN to train multi-class linear SVM classifier for actions recognition. One of the differences of our solution was using optical flow energy image to express human actions,

Fig. 8. Moving features of abnormal actions. Faint was showed in the first row, fighting was showed in the second row, robbery was showed in the third row, and smashing car was showed in the fourth row.

the other was that instead of using image features such as HOG or SURF, features were extracted using a CNN. Our experiment showed that the SVM classifier trained using CNN features provided close to 100% accuracy. The result was higher than the accuracy achieved using bag of features and SURF. Using our solution, the experiment performed on the Weizmann database achieved good recognition. The experiment result on the KTH database showed that the method was also effective when the lens was pulled closer, far away or the camera was moved slightly.

Applying the techniques studied in this paper, we designed and implemented a set of intelligent video surveillance system in the parking environment. The system was used for abnormal behaviors identification and alarm. In practice, the system worked well and the recognition result of abnormal behaviors was satisfactory.

9 Acknowledgment

Special thanks to my PhD supervisor, Prof. Li Li, School of computer science and technology, Changchun University of science and technology, she gave us lots of valuable advices in this study. The authors also would like to thank the financial support of National Social Science Foundation Project (17BSH135).

10 References

[1] Held C, Krumm J, Markel P, et al. Intelligent Video Surveillance[J]. Computer, 2012, 45(3):83-84. https://doi.org/10.1109/MC.2012.97

[2] Duque D. Prediction of Abnormal Actions for Intelligent Video Surveillance Systems[M]. 2007.

[3] Veeraraghavan A, Member S, Roychowdhury A K. Matching shape sequences in video with applications in human movement analysis[C] IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005:1896--1909. https://doi.org/10.1109/TPAMI.2005.246

[4] Ke S R, Thuc H L U, Lee Y J, et al. A Review on Video-Based Human Activity Recognition[J]. Computers, 2013, 2(2):88-131. https://doi.org/10.3390/computers2020088

[5] Chen C H. Handbook of Pattern Recognition and Computer Vision[M]. World Scientific Publishing Co. Inc. 2016. https://doi.org/10.1142/9503

[6] Bai L. Moving target detection method based on polarization characteristics under the condition of moving detector[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2014, 9301(19):2193-2200.

[7] May K, Krouglicof N. Moving target detection for sense and avoid using regional phase correlation[J]. 2013:4767-4772.

[8] Nixon M S, Aguado A S. Object description - Feature Extraction and Image Processing - 7[J]. Feature Extraction & Image Processing, 2002:247–290. https://doi.org/10.1016/B978-0-08-050625-8.50011-0

[9] Candamo J, Shreve M, Goldgof D B, et al. Understanding transit scenes: a survey on human action-recognition algorithms[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(1):206-224. https://doi.org/10.1109/TITS.2009.2030963

[10] Ke S R, Thuc H L U, Lee Y J, et al. A Review on Video-Based Human Activity Recognition[J]. Computers, 2013, 2(2):88-131. https://doi.org/10.3390/computers2020088

[11] Dalal N, Triggs B, Schmid C. Human Detection Using Oriented Histograms of Flow and Appearance[M]// Computer Vision – ECCV 2006. Springer Berlin Heidelberg, 2006:428-441.

[12] Barron J L, Fleet D J, Beauchemin S S, et al. Performance of optical flow techniques[J]. International Journal of Computer Vision, 1994, 12(1):43-77. https://doi.org/10.1007/BF01420984

[13] Bobick A F, Davis J W. The Recognition of Human Movement Using Temporal Templates[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001, 23(3):257-267. https://doi.org/10.1109/34.910878

[14] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436-444. https://doi.org/10.1038/nature14539

[15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[16] Karpathy A, Toderici G, Shetty S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision and Pattern Recognition. IEEE, 2014:1725-1732.

[17] Chapelle O. Training a Support Vector Machine in the Primal[M]. MIT Press, 2007.

[18] Noble W S. What is a support vector machine?[J]. Nature Biotechnology, 2007, 24(12):1565-1567. https://doi.org/10.1038/nbt1206-1565

[19] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C] Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893

[20] Bay H, Ess A, Tuytelaars T, et al. Speeded-Up Robust Features[J]. Computer Vision & Image Understanding, 2008, 110(3):404-417. https://doi.org/10.1016/j.cviu.2007.09.014
[21] Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He and Tieniu Tan. Robust View Transformation Model for Gait Recognition. Proceedings of the IEEE International Conference on Image Processing, 2011.

11 Authors

Honghua Xu, Li Li, and **Ming Fang** are with Changchun University of Science and Technology, Changchun, Jilin, China.

Fengrong Zhang is with Northeast Normal University, Changchun, Jilin, China.

Article submitted 16 October 2017. Resubmitted by the authors 29 November 2017. Final acceptance 23 March 2018. Final version published as submitted by the authors.