Spatio-temporal feature points detection and extraction based on convolutional neural network

Chaoyu Yang[1,2,a], Qian Liu[1,b], Yincheng Liang[1,c]

[1]1 School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, 100083, China

[2]School of Economics and Management, Anhui University of Science and Technology, Anhui, 232001, China

[a]yangchy@aust.edu.cn, [b]lq_nxdx@163.com, [c]liangyincheng111@163.com

Keywords: Convolutional neural network, spatio-temporal features, action recognition

Abstract. Convolutional neural network is a kind of deep learning model, but it only act on a single image generally. This paper expands the convolutional neural network, studies common spatio-temporal features detection and extraction algorithm, proposes a model for detecting and extracting spatio-temporal features based on convolutional neural network, applicates convolutional neural network in action recognition. This model use a plurality of consecutive video frames as input, extract image feature and time dimension information from sequent video frames, convolute and sub sampling alternately, extract a variety of advanced complex abstract features gradually. Experiments show that spatio-temporal convolution neural network has improved ability to classifing and learning.

1. Introduction

The traditional video surveillance system is highly dependent on the relevant staff. Supervisors observe and analysis happened suspicious behaviors through online or offline manually check the video recording, So it may cause huge security problems. In video, object's feature detection and extraction is the first step of human action recognition. Feature can be divided into global feature and local feature. The local feature of image and video can complete many recognition tasks, such as object recognition and scene recognition and human motion recognition.

2. The extraction and description of spatio-temporal feature points

The spatial-temporal feature mainly include spatial-temporal cube, spatial-temporal context, spatial-temporal interest point etc. At present, several commonly used feature point detector are Harris3D feature point detector proposed by Laptev, Cuboid feature point detector proposed by Dollar, feature point detector based on the minimum intensity change (MIC) and one-dimensional Gabor filter.

spatio-temporal feature points extraction based on 3-D Harris (STFPEH)

3D-Harris detection operator can find the most dramatic variety of image sequence in time and space.We assume that the video spatio-temporal is a function $f : R^2 \times R \to R$ at first, then we construct scale space for video by using the Gauss kernel function: $L\left(\bullet;\sigma_l^2,\tau_l^2\right) = g\left(\bullet;\sigma_l^2,\tau_l^2\right) * f\left(\bullet\right)$

The definition of Gauss kernel function:

$$g\left(x,y,t;\sigma_l^2,\tau_l^2\right) = \frac{\exp\left(-\left(x^2+y^2\right)/2\sigma_l^2 - t^2/2\tau_l^2\right)}{\sqrt{\left(2\pi\right)^3 \sigma_l^4 \tau_l^2}}$$

We find the sharp change point point in space and time as spatio-temporal interest point by defining the response function of H:

$$H = \det\left(\mu\right) - k * trace^3\left(\mu\right) = \lambda_1\lambda_2\lambda_3 - k*\left(\lambda_1+\lambda_2+\lambda_3\right)^3$$

In the formula, $\lambda_1, \lambda_2, \lambda_3$ are the three characteristic value of matrix.

Cuboid detection operator (CDO)

The Cuboid operator is the spatio-temporal interest point detection method based on Gabor filter and Gauss filter. Detection operator is used to solve the problem of the 3D-harris reliable interest points amount is less. Through the Gabor filter and Gauss filter, the corresponding function to construct the detection of interest points is as follows:

$$R=\left(I*g*h_{ev}\right)^2+\left(I*g*h_{od}\right)^2$$

In the formula, $g\left(x,y,\sigma\right)$ is two-dimensional space Gauss smoothing kernel fuction. h_{ev} and h_{od} is a set of orthogonal one-dimensional Gabor filter on time domain. It can be expressed as:

$$h_{ev}\left(t;\tau,\omega\right)=-\cos\left(2\pi t\omega\right)e^{-t^2/\tau^2},\ h_{od}\left(t;\tau,\omega\right)=-\sin\left(2\pi t\omega\right)e^{-t^2/\tau^2}$$

$\omega=4/\tau$, σ and τ control detector's spatial scales and time scales respectively. Finally in the same way, we screening spatio-temporal interest point by non maximum value suppression of each of point response value in the video space.

3. The extraction algorithm of spatio-temporal interest point based on convolutional neural network

3.1 Convolutional neural network structure based on spatio-temporal

Convolutional neural network combine artificial neural networks and deep learning, train the weights of the network by using an improved back-propagation algorithm based on gradient, achieve multilayer filter network structure and the global training algorithm combined of filter and classifier.

This paper proposes a spatio-temporal convolutional neural network model. The model can extract effectively image features and time dimension information from sequent video frames, it is obtained by convolution with a spatio-temporal convolution kernel from a plurality of stacked sequent video frames.

3.2 Convolution layer gradient calculation

Pre-order layer feature map convolute through learning nucleus in the volume of deposit, and then construct an output feature map by activation function. Each output map may contain multiple convolution of input map. In general:

$$X_j^1=f\left(\sum_{i\in M_j}X_i^{l-1}*Kernel_{ij}^l+B^l\right)$$

In the formula, X_j^l represents the first j characteristic map in the first l convolution layer, $\mathrm{f}\left(*\right)$ represents an activation function, such as sigmoid function, M_{j} represents the collection of the input map. All the weight is equal to a constant beta in drop sampling layer, we multiply by beta times to calculate δ^l, then repeat the steps to calculate each map j in the convolution layer, and correspond the result to the relative reduction sampling layer finally.

$$\delta_j^l=\beta_j^{l+1}\left(f^{'}\left(\mu_j^l\right)\bullet up\left(\delta_j^{l+1}\right)\right)$$

In the formula, $up(\bullet)$ represents up sampling operation. A simple realized method is as follows through the Kronecker product: $up(x)=x\otimes 1_{n\times n}$

We have the error signal of a given graph now, then we can calculate the gradient of the deviation according to sum all of the items in the δ_{j}^l: $\frac{\partial E}{\partial b_j}=\sum_{u,v}\left(\delta_j^l\right)_{uv}$

Finally, kernel function weight gradient need to calculate by back propagation. We sum all gradient involved with connection sharing all weight:

$$\frac{\partial E}{\partial K_{ij}^l}=\sum_{u,v}\left(\delta_j^l\right)_{uv}\left(P_i^{l-1}\right)_{uv}$$

3.3 Falling sampling layer gradient calculation

The result of down sampling layer'calculation input graph down sampling is if there are N input graph, there are N output graph also. $x_j^l = f\left(\beta_j^l down\left(X_j^{l-1}\right)+b_j^l\right)$

In the formula, $down(\bullet)$ represents down sampling function. It can be achieved effectively by the following formula:

$$\delta_j^l = f'\left(\mu_j^l\right)\circ conv2\left(\delta_j^{l+1}, rot180\left(k_j^{l+1}\right),' full'\right)$$

Now we can calculate beta and gradient of b. Additional deviation b is also adding all elements in error signal map: $\frac{\partial E}{\partial b_j}=\sum_{u,v}\left(\delta_j^l\right)_{uv}$

Multiplicator deviation beta is related with the original down sampling of the current layer in propagation apparently. $d_j^l = down\left(x_j^{l-1}\right)$ .The beta gradient is given by formula as follow: $\frac{\partial E}{\partial \beta_j}=\sum_{u,v}\left(\delta_j^l\circ d_j^l\right)_{uv}$

3.4 Spatio-temporal interest point extraction algorithm based on convolutional neural network (STEACNN)

In this paper, we will focus on the research of time dimension, aim to find the effect of time dimension size on action classification, and then proposed a spatio-temporal convolutional neural network model based on optimal time dimension.

Step1: In the current frame center, We take several consecutive frames of 32*32 as input of spatio-temporal convolutional neural network.

Step2: The C1 layer adopts 4*3*3 convolution kernel to obtain 24 feature maps from 7 consecutive input frames, it is extracting 24 different features by using 24 different learning kernels.

Step3: The S1 layer is a sub sampling layer. it zoom feature map scale which is obtained from the C1 layer.

Step4: The C2 layer is a convolution layer also, it use 24 feature map of S1 layer as input.

Step5: A sub sampling layer is similar to S1, which scaling factor is 2.

Step6: After the S2 layer there is a connection layer, we identify it with F layer.

Step7: The output layer is a all connection layer after the F layer, the number of its neurons is the target number of classification.

4. Experiment results and analysis

In public video data KTH, we compare action recognition rate of the method (STEACNN) proposed in this paper with action recognition rate of common spatio-temporal feature extraction method. We can see action recognition rate of the method (STEACNN) is higher than other common methods apparently.

Table 1: Comparison of action recognition rate for each algorithm

No.	Action categories	STFPEH	CDO	STEACNN
1	Boxing	93.1%	94.5%	95.6%
2	Clapping	91.4%	90.8%	96.7%
3	Jogging	94.8%	95.4%	95.8%
4	Running	93.6%	96.3%	96.6%
5	Walking	91.2%	91.1%	94.7%
6	Waving	92.5%	91.9%	95.2%
	Average	92.77%	93.33%	95.77%

We use different frame length in different KTH data set action, the different average recognition rate as shown in Figure 2.

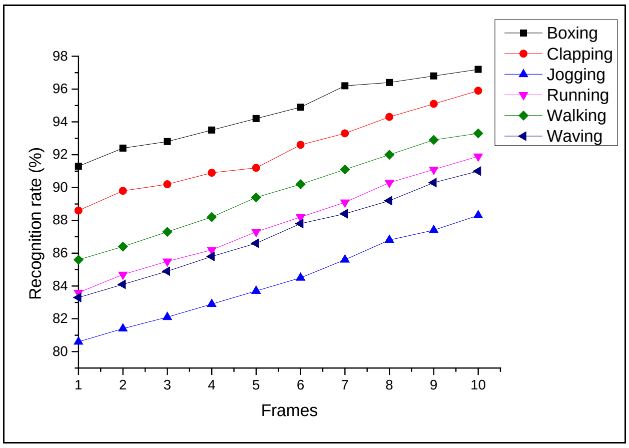

Fig.2: Average recognition rate of actions

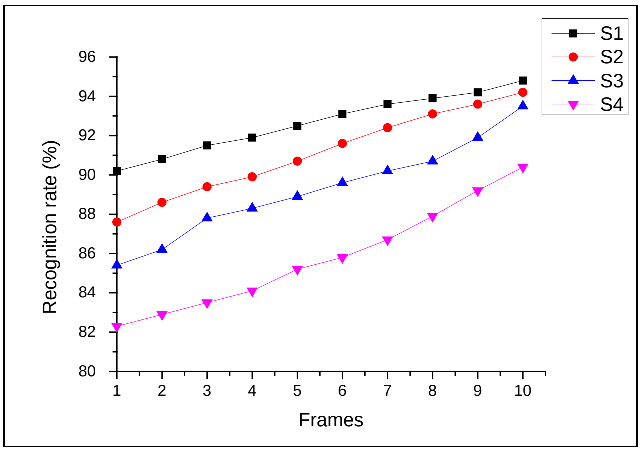

Fig.3: Average recognition rate of scenes

Figure 3 shows experimental result in different scenes of KTH data set. Result of outdoor scenes(S1, S2, S3) is worse than indoor scene(S4), because outdoor environment is variety but indoor environment is stable relatively.

5. Conclusions

We extract features from a video or sequent video images in the action recognition application. The convolutional neural network can only act on a single image. This paper proposes a spatio-temporal convolutional neural network model, it can extract effectively image features and time dimension information from continuous video frames. The feature is obtained by a spatio-temporal convolution kernel convoluting in a plurality of stacked consecutive video frames. In this structure, the feature mapping layer of convolution connect with a plurality of pre-layer consecutive video frames, so the structure can get continuous movement information and classify human actions.

Acknowledgements

The research work was supported by National Natural Science Foundation of China under Grant No. 51404007, National key foundation for exploring scientific instrument of China (No.E200038884G) and nuclear waste special scientific research project (No.FZ1402-08).

References

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504-507.

[2] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning,2009,2(1):1-127.

[3] Ahmed A, Yu K,Xu W,et al. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks[M]. Computer Vision-ECCV 2008. Springer Berlin Heidelberg,2008:69-82.

[4] Arel I,Rose D C, Karnowski T P. Deep machine learning-A new frontier in artificial intelligence research[J]. Computational Intelligence Magazine, IEEE,2010,5(4):13-18.

[5] Bengio Y. Deep learning of representations:Looking forward[J]. arXiv preprint arXiv:1305.0445,2013.

[6] Krizhevsly A, Sutskever I, Hinton G. Image net classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems 25.2012:1106-1114.