Research on Behavior Recognition in Infrared Video

Ruming Yang[1], Meng Ding[1, 2, a], Xu Zhang[1], Xinyan Jiang[1]
[1]School of Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China
[2]Key Laboratory of Civil Aircraft Health Monitoring and Intelligent Maintenance, Nanjing 210000, China.
[a]corresponding.nuaa_dm@nuaa.edu.cn

Abstract. We trained a neural network for behavior recognition task in infrared video and took a small amount of infrared video to evaluate performance. Limited by the quantity of data, we pre-trained the network on the visible datasets and fine-tuned on our infrared dataset. In order to explore the effect of the learning of the two datasets, we visualized the features. The experimental results demonstrated that our network has excellent learning performance for infrared video, and the learned features are efficient and generalized.

Keywords: Convolutional neural networks, infrared video, transfer learning, visualization.

1. Introduction

Video surveillance are used by security personnel to ensure the security of an area [1]. HD infrared cameras are one important part of the all-weather video surveillance system. The imaging mechanism of the infrared cameras by measuring the heat radiation of the object allows them to work in low visibility conditions. Although infrared images are not affected by light, they have the disadvantage of lack of color and texture information, which makes the processing algorithm of infrared images more complicated. In the past, due to the high cost and special conditions, HD infrared cameras are not common in ordinary environments. This also led to rare achievements in human behavior recognition for infrared video

In this paper, we study the behavior recognition in infrared video.

One of the main challenges of the behavior recognition is that it must find an effective descriptor which simultaneously describes the characteristics of the target in both spatial and time domains. An excellent video descriptor should have some properties: (i) It should be generic. In classification tasks, a good feature needs to have large among-class distance and small within-class distance. (ii) It should be simple. A good descriptor should work well even with a simple model (e.g. linear classifier), instead of using complicated feature encoding methods and classifiers. In the past, many hand-crafted descriptors have been proposed to solve the problems: Histogram of oriented gradient (HOG) [2], histograms of oriented optical flow (HOF) [3], optical flow, trajectory, etc. However, the performance of a single hand-crafted feature is not impressive. Researchers preferred to combine multiple descriptors and fuse them in the past, so as to maximize the use of spatial-temporal information. In recent years, inspired by the breakthrough of deep learning in the field of image, feature learning has made rapid progress in the past few years. S. Ji et al. proposed 3-dimensional convolutional network (3DCNN) [4] to apply convolutional neural network (CNN) [5] to video data. These network modern deep architectures to achieve the best performance on different types of video. However, the first layer of 3DCNN is a hard-coded convolution layer that extracts grayscale, gradients, and optical flow from the video, which greatly increase the cost of preprocessing. In this paper we propose to learn spatio-temporal features using deep 3DCNN with full video frames as inputs and does not rely on any preprocessing,

Another major problem faced by behavior recognition in infrared video is the lack of high-quality data. At present there is no public infrared dataset. Fortunately, we found that although there is very little infrared data available for training, the similarity in the content of visible video and infrared

video makes it possible to transfer the knowledge from visible video to infrared video. The differences between infrared video and visible video mainly exist in the spatial domain. However, for the behavior recognition, information in the time domain is more concerned. Both visible video and infrared video have abundant information in the time domain such as optical flow, trajectory, etc. It allows us to design a 3DCNN to learn powerful features in visible video and transfer the knowledge to infrared video. For simplicity, we call this net infrared-3D from now on.

To summarize, our contributions in this paper are:

We experimentally designed a deep 3-dimensional convolutional network based on 3DCNN to learn efficient spatial-temporal features of infrared video. The proposed features are generic and efficient.

2. Methodology

2.1 From 2D to 3D.

3D convolution is a spatio-temporal filter. The input is a cube with a size of c*l*h*w where c is the number of channels, l is length in number of frames, h and w are the height and width of the frame, respectively. The convolution kernel also has a depth dimension, so the convolution kernel slides through both the spatial (width and height) and temporal (depth) of the input cube.

Similar to the 2D pooling. 3D pooling injects invariance into features, including translation invariance, rotation invariance, and scale invariance. Retain the main features while reducing the parameters and calculations, prevent overfitting, and improve the model's generalization ability.

2.2 Datasets.

We used the UCF101 dataset for pre-training. The UCF101 dataset contains 101 types of sports video. There are approximately 12,000 video, covering a large number of individual sports and team sports. We used infrared cameras to take a small amount of infrared motion video under low visibility conditions, including 789 video in five categories of walking, running, jumping, waving and squatting.

Fig. 1 Infrared dataset

2.3 Network Architecture and Training Parameter Settings.

We set each non-overlapping 8 frames to a clips as input to the networks. We resized all video frames into 128*171. This is roughly half resolution of the UCF101 frames and our infrared video frames. Therefore, the size of the input cube in our network is 3*8*128*171. We also use jittering by using random crops with a size of 3*8*112*112 of the input clips.

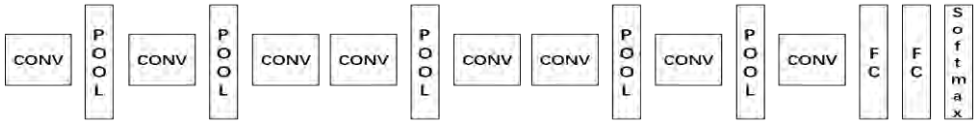


Fig. 2 Structure of infrared-3D

We design our infrared-3D to have 8 convolution layers, 5 pooling layers, followed by two fully connected layers, and a softmax output layer. The network architecture is presented in Fig. 2. All of 3D convolution filters are 3 * 3* 3 with stride 1 * 1* 1. The number of filters for 8 convolution layers from 1 to 5 are 64, 128, 256, 256, 512, 512, and 512, respectively. All 3D pooling layers are 2*2*2 with stride 2*2*2 except for pool1 and pool5 which have kernel size of 1 * 2 * 2 and stride 1 * 2 * 2. Each fully connected layer has 4096 output units.

Training: We use UCF101 dataset to pre-train infrared-3D networks and then use infrared datasets for fine-tuning. With one 1080Ti GPU, we pre-train from scratch, using mini-batches of 30 clips with initial learning rate of 0.001. The learning rate is divided by 10 after every 10,000 iterations. The

optimization is stopped at 100,000 iterations. When fine-tuning, we set the learning rate of 0.001, dividing every 10,000 iterations by 10, stopping at 50,000 iterations.

Results: Our infrared dataset includes five categories, each containing more than 130 video clips. We divide the dataset into 6 equal parts. Each time, we randomly select 5 for training and 1 for testing. Take the average of ten training results as the final accuracy. We have achieved an accuracy of 95.21%. The result is very surprising and shows how generic infrared-3D is on capturing spatial-temporal information in infrared video.

Table 1 presents classification results of infrared-3D compared with the IDT and C3D. Infrared-3D performs best among three methods.

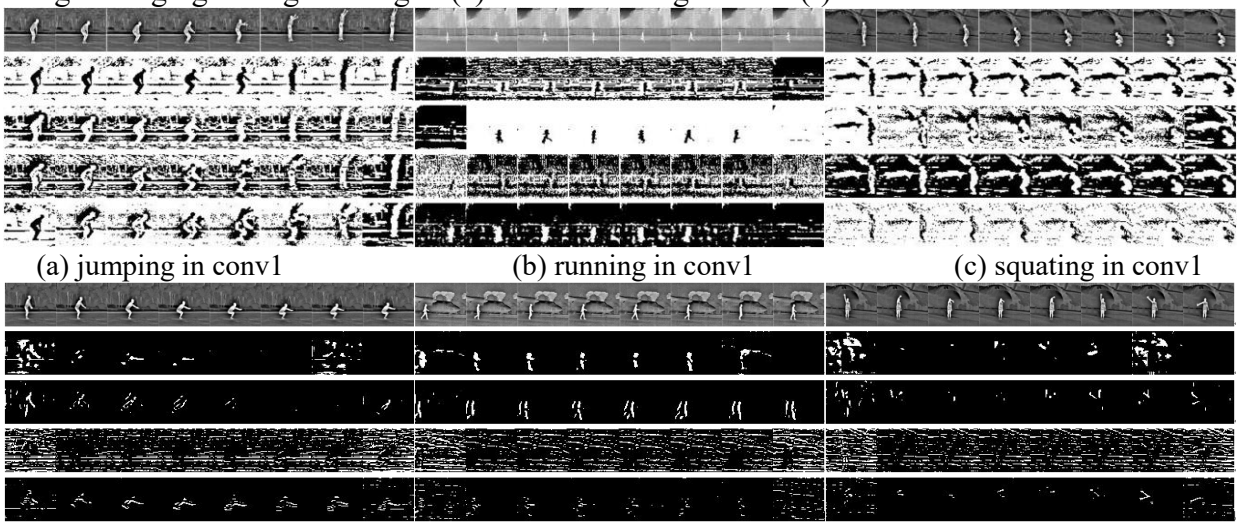
Table 1 Classification results on infrared dataset

Method	Accuracy
IDT	86.7%
C3D	92.8%
infrared-3D	95.2%
infrared-3D+svm	96.5%

3. Visualization

We use a deconvolution approach [6] to observe what infrared-3D has learned. We selected several representative video to visualize the 64 feature sequences obtained from the conv1 layer in infrared-3D and pick several easy to understand feature sequences shown in Fig. 3 (a) (b) (c). We found that almost all neurons responded to the difference in the brightness between the foreground (human) and the background of the infrared video and performed effective segmentation. At the same time, neurons also respond to signals in the time domain. As shown in the fourth row of (a), the second row of (b), and the third row of (c), in the first frame, the neurons focus on the difference in brightness, and successfully distinguish the foreground from the background. In the following frames, the neurons begin to pay attention to the areas where the brightness changes.

Then we continue to visualize the feature sequences obtained from the conv2 layer to see what infrared-3D learned. We noticed that after the second convolution, almost all neurons focus more on spatial information. We still select several easy to understand feature sequences shown in Fig. 3 (d) (e) (f). In the first row, neurons seem to be more concerned with the area where the brightness changes and almost filter out the background. In the second row, neurons are more concerned with the contours of the moving parts. In the third row, neurons seem to be more concerned with horizontal texture features in the background. In the fourth row, neurons are almost entirely concerned with the part where motion occurs, such as the arms swinging and knee bending during squatting in (d), arms and legs swinging during walking in (e) and the waving arms in (f).



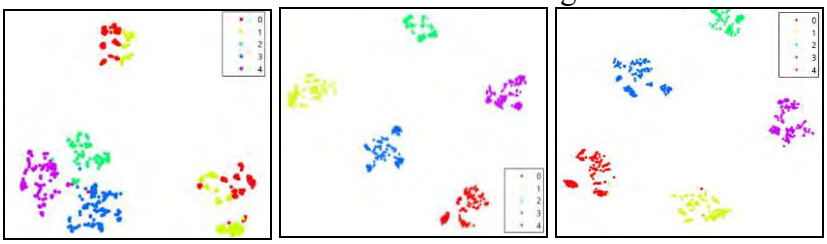
(a) jumping in conv1 (b) running in conv1 (c) squating in conv1

(d) squatting in conv2 (e) walking in conv2 (f) waving in conv2

Fig. 3 Visualization

From these feature sequences, we find that infrared-3D selectively attends to spatial-temporal information. Conv1 neurons extract a large amount of spatial information such as texture, brightness, and contour, and pass these underlying information to subsequent layers. The neurons of Conv2 begin to pay more attention to the information in the time domain, but retain the sensitivity to spatial information at the same time. All of this above shows that infrared-3D has good learning performance in the task of behavior recognition in infrared video.

We use the t-sne [7] to intuitively demonstrate the generalization ability of infrared-3D. Each point in the figure represents a feature sequence, and different colors represent different categories. We believe that powerful features with excellent generalization capabilities will have larger among-class distance and smaller within-class distance. Fig. 4 vely. It can be seen that features learned by the neural networks are better than feature extracted by IDT obviously. The feature learned by infrared-3D is slightly better: the points are closer and the within-class distance is smaller. Therefore, we believe that infrared-3D is more suitable for behavior recognition task in infrared video.



(a)IDT (b)infrared-3D (c)C3D

Fig. 4 Feature embedding

4. Summary

In this work we tried to solve problems in the behavior recognition in infrared video based on existing research. We proposed infrared-3D network that performs well on infrared datasets. We use visible dataset to pre-training and then use infrared datasets for fine-tuning. This across modalities transfer learning successfully avoided over-fitting and achieved impressive result. We visualized the feature sequences of infrared-3D and analyzed the area of interest in the task. We demonstrated that infrared-3D features perform well on infrared video datasets as an efficient feature.

References

[1]. Gowsikhaa D, Abirami S, Baskaran R.: 'Automated human behavior analysis from surveillance videos: A survey', Artificial Intelligence Review, 2014, 42(4), pp. 747-765.

[2]. Dalal N, Triggs B.: 'Histograms of Oriented Gradients for Human Detection', Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA: IEEE, 2005, pp. 886-893.

[3]. Laptev I, Marszalek M, Schmid C, et al.: 'Learning realistic human actions from movies', Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA: IEEE, 2008, pp. 1-8.

[4]. S. Ji, W. Xu, M. Yang, and K. Yu.: '3d convolutional neural networks for human action recognition', IEEE TPAMI, 2013, 35(1), pp. 221–231.

[5]. A. Krizhevsky, I. Sutskever, and G. Hinton.: 'Imagenet classification with deep convolutional neural networks', Proc. International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: ACM, 2012, pp. 1097-1105.

[6]. M. Zeiler and R. Fergus.: 'Visualizing and understanding convolutional networks'. Proc. European Conference on Computer Vision, Zurich, Switzerland: Springer, 2014, pp. 818-833.

[7]. Van der Maaten, L, Hinton, Geoffrey, Maaten, Laurens Van Der.: 'Visualizing data using t-SNE', Journal of Machine Learning Research, 2017, 9(2605), pp. 2579-2605.