2018 IEEE International Symposium on Multimedia (ISM)

3D Convolutional Network Based Foreground Feature Fusion

Hanjian Song
School of Software Engineering Xi’an
Jiaotong University Xi’an, China
hjsong1993@163.com

Lihua Tian
School of Software Engineering
Xi’an Jiaotong University Xi’an,China
lhtian@ xjtu.edu.cn

Chen Li*
School of Software Engineering Xi’an
Jiaotong University Xi’an,China
lylnnc@126.com

***Abstract*—With explosion of videos, action recognition has become an important research subject. This paper makes a special effort to investigate and study 3D Convolutional Network. Focused on the problem of ConvNet dependence on multiple large scale dataset, we propose a 3D ConvNet structure which incorporate the original 3D-ConvNet features and foreground 3D-ConvNet features fused by static object and motion detection. Our architecture is trained and evaluated on the standard video actions benchmarks of UCF-101 and HMDB-51, experimental results demonstrate that with merely 50% pixels utilization, foreground ConvNet achieves satisfying performance as same as origin. With feature fusion, we achieve 83.7% accuracy on UCF-101 exceeding original ConvNet.**

Keywords—Foreground Exaction; Feature Fusion; C3D;

I. Introduction

With the rapidly growing of multimedia on the internet, the requirement to understand and analyze unlimited number of videos has been more and more extensive and eager. The computer vision community has been working on video analysis for decades and tackled different problem such as action recognition [1], abnormal event detection [2]and active understanding [3]. Focused on action recognition, recognition of human actions in videos has received a significant amount of attention from the research community. Comparing to image classification, there are 3 major difficulties in processing large scale of video data. 1) Video is 3-dimension data clips, the approach to handle temporal information and fuse with spatial information will bring decisive influence on the final result. 2) Owing to video is a component of consecutive frames mostly with the camera moving and main target moving. There is a large of data redundancy and noise, and the decisive information usually occupies a small part of a whole one. 3) Videos are variable frame sequences, but most judging approach as classification, similarity calculation needs uniform feature, it means all approach needs to go through feature exaction, feature normalization, judging three stages.

In recent years, to solve above three problems, many researchers have made considerable efforts and progress. In traditional field, Wang et al. propose an improved Dense Trajectories (iDT) [4], a state-of-the-art hand-crafted feature. They achieved extraordinary performance comparing with deep-learning algorithm with excellent denoising, feature fusing and recoding. However, this method is computationally intensive and becomes unavailable in time consuming.

With the great success of the ConvNet in image classification field, using optical flow to capture temporal information to ConvNet is another attempt. Karen et al. proposes Two-steam[5] architecture using RGB image and optical flow image separately as input of the network. Two-stream achieve state-of-the-art performance but higher speed more than IDT, it set off a trend of research based on two-stream architecture as TSN[6], and TDD [7].

To unite spatiotemporal information, Du Tran et al. propose 3D ConvNets(C3D) [8] on the realistic and large-scale dataset. They use k×k×k convolutional kernel instead of k×k kernel to explore spatial and temporal connection at one adjacent area. Du Tran’s research shows great potentiality of 3D ConvNet, but 3D ConvNet has greatly more parameters than 2D ConvNet causing inevitable over-fitting especially on the smaller dataset as UCF101[9], HMDB51[10] with less sample diversity. The result depends on a very large-scale of datasets seriously.

In this work, on the basis of 3D ConvNet, we aim at alleviating dependence of 3D ConvNet to multiple large scale dataset and enhance performance in a single dataset.

Inspired by Two-stream and IDT, in this paper we propose a method based on foreground feature of ConvNet, first we extract the foreground image of dataset through an extraction method called saliency and motion object foreground extraction(SMOF) after we investigate several representative foreground detection, then we conduct feature combination with two input stream(full RGB, foreground) to generate final video descriptor as input of linear classifier, after that, we use a linear svm classifier to conduct action recognition. In this paper, we show that with partial crucial area utilization, the foreground ConvNet is able to achieve almost same performance as the original ConvNet, less information push ConvNet to emphasis generalization of crucial area and increase complementary when feature fuse. And finally after feature fusion with two ConvNet, we show the promotion of performance in two baseline dataset UCF101 and HMDB51.

The rest of the paper is organized as follows. In sect 2 we review the related work on action recognition and analyze the problem of C3D. In sec 3 we introduce our foreground extraction and feature fusion method. The experimental details and result are given in sec 4. Sec 5 shows our final conclusion.

II. Related works

Video recognition research is spring up with rising of image recognition methods. Early when convolutional network did not become main trend of image recognition. Video recognition mainly relied on a family of image recognition methods. For instance, the method of [19] adopts two classical image features, Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF), then encodes with Bag of Feature (BoF) as input of SVM classifier. Some of researchers focus on altering 2D feature to 3D such as SIFT-3D[20] and HOG-3D[21]. Among them, the most representative method is Improved Dense Trajectories(IDT)[4], it uses autocorrelation matrix to eliminate background point and introduces dense trajectories to represent motion information, then combines with

978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00036

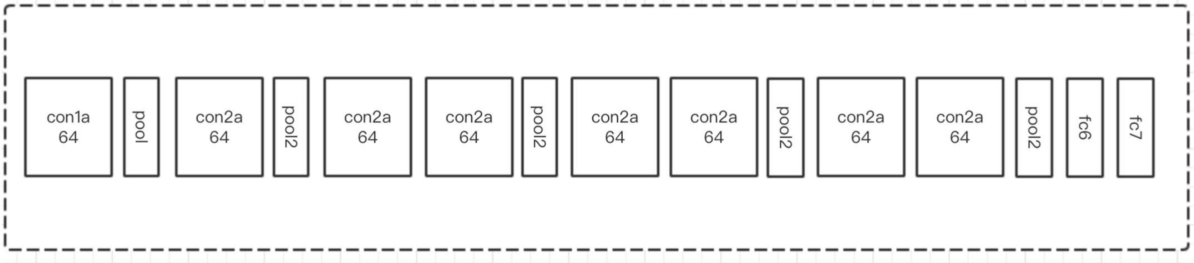

Fig 1 C3D structure : consist of 5 convolution layers and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), 2 fully-connected layers and a softmax loss layer

several spatial feature as HOG, HOF, MBH to further boost the accuracy. Although adoption of more feature brings the amount of calculation and the 220-thousands long descriptor is huge to be unavailable in time consuming, the thought of IDT has had a far-reaching impact.

With the rise of ConvNet in image recognition, the ConvNet in video recognition has grown up to be an important subject. After a number of attempts by large researchers, in 2015. Two-stream [5] overlooked the customary methods and pioneered the union of optical flow with ConvNet. They use RGB frames to input ConvNet as spatial stream and optical flow to input ConvNet as temporal stream. The result indicates the extraordinary compensation between two-stream. Two-stream opened a new thought to utilize temporal information, after that, many outstanding achievements are based on it such as TSN [6], a method which extends the input scale from clip level to video level, and TDD[7], a method which extracts the two-stream ConvNet feature of dense feature points generated by IDT.

Our research starts with 3-dimension deep ConvNet which is proposed by Du Tran. Using a set of frames as input into ConvNet to represent video, C3D is a true meaning 3D convolutional network because of the new convolutional kernel. The kernel size is adjusted to k×k×k instead of k×k in whole convolutional progress. The network is designed to consist of 5 convolution layers and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), 2 fully-connected layers and a softmax-loss layer(Fig 1). Through kernel size experiment, Du Tran identified that 3×3×3 kernel is the best kernel, it is small enough but represents the best generation capacity. C3D has better generalization than 2d ConvNet, and the network is enough convergent in experiment, but the result of the test set is unsatisfied, the accuracy is only higher a little than RGB network in two-stream, the result shows that the network exists over-fitting seriously. To solve over-fitting, considering the UCF-101 dataset is a small dataset, they use sports-1M[11], a dataset consists of 1. 1 million sports videos and 487 categories and another large scale dataset I380K as pretrained dataset, combine feature from network finetuned from different dataset, finally achieve huge promotion. From the progress, it demonstrates that the dependency on dataset exceeds more than network itself. Due to videos exist large data redundancy and noise to cover the decisive information, it is hard to find the key information in network. Compared with IDT and Two-stream, the background points in IDT are filtered by computing Autocorrelation matrix and camera motion estimation so these pixels which are believed as low value but higher possibility to be noise was eliminated, as same as two-steam do because optical flow is an effective approach which will normalize static pixels and highlight the moving part.

Among these approaches, we propose a hypothesis that extracting the foreground part of frames as input will guide the network to convergence to a more precise direction. We investigate several foreground extraction methods and design several fusion strategies based on methods of[12],[13],[15], then extract foreground ConvNet feature to fuse with C3D feature. We discuss more detail in Sec 3.

III. Proposed Method

In this section, we explain in detail and the basic operation of our method, discuss different foreground extraction, analyze different foreground fusion strategies and feature combination strategies.

Our method is mainly split into 3 steps. 1) Through current relatively effectual foreground extraction, we design several different foreground fusion methods to extract foreground images as a new stream to be input to ConvNet. 2) We input the foreground dataset generated by step 1 to ConvNet to pretrain with sports-1M pretrained model, and get a foreground model. We use the foreground model to extract innerproduct feature fc6 from foreground ConvNet. Every 16 consecutive frames are extracted to be a 4096-dim vector as same as original C3D. 3) We fuse the foreground fc6 features with original C3D fc6 in corresponding frames. The dimension of feature expands to 8192 to symbolize 16 frames in a video. We average the fusion features with L2 normalization to generate video-level descriptors. Then we input them into linear SVM classifier and conduct action recognition with these video-level descriptors. The principal process is shown as Fig 2.

A. Foreground Extraction

To extract foreground part of frames, we conduct two aspects of consideration. We try to subtract the background part from a static image and a set of consecutive frames of an object moving trajectories. We choose to use PBCN[12], SSD[13], two salient object detection approaches and PGBSUB[15], a motion detection approach to subtract background area.

PBCN and SSD are suggestive methods in object detection based on ConvNet. PBCN focus on image segmentation and SSD learns how to object detecting. There are 3 common advantages for two methods. First two methods are static extractions. It implies that the interference of camera moving is little. Second detection results of two methods are binaried image and almost noiseless. We can use the detection results to conduct mask operation directly without any post-treatment. Third two methods are relatively stable in video level, frame, previous frame is a successful detection but failure on next

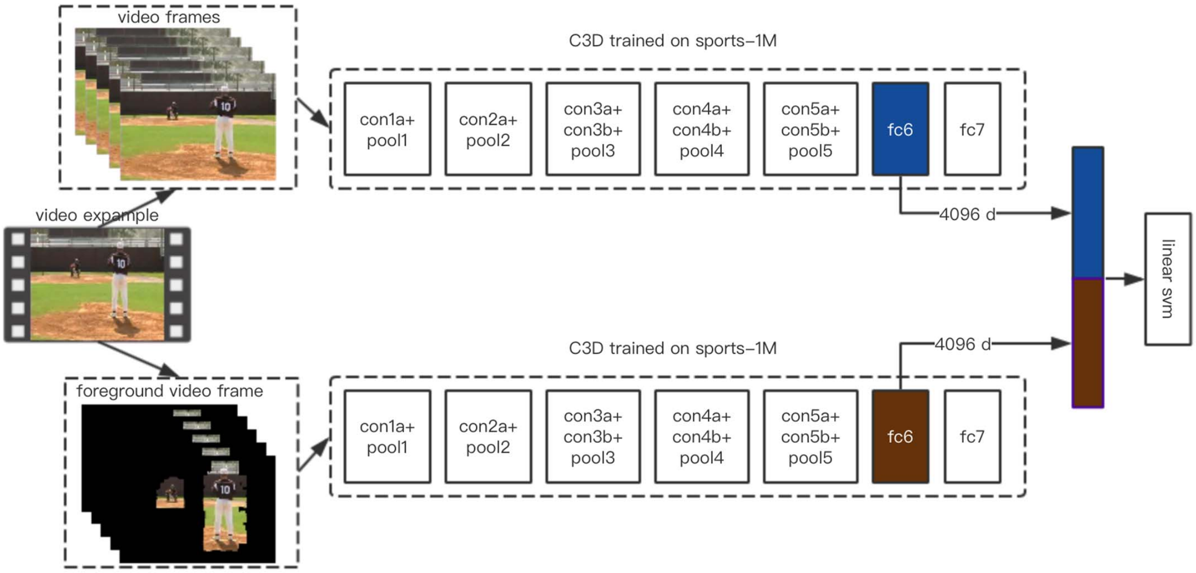

Fig 2 use original RGB frames and foreground frames to train C3D with sports-1M pretrained model, extract the fc6 feature to form 8192 dimensions feature and input them to linear svm

this situation is rarely happening. But the defects are noteworthy as well. PBCN is sensitive to the distance from object to camera, the long distance will blur the difference between object and background in low level feature, and PBCN have a high possibility to confuse the object and background such as basketball, cricket bowing. Due to the gradual of video frames, a failed detection in one frame will possibly fail all in a video, failed detection will cause chain reaction because they are the input of ConvNet, not only influence the test dataset but also confuse the ConvNet to a wrong way when it appears in train set. Turning to SSD, training dataset of SSD does not aim at pedestrian or people, the network is sensitive to gesture of person precisely the dataset contains of large different actions, the prediction is unstable even in a video especially the gestures have large difference with standing, sitting and walking such as swimming, the detection result is unsatisfied.

Traditional motion detection approaches are maturely developed, such as approaches based on RGB difference(PGBSUB) [15], Gaussian model, optical flow and background samples model(ViBe)[14]. All of these approaches depend on object movement between two consecutive frames. Inevitably, affected by camera movement, shaking and the movement of the background object such as water flow, shaking swings, these traditional approaches produce more noise than static extractions. But the robustness of motion object detection is what above two approaches are not sufficient. We choose PGBSUB to be our motion extractions because it represents better on complex scenes, and PGBSUB is fastest among them.

Relying on one approach is hard to be robust. Static extractions own higher accuracy but are unstable with complex scenes. Motion extractions are adaptive in all scenes but only when objects move, and accuracy is lower than static extractions with more noise. Due to existing complementarity of these approaches, we choose to combine the mask generated by above three extractions, we propose three fuse strategies.

1) PBCN+SSD

The fusion strategy of PBCN and SSD is easiest because the masks of PBCN, SSD are binarized images and almost without noise, so we simply add the masks of PBCN and SSD in corresponding frames, but the complementarity of two approach is not strong, because it has high probability of both two approaches fail in far camera distance with complex background.

2) PBCN+PGBSUB

The fusion of PBCN and PBGSUB is more complex than PBCN+SSD, considering there is only object contour in most situations of PBGSUB masks, we need to conduct denoising and detected area expansion to masks of PBGSUB. Due to noise points are dispersed in PBGSUB, we choose to use an average convolutional filter and a low pass filter to decrease noise:

contour
轮廓

$$P_{denoise}(x,y)=\frac{\sum_{i=-k}^{k}\sum_{j=-k}^{k}P_{PGBSUB}(x+i,y+j)}{k\times k} \quad (1)$$

where P_{PBGSUB}(i, j) is a pixel of PBGSUB masks of corresponding frame at coordinates i, j as the same definition as P_{denoise}. We use a 9×9 kernel trying to dilute the gray scale value of dispersed noise points, then we need to binary the denoise masks, we use a low pass filter with a threshold:

$$P_{binary}(x,y)=\begin{cases}0 & if\ P_{denoise}(x,y)<T\\ 255 & otherwise\end{cases} \quad (2)$$

with amount of experiment, setting the threshold T to 60 is most suitable, after denoising and binarization(2), we use a convolutional kernel which all the parameter is 1 to avoid only contour to be detected. In expansion, we use a 3×3 kernel to

$$P_{\exp and}(x,y)=\sum_{i=-k}^{k}\sum_{j=-k}^{k}P_{binary}(x+i,y+j) \qquad (3)$$

expand masks of PBCN and a 9×9 kernel expand masks of PBGSUB binaries, we call it PBCN-exp and PBGSUB-exp, finally we add PBCN-exp and PBGSUB-exp to a final mask of every frame, and conduct a mask operation to extract foreground image.

3) SMOF

In the process of improving the fusion methods, we realize that there is a point to be worth concerning, we notice that three extractions are unstable enough in one video, there is a high probability to occur successful detection on previous frame but failure on next several frames, to solve the flaw, we choose to utilize the correlation and gradual on video, considering most moving of objects is traceable and limit in a minor area, we design two approach to enhance the mask stability of adjacent mask frames.

To SSD and PBCN which the phenomenon occurs more seriously, we denote that Given a video *V*, we obtain a set of mask matrix extracted by PBCN, SSD, PBGSUB:

$$\mathbb{S}(V)=\left\{S_1,S_2...,S_M\right\} \qquad (4)$$

where *S* is a gray scale matrix of frame in video *V*, M is number of frames of *V*, due to noise of PBCN, SSD is little enough that we can ignore, we choose to use a mask which sums all mask in video V to represent the video-level mask:

$$S_v=\sum\mathbb{S}(V)=\sum_{i=1}^{M}S_i \qquad (5)$$

Because gray scale value is between 0 to 255, if result of sum function overflow, we set it to 255. But simple sum of mask is not adaptive in PBGSUB, the threshold of low pass filter is set by handcraft, in experiment, it shows a satisfied result but not effective in every sample, the pure sum will enhance the impact of noise seriously, so we introduce momentum, use momentum to combine the preview historical and current information. First we use formula (1), (2), in experiment, we set average kernel k to 9 and threshold T to 70 to extract the mask of PGBSUB, we call the set of masks as $S_{\text{PGBSUB-exp}}$, for a new mask, we use:

$$\widetilde{S_i}=\begin{cases} S_i & if\ \ i=1\\ (1-d)\cdot\widetilde{S_{i-1}}+d\cdot S_i & otherwise\end{cases} \qquad (6)$$

where S_i is number i of mask frames in video V appeared in formula (4), d is the parameter of momentum between 0 and 1 to decide the rate of fusion, the new mask is a combination of preview mask and current mask, so when fail detecting on current frame, we can use previous mask to estimate the object area, in experiment we set d to 0.5, and then we use a low pass filter(2) with T=25 to rebinary, finally we sum the video-level masks $S_{\text{v-ssd}}$, $S_{\text{v-pbcn}}$ handled by (5) and the new set of PBGSUB masks to rebuild a new set of masks, from now on, we call it saliency and motion object foreground extraction(SMOF), we use the set of SMOF to generate foreground dataset corresponding with original dataset through mask operation, and send it to C3D as input with sports-1M pretrained model to finetune.

B. Feature Choosing

For a video example, it is composed by several adjacent 16 frames input unit, and as Fig 1, all the features of every layer are the output of 16 frames, because the features of fore layer mainly concentrate on the partial information, mostly we use the innerproduct features as fc6, fc7 and probability of final layer as fc8, prob. In C3D, the author use fc6 feature as their video descriptor, with his experience, we design a series of experiments to explore which fusing way is the best, we exact the fc6, fc7, prob features as clips-level descriptors to represent 16 adjacent frames. Fc6, fc7 is the last two innerproduct layers consisted of 4096 dimensions, it is the one-dimensional representation which is between the spatio-temporal feature and probability of ConvNet, the prob is the probability vector to represent probability of every label. To fc6, fc7 feature, we design three ways, fc6, fc7 alone, and combine them as a 8192-dimensions vector, then we average these clips-level descriptors after L2 normalization to generate video-level descriptors and input them to a multi-class svm classifier, to prob, because they are probability vector, we only need to sum prob vector in a video and calculate the Max probable label. We test the 4 types of features on UCF101, and find that fc7 does not assist to promote the performance of result and the performance of prob is not good as fc6, finally we choose fc6 alone with 4096-dimensions vector as features of 1 net.

C. Feature Fusion

Now we have two 3D ConvNet, one is ConvNet trained on original dataset, the other one is trained on corresponding foreground dataset, referring to the result of Table 1, we adopt the foreground extraction method of PBCN+PBGSUB and SMOF, we extract both fc6 feature of PBCN+PBGSUB and SMOF and average them after L2 normalization to form video-level descriptors. We combine them with original RGB feature separately as 8192-dimensions features and combine them all as 12288-dimensions features and input them to linear svm, but the result of experiment shows that the difference of 3 net and 2 net is very close, finally we choose 2 net fusion with 8192-dimensions features to be our final video-level descriptors.

IV. Experiment

A. Dataset

In order to verify the effectiveness of SMOF, we conduct experiments on two public large datasets, namely HMDB51 [10] and UCF101[9]. The HMDB51 dataset is a large collection of realistic videos from various sources, including movies and web videos. The dataset is composed of 6, 766 video clips from 51 action categories, with each category containing at least 100 clips.

The UCF101 dataset contains 101 action classes and there are at least 100 video clips for each class. The whole dataset contains 13,320 video clips, which are divided into 25 groups for each action category.

B. Implement Detail

Because the two data input is RGB channel, both ConvNets we adopt are C3D, which consist of 5 convolution layers and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), 2 fully-connected layers and a softmax loss layer, the pretrained model of foreground dataset is hard to solve, the best action pretrained dataset is sports-1M which has 1.1 million videos, it means we need at least a month to conduct pretrainning, so with excellent performance in C3D, we adopt sports-1M pretrained model trained on C3D in original dataset to be the pretrained model of foreground data. Both ConvNet we set base_lr to 10^{-3}, momentum to 0.9, weight_decay to 10^{-3}, we set step method to be our strategy of learning rate declining, and we set learning rate to decline to0.1 of current base_lr every 10000 iteration, we conduct 20000 iteration at all, then we use the two train model to extract and combine to 8192-dimensions features with L2 normalization after averaging. For foreground dataset choosing, as experiment in Table 3, the difference between PBCN+PBGSUB and SMOF is not as obvious as we think, we use both foreground datasets generated by PBCN+PBGSUB and SMOF to conduct experiment.

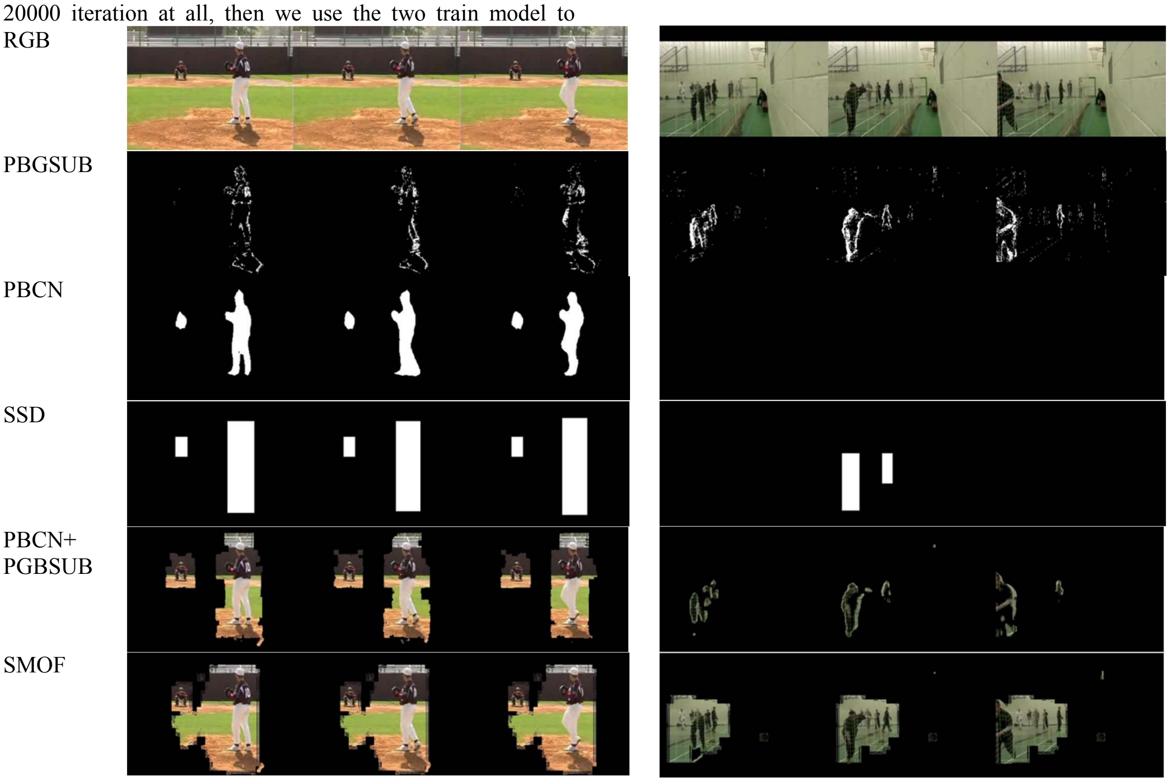

Fig 3 There are 2 examples to be chosen of 5 different foreground extraction approach, right example shows a simple scene that object is significant different with background, static extraction approach (PBCN, SSD) represent a precise distinction, but in a complex scene that object is faint different with background, static extraction approach is complete failure, motion exaction approach shows the better robust.

C. *Result*

We first test the performance of foreground image extracted by PBCN, PBCN+PBGSUB and SMOF in single C3D, **Table 1** is the result and shows that with only less than 50%-pixel information, the foreground achieves not bad performance only lower than original C3D 4.4%, the C3D uses 16 consecutive frames as input unit without overlap, after we change the train input to 16 consecutive frames with 12 frames overlap, the result of foreground dataset generated by SMOF rise to 78.8% which is close to original C3D, meanwhile the original dataset result with 12 frames overlap is only rise 1%. Then we test the feature extract from 1 Net using original RGM frames to find which feature is suitable to be our video-level descriptor, Table 2 is the result of feature choosing in UCF-101. The result shows that the differences of innerproduct feature is little, actually **fc6 feature** with linear svm is enough.

Table 1 Explore of different fusion strategy for C3D on the UCF101 dataset

Strategy	Performance
PBCN	58.9%
PBCN+SSD	66.5%
PBCN+PBGSUB	72.6%
SMOF	75.6%
SMOF(4 frame overlap)	78.8%

Table 2 feature choosing result on UCF- 101

Method	Accuracy(%)
fc6+svm	82.3
fc7+svm	82.2
fc6,fc7+ svm	81.9
prob	82.1

Table 3 is the result of feature fusion, the result shows that the advantage of fusion 3 net (RGB, PBCN+PBGSUB, SMOF) is not obvious, only original RGB features and one foreground feature is enough. Finally, fusing RGB feature and one foreground feature is our final structure as Fig 2.

Table 3 feature fusion result on UCF- 101, HMDB 51

Method	UCF101	HMDB 51
C3D (PBCN+PBGSUB)+C3D	83.4	50.9
C3D(SMOF)+C3D	83.7	49.6
C3D(PBCN+PBGSUB)+C3D(SMOF)+C3D	83.7	50.6

Table 4 represents action recognition accuracy of C3D compared with original C3D and other action recognition approaches on dataset UCF101. The reason C3D(3 net) is higher than us is that C3D(3 net) use three datasets and three different pretrained models， different pretrained models enhance complementarity between features, but it's a higher cost approach especially in dataset collecting and training. We keep one dataset, same network structure, pretrained model, and have a 1.4% promotion just relying on reducing background area.

Table 4 Action recognition results on UCF101

Method	Accuracy(%)
Deep networks[11]	65.4
Spatial stream network[5]	72.6
LRCN[22]	71.1
LSTM composite model [16]	75.8
C3D (1 net) + linear SVM[8]	82.3
C3D(SMOF)+C3D+ linear SVM	**83.7**
C3D(3 net)[8] + linear SVM[8]	85.2
Two Stream[5]	88.0

The result on HMDB51 is unsatisfied because HMDB51 is a small dataset, and huge number of parameter of C3D will occur serious overfitting, Table 5 represents the action recognition accuracy of our method, C3D and other methods.

Table 5 Action recognition results on HMDB51

Method	Accuracy(%)
STIP+BoVW[10]	23.0
Motionlets [17]	42.1
DT+BoVW[18]	46.6
C3D(1 net)+linear svm	49.9
C3D (PBCN+PBGSUB)+C3D	**50.9**
Two stream[5]	59.4
TDD+FV[7]	64.2

In Table 5, C3D does not show an advantage, even highest accuracy has achieved 50%, we believe it is led by unbalance of heavy network structure in small scale of dataset. The noteworthy question is that SMOF enhance stability which perform better than PBGSUB +PBCN with single net and fusion in UCF101, but failed to gain profit for C3D when fusion, but PBGSUB+PBCN have a 1% promotion, we believe that because we adopt same network structure and same pretrained model. To ensure object area to be extracted, we use an extensive object expansion strategy meanwhile it will increase the risk of homogenization between foreground network model and original network model and decrease complementarity of two model, In small scale dataset，we believe the complementarity is more important than stability.

V. Conclusion

This paper proposes an effective video presentation, which integrates original ConvNet feature and foreground ConvNet feature based on fusion of static object and motion extraction. Currently it appears that foreground images have almost the same expressive capacity as full RGB images on 3D-ConvNet and it does enhance the complementarity through ConvNet feature fusion. The experimental results show that our features achieve satisfactory performance on two dataset for action recognition.

References

[1] I. Laptev and T. Lindeberg. Space-time interest points. In ICCV.

[2] O. Boiman and M. Irani. Detecting irregularities in images and in video. IJCV, 2007. 1, 2

[3] Zhang T, Wang R D. Copy-Move Forgery Detection Based on SVD in Digital Image[C]// International Congress on Image and Signal Processing. IEEE, 2009:1-5

[4] H.Wang and C.Schmid.Action recognition with improved trajectories. In ICCV, 2013.

[5] Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recog- nition in videos. In: NIPS. (2014) 568–576

[6] L Wang, Y Xiong, Z Wang, Y Qiao, D Lin: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: ECCV. 2016 , 22 (1) :20-36

[7] Limin Wang; Yu Qiao; Xiaoou Tang: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR, 2015: 4305 - 4314

[8] Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotem- poral feature with 3d convolutional networks. In: ICCV. (2015) 4489–4497

[9] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recogni- tion. In ICCV, 2011.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[12] https://github.com/w865194269/Pixelwise-Binary-Classification-for-Salient-Object-Detection.git

[13] W Liu,D Anguelov,D Erhan,C Szegedy,S Reed :S Single Shot MultiBox Detector. European Conference on Computer Vision , 2016 :21-37

[14] MV Droogenbroeck,O Paquot: Background subtraction: Experiments and improvements for ViBe. Computer Vision & Pattern Recognition Workshops , 2012 , 71 (8) :32-37

[15] https://github.com/sagi-z/BackgroundSubtractorCNT

[16] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In ICML, 2015.

[17] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3D parts for human motion recognition. In CVPR, 2013.

[18] H. Wang, A. Kla¨ser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. IJCV, 103(1), 2013.

[19] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. CVPR, 2008.

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In ACM MM, 2007.

[21] A. Kla¨ser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In BMVC, 2008.

[22] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convo- lutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.