Computer Communications 147 (2019) 35–39

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Crowd Video Event Classification using Convolutional Neural Network

S. Jothi Shri [a,*], S. Jothilakshmi [b]

[a] *Department of Computer Science and Engineering, Annamalai University, Annamalainagar, India*
[b] *Department of Information Technology, Annamalai University, Annamalainagar, India*

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural network (CNN)
Crowd event classification
Support vector machine (SVM)
Deep neural networks

ABSTRACT

Crowd Event Classification in videos is an important and challenging task in computer vision based systems. The Crowd Event Classification system recognizes a large number of video events. The decisive of the model is a difficult task in the event classification. The event classification model has generalization capability on works with a higher number of videos. The embodiment of Deep Learning in video event classification derives powerful and distinguishes feature portrayals. The features of events are extracted from raw data through massive videos with effective and efficient detection. The Convolutional Neural Network (CNN) has been established as a powerful classification model for event recognition problems. A higher quality of new dataset of 3000 frames collected from youtube videos belonging to four classes of crowd events namely Marriage, Cricket, Jallikkattu, and Shopping mall The system has used two Deep CNN infrastructures are namely baseline and VGG16, which detects predefined events and provides temporal evidence. The CNN Model automatically tests input video frames and detects the events of centrality at the video. The CNN extracts the video events features from the video input frames and distinguishes the events name correctly. The system shows more improved 100% results compare with each other models.

1. Introduction

Crowd Event Classification in videos has great potential in many applications and it is useful to detect pertained crowd events from the input videos and classify them. An event is a recognizable occurrence in scenes of videos and is characterized by the subject. The Crowd Event Classification is a difficult process due to various subject contained in the input videos. The subject of the event can be human action recognition, motions, and places. Deep Convolutional Neural Networks have significantly advanced in computer vision and natural language processing. The CNN network is very operative in the performance of describing the high level and difficult features of video input abstraction data over a hierarchical learning method. Deep Neural Network achieves superior performance in visual object recognition [1]. Under these conditions, the Convolutional Neural Network (CNN) learns the potent and interpretable event of video features. Stimulated with good results in the domain of video, the system studies the involvement of CNN in video event categorization. The CNN model [2] has access to not only for the appearance information present in single static frames but also their composite time-based development. The temporal gets a sequence of frames of video as input. There are various testing challenges in expanding and applying CNN Model.

Many vision approaches have been developed for events classification under different environment and activities. The existing system [3], automatically classifies such videos into corresponding classes. The large scale events are taken as the classes for evaluating statistics of features. The explorations of objects are identified and tracked their pose from the events. The process of detecting entities is a difficult and challenging task. The main focus of the event classification has been on techniques based only by using statistical features.

In [4], VGG-16 network model is used to describe the video event features with 13 Convolutional, 5 Max pooling and 3 Fully connected layers with another Soft-Max layer. The system is trained by Imagenet network model with 1000 categories to describe the video event features. The classifier learned to classify the extracted features. The DevNet system [5] contains 9 convolutional layers and 3 fully connected layers. A spatial pyramid pooling layer is implemented in between these two measures of the system.

Without appropriate data and its parameters, DevNet network model is very problematic to attain and recognize the events. Thus spatial–temporal key evidence used for event recounting of DevNet. Thus a spatial–temporal key used for event recounting in Devet.

The model of spatial stream ConvNet [6] is trained from the two different datasets. The datasets ILSVRC-2012 and UCF-101 are used for pretraining and fine-tuning respectively. The Convert based event classification system performance is 72%. In contrast to the variety of video features, Support Vector Machine was the lead classifier for over a decade. In recent times, The Deep Learning based approach is very popular and neural networks have also been implemented for

[*] Corresponding author.
E-mail addresses: sri.sankar6@gmail.com (S.J. Shri), jothi.sekar@gmail.com (S. Jothilakshmi).

https://doi.org/10.1016/j.comcom.2019.07.027
Received 11 June 2019; Received in revised form 10 July 2019; Accepted 30 July 2019
Available online 8 August 2019

video classification. From a practical point of view [7], the video classification benchmarks are not presently available to match the scale and existing video datasets because videos are extensively additional problematic to gather and stock.

The proposed system of Crowd Event Classification performs as a necessary and indispensable phase in the process of evaluating the video contents (Marriage, Cricket, Shopping mall, and Jallikkattu). The system will be immensely cooperative if the enormous event videos on the web can be routinely categorized into predefined classes. The video contains visual information of the event can be processed on a frame basis using Convolution Neural Network (CNN). The system event has randomly initialized Deep CNN and performed on high-resolution video frames. The huge amount of labeled statistics data used for working out the model.

The classifier learns the relative significance of the features from the training set of data. The CNN introduces two approaches namely baseline and VGG-16 feature representations to improve the effectiveness of the model. The model introduces temporal (video frames) connectivity of architecture to improve the computational time. The significant improvements are identified from the results of two different representations of the CNN model. Fig. 1 shows the steps involved in the network proposed system.

The rest of the paper is prepared as follows: Section 2 discusses a more comprehensive survey of correlated mechanism. Section 3 elaborates on the CNN approach of the proposed system for classifying events. Evaluation of the framework methods and experimental results of the proposed system are provided in Section 4, where the system also briefly introduces the new dataset. Finally, Section 5 concludes the performance results of the paper.

2. Related work

A model of ImageNet Classification [1] needs to be trained by large scale objects from lots of images for a high knowledge power. The difficult task of object identification is handling a large dataset. Thus, the model of object recognizer should also collect lots of knowledge about data. Convolutional Neural Network constitutes in such models for classification The system vary their depth and breadth to control their capacity. CNN has some connections and parameters. These are very easier to train compared with standard feed-forward neural network layers.

The two-stream Convolutional Neural Network (CNN) [2] has the architecture same as a general CNN for video classification. The system takes individual frames of video as network inputs surveyed by many convolutional layers, pooling layers, and fully connected layers. Finally, after a Soft-max, the output ranges [0,1] are the predicted output probabilities from the trained model. The selected frames are tested from each video. The model processes each frame and individual frames are combined by an average of probability mixture for the predictions

The dataset EK0 [3] is automatically pruned for event-related training dataset from the web site of youtube with an algorithm. The text and vision-based features are used to retain only the most related video content. The selected video is represented by features and event classes with Convolutional Neural Network, together with dense lines. The system analyses the different steps of the algorithms namely, query generation, expansion and pruning on the test videos, and the methods are compared to the TrecVid MED 2013 EK0 dataset. The dataset contains 24,957 test data. The system shows that the model attains good performance with some improvements. The dataset gets the result of more than 30% mean average precision (MAP) over recent results.

A Deep Convolutional Network infrastructure [4] is specified as namely Deep Event Network (DevNet) that simultaneously detects events and evidence. The keyframes as input, the model detect the correct event from video by collecting the features. The evidence recount the event detection results. The results are inevitably limited by the methods of temporally and spatially. Videos have key evidence at each frame levels usually. The CNN properties first generate a spatial–temporal saliency map by back passing through DevNet. The DevNet finds the keyframes and most suggestive to the event.

In static image analysis, deep learning techniques such as CNN have been efficient and are now also being used for temporary information. For a large dataset sports videos demonstrated their efficiency for video classification [7]. By incorporating appearance and movement data suggested a completely convolutionary neural network for crowd segmentation. Previously Some video classification system suggested and assessed several profound neural network architectures and studied multiple temporal convolutionary features pooling architectures but did not concentrate on any specific assignment of video analysis. Thus the system used CNNs to determine the actor, action, and location of videos using multi-label classification suggested for this purpose a big dataset of 10,000 videos and also used profound networks to recognize action by extracting data about movement and appearance and finally concatenating their outcomes. The system explored different strategies for doing event detection in videos using CNNs trained for image classification.

3. Deep learning approach

In this system, the concept of basic deep neural network model has been widely adopted for event classification. In general terms, DNN is a mathematical model that studies from a huge amount of categorized video input frames. The essential nerve system in the brain of animals is inspired to create neural network architecture.

3.1. Convolutional Neural Network (CNN) baseline

Convolutional Neural Network [7,8] is essentially like a fully connected multilayer perceptron. Weights and biases are network neurons. A dot product between input and neuron weights is the main functionality in the network.

Convolutional layers, pooling layers, and fully connected layers are contained in the CNN architecture as shown in Fig. 2. The convolutional layers are performing in the main function of CNN. The convolutional layer is connected nearby each neuron to a small region of the video input data at the same time. The region is called a receptive field.

Each neuron of CNN layers is considered by a dot product among its weights and the small input region. The neurons are described by the whole area of the image and it is covered by the respective fields of every neuron. Every neuron performed by sliding the layer weights. Both The resulting convolution output and the input image are the same size $(224 \times 224 \times 3)$. The receptive field depth gets all the color information from the input image. The receptive field is the same as the input color channel.

The baseline architecture is used to contribute to the static appearance to the event classification accuracy. The video inputs size $(224 \times 224 \times 3$ pixels) is given to the network. The convolutional layer specified by $C(d, f, s)$ with d filters of spatial size 3 × 3, and used to the input with strides. $FC(n)$ F is a fully connected layer with n nodes. All pooling layers Pool spatially in non-over lapping 2 × 2 regions and all hidden weight layers use the rectification (ReLU) activation function [9,10] with parameters $k = 2, n = 5, \alpha = 10, \beta = 0.5$. The final layer is connected to a softmax classifier with dense connections.

The system gives the convolution output is known as a feature map. The number of feature maps is specified by the layer depth K [11]. The convolution layers computed data in an array which creates the model with three dimensional. The output layer gives feature maps and K specified by use case.

A collection of the neuron is called as one feature map. Each neuron in the K network layer has the same amount of weights. The convolution layer used the method of parameter sharing to limit the number of weights. Each neuron in K network layers has shared the same parameters as shown in Fig. 3. Every network layer K gives

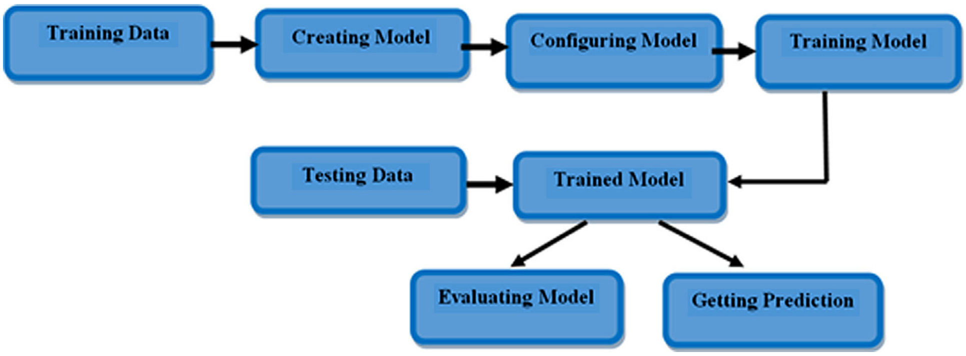

Fig. 1. Steps involved in the proposed system.

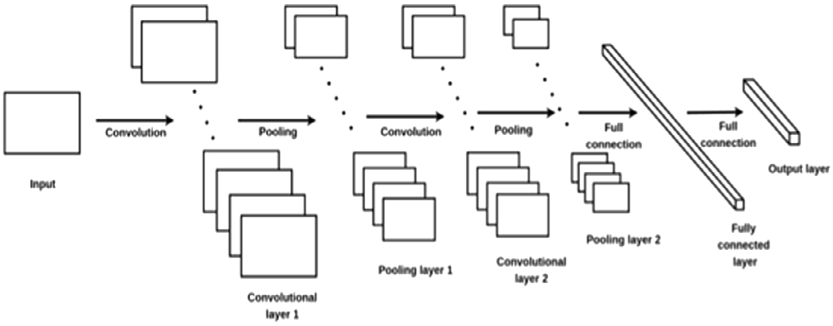

Fig. 2. Architecture of CNN.

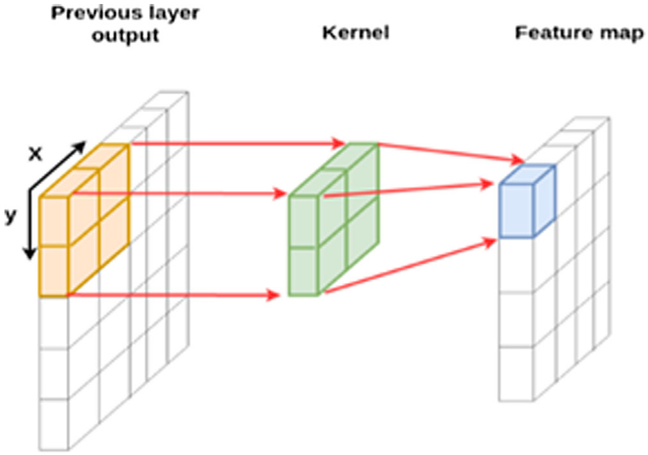

Fig. 3. Kernel of network layer.

a feature map and it which needs one set of weights to reduce the memory requirements. The convolution kernels are known as filters defined by a set of weights. The reason for the shared weights the filter is implemented and convolved through the entire three-dimensional space of the input tracked by certain non-linearity. The output feature map is two-dimensional filters applied at each position of the given input. The relatively small number of weights is one of the main reasons why The CNN model [12–15] is good for creating general features because the main reason is related to a small number of weights.

3.2. Convolutional Neural Network VGG 16

The proposed system implements the VGG16 model to compare the performance with the baseline model. In VGG-16 model has the input to the convolution layer is 150 × 150 image in a fixed size. The picture is carried through a stack of convolutionary layers where the filters were used with a very tiny receptive field 3 × 3, the smallest size to capture the concept of left, right, up, down, middle. It also uses 1 × 1 random convolution filters in one of the configurations, which can be viewed as a linear input channel transformation. The convolution step is set at 1 pixel; convolution spatial padding. The input of the layer is such that after convolution the spatial resolution is maintained, i.e. the padding is 1-pixel for 3 × 3 sequence layers. The convolution step is set at 1 pixel; conv spatial padding. Spatial pooling is performed by five layers of max pooling that follow some of the convolution layers. Max-pooling takes place over a window of 22 pixels, with step 2. All concealed layers are fitted with non-linearity of rectification (ReLU). It is also observed that none of the networks (except one) contain Local Response Standardization (LRN), such standardization does not enhance the efficiency of the ILSVRC dataset, but results in enhanced memory consumption and computation time.

4. Crowd event classification

The proposed system has used different event videos to classify correctly for identifying video events easily. Many surveillance systems have used the crowd event classification system to monitor and maintain the crowd and traffic.

In the proposed system, the success of CNN model [16–18] features on video exploration process has stimulated of deep features for video event classification. The features are extracted for each frame collected from videos. The representation of features could be derived by successively a feed-forward pass of the fully connected layer. The model is trained on a new dataset of videos. The frames of features are averaged into video representations as inputs of standard classifiers for recognition. The model studied the performance of frame-based video recognition using features from the layers of a deep convolutional model organized with various kernels for event classification. The system has collected new dataset, which consists of 3000 images from a variety of crowd event videos from youtube belonging to four classes of events namely Marriage, Cricket, Jallikkattu, and Shopping Mall. Fig. 4

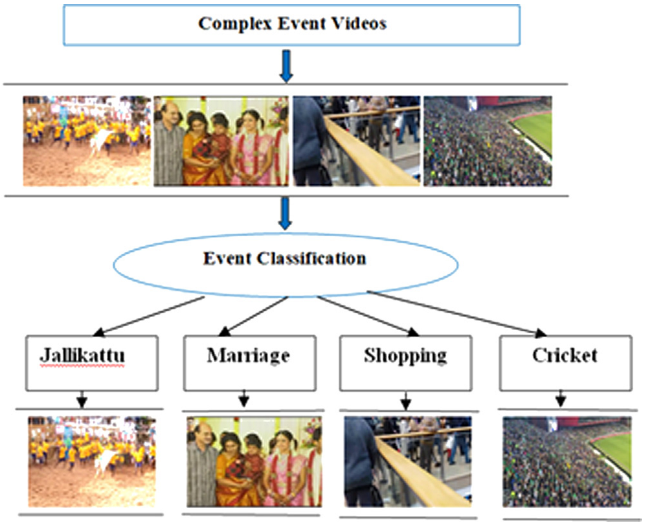

Fig. 4. Event classification.

shows event videos classified into four event classes by the proposed system.

CNN [19–21] requires extensive periods of training time to effectively enhance ten thousands of parameters that parameterize the model. The issue is compounded when extending the connectivity of the architecture in time because the model must process not just a single frame but several frames of video at a time. To moderate this computational issue, the system shows that an effective approach to rapid up the runtime performance of CNN to modify the architecture to contain two separate models of processing stream namely CNN baseline and VGG16 model that learns features on the different resolution of frames that only operates on the middle portion of the frames. The performance of the network in runtime increases due to the reduced dimensionality of the input while retaining the classification accuracy.

The system analyzes the performance of two visual feature representations namely baseline and VGG-16 of CNN model. The baseline of CNN shows a better performance for four events. The CNN architecture model takes a different approach to combine information across the time domain for testing videos. The performance improved by temporal connectivity pattern in CNN architecture [22–24] and it predicts information in the video. The framework performs significantly on event detection task and achieves satisfactory results. The system confirms the importance of learning and representation of the event detection with keyframes.

5. Experiment and results

5.1. Dataset

The proposed system applies Convolutional Neural Network (CNN) to classify 4000 data samples and reports major developments in the result. The results of baselines and VGG-16 gives the result based on Features established by training networks. A Deep Learning is established on event classifier trained through 3000 frames of videos. First, randomly selected 1000 images per event category are a training set and 1000 images are a validation set for 4 categories. The deep CNN achieved 100% event classification accuracy on the validation set after training.

5.2. Implementation work

Deep learning algorithm has been implemented by Python 3.5 version along with Anaconda Library. OpenCV (version 3.3.0) library imported for handling video and image files easily. Jupyter Notebook is very helpful for compiling Crowd Event Classification system of python code.

Table 1
Computation result of the experiment.

Epoch	Baseline training accuracy	VGG 16 training accuracy
1	73	82
2	100	90
3	100	96
4	100	82
5	99	88
6	100	92

Table 2
Performance of crowd event classification system.

Model	No. of dataset			
	True positive	True negative	False positive	False negative
CNN baseline	100	40	0	0
VGG16	100	40	40	8

5.2.1. Parameters

The CNN network of the proposed system has some parameters to design the architecture. There are three convolution layers followed by the activation function of Relu and max-pooling layer [25]. The filter kernel size is 2 × 2. There are four neurons in the output layer since the model is trained for four classes. For classifying the dataset of events, especially one activation function of categorical-cross-entropy is used.

5.2.2. Training

In the training phase, totally 6 epochs and 4000 training samples are used to generate the model for extracting features and training. The event dataset is stored in a stack array and resized into 150 × 150. The array is changed into the batch file to pass data to the model for training. The model passes features through epochs and classifies data using a separate label (0, 1, 2, and 3) for each event.

5.2.3. Testing

In the testing phase, different video events are taken and converted these into frames. Totally 100 true datasets of events are stored in an array of a batch file and resized into 150 × 150. The frames are collected from different events namely Jallikattu, Marriage, Shopping, Cricket. From each event, 25 frames are collected and stored in the test folder. 10 false datasets are collected from other videos and stored in the test folder. The batch file of testing data is passed to the trained model. The model classifies the dataset into four categories of events and shows label name for each class correctly.

Table 1 compares the validation results of the CNN baseline and VGG16 model. The system found the variations of each epoch in the model of CNN baseline and VGG 16 Model. The validation resulting variations of CNN Model are shown in Fig. 5. The proposed system implements both models CNN baseline and VGG 16.

Fig. 5 shows the accuracy of both models in each epoch. The baseline accuracy is in the first epoch 68%. Since it shows 100% in next 5 epochs constantly. But the VGG 16 model gives less accuracy 92% and not constantly in continuous epochs compare to the baseline model. The CNN baseline gives 100% validation accuracy.

The main goal of the proposed system is used to acquire in what way to recognize events on various crowd videos. The proposed system has applied CNN baseline and VGG-16 for crowd video event classification. The network model predicts the scenes of the test images at that time. The performance of the CNN baseline and VGG-16 model is evaluated and calculated as true positive, true negative, false positive and false negative in Table 2. Totally 100 true datasets and 40 false datasets are taken to classify the events from crowd video.

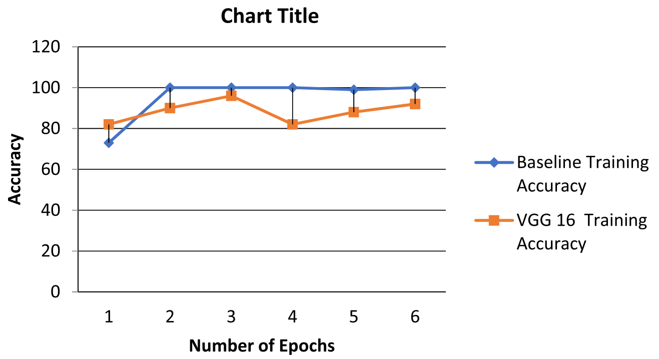

Fig. 5. CNN evaluation of accuracy on training.

Table 3
Computation results of the experiment.

Model	Precision	Recall	F1score
CNN baseline	100%	100%	100%
VGG16	71%	92%	80%

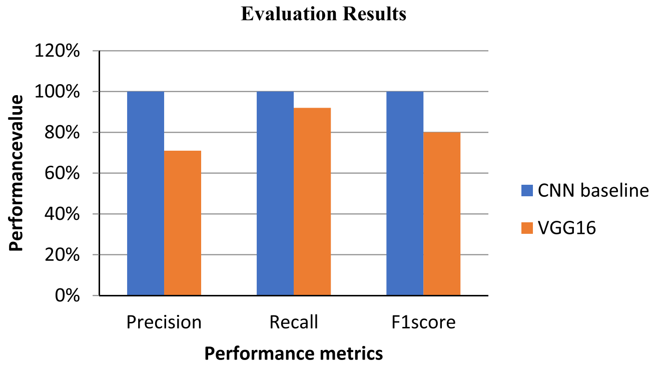

Fig. 6. Performance valuation of results . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The computation result of the event classification system is analyzed and quantified as precision, recall, and f1score. The results of CNN baseline and VGG-16 are given in Table 3.

The event classification system performs test and evaluation. The evaluation results are represented graphically in Fig. 6. Test samples are classified into four classes. The test results are compared and showed in different colors. The CNN baseline method gives 100% percentage of accuracy.

6. Conclusion

The Crowd event classification system provides a Convolutional Neural Network model to classify the crowd events collected from youtube videos. The proposed system trains the model of CNN with 4000 video frames of four categories. There is two feature representation of approaches are implemented namely baseline and VGG16 model to classify the video frames. The performance of the two approaches is compared and showed. The VGG16 model gives the 82% result and it shows inconstant accuracy in continuous epochs. The proposed system has found the best result of 100% constant in continuous epochs in this work suggests a baseline approach for crowd event classification. The baseline model shows improved runtime performance at low cost. In future work, the high number of crowd video event dataset will be considered to classify video events in the best way. Mainly the proposed may use this work in video survelliance applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.comcom.2019.07.027.

References

[1] E. Hinton, IlyaSutskever, ImageNet Classification with Deep Convolutional Neural Networks, Science Department, University of Toronto, 2009.
[2] K. Simonyan, Two- stream convolutional networks for action recognition in videos, CVPR, 2016.
[3] Chen Sun, Ram Nevatia, Large Scale Web Video Event Classification by use of Fisher Vectors, University of Southern California, Institute for Robotics and Intelligent Systems Los Angeles, USA, 2013.
[4] Nicolas Chesneau, Karteek Alahari, Learning from Web Videos for Event Classification, IEEE and Cordelia Schmid, Fellow, IEEE, 2017.
[5] Chuang Gan1, Naiyan Wang, DevNet: A Deep Event Network for multimedia event detection and evidence recounting, ConferencePaper, 2015.
[6] Bingbing Ni, Yang Song, Ming Zhao, YouTube Event: On large- scale video event classification, Published 2011 in. Tran D, Bourdev L D, Fergus R, C3D: generic features for video analysis, 2014.
[7] A. Karpathy, G. Toderici, S. Shetty, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[8] B. Fernando, S. Gould, Learning end-to-end video classification with rank-pooling, in: International Conference on Machine Learning, 2016, pp. 1187–1196.
[9] D. Tran, L.D. Bourdev, R. Fergus, C3D: generic features for video analysis, CoRR, 2014.
[10] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, 2016.
[11] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[12] S. Zha, F. Luisier, W. Andrews, Exploiting Image-Trained CNN Architectures for Unconstrained Video Classification, Computer Science, 2015.
[13] J. Sánchez, F. Perronnin, T. Mensink, Image classification with the fisher vector: Theory and practice, Int. J. Comput. Vis. (2013) 222–245.
[14] S. Ji, W. Xu, M. Yang, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2012) 221–231.
[15] L. Sun, K. Jia, D.Y. Yeung, et al., Human action recognition using factorized spatiotemporal convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.
[16] H. Ye, Z. Wu, R.W. Zhao, Evaluating two-stream CNN for video classification, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 435–442.
[17] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.
[18] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[19] L. Wang, Y. Xiong, Z. Wang, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer International Publishing, 2016, pp. 20–36.
[20] B. Zhang, L. Wang, Z. Wang, Real-time action recognition with enhanced motion vector CNN's, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2718–2726.
[21] X. Wang, A. Farhadi, A. Gupta, Actions transformations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2658–2667.
[22] W. Zhu, J. Hu, G. Sun, A key volume mining deep framework for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1991–1999.
[23] H. Bilen, B. Fernando, E. Gavves, Dynamic image networks for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3034–3042.
[24] B. Fernando, P. Anderson, M. Hutter, Discriminative hierarchical rank pooling for activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1924–1932.
[25] B. Fernando, E. Gavves, J. Oramas, Rank pooling for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2017) 773–787.