THINK INDIA JOURNAL ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Current Perspective of Deep Learning Based Human Activity Analysis

Chandni, Gagan Amrit Kaur
School of Computer Science & Engineering
Lovely Professional University, Phagwara, India
chandnikathuria5@gmail.com, gaganamrit27@gmail.com

Abstract

Activity analysis from videos is a highly active area of research with applications over a wide spectrum including health monitoring, intelligent surveillance systems, human computer interaction, robot learning, sports analysisand context based retrieval. As with other research areas of computer vision, recent work has focused on applying deep network models to accomplish the task of activity recognition as it dramatically improves the recognition performance by exploiting the video frame structure. The survey basically captures the current perspective of deep model based solutions for activity recognition from video sequences and imparts potential directions for future research. It also provides comparative view of video datasets usedin recent literature.

Keywords: Deep learning, Activity recognition, Handcrafted features, spatio-temporal

I.Introduction:

A. Taxonomy of activity recognition

Identifying human activities from video streams is an area of concern for computer vision researchers. Before delving into current trends in human activity recognition, we need semantic awareness of its terminology as given in Fig.1. According to Thomas B. Moeslund et al.[1] an action primitive is the most basic movement like moving hand whereas as an action consists of different action primitives and finally, an activity contains a number of subsequent actions with particular semantic meaning, like clapping. Another novel hierarchy is given by Aggarwal et al.[2] presented a viewbased on complexity level as shown in Fig.1(b).

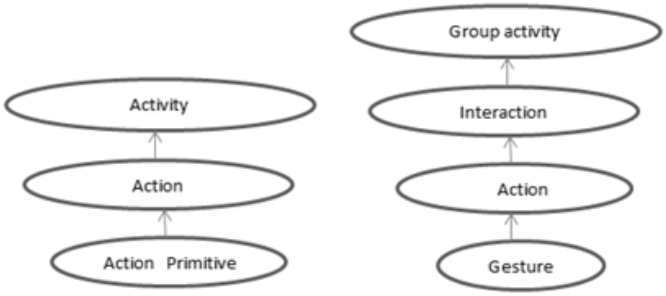


Fig-1: Taxonomy a)Thomas B.Moeslund et al.[1] b)J. K. Aggarwal et al.[2]

 Copyright © 2019Authors

B. Challenges of the domain

Despite of tremendous, activity recognition methods may still misclassify videos due to various difficulties [3]:(i) different styles of human motion, action rate and environmental variations.(ii)Overlap between activity classes(iii) illumination conditions.(iv)Insufficient datasets which are quite difficult to generalize to real world applications. Hence activity recognition methods should be robust enough to be invariant to such harsh real worldscenarios.

C. Activity recognition as feature representation and classification

The overall task of human activity recognition consists of two main steps, feature representation followed by classification both of which can be accomplished in different ways. This paper primarily captures recent trends in application of deep learning approach to feature extraction and classification for activity recognition. However earlier work on activity recognition mainly involves **global and local feature representations**. Global features capture holistic information of whole scene comprises of geometric structure, motion, appearance, common examples are Motion energy image(MEI),Motion history image(MHI) introduced in in Bobick et al.[4].This global representations are variant to noise and cluttered background.**Local features** such as spatio-temporal interest points(Ivan Laptev[5]) describe human motion in space –time regions. Main advantage of spatio-temporal interest points is robustness to translation and appearance variation. But they captures only short-term duration information, trajectories are useful at capturing long-term information. Despite of the great success achieved by local and global feature methods, hand crafted features are very time taking and demands for the domain knowledge. Therefore feature extraction and classification using **deep learning** is receiving an increasing attention.

II: Datasets

Due to ever increasing importance of activity recognition, various video datasets dedicated to action and activity recognition have been created. These datasets are either heterogeneous or specific in terms of type of activities they involve. Heterogeneous datasets involve variety of activities where as specific dataset are meant to learnspecific activities. These datasets vary in terms of features also, like number of views, number of actors, scenes, camera movement etc. Table:1 provide summary of some commonly used datasets in the literature.

III. State-of-art deep learning methods

A.Qualitative Analysis

 Copyright © 2019Authors

Activity recognition from videos involves learning classification model or decision function using action features extracted from available video dataset, so as to identify the activity class label. This sectioncomprehensively discusses the insights of potential deep learning models that have been used in recent literature to address the problem of activity recognition from videos. Deep neural network architectures address this by applying multiple levels of non-linear operations.Popular deep architecture useful for both action feature extraction and classification is Convolution Neural Network, which mainly consists of set of convolutional and pooling layers followed by set of fully connected layers for classification.For activity recognition CNN must be able to learn along temporal dimension.Simplest approach by ShuiwangJi et al. [6] makes use of 3D convolution filters to learn action features from spatial as well as temporal dimension. Activity recognition is actually feature extraction followed by classification; both of them can be achieved by different ways using deep networks. Features extracted using skeleton information are classified using CNN in Mo et al.[7] . Parka et al.[8] uses joint angle features with recurrent network.In Baccouche et al. [9] learnt features are processed by LSTM recurrent architecture for identifying activity class.Dobhal et al. [10] implemented CNN model with BMI (Binary Motion Image) input that is better at handling background occlusions. Fusion based approaches[11-13] also boost classification precision. Ensemble learning[14] and chained multi-stream model[15] each working with different form of input also boost classification accuracy. In addition to labeling the activity class, Alberto monte et al.[16] proposed temporal activity detection model that can also find the temporal extent of activity in untrimmed video streams. Pipeline architecture for temporal activity detection splits video into 16-frame clips, each being fed to C3D for feature extraction which in turn acts as input to RNN for predicting class probabilities. Class probabilities are then averaged over all clips and maximum probability class is considered as predicted class. Multi-stream fusion based approaches attain better results thantraditional and deep networks with RGB/OFL alone as evidenced by literature study.

B.Quantitative Analysis

In this section we provide analysis of accuracy obtained by the aforementioned deep learning based solutions on their respective dataset and mention capabilities and limiting factor. It is clear that we cannot consider one method as universal solution based on degree of accuracy obtained because of the fact that datasets experimented with methods vary in terms of complexity or characteristics, volume of data available for training .For example it is clearly not reasonable to compare two approaches one working upon UCF101 with 101 action

 Copyright © 2019Authors

classes from five different activity categories and other on simple dataset like KTH which consist of just six activity classes. However we can compare methods working upon similar datasets easily. Considering multi-stream approaches for UCF dataset, Feichtenofer et al. [12] is working better than Simonyan and Zisserman [11],suggesting that fusion at convolution layerscan give better performance compared to only decision fusion. Approach ofIjjina et al.[17] on UCF50 achieved much higher degree of accuracy; however we cannot expect to give exactly same result on other dataset.Similarly methods of Dobhal et al.[10] and Liu et al.[13] on MSR-Action3D gave comparable results.

Performance of [11,12,18]clearly signify the importance of motion information offered by optical flow for activity recognition. Method of [10, 13] can not only give high accuracy but are also invariant to holes and viewpoint variations. Also its clear that multi-stream, fusion based approaches based on different cues offer better results than single stream approaches.

IV. Future research directions

Training of deep networks is resource consuming as it requires very high speed multi-core systems, therefore benefitting from pre-trained models is an avenue to explore, which will save time as well as money. Much of research is focused on attaining high accuracy, for real time response faster analysis is required. Therefore future research needs to be focused on designing efficient as well as fast video analysis systems. Performance of neural network based solutions for activity recognition is conditional on the amount of training data. Especially for videos, it needs enough data to train the network to learn decision function. Standard datasets limit training opportunities due to environmental restrictions like capturing particular view, limited participating subjects in the activity, hidden minute descriptions of actions. Carefully designed data augmentation techniques are thus required. Due to such reasons precision of activity analysis is high for simple actions only. Benefitting from easily available annotated image data needs consideration. Also datasets are built under constraints which real scenarios may not fulfill. Therefore methods which are invariant to holes, shadows, occlusions, view-point variations need attention. To be useful for application to real life scenarios future research should also target more refined and realistic activities.

V. Conclusion

Human activity recognition is the fundamental video analysis problem in computer vision community. Despite of great success achieved by local and global feature learning approaches, deep learning is receiving increasing attention to activity level analysis of video streams. In addition to various obstacles such as inter-class, intra-class diversities, dynamic

 Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

and cluttered backgrounds, activity class labeling also involves processing voluminous video data. Deep models are capable achieving great success at overcoming these obstacles. This is due to heirarchical feature learning capability without even need of domain knowledge to design feature extraction framework and labour intensive feature engineering as with handcrafted features. The deep network models have also dramatically improved the performance. This is well evidenced by the qualitative and quantitative analysis of deep methodologies used in recent literature for the analysis of activities provided in this paper.

Table-1: Description of various public datasets usedfor computer vision

Sr. No	Dataset	Content
1	UCF 50	Realistic videos taken from youtube
2	UCF101	13320 action videos collected from YouTube, gives the largest diversity in terms of actions
3	HMDB51	6849 movie clips with some social videos
4	CAD60	60 RGB-D video sequences of humans performing activities recorded using the Microsoft Kinect sensor.
5	Volleyball	6 videos from professionals in the Austrian Volley League
6	KTH-1	599 videos with different actions and subjects.
7	KTH-2	as KTH-1 but 2, 391 video sequences of duration 1sec
8	MSR Action3D	567 depth map sequences, actions performed by 10 subjects recorded using a depth sensor similar to the Kinect device
9	MSR DailyActivity3D	Daily activity dataset captured by a Kinect device.
10	UT KinectAction	Ten types of human actions were performed twice by 10 subjects. Three channels recorded: RGB, depth and skeleton joint locations
11	MSRC12	Two types of gestures(including whole body movement): Iconic gestures and Metaphoric gestures, captured using Kinect sensor
12	DogCentric	It consists of videos of dog activity recorded by camera attached to back of dog.
13	JPL-Interaction-E	Provide first-person videos of interaction-level activities recorded by camera attached to humanoid model.
14	Thumos 14	Temporally untrimmed videos collected from Youtube, include pre-extracted low-level features (Improved Dense Trajectory Features)
15.	(TRECVID)	TREC Video Retrieval Evaluationdata, which consist of surveillance video data recorded at London Gatwick Airport (10 days × 2 hours/day × 5 cameras)
16.	JPL-Interaction-E	Provide first-person videos of interaction-level activities recorded by camera attached to humanoid model.

Copyright © 2019Authors

REFERENCES

[1] T. B. Moeslund, A. Hilton, V. Krüger, "A survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding (CVIU), 104 ED-2, pp. 90–126, 2006.

[2] J. K. Aggarwal, M. S. Ryoo, "Human activity analysis: A review", ACM Computing Survey 43, 3, Article 16 , 43 pages, April 2011.

[3] R. Poppe, "A survey on vision-based human action recognition",Image and Vision Computing ,volume 28 ED- 6, pp. 976-990, June,2010.

[4] A. F. Bobick, J.W Davis,"The recognition of human movement using temporal templates",IEEE Trans. Pattern Analysis and Machine Intelligence 23 (3), 257–267, 2001

[5] I. laptev ,"On space-time interest points", Int. Journal of Computer Vision,vol. 64 (2),pp.107–123,2005.

[6] S. Ji, M. Yang,K. Yu,"3d convolutional neural networks for human action recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.35 (1), pp.221–231, Jan 2013.

[7] L. Mo, F. Li, A. Huang, "Human Physical Activity Recognition Based on Computer Vision with Deep Learning Model", Instrumentation and Measurement Technology Conference Proceedings (I2MTC),IEEE International,2016.

[8] S. U. Park, J. H. Park, M. A. Al-masni, M. A. Al-antari, Md. Z. Uddin, T. -S. Kim,"A Depth Camera-based Human Activity Recognition via Deep Learning Recurrent Neural Network for Health and Social Care Services", Procedia Computer Science vol. 100 ,pp. 78 – 84,2016.

[9] M. Baccouche, F. Mamalet, C.Wolf, C.Garcia, A. Baskurt, "Sequential Deep Learning for Human Action Recognition", A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 29–39, 2011.

[10] T. Dobhal,"Human activity recognition using binary motion image and deep learning", Procedia Computer Science vol. 58,pp.178 – 185,2015

[11] K.Simonyan, A. Zisserman, A,"Two-stream convolutional networks for action recognition in videos", In: Proc. Advances in Neural Information Processing Systems (NIPS). pp. 568–576, 2014.

 Copyright © 2019Authors

[12] C. Feichtenhofer, A. Pinz, A. Zisserman,"Convolutional two-stream network fusion for video action recognition", In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1933–1941, 2016.

[13] Z. Liu, C. Zhang, Y.Tian, "3D-based Deep Convolutional Neural Network for Action Recognition with Depth Sequences",Image and Vision Computing,2016.

[14] R. Granada, J. Monteiro, R. C. Barros, F. Meneguzziy, "A Deep Neural Architecture for Kitchen Activity recognition", Association for the advancement of Artificial Intelligence,2017.

[15] M. Zolfaghari, G. Oliveira, N. Sedaghat, T. Brox, "Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection", https://arxiv.org/abs/1704.00616,2017.

[16] A. Monte, A. Salvador,X. G. Nieto, "Temporal activity detection in untrimmed videos using recurrent neural network", Neural Information Processing Systems (NIPS), 2016.

[17] E. P. Ijjina and C. K. Mohan, "Human action recognition using genetic algorithms and convolutional neural networks", Pattern Recognition" , http://dx.doi.org/10.1016/j.patcog.2016.01.012, 2016.

[18] G. Varol, I. Laptev, C. Schmid, "Long-term Temporal Convolutions for Action Recognition", arXiv:1604.04494,2016.

 Copyright © 2019Authors