IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

A Survey on Neural Network based Approaches and Datasets in Human Action Recognition

C. Indhumathi[1*], V. Murugan[2]

[1] Department of CSE, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, India
[2] Department of Computer Science, MSU Constituent f Arts and Science College, Kadayanallur, India

[*]*Corresponding Author: indhuinfo2013@yahoo.com*

Available online at: www.ijcseonline.org

Accepted: 17/Jun/2018, Published: 30/Jun/2018

Abstract— Vision-based human action recognition has an increasing importance among the computer vision community with applications to visual surveillance, video retrieval, Video Indexing, Robotics and human-computer interaction. This paper presents a survey on human recognition using neural networks and the popular datasets used for it. A detailed survey of learning based approaches for human action representation is presented in this paper which is the core of action recognition. The Experimental Evaluation of various papers are analyzed efficiently with the various performance of recent methods using KTH and UCF sports action dataset are also analyzed.

Keywords: action recognition, convolution neural network, action representation

I. INTRODUCTION

Human action recognition system identifies the actions of human like walking, jumping etc. from a video. There are many kind of research in this area in the last two decades. Human action recognition plays a vital role in visual surveillance, video retrieval and human-computer interaction among others. The general architecture of Human action recognition system consists of image capture, segmentation, tracking, identification, and classification.

Human perception consists of organization, identification, and interpretation of feelings that humans acquire from the surrounding environment [1]. The identification of human actions is challenging topic in many modern science subjects, including psychology, cognitive science, neuroscience and biology. It can be identified by a series of observations of body movements. Such observations are then used to predict the category of movement such as running, waving the hand, and jumping. The daily activities of human are acquired from a large number of visual perceptions. While one of the most important tasks of computer vision is to mimic the way how humans perceive the surrounding environment and make predictions accordingly. Some existing computer vision technologies such as the Convolution Neural Network (CNN) [2] have gained a remarkable reputation for image-based categorization tasks in terms of performance. Similar to image classification, the ability of correctly recognizing human actions is a basic component of human perception system. For robust action recognition system, computer vision researchers have paid significant efforts in the past decades. But due to the challenging issues, such as high environment complexity and high intra-class action variations, it has not been achieved today. Traditionally, the action representations mainly rely on statistics of gradients, combinations of global filters, depth images, skeletons, etc. which can be together referred to as handcrafted features. The learning-based approaches dominate the handcrafted features.

This paper gives a survey of action representation, which is the core of action recognition. The term “action” is always confused with similar terms “gesture”, “interaction” and “activity”. To clarify the action is defined as an intentional, purposive, conscious and subjectively meaningful activity, where the above stated four terms can be associated with an ascending order of complexity levels: “gesture”, “action”, “interaction” and “activity”[3]. In the literature, there are several existing surveys on the topic of vision-based action recognition.

The handcrafted action representation includes four subcategories. They are spatial-temporal volume-based approaches, depth image-based approaches, trajectory-based approaches and global approaches. In recent years, the learning based approaches increase the performance of image classification on challenging tasks, such as ImageNet [4]. This inspired researchers to follow learning based methodology to extract robust representations from action videos. Hence, this paper focuses on action representations and covers learning-based approaches.

In this survey, a comprehensive investigation of existing learning-based action representation approaches is provided. It includes both non-neural network-based approaches and neural network-based approaches, where the

focuses allocated to the latter. In non-neural network-based approaches, genetic programming-based action representations and dictionary learning-based action representations are included. In neural network-based approaches, a finer taxonomy is employed to review these approaches from the aspects of static frames-based approaches, frame transformation-based approaches, handcrafted features-based approaches, 3D CNN-based approaches and hybrid models, where certain overlapping may exist between these taxonomies.

The remaining of the paper is given as follows: Section 2 gives the survey of neural network based approaches. Section 3 common datasets used for action recognition followed by analysis of the results of various researches introduced in the recent years. Finally, a conclusion is given in Section 4.

II. NEURAL NETWORK-BASED APPROACHES

The most popular machine learning algorithm is based on neural networks. It aims to model high-level abstraction of data using hierarchical structures. In contrast to handcrafted approach, deep learning approach performs more intellectual learning and contains hierarchical feature extraction layers. As deep learning approach succeeded in image classification, some recent works have been using similar learning-based representations for action recognition, where these works are summarized according to the following directions: 1) learning from video frames, 2) learning from frame transformations, 3) learning from hand-crafted features, 4) three-dimensional convolutional networks and 5) hybrid models.

2.1. LEARNING FROM STATIC FRAMES

There are several existing deep learning methods for image feature extraction. These methods can be applied to each frame within an action video to extract action features. Ning *et al* [5] decompose the video-based analysis of embryos development problem into frame-level 2D images and apply two-dimensional CNN to various stages (from fertilization to four-cell stage) of the development process.

2.2. LEARNING FROM FRAME TRANSFORMATIONS

A Restricted Boltzmann Machine (RBM) [6] is a more generative artificial neural network that can provide a deep architecture by successively composing several RBM. Using RBM, Taylor *et al.* [7] developed an unsupervised approach for learning spatio-temporal features using a Gated RBM (GRBM) to extract motion features from neighboring frames. The GRBM architecture is first introduced by Memisevic and Hinton [8] to describe the probabilistic model of learning distributed representations of image transformations. It tried to predict the next frame in the video based on the current frame. This is called as "mapping" which extract features based on frame transformations. This model is extended to image patches at identical spatial locations in sequential frames. Hence, GRBM can capture the transformations of successive frames. In order to avoid the limitations of training GRBMson isolated patches, Taylor *et al.* extended the GRBM model to a convolutional GRBM (convGRBM) model by incorporating the convolutional architecture [8, 9]. It operates in a multi-stage architecture where weights at multiple locations within an image are shared by a convGRBM.. At the lowest layer, convGRBM extracts motion features from nearer frames. At the intermediate layer, spatio-temporal cues are obtained by 3D spatio-temporal filters. Then normalization is done which is followed by average spatial pooling and max-pooling in the temporal dimension. Action representations can be obtained by the fully-connected layers and action label is obtained at the topmost layer. The lower layer of conGRBMis trained separately without using upper layers. Back propagations performed on upper layers.

2.3. LEARNING FROM HANDCRAFTED FEATURES

The next choice of computing learning-based action representations is to learn the top of handcrafted features. Most of the early attempts aim to address action recognition using learning-based representations lie in this category. This category is different from the previous category; "handcrafted representations-based action recognition". The former discusses how the learning network can be established based on existing low-level features while the latter discusses how low-level features can be extracted from raw pixel-level video data. Kim *et al.* [10] developed a modified CNN-based action feature extraction and classification framework. At the lower level, action information is captured by handcrafted features. When an action performed by an agent is presented in a 3Dvideo, a sequence of the agent's outer boundary, which can be considered as a 2D contour in the spatial plane, generates a spatio-temporal volume, where the outer boundary information are extracted using three-dimensional Gabor filters [11]. Thus, in each spatio-temporal volume, actions are presented in a view-invariant form. In order to reduce the post-normalization location variances, the extended 3DCNNis applied to each spatio-temporal volume, and subsequent action features are extracted from a set of hierarchical layers based on the agent's outer boundary features. A 3DCNN consist two convolution layers and two sub-sampling layers, where each convolution layer or sub sampling layer consists of two sub-layers. The extracted features are then fed into a discriminative classification model.

Jhuang et al [12] presented a system that utilizes a feed-forward method of spatio-temporal feature detectors to measure motion-direction sensitive units, which lead to

position-invariant spatio-temporal feature detectors. Using the scale and position invariant features [13], a vector of scale and position invariant features is obtained by computing a global maximum for each feature map at the top of the hierarchy. The disadvantage of this method is that it requires handcrafted spatio-temporal feature detectors. Wu and Shao [14] developed a Hierarchical Parametric Networks (HPN) based on skeleton features. By replacing the RBM in [15] with a multi-layer network, the HPN approach can serve as a better model for estimating emission probability of hidden Markov models [16].

2.4. THREE-DIMENSIONAL CONVOLUTIONAL NETWORKS

The first attempt to develop a 3D CNN and performs 3D convolution along both spatial and temporal dimensions at the pixel level is introduced in [17]. By applying multiple distinct convolutional operations at identical input locations, features are extracted subsequently from multiple information channels. Action representations obtained by such a 3D CNN approach contain a variety of information. In 2D CNN [8], convolution is performed in the spatial domain, where features are extracted from neighbouring units that share the same feature map in the previous layer. Whereas in 3D CNN, convolution is performed in both spatial and temporal dimensions using 3D cubes, which are generated by stacking multiple contiguous frames. The 3D CNN consists of 7 layers including the input layer, which is hardwired to three convolution layers and two sub sampling layers in with an alternating order. The last layer consists of 128 feature maps and is fully connected to all feature maps in the previous layer. The main idea of the 3D CNN architecture comes from the feed-forward nature, which enables efficient feature extraction in the recognition phase.

Le et al. [18] introduced a Hierarchical Invariant Spatio-Temporal (HIST) action features. The feature learning is done based on the Independent Subspace Analysis (ISA). The ISA training process is less efficient when handling large-scale video data. Hence Principle Component Analysis (PCA) is used for feature learning. The HIS framework consists of several ISAs, where each ISA is initially trained on small input patches and the outputs are propagated to the next-layer ISA with reduced dimensions by PCA. HIST operates in an unsupervised manner over action videos, as it inherits ISA.

Baccouche *et al.* [19] presented Sequential Deep Learning (SDL) method which extended CNN to a three-dimensional scenario as in [11]. Both the methods are different in a way that the CNN architecture directly operates on raw video pixels in the former scenario while the CNN framework is established on the top of handcrafted lower level features in the latter. The construction of the CNN architecture in[19] is also different from [11] by following the order of two alternating convolutional layers, a rectification layer, a sub-sampling layer, a second alternating convolutional layer, a second sub-sampling layer, a third convolutional layer and two neuron layers. The training process uses a standard back propagation with momentum algorithm. Once action features are extracted using 3D CNN, a sequential action labelling scheme is utilized. Instead of utilizing small sized spatio-temporal volumes to generate three-dimensional regions for CNN learning, the temporal features are captured by adapting CNN to sequential data. Here, the Recurrent Neural Networks (RNN) [20] is used with one hidden layer of Long Short-Term Memory (LSTM). There are various RNN-based approaches for human action recognition, which includes the hierarchical recurrent neural network for skeleton-based representation and the differential recurrent neural networks [21]. When comparing to the LSTM model, regular RNN is incapable of capturing long-term dependencies between frames, so that these models perform well. Karpathy*et al.* [22] conducted comprehensive study of3DCNN-based action recognition and evaluated all these approaches. The main focus of the study is to investigate the best approach for incorporating the motion information for recognizing actions, and how much improvement is needed to improve CNN on static video frames. From their study, it is proved that normally most videos present in a highly inconsistent nature, and thus cannot be easily processed with fixed-sized architectures. Consequently, the network is designed to learn spatio-temporal features by connecting several contiguous frames in time and plugging in the network. Based on their study, single frame architecture is equivalent to applying CNNs to 2D images.

In their study, they also presented three CNN-based learning strategies: Early Fusion, Late Fusion and Slow Fusionto better analyze the benefits that come from the motion information. Late Fusion uses two separate CNNs on two apart frames (e.g., from a distance of15 frames) and connects both networks in the first fully connected layer, so that action motion characteristics can be captured at a global level. Early Fusion modifies the 2D single-frame based convolution to include the temporal dimension, and feed these 3D cubes to the first convolutional layer. As a compromise between Early Fusion and Late Fusion, Slow Fusion (SFCNN) progressively connects adjacent frames from convolutional layers in both spatial and temporal dimensions. All the three types of fusion strategies can be generalized to a common learning-based action representation approaches. Due to the high computational requirements of CNN-based approaches, especially when extending to the third dimension, efforts have been paid on how to speed up the training process.

III. DATASETS

In this section, two common datasets used for action recognition are discussed. The results of the recent methods which use these two datasets are also analyzed in this section.

3.1. THE KTH DATASET

This database covers six actions - walking, jogging, running, boxing, hand waving and hand clapping. All are performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. It contains a total of 2391 sequences. All sequences are taken with a static camera with 25fps frame rate, down sampled to the spatial resolution of 160x120 pixels. This dataset does not provide background models and extracted silhouettes.

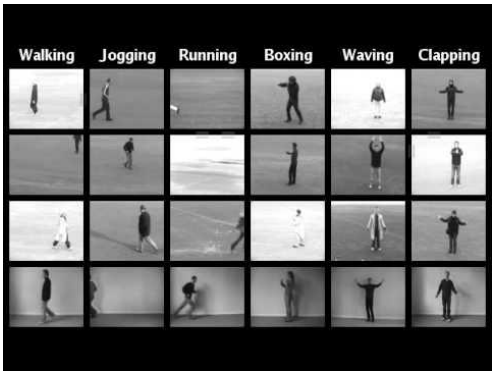

Figure 1 KTH action dataset

3.2. UCF SPORTS ACTION DATASET

The UCF sports action dataset contains 150 sequences of sports motions (diving, golf swinging, kicking, weightlifting, horseback riding, and running, skating, swinging a baseball bat and walking). Bounding boxes of the human figure are provided with the dataset. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and background.

Figure 2 UCF action dataset

Table 1.1: Results of recent methods using KTH and UCF sports action dataset

Dataset	Year	Author	Method	Accuracy (%)
KTH	2017	Allah BuxSargano et al. [23]	SVM-KNN	98.15
		Allah BuxSargano et al. [23]	SVM	94.83
		Allah BuxSargano et al. [23]	KNN	89.91
	2016	Charalampous and Gasteratos [24]	Online Deep learning	91.99
		Ahad et al. [25]	Action and history and histogram	86.7
		Ding and Qu [26]	Interest Point Detector	95.58
	2015	Veeriah et al. [27]	Differential RNN	93.96
		Shi et al. [28]	DTD, DNN	95.6
UCF sports Action	2017	Allah BuxSargano et al. [23]	SVM-KNN	91.47
		Allah BuxSargano et al. [23]	SVM	89.60
		Allah BuxSargano et al. [23]	KNN	82.75
	2016	Tian et al. [29]	Local Consistent Group Sparse Coding	90.0
		Charalampous and Gasteratos [24]	Online Deep Learning	88.55
		Ballas et al. [30]	GRU-RCN	80.7
	2015	Atmosukarto et al. [31]	Discriminative Structured Trajectory groups	82.6
		Sun et al. [32]	Factorized spatio-temporal CNN (SFTCN)	88.1
		Wang et al. [33]	TDD, CNN	95.1

From Table 1, it is observed that SVM- KNN method achieves better accuracy rate than other methods in KTH dataset. Also, in UCF sports action dataset, CNN method outperforms other methods by more than 3%. In KTH dataset, deep learning method does not achieve better performance than SVM.

IV.CONCLUSION

Action recognition plays a vital role in video surveillance, video retrieval and human-computer interaction. This paper gives a brief survey on action representation in deep learning methods such as convolution neural network. The common datasets used for action recognition such as KTH and UCF sports action dataset are also discussed. The experimental results of the recent methods in the aforementioned datasets are also analyzed. From the survey, it is clear that SVM with KNN (K-Nearest Neighbors) method achieves a better accuracy of 98.15% when compared to other methods in KTH dataset. In UCF sports action dataset, the deep learning method (CNN) and TDD achieves 95.1% when compared to other methods.

ACKNOWLEDGMENT

At the very outset, I express my thanks to GOD ALMIGHTY and my parents who blessed me with a healthy constitution and bestowed up onus the required skills to complete this work. The authors would like to thank the anonymous reviewers whose insightful comments have helped to improve the presentation of this paper significantly.

REFERENCES:

[1] D. G. Daniel Schacter, D. Wegner, Psychology, New York: Worth, 2011.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11), (1998) 2278–2324.

[3] J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, ACM Computing Surveys (CSUR) 43 (3) (2011) 16.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2009.

[5] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, P. E. Barbano, Toward automatic phenotyping of developing embryos from videos, IEEE Transactions on Image Processing 14 (9) (2005) 1360–1371.

[6] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[7] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: European Conference on Computer Vision, Springer, 2010.

[8] R. Memisevic, G. Hinton, Unsupervised learning of image transformations, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2007.

[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11)(1998) 2278–2324.

[10] H.-J. Kim, J. S. Lee, H.-S. Yang, Human action recognition using a modified convolutional neural network, in: Advances in neural information processing systems, Springer, 2007.

[11] J. P. Jones, L. A. Palmer, An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, Journal of neurophysiology 58 (6) (1987) 1233–1258.

[12] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: IEEE International Conference on Computer Vision,, 2007.

[13] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological cybernetics 36 (4) (1980) 193–202.

[14] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[15] G. W. Taylor, G. E. Hinton, S. T. Roweis, Modeling human motion using binary latent variables, in: Advances in neural information processing systems, 2006.

[16] L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, The annals of mathematical statistics.

[17] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 221–231.

[18] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[19] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: Human Behavior Understanding, Springer, 29–39, 2011.

[20] A. Graves, M. Liwicki, S. Fern´andez, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained hand-writing recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 855–868.

[21] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1110–1118, 2015.

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei- Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[23] Allah BuxSargano, Xiaofeng Wang, PlamenAngelov, and ZulfiqarHabib, "Human Action Recognition using Transfer Learning with Deep Representations, IEEE, 2017, pp. 463-469.

[24] Charalampous, K. and A. Gasteratos, On-line deep learning method for action recognition. Pattern Analysis and Applications, 2016.**19**(2): p. 337-354.

[25] RahmanAhad, M.A., M.N. Islam, and I. Jahan, Action recognition based on binary patterns of action-history and histogram of oriented gradient. Journal on Multimodal User Interfaces, 2016.

[26] Ding, S. and S. Qu. An improved interest point detector for human action recognition.in Control and Decision Conference (CCDC), 2016 Chinese. 2016. IEEE.
[27] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4041–4049.
[28] Y. Shi, W. Zeng, T. Huang, and Y. Wang, "Learning deep trajectory descriptor for action recognition in videos using deep neural networks," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015, pp. 1–6.
[29] Tian, Y., Ruan, Q., An, G., Fu, Y., Action Recognition Using Local Consistent Group Sparse Coding with Spatio-Temporal Structure. in Proceedings of the 2016 ACM on Multimedia Conference. 2016. ACM.
[30] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *International Conference of Learning Representations*, 2016.
[31] Atmosukarto, I., N. Ahuja, and B. Ghanem. Action recognition using discriminative structured trajectory groups. in 2015 IEEE Winter Conference on Applications of Computer Vision. 2015. IEEE.
[32] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognitionusing factorized spatio-temporal convolutional networks," in *Proceedingsof the IEEE International Conference on Computer Vision*, 2015,pp. 4597–4605.
[33] L. Wang, Y. Qiao, and 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.

Authors Profile

C.Indhumathi is currently working as an Assistant Professor in VPMM Arts and Science Collge , Srivilliputhur. And also she is pursuing her Ph.D in Manonmaniam Sundaranar University. She completed her M.phil degree in Madurai Kamaraj University and M.Sc degree in Manonmaniam Sundaranar University. Her research interests include Image Processing and Neural Network.

Dr.V.Murugan is working as an Assistant Professor in the Department of Computer Science, Manonmaniam Sundaranar University Constituent Arts & Science College, Kadayanallur. He has completed his Ph. D in Computer Science & Engineering from the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli in the year 2016. He has completed his M.E Degree from the Department of computer science and Engineering, Manonmaniam Sundaranar University, Tirunelveli in the year 2012. He has completed his MCA Degree from the Department of computer science and Engineering Manonmaniam Sundaranar University, Tirunelveli in the year 2010. His research interests include Image Processing and Data Mining.

