A Survey on Deep Neural Networks for Human Action Recognition based on Skeleton Information

Hongyu Wang(✉)

Department of Applied Mathematics,
Northwestern Polytechnical University, Xi'an, China
whyer@mail.nwpu.edu.cn

Abstract. Human action recognition has been a significant topic in the field of computer vision. As deep learning develops, the application of deep neural network in related research is gradually more prevalent. This paper provides a survey of deep neural networks for human action recognition based on skeleton information. The detailed description about each method is explained and several related main datasets are briefly introduced in this paper, all papers are published ranging from 2013 to 2015, which provides an overview of the progress in this area.

Keywords: Action recognition · Deep learning · Skeleton information

1 Introduction

Human action recognition has been drawing more and more attention in computer vision, which is partially due to the development of vision sensors. The information from vision sensors is so abundant and comprehensive that it is qualified to be analyzed for human action recognition [1]. Motion capture data (Mocap) is considered as a main kind of source data. As the low-cost and high-mobility sensors (such as Kinect) appear, more and more researchers have shifted their attention to human action recognition based on the human skeleton information [2] and the related researches become significantly worthy.

Nowadays, the study on Human Action Recognition in general involved four aspects: gestures, actions, interactions and group activities [3]. Gesture refers to a static state which is about a certain movement of body such as standing or bowing. Action generally consists of several sequential gestures such as running, waving. Interaction indicates a correlative action involving two persons or a person and an object such as brawling. Group activity always involves many persons such as meeting. The overview process of handling these tasks consists of feature extraction, action representation learning and classification [4].

In the recent years, many researches has published to find many ways to solve the problems on human action recognition and many methods are proposed on skeleton data [5]. Moreover, deep neural networks are especially and widely employed. There are several frequent and classical deep neural networks such as DBN (Deep Belief

F. Xhafa et al. (eds.), *Recent Developments in Intelligent Systems and Interactive Applications*, Advances in Intelligent Systems and Computing 541, DOI 10.1007/978-3-319-49568-2_47

Network) [6], RNN (Recurrent neural Network) [7], Denoising Autoencoder [8]. As known, DBN could be viewed as a probabilistic generative model. The component of DBN is RBM (Restricted Boltzmann Machines) which is a stochastic neural network to learn probability distribution from data. Unlike feedforward neural network, RNN works on extract temporal feature in data effectively because the output of each neuron in RNN is not only as input for next neuron but also acts on itself. Denoising Autoencoder is intended to simulate animal's vision and is designed in order to cope with unlabeled data. It performs feature extraction in form of unsupervised learning.

This paper presents the state-of-the-art methods about human action recognition skeleton information in recent years. All involved papers are published from the year 2013 to 2015. Most of them are from CVPR, ECCV, etc. Through the review of these papers, a general framework of the methods is shown in Fig. 1. A detailed discussion on relevant methods is provided in following sections.

The rest of this paper are organized as follows. In Sect. 2, the datasets employed in experiments are reviewed briefly. Sections 3 and 4 describe the methods on single networks and hybrid networks in detail respectivly. The conclusion and future works are presented in Sect. 5.

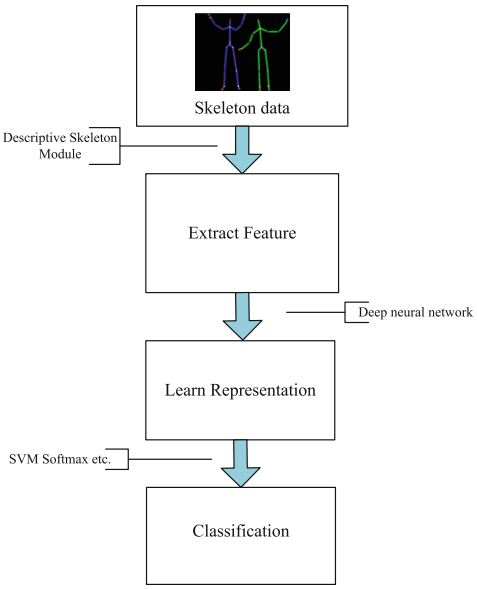

Fig. 1. A general framework of the methods

2 Datasets

In this section, it provides the description of datasets that are regarded as a benchmark in the current research (as shown in Table 1). Some datasets are classic and acknowledged such as MHAD. In addition, there are also some new datasets and several datasets for specific problems.

Table 1. A comparison of datasets

	Action types	Resolution	# of sequences	Format
HDM05	70	N/A	1500	C3D
MSR Action3D	20	320 × 240	567	ASCII
MHAD	11	640 × 480	660	PGM/ASCII
CMU Mocap	109	320 × 240	2605	ASF

HDM05 Database. HDM05 [9] contains more than three hours of recorded and well-documented motion capture data in the C3D as well as in the ASF/AMC data format. Specifically, HDM05 is comprised of more than 70 types of action executed by 10 to 50 actors (in Fig. 2). In this database, most of the sequences have been performed several times by all five actors according to the guidelines fixed in a script. The script is divided into five parts and each part is subdivided into several scenes.

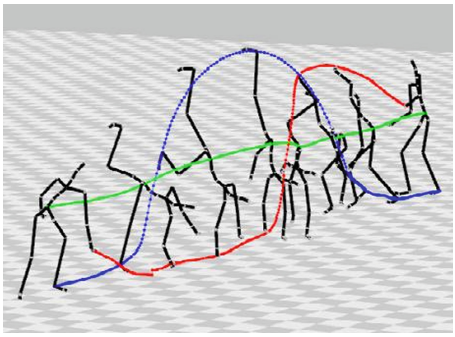

Fig. 2. Example from HDM05 dataset

Berkeley Multimodal Human Action Database (MHAD). This database [10] includes 11 classes of actions performed by 12 youths. And all the persons performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 min of recording time. The set of actions comprises of the following: (1) actions with movement in both upper and lower extremities, e.g., jumping in place, jumping jacks, throwing, etc., (2) actions with high dynamics in upper extremities,

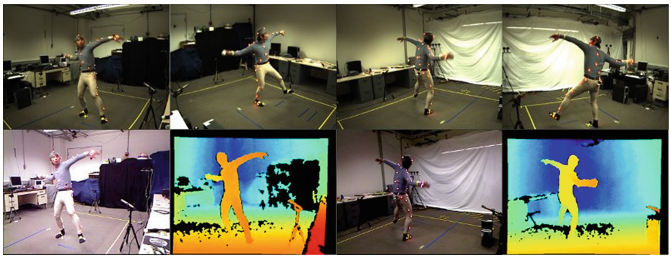

Fig. 3. One class of action in MHAD dataset

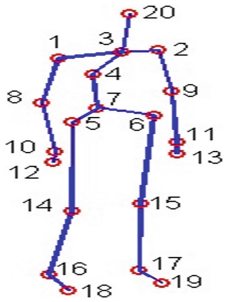

Fig. 4. Skeleton model in MSR Action3D

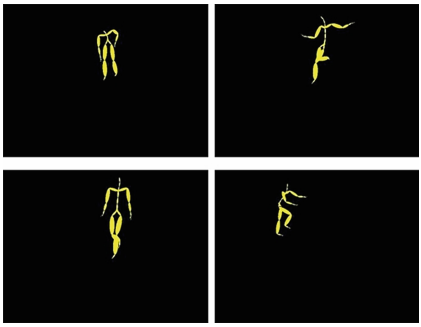

Fig. 5. Four classes of actions in CMU Mocap

e.g., waving hands, clapping hands, etc. and (3) actions with high dynamics in lower extremities, e.g., sit down, stand up. A sample is shown in Fig. 3.

MSR-Action3D Dataset. MSR-Action3D dataset [11] contains twenty actions: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. As shown in Fig. 4. A skeleton has 20 joint positions. The resolution is 320 $\times$ 240 in this dataset.

CMU Motion Capture Database. The CMU Motion Capture Dataset [12] is collected by Carnegie Mellon University. It is a free database for all use and contains various actions ranging from locomotion to sports and pantomime. There are 109 types of action and more than 100 subjects in it. The data is stored as ASF format and the resolution is 320 $\times$ 240. Some samples are showed in Fig. 5.

3 Methods Based on Single Network

The structure, algorithm, model of all reference articles will be presented in the lines below.

The method that is proposed by D. Wu and L. Shao in this paper is called Deep Dynamic Neural Networks (DDNN) [13]. It is a semi-supervised hierarchical dynamic framework taking both skeleton and depth images as input but this survey only focuses on its skeleton module. This paper only considers 11 upper body joints based on the assumption that recognition tasks are just relevant to upper body. According to 3D coordinates of joints, posture features and motion features are defined. These two features are concentrated and are inputted into a deep belief networks to extract high level features. Taking into consideration that the skeletal features are continuous, the Gaussian RBM is used to model the energy term in the above DBN [14]. Then the emission probabilities of the hidden states are outputted after the high level features are extracted through the DBN. Unlike traditional hidden markov model (HMM), the method productively adds an ergodic states to the HMM in order to perform both action segmentation and recognition. Afterwards the representations can be learned from both skeletal data and depth images. The framework in this paper is a data-driven approach, which brings about more discriminative information. The HMM is extended by introducing an ergodic states so that it can be capable to segment and to recognize action sequences simultaneously.

J. Wang et al. has proposed a novel model called actionlet ensemble [15] for action recognition using depth images and skeletal data. There are two types of skeletal features in this paper. The 3D joint positions are employed to characterize the motion of the body and the local occupancy pattern is to describe the interaction between the human subject and the objects. Inspired by the Spatial Pyramid approach in [16], a descriptive representation called Fourier temporal pyramid is designed which is used to represent the temporal patterns of actions. Based on the above features, the actionlet ensemble approach is proposed to deal with the errors of the skeleton tracking and characterize the intraclass variations. An actionlet is defined as a conjunctive structure of features for a subset of the joints. For increasing the number of the discriminative actionlets, a data mining technique is designed to select them. Afterwards an SVM is trained on each selected action let as an action let classifier and an actionlet ensemble is established by combining these classifiers linearly. With a joint feature map defined on data and labels, an actionlet ensemble could be learned by applying multiple kernel learning approach. This method is insensitive to noise as well as translation and is capable of handle view changes. Moreover, human actions with subtle differences can be discriminated.

For modeling the long-term contextual information of temporal dynamics of human skeleton, Y. Du et al. has established an end-to-end hierarchical recurrent neural network (RNN) [17] whose basic module is a bidirectional recurrent neural network (BRNN) as subnet. Considering human physical structure, the human skeleton is divided into five parts according to arm, trunk and leg, which is used for input. As an innovative improvement, this paper replaces the nonlinear units with LSTMs [18] in order to vanishing gradient and error blowing up problems. As for architecture, the framework in this paper is composed of nine layers. Besides input layer, there are three sets of BRNNs and fusion layer arranged alternately followed by a BRNN with LSTM, a fully-connected layer and a softmax layer. With the skeletal data inputted, the features could be extracted and fused through the hierarchical BRNN architecture. The fully-connected layer and a softmax layer perform classification using the learned

representation from the former layers. This is the first paper to apply hierarchical RNN for skeletal based action recognition. The method can capture the spatiotemporal features of action sequences without complex preprocessing.

4 Methods Based on Hybrid Networks

The model proposed by Z. Yu and M. Lee is a hybrid of multiple timescale recurrent neural network (MTRNN) [19] and deep learning neural network (DN) for recognizing walking, running and swinging. MTRNN is adopted for performing dynamic action recognition while DN is capable of static posture recognition. The two major components of MTRNN are slow context layer and fast context layer, which are modeled by a special type of RNN. Utilizing self-organizing map, MTRNN can accept vision signals to predict without supervision. The prediction of MTRNN rely on a prior knowledge of human action so the initial state needs to be modified when the action change. In order to handle the issue, a DN is used to choose the proper initial state, The DN receive the visual information of the current time step from Kinect and the same dimensional data from MTRNN in the adjacent time step. This process has the capability to capture the dynamic features and correct initial states of the current action sequences. This hybrid method gives a compensation for MTRNN. The combination of MTRNN and DN can extract static and dynamic features simultaneously.

K. Cho and X. Chen proposed a novel method [20] to archive human action recognition from skeleton data by introducing deep neural networks. The method in this paper is based on joint distribution model of feature in each frame, which consists of the relative positions of joints (PO), temporal difference (TD) and the normalized trajectory of the motion (NT). For utilizing more information in layers, this paper proposed a hybrid multi-layer perceptron containing an MLP [21] and a deep autoencoder [22]. The MLP and the deep autoencoder are trained with the same set of parameters, which is used to classify and reconstruct respectively. By introducing a hyperparameter λ ranging from 0 to 1, the supervised learning and the unsupervised learning, namely MLP and deep autoencoder are combined. Given a frame, the hybrid MLP can model the posterior probability distribution of classes. With the assumption that the class of each frame only depends on the features of the frame, the classifier is just trained for frame-level classification. This method combines supervised learning and unsupervised learning, which perform the reconstitution of features and classification simultaneously. The deep autoencoder visualizing the features, more distinctive information can be extracted and it makes it possible to study what deep neural networks learned.

In order to understand human intension by analyzing actions effectively, Z. Yu and M. Lee proposed a novel method [23] in 2015 which combine multiple timescale recurrent neural networks (MTRNN) and stacked denoising auto-encoder (SDA). In this paper, supervised MTRNN, an extension of MTRNN has been applied. With the context layers modeled by CTRNN introduced, MTRNN is capable for dynamic action classification. In addition, SDA aims to predict human intention by analyzing the distance between the human's hand and the objects. For catching the scale-invariant features of the object, speeded up robust features (SURF) [1] is employed to find proper

matches between image and the object and the output of SURF is used as the input of SDA. When supervised MTRNN and SDA work cooperatively, SDA needs to be trained before supervised MTRNN because the output of the code layer in SDA is fed into the slow context layer in supervised MTRNN. By this way, skeletal data is considered as compensation for dynamic characteristics. The method takes both the action signals and the information of the objects into account. With the unsupervised learning and the supervised learning complementary to each other, it is able to recognize the human intention more accurately.

5 Conclusion

As the application of human action recognition becomes more and more widespread, it has not only been an active area but also a challenging task in computer vision. In recent years, a trend of the research on human action recognition is popular, which address the issue by establishing deep neural networks as deep learning neural networks can learn hierarchical nonlinear function relation in order to model the vision system.

Most of conventional methods construct handcrafted features for recognition relying on domain knowledge, which is time-consuming and inefficient. As one of the most significant type of data for recognizing human actions, skeleton data shows the effectiveness. To tackle with the deficiencies of the handcrafted feature, the methods with deep learning architecture are driven by data, which are capable to extract features from original data automatically.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
2. Wei, Z.Q., Wu, J.A., Wang, X.: Research on applied technology in human action recognition based on skeleton information. Adv. Mater. Res. **859**, 498–502 (2013)
3. Saad, A., Mubarak, S.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **32**(2), 288–303 (2010)
4. Tanaya, G., Rabab Kreidieh, W.: Learning sparse representations for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**(8), 1576–1588 (2012)
5. Presti, L.L., Cascia, M.L.: 3D skeleton-based human action classification: a survey. Pattern Recogn. **53**, 130–147 (2015)
6. Salama, M.A., Ella Hassanien, A., Fahmy, A.A.: Deep belief network for clustering and classification of a continuous data. In: The IEEE International Symposium on Signal Processing and Information Technology, pp. 473–477. IEEE Computer Society (2010)
7. Gers, F.A., Schraudolph, N.N., Schmidhuber, J., et al.: Learning precise timing with LSTM recurrent networks. J. Mach. Learn. Res. **3**(1), 115–143 (2003)
8. Memisevic, R.: Gradient-based learning of higher-order image features. In: IEEE International Conference on Computer Vision, pp. 1591–1598. IEEE (2011)
9. Müller, M., Röder, T., Clausen, M., et al.: Documentation mocap database HDM05. Computer Graphics Technical Reports (2007)

10. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: a comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, pp. 53–60 (2013)
11. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points, pp. 9–14 (2010)
12. Fernando, D.L.T., Hodgins, J., Bargteil, A., et al.: Guide to the Carnegie Mellon University Multimodal Activity (CMUMMAC) Database. Carnegie Mellon University (2009)
13. Wu, D., Shao, L.: Deep dynamic neural networks for gesture segmentation and recognition. In: Agapito, L., Bronstein, Michael M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 552–571. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16178-5_39
14. Wu, D., Shao, L.: Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724–731. IEEE (2014)
15. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 914–927 (2014)
16. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. CVPR **2**, 2169–2178 (2006)
17. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Computer Vision and Pattern Recognition. IEEE (2015)
18. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 29–39. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25446-8_4
19. Yu, Z., Lee, M.: Continuous motion recognition using multiple time constant recurrent neural network with a deep network model. In: Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., Yao, X. (eds.) IDEAL 2013. LNCS, vol. 8206, pp. 118–125. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41278-3_15
20. Cho, K., Chen, X.: Classifying and visualizing motion capture sequences using deep neural networks. In: International Conference on Computer Vision Theory and Applications, pp. 122–130 (2013)
21. Haykin, S.S.: Neural Networks and Learning Machines. China Machine Press, China (2009)
22. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2015)
23. Yu, Z., Kim, S., Mallipeddi, R., et al.: Human intention understanding based on object affordance and action classification. In: International Joint Conference on Neural Networks. IEEE (2015)