State-of-the-art video recognition of human action under natural scenes

Yang Yi, Qian Mo*, Yihua Chen
School of Information and Intelligence Engineering, Guangzhou Xinhua University, Guangzhou, Guangdong, China

ABSTRACT

Human action video recognition focuses on discussing the theories and new algorithms to make computers understand the meanings of variety videos, such as those from TV, sports or family videos, and recently it plays an important role in advanced human-machine interface, security monitor, video retrieval and so on. The step-by-step process of traditional computational algorithm for human action recognition from videos are described in this paper, and some important deep learning models are also addressed by which some advantages and disadvantages of each approach in different tasks are analysed. Finally, some current trends and prospects of video recognition models are concluded.

Keywords: Human action recognition from video, feature retrieval, cluster and classification, computational framework, deep learning

1. INTRODUCTION

Human action recognition from video (HARfV) is one of the most popular tasks of computer vision, which refers to extracting features from video clips and recognizing the behaviours that the video shows, and this moment it plays a significant role in many aspects such as human action detection, pedestrian detection, AI-assisted healthcare, sports live streaming and so on[1-3].

The computational framework of HARfV often is composed of three steps: feature extraction, feature encoding, and the computing of cluster and classification. The approaches of feature extraction are usually classified into two different aspects: global feature-based extraction and local feature-based extraction. Global features refer to those overall features such as the shape of different objects, while local features refer to local points of interests, such as local spatio-temporal areas where their grayscale information changes drastically. The feature encoding step normalize those extracted feature vectors so as to be used in further computing. Thus, the HARfV problem could be transformed into a cluster and classification problem in which features play a crucial role in final recognition, so the quality of features might determine the quality of HARfV results.

Traditional methods for HARfV focuses on obtaining coding features of the appearance of objects and extracting the features of motion information by using hand-made features to describe the shape of the main object to be recognized. Commonly used methods include Histograms of Oriented Gradients (HOG)[4] and Histogram of Optical Flow (HOF)[5]. The Motion Boundary Histogram (MBH)[6], which is the trajectories of dense points are obtained through optical flow-based tracking in a dense grid. A global video-level descriptor is generated for the acquired features through the method of Bag of Words[7] or Fisher vector-based coding[8]. With the feature extraction task completed, the pattern classification methods such as Support Vector Machine (SVM), Hidden Markov Model (HMM) and Dynamic Bayesian Network (DBN) are usually used to classify and recognize the videos.

In recent years, helped by the rapidly boosted computing power, deep learning-based models could improve the overall accuracy greatly compared to most of traditional methods, though there are also some problems and challenges remaining.

The following are some commonly used datasets or benchmarks in HARfV listed in Table 1.

* moqian@xhsysu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125063A
 · 0277-786X · doi: 10.1117/12.2662221

Table 1. Benchmark datasets of HARfV.

Dataset	Introduction	Categories	Samples
ImageNet	Static images for pretraining	21,000+	14 million
Sports-1M	Sports-related video clips	487	1113158
Sth-Sth-v2	Videos of human actions on objects	174	220847
UCF101	Videos of real-world actions	101	13320
Kinetics-400	Videos of real-world actions	400	~300000
Kinetics-600	Videos of real-world actions	600	~500000
Kinetics-700	Videos of real-world actions	700	~650000
Kinetics-700-2020	A refined version of K-700	700	~650000
AVA actions dataset	Video clips from movies, for action recognition	80	430
AVA-Kinetics	AVA protocol for K-700 videos	80	~230000
HMDB51	Collected from public videos	51	6849
Charades	Indoor activities videos	157	9848
Charades-Ego	Newer Charades dataset with both first and third-person videos	157	7860
EPIC-Kitchen-55	55 hours video taken in Kitchen	-	35594
EPIC-Kitchen-100	100 hours video taken in Kitchen	-	~90000
YouTube-8M	Collected from YouTube	4716	7 million+
YouTube-8M segments	Human validated videos from YouTube-8M	1000	237000

The following content of this survey will be arranged in this order: Section 2 will introduce the traditional HARfV methods and their research progress. In Sections 3, several deep learning networks for HARfV are addressed, and they are categorized according to their types, respectively. Different models are compared horizontally not only in the aspect of their accuracy but also their inference latency, and their characteristics are analyzed and summarized. The current situation of research in HARfV is outlined, and the outlook on future research is also proposed in the conclusion of Section 4.

2. TRADITIONAL COMPUTATIONAL FRAMEWORK OF HARFV

Traditionally, local feature descriptors are usually be used in dealing with HARfV, and the feature of local spatio-temporal regions is extracted first and used in the further steps. It is pointed out that traditional feature extraction methods generally be based on artificially designed features, and HOG is a kind of typical feature in describing an image in video frame by counting value of gradient which is calculated the value of dense grid, so as to improve total accuracy.

HOF and MBH on the optical flow are often considered as temporal information. HOF calculates the direction of optical flow in each frame and obtains the histogram information, which can be a good feature by not only characterizing the motion information, but also being insensitive to the scale of the image and the direction of the motion. MBH is also a kind of feature convenient and fast in calculating what decomposes the optical flow image into grayscale images in both x-direction and y-direction, extracts their gradient histograms, and obtains the video descriptor by extracting the boundary information of the moving object.

To further increase the effectiveness of HOG, Kläser et al.[9] extended HOG features to 3-dimensions, called HOG3D, in order to sample spatio-temporal blocks quickly at multiple scales and a linear SVM classifier is used in the following classification step.

The efficiency of traditional method achieved a great improvement by the approach of DT (Dense Trajectory)[10,11] as well as the iDT (improved Dense Trajectory). DT analyzes the optical flow information to obtain the motion trajectory of the video and extracts features along the motion trajectory, while iDT mainly uses the three main features: HOG, HOF, and MBH. iDT also compensates for camera motion and uses additional detectors to detect human motion. Since iDT has high feature dimensionality that leading the achieved feature to a much larger data size than the original video, thus the time cost might be heavier in some conditions. However, iDT is still one of the best and most effective methods till deep learning framework appeared. According to the results of data test[12], using iDT and BoW together with the linear SVM, it can get a 76.2% average accuracy on UCF101. When using iDT and Fisher vector, the accuracy is 87.9%. Those above methods provide excellent prior knowledge and features processing calculation that still have a chance to improve the performance as part of the deep learning models[13].

3. SOME DEEP LEARNING MODELS TO DEAL WITH THE PROBLEM OF HARFV

3.1 2D CNN-based models with temporal information

2D-CNN is fast but it cannot use temporal information, a representative method is the two-stream CNN[14] that processes spatial and temporal information with two networks separately. It is trained with frames and optical flows what could add the knowledge about how the motion flows in the frames to learn motion features better. The two-stream network requires two independent and separated backbone networks, and its structure, compared with the single-backbone network, the number of calculations increases, resulting in heavier calculations when training the model, but it performs well on the UCF101 dataset and can have an accuracy of 88.0%.

Feichtenhofer et al.[15] presented a two-stream network fusion strategy to improve the performance of the two-stream network, and when the spatio-temporal convolutional network uses VGG16[16] as the backbone network, its accuracy on UCF101 can go up to 92.5%, while if it is combined with iDT, its accuracy can be further increased to 93.6%.

Wang et al.[17-22] also proposed a series of lightweight models based on temporal sequence information, such as Temporal Segment Network (TSN), Temporal Enhancement and Interaction Network (TEINet), Temporal Excitation and Aggregation (TEA), Temporal Adaptive Module (TAM), and Temporal Difference Network (TDN).

TSN is an effective model for recognizing actions in the video which proposed a segment-based sampling method. It can learn motion features from the entire video and uses an aggregation module to model the long-range temporal structures. TSN acts well on multiple datasets such as HMDB51 and UCF101, additionally, TSN has a higher efficiency and can process video with a higher frame rate effectively.

Table 2. Comparison results of average accuracy on the UCF101 dataset.

Method	Accuracy (%)
iDT+MVSV[11]	83.5
iDT+FV[11]	85.9
iDT+HSV[11]	87.9
Two-Stream CNN[14]	88.0
TAM[20]	91.3
Two-stream CNN+ iDT[15]	93.6
TSN[17]	94.9
TSM[21]	96.0
TEA[19]	96.9
TDN[22]	97.4

The TDN is effective and it explicitly extracts the changes in the timing of motion and adds them to the network, by proposing the Long-term Temporal Difference Module (L-TDM) and the Short-term Temporal Difference Module (S-TDM), which can extract the motion changes from the temporal information through the difference operation.

Zhang et al.[23] proposed a method to calculate the parameter's value of motion vector as the input of CNN which inference speed can reach 390.7 frames per second. They transferred the features and characteristics learned from results of optical flow to action vector CNN. It is expected for motion vectors to improve the performance of recognizing fine-grained motions and noise processing. The comparison results are shown in Table 2.

3.2 3D CNN-based HARfV models

Ji et al.[24] proposed 3D CNN-based model to describe the human motion in multiple adjacent frames by extracting attributes from both spatial and temporal dimensions. Carreira et al.[25] presented a two-stream Inflated 3D ConvNet (I3D) and a large-scale video dataset: Kinetics. Pre-training a model on this dataset can make those 3D convolution models get better results on different datasets. They pointed out that it is better to use a two-stream network to capture information effectively. Compared with the two-stream network, the approaches based upon Long Short-Term Memory has lost some temporal information during processing, and the training time in the backpropagation step is also very long. By pre-training through the Kinetics dataset, and then training on the HMDB51 and UCF101 datasets, I3D can achieve an average accuracy of 66.4% on HMDB51 and an average accuracy of 93.4% on the UCF101 dataset.

Many videos usually have a slow or unchanged background and a quickly changing area caused by the motion. Feichtenhofer et al.[26] presented a SlowFast network what take into calculation from spatial and temporal data respectively. They applied two parallel CNN to the same video clip, one of which is a CNN used for the slow but high-resolution background (slow channel), which is used to analyze the static content in the video. The other is a fast and low-resolution CNN (fast channel), which is used to recognize moving parts in a video. They all use a 3D residual network model and use a 3D convolution operation immediately after capturing several frames. The data from the fast channel is sent to the slow channel through the side connection. Compared with the C3D, SlowFast processes spatial information and time sequence information separately and significantly improves the recognition accuracy on those videos with faster motion. The test results are shown in Table 3.

Table 3. 3D CNN-based networks on Kinetics-400.

Method	Accuracy (%)
C3D[12]	62.8
I3D[25]	71.1
X3D[24]	79.1
ResNet-101[26]	79.8

Models based on 3D CNN usually have many parameters, which result in a heavy computational overhead, longer training time, and the delay in inferencing is high[27]. Feichtenhofer[28] also proposed an eXpand 3D (X3D) model. The intent was to reduce this problem. This method can gradually expand in the dimension of time, frame rate, space, width, bottleneck width, and depth and classify small-scale 2D images.

X3D achieved similar accuracy to the prior work but significantly reduced the amount of calculation during network training. But if compared with 2D CNN-based models, X3D often achieves better results in large-scale scene datasets such as Kinetics. On Kinetics-400, X3D can get a top-1 accuracy of 79.1%, a top-5 accuracy of 93.9%. It is comparable to the SlowFast model, which has top-1 accuracy of 79.8% and a 93.9% top-5 accuracy. However, the X3D network cannot learn temporal changes, so it cannot achieve very good results on video datasets such as Something-Something, which are more sensitive to the temporal sequence.

3.3 3 RNN-based HARfV models

The recurrent neural networks (RNN) can record time knowledge from the current and past, thus it can be a pretty good solution to calculate time and order-dependent data like video frames[29]. LSTM (long short-term memory)[30] model was

introduced to RNN to mitigate this problem that RNN architecture can only handle short-term dependencies in dealing with HARfV[31].

Hochreiter et al.[32] proposed a method that combines CNN and LSTM, and Ng et al.[33] proposed five convolutional temporal feature pool architectures, namely: (1) Conv Pooling, which can retain spatio-temporal information; (2) Late Pooling, which can combine the temporal information of high-dimensional abstract features; (3) Slow Pooling, which combines local temporal motion information before combining high-dimensional abstract features; (4) Local Pooling, which does not combine global motion information, thus can reduce the loss of temporal information; (5) Time-Domain Convolution, which adds a time convolution layer, which can combine local temporal information in a shorter time window before calling the max-pooling layer.

By using RNN with LSTM, a model can usually get better training results after training on a longer video. The test results show that on Sports-1M, this solution is significantly improved compared to the best model of the same period, achieving a Top-1 accuracy of 73.1%, and on UCF101 too, achieving a Top-1 accuracy of 88.6%.

Donahue et al.[34] presented a Long-term Recurrent Convolutional Network (LRCN) based on LSTM. LRCN uses a union of CNN and LSTM for HARfV and video description. 2D convolution can only process a single frame, and LSTM can fuse that information from single frames. It passes each frame of the video to the CNN and uses the output of the CNN as the input of the LSTM. Then use the output of the LSTM as the final network output. The parameters of the CNN and LSTM are shared along with time. Research also shows that the use of LSTM for fusion usually can achieve better results than only using a 2D CNN.

Most methods of the above are implemented in the form of CNN combined with LSTM, use CNN to extract features from a video frame and use LSTM to aggregate multiple video frames directly to obtain temporal dependencies of the video. However, the motion learned in this way implicitly assumes that the motion in the video is static at different spatial positions.

4. CONCLUSIONS

This paper summarizes different HARfV methods based on traditional methods and deep learning. Traditional HARfV methods rely heavily on feature designing and its selection. Human intervention is usually required in creating those features. On the other side, deep learning-based methods can automatically train and learn features from labeled video datasets. They usually have better accuracy compared to traditional methods.

Among traditional methods, the DT and iDT are effective yet accurate methods. As the DT and iDT act as a certain degree of prior expert knowledge, when adopting DT and iDT into deep learning-based methods, some can have a better performance.

3D CNN-based models usually surpass 2D CNN-based models in HARfV tasks. However, it has the problems of many parameters and a long training time. Therefore, some use 2D CNN-based models simulating 3D CNN, resulting in reduced parameters but they have similar accuracy and improved the overall performance.

Recently, transformer-based models have quickly become a hot topic in the field of HARfV. Although they have more parameters than 3D CNN-based models, they usually act better in those HARfV tasks. The attention mechanism can be used to capture the overall characteristics of the video flexibly and enabled them to learn from a longer video. With further understanding and the improvement of the efficiency of those models, more deep learning models for video and image recognition tasks will be proposed in the future, and it is expected that the current results will be further advanced.

However, in the real world, the performance of video recording devices is increasing, and the demand for science and technology is also changing rapidly. This poses many challenges to the existing research, such as the inference delay and accuracy of video with a high frame rate. This is particularly important in some scenarios such as car autopilot. Other smart IoT devices, such as the smart doorbell, often have limited resources. This restriction put forward stricter requirements on the accuracy and resource consumption of HARfV models. The solution to these problems is to be supported by further research and new technical solutions by optimizing models.

ACKNOWLEDGMENTS

This work was supported by Guangzhou Science and Technology Project (Number: 202002030273 & 202102080656), and the Key Discipline Project of Guangzhou Xinhua University (number: 2020XZD02), and Guangdong Key Platform for University and Major Scientific Research Project with No. 2018KQNCX361.

REFERENCES

[1] Li, Y., He, H. and Zhang, Z., "Human motion quality assessment toward sophisticated sports scenes based on deeply-learned 3D CNN model," Journal of Visual Communication and Image Representation 71, 102702 (2020).
[2] Yi, Y., Hu, P. and Deng, X., "Human action recognition with salient trajectories and multiple kernel learning," Multimedia Tools and Applications 77(14), 17709-17730 (2018).
[3] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," Proc. CVPR, 886-893 (2005).
[4] Chaudhry, R., Ravichandran, A., Hager, G. and Vidal, R., "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," Proc. CVPR, 1932-1939 (2009).
[5] Wang, H., Kläser, A., Schmid, C. and Liu, C. L., "Dense trajectories and motion boundary descriptors for action recognition," International Journal of Computer Vision 103(1), 60-79 (2013).
[6] Lazebnik, S., Schmid, C. and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. CVPR, 2169-2178 (2006).
[7] Yang, M., Zhang, L., Feng, X. and Zhang, D., "Sparse representation based fisher discrimination dictionary learning for image classification," International Journal of Computer Vision 109(3), 209-232 (2014).
[8] Hinton, G. E., "Learning multiple layers of representation," Trends in Cognitive Sciences 11(10), 428-434 (2007).
[9] Huang, K., Delany, S. J. and Mckeever, S., "Human action recognition in videos using transfer learning," Proc. IMVIP, (2019).
[10] Hinton, G. E., "Deep belief networks," Scholarpedia 4(5), 5947 (2009).
[11] Bengio, Y., "Learning deep architectures for AI," Foundations and Trends in Machine Learning 2(1), 1-127 (2013).
[12] Taylor, G. W., Hinton, G. E. and Roweis, S., "Modeling human motion using binary latent variables," Advances in Neural Information Processing Systems 19, 1345-1352 (2006).
[13] Taylor, G. W. and Hinton, G. E., "Factored conditional restricted Boltzmann machines for modeling motion style," Proc. ACML, 1025-1032 (2009).
[14] Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F. and Lu, T., "Teinet: Towards an efficient architecture for video recognition," Proc. AAAI, 11669-11676 (2020).
[15] Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B. and Wang, L., "Tea: Temporal excitation and aggregation for action recognition," Proc. CVPR, 909-918 (2020).
[16] Liu, Z., Wang, L., Wu, W., Qian, C. and Lu, T., "Tam: Temporal adaptive module for video recognition," Proc. ICCV, 13708-13718 (2021).
[17] Lin, J., Gan, C. and Han, S., "TSM: Temporal shift module for efficient video understanding," Proc. ICCV, 7083-7093 (2019).
[18] Wang, L. M., Tong, Z., Ji, B. and Wu, G. S., "TDN: Temporal difference networks for efficient action recognition," Proc. CVPR, 1895-1904 (2021).
[19] Zhang, B., Wang, L., Wang, Z., Qiao, Y. and Wang, H., "Real-time action recognition with enhanced motion vector CNNs," Proc. CVPR, 2718-2726 (2016).
[20] Cai, Z., Wang, L., Peng, X. and Qiao, Y., "Multi-view super vector for action recognition," Proc. CVPR, 596-603 (2014).
[21] Wang, H. and Schmid, C., "LEAR-INRIA submission for the THUMOS workshop," Proc. ICCV Workshop on THUMOS Challenge, 1-3 (2013).
[22] Zhu, Y. and Newsam, S., "Depth2action: Exploring embedded depth for large-scale action recognition," Proc. ECCV, 668-684 (2016).
[23] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition," Proc. CVPR, 770-778 (2016).
[24] Feichtenhofer, C., "X3d: Expanding architectures for efficient video recognition," Proc. CVPR, 203-213 (2020).
[25] Lee, Y., Kim, H. I., Yun, K. and Moon, J., "Diverse temporal aggregation and depthwise spatiotemporal factorization for efficient video classification," IEEE Access 9, 163054-163064 (2021).
[26] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M., "A closer look at spatiotemporal convolutions for action recognition," Proc. CVPR, 6450-6459 (2018).
[27] Qiu, Z., Yao, T. and Mei, T., "Learning spatio-temporal representation with pseudo-3d residual networks," Proc. ICCV, 5533-5541 (2017).
[28] Yang, X., Molchanov, P. and Kautz, J., "Multilayer and multimodal fusion of deep neural networks for video classification," Proc. ACM Multimedia, 978-987 (2016).
[29] Nie, W., Wang, K., Wang, H. and Su, Y., "The assessment of 3D model representation for retrieval with CNN-RNN networks," Multimedia Tools and Applications 78(12), 16979-16994 (2019).

[30] Hu, M., Wang, H. W., Wang, X. H., Yang, J. and Wang, R. G., "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks," Journal of Visual Communication and Image Representation 59, 176-185 (2019).
[31] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," Neural Computation 9(8), 1735-1780 (1997).
[32] Li, Z., Gavrilyuk, K., Gavves, E., Jain, M. and Snoek, C. G., "Videolstm convolves, attends and flows for action recognition," Computer Vision and Image Understanding 166, 41-50 (2018).
[33] Sun, L., Jia, K., Chen, K., Yeung, D. Y. and Savarese, S., "Lattice long short-term memory for human action recognition," Proc. ICCV, 2147-2156 (2017).
[34] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S. and Shah, M., "Transformers in vision: A survey," ACM Computing Surveys (CSUR), (2021).