*Computers, Materials & Continua* CMC, vol.61, no.3, pp.1189-1204, 2019

# Research on Action Recognition and Content Analysis in Videos Based on DNN and MLN

**Wei Song[1, 2, *], Jing Yu[3], Xiaobing Zhao[1, 2] and Antai Wang[4]**

**Abstract:** In the current era of multimedia information, it is increasingly urgent to realize intelligent video action recognition and content analysis. In the past few years, video action recognition, as an important direction in computer vision, has attracted many researchers and made much progress. First, this paper reviews the latest video action recognition methods based on Deep Neural Network and Markov Logic Network. Second, we analyze the characteristics of each method and the performance from the experiment results. Then compare the emphases of these methods and discuss the application scenarios. Finally, we consider and prospect the development trend and direction of this field.

**Keywords:** Video action recognition, deep learning network, markov logic network.

## 1 Introduction

Video action recognition and understanding is closely related to people's intelligent life, covering a variety of application areas including intelligent home, human-computer interaction, automatic driving, video monitoring and so on. Many types of problems related to video have also been studied for a long time, such as video segmentation [Song, Gao, Puscas et al. (2016)], video retrieval [Liu, Li, Deng et al. (2017); Song, Gao, Liu et al. (2018)], motion recognition [Wang and Schmid (2013); Ng, Hausknecht, Vijayanarasimhan et al. (2015); Peng, Wang, Wang et al. (2016)] and so on [Pan, Lei, Zhang et al. (2018); Zhang, Meng and Han (2017)]. Among them, video action recognition and content analysis are widely studied, because they are related to People's Daily life and safety. With the development of deep learning, probabilistic graph model, logical reasoning, new remarkable breakthroughs have been made in this field.

In the ImageNet 2012 competition, Alex et al. [Krizhevsky, Sutskever and Hinton (2012)] used the deep learning framework Alex Net to reduce the top-5 error rate of image content recognition by 10 percentage points, which enabled deep learning to be rapidly applied to all fields of computer vision. Since then, deep learning has developed

[1] School of Information Engineering, Minzu University of China, Beijing, 100081, China.

[2] National language resource monitoring &Research Center Minority Languages Branch, Minzu University of China, Beijing, 100081, China.

[3] School of Electronic Information Engineering, Beijing Jiaotong University, Beijing, 100044, China.

[4] New Jersey Institute of Technology, 323 Dr Martin Luther King Jr Blvd. Newark, NJ, 07102, USA.

[*] Corresponding Author: Wei Song. Email: songwei@muc.edu.cn.

rapidly. Researchers have explored a variety of new effective deep network structures from multiple perspectives, such as the depth and width of network structure and the microstructure contained therein, including VGG-Net, GoogleNet, NIN, ResNet [Simonyan and Zisserman (2014); Szegedy, Liu, Jia et al. (2015); Lin, Chen and Yan (2013); He, Zhang, Ren et al. (2016)], etc. Later, Yu kai et al. proposed 3D convolution network (3DCNN) for video analysis [Ji, Xu, Yang et al. (2013)], which is different from the two-dimensional convolution network. First, the manual design of convolution kernels is act on the input frame, then five features are extracted, including the grey value, vertical gradient, horizontal gradient, horizontal flow and vertical flow. And then the convolution kernel is used on the adjacent three consecutive frames to extract temporal features from the video. Finally, a 128-dimensional feature vector contained movement information from the specific frame. The proposed 3DCNN model provides a new direction for the research of action recognition and content analysis.

On the other hand, deep learning [Xiong, Shen, Wang et al. (2018); Zhou, Liang, Li et al. (2018)] is extremely dependent on the size of the dataset and it is not yet possible to explain its mechanism of action. Because of the full use of data correlation and a rigorous logical reasoning process, the probabilistic graph model and the first-order logical structure has been highly praised by some scholars and widely used. Specifically, in earlier integration of multiple local features of multimedia content, the probabilistic map model and first-order logic was introduced in consideration of temporal coherence, and these methods combined the other techniques, such as the bag-of-visual-words model, spatio-temporal dictionary learning, and sparse coding. And the most classic methods among them are Hidden Markov Models (HMMs) [Rabiner (1989)] and Conditional Random Fields (CRFs) [Lafferty, McCallum and Pereira (2001)]. Subsequently, Domingos et al. [Richardson and Domingos (2006)] proposed Markov logic network in 2006, and it combines first-order logic and probabilistic graph models, which can soften its hard constraints effectively, and it can be used when many compactly representing relationship is uncertain.

This paper will discuss the methods of video action recognition and content analysis from two aspects: deep neural network and Markov logic network. The remaining of the article is as follows: The second section reviews and compares the video action recognition methods based on deep neural networks, and analyzes the main features of each method. The third section makes a deep analysis of the application of Markov logic network in the video field. The future development of the field will be given in the conclusion.

## 2 Action recognition based on deep neural network

The development of computer hardware technology has greatly promoted deep learning, and video action recognition as an important field in computer vision is also very dependent on deep learning. With the rapidly development of deep neural networks in recent years, researchers have tried to use deep learning in the video field mainly from two aspects: deep features and end-to-end networks.

### *2.1 Deep feature*

Video content representation, that is feature extraction, is the core of video action

recognition [Yao, Lei and Zhong (2019)]. Then, whether the feature extraction and effective characterization of the video content can be better realized will directly determine the motion recognition effect.

A robust and efficient three-dimensional convolutional network model (C3D) proposed by Tran et al. [Tran, Bourdev, Fergus et al. (2015)] extends the convolution kernel of the convolutional layer and the pooled layer in the network into three-dimensional, which can simultaneously extract the spatial and temporal features of the video. And this method simply makes the resulting features be the input of a linear classifier, and it got good results. Further, using the deep residual learning framework to deploy the 3D architecture, the resulting Res3D [Tran, Ray, Shou et al. (2017)] not only improves the recognition accuracy, but also reduces the model size by more than 2 times and shortens the running time by more than 2 times. The whole model is more compact than C3D. Xu et al. [Xu, Das and Saenko (2017)] added two modules based on C3D, namely Proposal Subnet and Classification Subnet, where the features through C3D network separately extract the time dimension feature information by pooling. Zisserman et al. replaced the basic network of the three-dimensional convolutional neural network with Inception-V1, obtained the I3D depth feature extraction mechanism [Carreira and Zisserman (2017)], and pre-trained on the newly constructed dataset Kinetics, and achieved good recognition results.

Shi et al. [Shi, Tian, Wang et al. (2016)] defined an effective long-term descriptor: sDTD. Specifically, dense trajectories are mapped into the binary image space, and then CNN-RNN is used to perform effective feature learning for long-term motion. Currently, video frames, dense trajectories, and sDTDs are effective complementary to video characterization as spatial, short-term, and long-term features. The mapping formula of the dense trajectory extracted to a series of trajectory texture images is as follows:

$$I_j(x,y)=\begin{cases}\sqrt{(\triangle x_l^k)^2+(\triangle y_l^k)^2}, & if\ x_l^k=x\ and\ y_l^k=y\\ 0, & otherwise.\end{cases} \tag{1}$$

where, $\triangle x_l^k$, $\triangle y_l^k$ denotes the displacement of the $k$th track between $l$ frame and $l$+1 frame, $x_l^k$ and $y_l^k$ denotes the position of track $k$ at timestamp $l$. Subsequently, the converted trajectory texture image is input to the CNN to obtain a DTD, and then the LSTM is input to obtain sDTD.

Wang et al. [Wang, Gao, Wang et al. (2017)] realized the feature of extracting the same dimension for any size scene by spatial temporal pyramid pooling (STPP), and set multi-level pooling to change the feature size. And this method solved the previous network structure limitation that the input of the former structure is the fixed number of frames of video data and the fixed size of the frame. However, the number of frames with the same action in the video is variable, and the fixed frames will destroy the integrity of the whole action.

FlowNet [Dosovitskiy, Fischer, Ilg et al. (2015)] and FlowNet 2.0 [Ilg, Mayer, Saikia et al. (2017)] are a set of work based on convolutional neural networks to predict optical flow. Based on the above two methods, Ng et al. [Ng, Choi, Neumann et al. (2016)] proposed a feature multitask learning method for small number of labeled samples. The features of unlabeled video action information can be effectively learned. The recognition accuracy is greatly improved (23.6%) without relying on extra massive data and incidental optical flow.

Unlike the previous external input with optical flow, only the video frame is used as input to simultaneously acquire the optical flow and category labels.

$$MT-Loss_{j,t}=-I(y_j=\hat{y}_j)\log p(\hat{y}_j)+\lambda\sum_{r=1}^{4}\alpha_r\sum_{p}\left\|o_{j,t,p}^{(r)}-\hat{o}_{j,t,p}^{(r)}\right\|_2 \tag{2}$$

where, $I$ denote indication function, and $j$ denotes video index number.

In order to compensate for the gap between artificial optical flow characteristics and end-to-end learning, which are independent of each other and cannot adjust each other. Fan et al. [Fan, Huang, Chuang et al. (2018)] proposed a new neural network TVNet to obtain similar optical flow characteristics from data instead of artificial features, mainly to solve the problem of disconnection between deep network and artificial optical flow, and the problem of space and time consumption caused by calculating and storing optical flow. For these two problems, the network structure is based on the TV-L1[Zach, Pock and Bischof (2007)] method. As shown in Eq. (3), the simulation and expansion are carried out, and the iteration is converted into a TVNet module, which can be integrated with other specific task networks to build a model to avoid pre-training and storage features.

$$\min_{u(x),x\in\Omega}\sum_{x\in\Omega}(\left|\nabla u_1(x)\right|+\left|\nabla u_2(x)\right|)+\lambda\left|\rho(u(x))\right| \tag{3}$$

where, the first term accounts for the smoothness condition, and the second term corresponds to the brightness constancy assumption, which is equal to the difference in brightness of the pixel $x$ between the two frames.

Wang et al. [Wang and Cherian (2018)] proposed a method to improve the robustness of video features. The original data feature of the video sequence and its corresponding perturbation are treated into two packets. By modeling the two classifications problem, a set of hyper-planes can be acquired to separate the two classes of packets, and the obtained hyper-plane is used as a descriptor of the video, which is called discriminant subspace pooling. The descriptors obtained above are relative to the corresponding sequences and are not compatible with other videos, so it is necessary to regularize the subspace by adding orthogonal constraints.

Girdhar et al. [Girdhar, Ramanan, Gupta et al. (2017)] proposed a local aggregation descriptor for video motion recognition, which is obtained the global feature of the video level by softening the sub-actions in the video. Compared with the traditional maximum pooling and average pooling, this feature can fully obtain the distribution of sub-features.

The comparison of video features based on deep networks is shown in Tab. 1:

**Table 1:** Comparison of depth characteristics

| Feature | Dataset | | Characteristics | Classifier or System architecture |
|---|---|---|---|---|
| | UCF101 | HMDB51 | | |
| C3D | 82.3 | - | The most successful video 3D features earlier | SVM |
| I3D | 97.9 | 80.2 | Pretraining on larger dataset Kinetics is required | Linear Classifier |
| sDTD | 92.1 | 63.7 | Converting dense features into binary images to obtain long-term features through deep networks | Two-stream |
| STPP | 92.6 | 70.5 | Uniform dimensional features can be obtained from arbitrary length and image size video | LSTM & CNN-E + Two Stream |
| TVNet | 94.5 | 71.0 | Using deep networks instead of iterations to obtain more discriminative flow characteristics | CNN |
| Discriminative subspace pooling | - | 72.5 | More robust by perturbation addition | Two-stream |
| ActionVLAD | 92.7 | 66.9 | Aggregation of subspaces in video | Linear Classifier |

For the depth feature, the existing methods mainly come from the following two aspects: The first one is to design the three-dimensional convolution kernel and construct the three-dimensional deep network to realize the synchronous extraction of video spatial features and temporal domain features, and then can preserve the internal correlation. The second one is to use deep neural network instead of manual design to obtain the action information features, such as optical flow, and the features make the features more compact and more robust, and thus achieve the purpose of improving the recognition accuracy.

### *2.2 Multi-channel end-to-end network*

Compared to still images, video content can be thought of as a collection of ordered still images, but at the same time contains more extensive action and time domain information. How to effectively represent and analyze the spatial and temporal characteristics of video content at the same time is the key to recognize the action and the content.

In 2014, Two-Stream Convolutional Networks [Simonyan and Zisserman (2014)] is proposed, which opened a new door for video action recognition. The article first points out that one of the challenges of video motion recognition is how to extract complementary appearance and motion information from still frames and multiple frame images. And aims to generate the best artificial features under a data-driven framework. The paper proposes a two-channel network architecture combining time and space networks. Later, in 2016, CVPR proposed a two-channel fusion [Feichtenhofer, Pinz and Zisserman (2016)], which increased the information interaction between the two channels. For fusion strategies, convolutional layer fusion can reduce parameters without loss of performance compared to softmax layer fusion.

In 2016, the Swiss Federal Institute of Technology Zurich proposed a new deep structure

called semantic domain-based dual-stream deep network (SR-CNNs) by combining the detection results of people and objects [Wang, Song, Wang et al. (2016)]. Semantic cues are used in addition to the dual stream network structure used in Simonyan et al. [Simonyan and Zisserman (2014)]. Based on the basic two-stream channel architecture, the final layer is replaced with the RoIPooling layer to isolate three different semantic cues and generate a score. In order to process multiple items, a multi-instance learning layer is employed.

In 2017, the spatio-temporal pyramid network for video action recognition [Wang, Long, Wang et al. (2017)] designed a new spatio-temporal compact bilinear (STCB) fusion mechanism to fuse the feature information of two channels in time and space. In addition, the pooling operation based on the attention mechanism is used, and the effect is better than the average pooling and the maximum pooling. It reached 94.6% on UCF101 and 68.9% on HMDB51. Also, a novel space-time multi-layer network for video motion recognition [Feichtenhofer, Pinz and Wildes (2017)] is proposed, which is also intended to link two separate channels, using a motion gating strategy.

Feichtenhofer et al. [Feichtenhofer, Pinz and Wildes (2016)] combined the best-performing ResNet and Two-Stream frameworks in the field of still image recognition and established a connection between two channels, preserving the correlation between space and time domain in video features, i.e. ST-ResNet, the recognition effect has been greatly improved compared to the basic Two-Stream framework.

Wang et al. proposed a video timing segmentation network TSN [Wang, Xiong, Wang et al. (2016)], which divides the full-length video into several video segments and inputs them into the temporal and spatial feature extraction network. Finally, the spatial and temporal decisions are merged to obtain the final category. Inspired by Wang et al. proposed a time-domain difference network [Ng and Davis (2018)]. For multiple consecutive frames, the Euler distance-based differential calculation is performed on each output of the convolutional network. The motion feature and the image feature are collaboratively calculated to achieve efficient analysis of the video.

Zolfaghari et al. designed a chained multi-stream network [Zolfaghari, Oliveira, Sedaghat et al. (2017)] that integrates action information, motion information and original images. For integration, a Markov chain model was introduced to enhance the continuity of the clues. Through Markov chain integration, the action tag is refined. This strategy is not only superior to the independent training of the channel, but also imposes an implicit regularization, which is more robust to over-fitting. Using Markov networks for multi-channel fusion, unlike previous work, multiple channels are sequentially connected: first the action flow, then the optical flow refinement, and finally further refined using the RGB stream. Based on the assumption that each category of prediction conditions is independent, the joint probability of all input streams can be decomposed into the conditional probability of each independent stream. In the model, the prediction for each phase is the conditional probability of the previous prediction and its new input.

Jiang et al. [Wu, Jiang, Wang et al. (2016)] further extended the dual-channel architecture to multiple channels. First, three convolutional neural networks are trained to model spatial, short-term action and audio features, respectively, and then use Long Short Term Memory Networks (LSTM) to explore long-term dynamics for spatial and short-term

channels. Based on the above five characteristics, a five-channel video content analysis framework is constructed.

The system structure pair based on the dual channel architecture is shown in Tab. 2:

**Table 2:** Multi-channel structure comparison

| Framework | Date Set | | Characteristic |
|---|---|---|---|
| | UCF-101 | HMDB-51 | |
| Basic Two-Stream | 88.0 | 59.4 | The earliest architecture to handle RGB and Flow as two channels |
| ST-ResNet | 93.4 | 66.4 | Embed the best-performing ResNet into the base frame with connections between the channels |
| TSN | 94.0 | 68.5 | Segmenting the video and then sparsely sampling to obtain video-level features |
| SR-CNNs | 92.6 | - | Using Faster R-CNN as a detector, only the detected semantic regions are processed |
| Multi-Stream Multi-Class Fusion | 92.6 | - | Up to five feature channels, enhanced multi-modal feature understanding |

For dual or multi-channel system frameworks, information interaction and fusion between channels is important. Processing the spatial and temporal components of the video separately can make the video content analysis simple, but at the same time destroy the strong correlation between the spatial-temporal in the video. How to reestablish a reasonable fusion mechanism and interaction between the channels is a focus research direction. In addition, the current selection of spatial domain action information by such methods generally depends on optical flow characteristics, and the calculation and storage of optical flow information requires a large amount of resources. Therefore, how to process the video action information is a bottleneck of these methods.

### *2.3 Other research for deep networks*

The three-dimensional convolution network faces the problem of high computational complexity and large demand for training sample size. Several recent studies have proposed factoring three-dimensional spatial-temporal convolution kernels to achieve faster and more efficient processing. Specifically, the three-dimensional convolution kernel is decomposed into a two-dimensional spatial convolution kernel followed by a one-dimensional time-axis convolution kernel.

Tran et al. [Tran, Wang, Torresani et al. (2017)] proposed an R(2+1)D structure for video action recognition, which is to factorize the previous ResNet-3D (R3D) into 2D+1D. There are two advantages of this: first, the increasement of the nonlinear activation function is to improve the power of nonlinear representation; secondly, promotes the model optimization ability and obtains lower Training loss and test loss.

This structure is related to Factorized Spatio-Temporal Convolutional Networks (FSTCN) [Sun, Jia, Yeung et al. (2015)] and P3D (Pseudo-3D network) [Qiu, Yao and Mei (2017)], but it has its own unique advantages: FSTCN focuses on the factorization of the network, and it is deployed at the bottom of the spatial layer and at the top of the two parallel time domain layers, while R(2+1)D focuses on the factorization of the layer, decomposing each space-time convolution kernel; P3D combines a single spatial convolution kernel and a time convolution kernel in series, and R(2+1)D uses only an overall spatio-temporal residual convolution kernel, making the model more compact. Another benefit of factoring decomposition of 3D deep network is that the system can be pre-trained in a static image dataset.

The three-dimensional deep network factorization strategy and its characteristics are shown in Tab. 3:

**Table 3:** 3D depth network factorization comparison

| | |
|---|---|
| P3D | It is characterized by a separate combination of spatial convolution and time convolution |
| FSTCN | The feature is to decompose at the network level, deploy the 2D space layer at the lower layer of the network, and deploy the 1D time layer at the upper layer of the network |
| R(2+1)D | Simply decompose all or part of the 3D convolutional layer in the existing 3D CNN (for example, R3D) into a 2D+1D structure |
| S3D [Xie, Sun, Huang et al. (2017)] | Separable 3D CNN, which decomposes the 3D convolutional layer in the best-performing 3D CNN structure I3D, because it is separable in time and space information, called S3D |

In addition, the field of video analytics seems to have reached the bottleneck stage for building a new end-to-end network architecture or designing new video deep learning features. Researchers are more inclined to design a small module that can be embedded into existing networks to improve computational efficiency and characterization capabilities.

Diba et al. [Diba, Fayyaz, Sharma et al. (2018)] proposed a Spatio-Temporal Channel Correlation (STC) that can be embedded in existing networks. For STC, it is divided into Temporal Correlation Branch (TCB) and Spatial Correlation Branch (SCB). The independent information extraction of different dimensions is realized by pooling in different dimensions.

$$z_{tcb} = \frac{1}{W \times H \times T} \sum_{i}^{W} \sum_{j}^{H} \sum_{t}^{T} x_{ijt} \ , z_{scb} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} x_{ij} T \tag{4}$$

Yan et al. [Chen, Kalantidis, Li et al. (2018)] proposed a sparse connection module embedded in an existing network to reduce the number of parameters in large quantities. Wherein, a plurality of separate lightweight residual units is connected by a multiplexer to ensure information interaction between the paths, and the multiplexer is composed of two 1×1 linear mapping layer. Due to the uniform input and output dimensions of the module, it can be embedded anywhere in the network to increase the cost of a very small number of parameters to deepen the network.

## 3 Event recognition based on logical reasoning

In daily behavior videos, especially in security surveillance videos, a series of noise interferences such as occlusion, illumination changes, and viewing angle changes often occur. At the same time, video content analysis inevitably needs to combine existing experience and knowledge, whereas the existing machine learning algorithms lack the use of background knowledge and the treatment of uncertainty [Katzouris, Michelioudakis, Artikis et al. (2018)]. In summary, event recognition often needs to deal with data such as incompleteness, error, inconsistency, and situational changes. At this time, the causal and uncertainty analysis of the probability graph model, and the correlation reasoning of the first-order logic can precisely apply to the processing of this kind of data.

Domingos et al. [Richardson and Domingos (2006)] proposed the Markov logic network in 2006. This model is a method that combines first-order logic and probability graph models. It can effectively soften its hard constraints and deal with uncertainties while compactly representing many relationships. First-order logic is a knowledge base consisting of a series of sentences or rules [Onofri, Soda, Pechenizkiy et al. (2016)], and the Markov logical network gives weights to each rule, softening the hard rules. From the perspective of probability, Markov logic network can be flexibly and modularly combined with a large amount of knowledge. From the perspective of first-order logic, Markov logic network can deal with uncertainty robustly, allowing for a few flaws and even contradictory knowledge bases, to reduce vulnerability. Specifically, the Markov logic network probability distribution is as follows:

$$P(X=x)=\frac{1}{Z}\exp(\sum_i \omega_i n_i(x))=\frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)} \tag{5}$$

Among them, $n_i(x)$ is the basic number of rules for truth of first-order logic rule $F_i$, and $x_{\{i\}}$ is the true atom of $F_i$, and with $\phi_i(x_{\{i\}})=e^{\omega_i}$.

Based on the above characteristics, Markov logic network is widely used in complex event recognition, and can perform automatic reasoning based on partial or incomplete information, mainly for activities of daily living (ADLs). In addition, because ADLs related data sets are expensive and subjective, Markov logic networks have become an indispensable technology.

Luo et al. [Song, Kautz, Allen et al. (2013)] proposed a universal framework for integrating information that varies in detail and granularity, and then using multimodal data to recognize hierarchical sub-events in complex events. The framework's deploying system input includes both visual and linguistic parts. By detecting the two categories of items and characters, and location of the relationship between them and matching of specific rules, it analyzes and generates low-level events. Experiments verify the importance of linguistic information when visual information is full of noise or incompleteness.

Deng et al. [Liu, Deng and Li (2017)] proposed a multi-level information fusion model to process dynamic data and contextual information of monitoring events, and simultaneously used the corresponding method based on Markov logic network to deal with uncertainty. Among them, the rules are obtained through information fusion, and the related weights are obtained through statistical learning of historical data. The MLN-

based method for event recognition is mainly composed of three parts: a. Multi-level information fusion module, specifically the monitoring layer, the contextual layer and the event layer, wherein the context layer uses the probability map to fuse the lower layer and the domain knowledge, and then forms a series of rules. b. The rule weight is obtained by statistical learning method, and the specific method is Newton iteration method. c. Dynamic weight update, which aims to update the unsuitable weights in the event recognition process. When the incorrect event exceeds the threshold, correct its correlation weights. Experiments show that the proposed algorithm has higher accuracy than the traditional HMM algorithm, and the dynamic weight update is very impressive for the performance of traditional MLN.

Civitarese et al. [Civitarese, Bettini, Sztyler et al. (2018)] proposed a knowledge-based collaborative action learning recognition method to improve the correlation between sensory events and behavior types. Firstly, the semantic integration layer is set to pre-process the original sensory signals. Secondly, numerical constraints are imposed on the Markov logic network, and the behavior is recognized by modeling and reasoning the detected events and semantic correlations. At the same time, the parallel rule-based online segmentation layer splits the continuous data flow of the sensory event. Finally, the cloud server is used to perform collaborative calculation and feedback on the above two modules.

Bellotto et al. [Fernandez-Carmona, Cosar, Coppola et al. (2017)] collected information from both local and global levels. The local information is obtained by the RGB-D camera, and the global information is defined by the normalized entropy. Specifically, by providing a specific position under a plurality of time stamps, information entropy is used to define a probability distribution of various activities. Finally, a hybrid Markov logic network is used to fuse the two types of information. Experiments have shown that MLN detectors can greatly improve the detection accuracy compared to a single rule-based detector. On the other hand, providing a confidence value can greatly improve the robustness of the system.

Tran et al. [Tran and Davis (2008)] proposed a method based on Markov logic network for the modeling and recognition of surveillance video events, combining traditional computer vision algorithms with common reasoning to compensate for the uncertainty in recognition. Uncertainty specifically refers to logical singularity and detection uncertainty. This method naturally combines the uncertainty of computer vision with logical reasoning, and the rules consisting of first-order logic can be further combined with an easy deductive algorithm to construct the network.

Cheng et al. [Cheng, Wan, Buckles et al. (2014)] analyzed the reasons for the effectiveness of MLN in video behavior analysis applications. Among them, the usual logic rules define behaviors by intersecting multiple low-level actions. The experimental results in the Weizmann dataset show that MLN is effective in video behavior recognition, but it is unsatisfactory for similar actions, and has strong dependency on trajectory detection.

Gayathri et al. [Gayathri, Elias and Ravindran (2015)] used four factors, object perception, location, timestamp and duration to build a hierarchical structure for the detection of anomalous events. The innovation is the use of MLN to combine data-driven

and knowledge-driven methods, hard rules and soft rules to give a hybrid approach. The experimental results in the UCI machine learning repository [Fco, Paula and Araceli (2013)] show that the MLN method has better generalization performance than the hidden Markov model, and the F measure performs better; The hierarchical structure has a faster response compared with the non-hierarchical structure.

For sports, such occlusions and motion events with severe dynamics and cluttered backgrounds are difficult to recognize. William Brendel et al. [Brendel, Fern and Todorovic (2011)] proposed a probabilistic event logic to address the following three problems in the above field: identifying each event, locating the time and location of events; interpreting time and space relationship and semantic constraints from the perspective of domain knowledge.

Gayathri et al. [Gayathri, Easwarakumar and Elias (2017)] used the ontology model to deal with issues such as action granularity, contextual knowledge, and activity diversity. And simultaneously Markov logic network is used to respond to problems like action diversity and data uncertainty by probabilistic reasoning of the represented domain ontology. Experiments on the WSU CASAS dataset [Singla, Cook and Schmitter-Edgecombe (2009)] show that the proposed method has a lower F measure and higher recognition accuracy than traditional Neural Networks, Support Vector Machines, Bayesian Networks, and hidden Markov Networks, etc.

Markov Logic Network can benefit from formal declarative semantic information, and then construct various inference mechanisms for complex event data, enabling efficient management of complex event characteristics, and make the results verifiable and traceable [Stojanovic and Artikis (2013); Laptev and Lindeberg (2005)]. At the same time, compared with the hidden Markov model, Markov Logic can integrate rich time domain contextual information without re-updating the model every time, and only needs to add rules. And this method overcomes the non-reusable problem.

Markov Logic Network also has some shortcomes. The basic idea of Markov Logic Network in reasoning is to divide complex events into multiple simple actions, and to infer complex event categories through logic rules with weights. However, sub-component trajectories with different objects in the video content are needed to construct rules for reasoning. Therefore, data pre-processing before recognition may become the bottleneck of this type of method. In addition, although the behavior recognition based on Markov Logic Network is highly accurate in some specific contexts, and the first-order logic has no limitation on the domain knowledge representation of time domain and composite activities. The first-order logic cannot automatically find the inconsistencies between the represented knowledge, and cannot achieve the hierarchical association of domain knowledge organization and related concepts. These limitations have led to the inability of Markov Logic Network to model the granularity and diversity of activities.

## 4 Conclusion

Although video action and behavior recognition can be regarded as a series of continuous static image understandings, the analysis process is still very complicated. As a data-driven method, deep neural network builds a model through statistical machine learning mechanism, which is good at dealing with uncertainty and time-domain data. Markov Logic Network is a

knowledge-driven method, which constructs models through knowledge representation. And it is good at reusability and analyzing the content based on context.

At present, for deep neural networks, it is better at using a large amount of data to address a single type of video action and content. When the angle of view in the video changes, the noise such as illumination becomes larger, the recognition performance is greatly affected. At the same time, the network is biased towards the black box model application, which is poorly interpretable. Therefore, exploring the mechanism of deep learning and finding a set of design guidelines for system network structure may be the next step in the field. For Markov Logical Network, its powerful reasoning ability makes it more suitable for the field of event reasoning in daily life.

In addition, compared with deep learning, Markov Logic Network has small dependency on data and lower cost of sensor equipment. So, the model of smart home is often based on probabilistic logic framework, and it is difficult to embed deep network into it. However, how to make rational use of domain knowledge and establish a complete knowledge rule base are the key and difficult points in this field.

**Acknowledgement:** This work was supported in part by National Science Foundation Project of P. R. China (Grant Nos. 61503424, 61331013), Promotion plan for young teachers' scientific research ability of Minzu University of China, Youth Academic Team Leadership Project, MUC111 and the First-class University and First-class Discipline of Minzu University of China ("intelligent computing and network security"). Our gratitude is extended to the anonymous reviewers for their valuable comments and professional contributions to their improvement of this paper.

## References

**Brendel, W.; Fern, A.; Todorovic, S.** (2011): Probabilistic event logic for interval-based event recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3329-3336.

**Carreira, J.; Zisserman, A.** (2017): Quo vadis, action recognition? a new model and the kinetics dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724-4733.

**Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J.** (2018): Multi-fiber networks for video recognition. *Proceedings of the European Conference on Computer Vision*, pp. 352-367.

**Civitarese, G.; Bettini, C.; Sztyler, T.; Riboni, D.; Stuckenschmidt, H.** (2018): NECTAR: knowledge-based collaborative active learning for activity recognition. *IEEE International Conference on Pervasive Computing and Communications*, pp. 1-10.

**Cheng, G.; Wan, Y.; Buckles, B. P.; Huang, Y.** (2014): An introduction to Markov logic networks and application in video activity analysis. *International Conference on Computing, Communication and Networking Technologies*, pp. 1-7.

**Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.** (2015): Flownet: learning optical flow with convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2758-2766.

**Diba, A.; Fayyaz, M.; Sharma, V.; Mahdi Arzani, M.; Yousefzadeh, R. et al.** (2018): Spatio-temporal channel correlation networks for action classification. *Proceedings of the European Conference on Computer Vision*, pp. 284-299.

**Feichtenhofer, C.; Pinz, A.; Zisserman, A.** (2016): Convolutional two-stream network fusion for video action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1933-1941.

**Feichtenhofer, C.; Pinz, A.; Wildes, R. P.** (2017): Spatiotemporal multiplier networks for video action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7445-7454.

**Feichtenhofer, C.; Pinz, A.; Wildes, R.** (2016): Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, pp. 3468-3476.

**Fernandez-Carmona, M.; Cosar, S.; Coppola, C.; Bellotto, N.** (2017): Entropy-based abnormal activity detection fusing RGB-D and domotic sensors. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 42-48.

**Fan, L.; Huang, W.; Gan, C.; Ermon, S.; Gong, B. et al.** (2018): End-to-end learning of motion representation for video understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6016-6025.

**Fco, J. O.; Paula, D. T.; Araceli, S.** (2013): Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, vol. 13, no. 5, pp. 5460-5477.

**Gayathri, K. S.; Easwarakumar, K. S.; Elias, S.** (2017): Probabilistic ontology-based activity recognition in smart homes using Markov logic network. *Knowledge-Based Systems*, vol. 121, pp. 173-184.

**Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B.** (2017): Actionvlad: learning spatio-temporal aggregation for action classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 971-980.

**Gayathri, K. S.; Elias, S.; Ravindran, B.** (2015): Hierarchical activity recognition for dementia care using Markov logic network. *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 271-285.

**He, K.; Zhang, X.; Ren, S.; Sun, J.** (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

**Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A. et al.** (2017): Flownet 2.0: evolution of optical flow estimation with deep networks. *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 6.

**Ji, S.; Xu, W.; Yang, M.; Yu, K.** (2013): 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231.

**Katzouris, N.; Michelioudakis, E.; Artikis, A.; Paliouras, G.** (2018): Online learning of weighted relational rules for complex event recognition. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 396-413.

**Krizhevsky, A.; Sutskever, I.; Hinton, G. E.** (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

**Lin, M.; Chen, Q.; Yan, S.** (2013): Network in network. *arXiv*: 1312.4400.

**Lafferty, J.; McCallum, A.; Pereira, F. C. N.** (2001): Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289.

**Liu, X.; Li, Z.; Deng, C.; Tao, D.** (2017): Distributed adaptive binary quantization for fast nearest neighbor search. *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5324-5336.

**Laptev, I.; Lindeberg, T.** (2005): On space-time interest points. *IEEE International Conference on Computer Vision*, vol. 64, pp. 107-123.

**Liu, F.; Deng, D.; Li, P.** (2017): Dynamic context-aware event recognition based on Markov logic networks. *Sensors*, vol. 17, no. 3, pp. 491.

**Ng, J. Y. H.; Choi, J.; Neumann, J.; Davis, L. S.** (2018): Actionflownet: learning motion representation for action recognition. *IEEE Winter Conference on Applications of Computer Vision*, pp. 1616-1624.

**Ng, J. Y. H.; Davis, L. S.** (2018): Temporal difference networks for video action recognition. *IEEE Winter Conference on Applications of Computer Vision*, pp. 1587-1596.

**Ng, J. Y. H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R. et al.** (2015): Beyond short snippets: deep networks for video classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694-4702.

**Onofri, L.; Soda, P.; Pechenizkiy, M.; Iannello, G.** (2016): A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Systems with Applications*, vol. 63, no. 97-111.

**Peng, X.; Wang, L.; Wang, X.; Qiao, Y.** (2016): Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. *Computer Vision and Image Understanding*, vol. 150, pp. 109-125.

**Pan, Z.; Lei, J., Zhang, Y.; Wang, F. L.** (2018): Adaptive fractional-pixel motion estimation skipped algorithm for efficient HEVC motion estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 12.

**Qiu, Z.; Yao, T.; Mei, T.** (2017): Learning spatio-temporal representation with pseudo-3D residual networks. *IEEE International Conference on Computer Vision*, pp. 5534-5542.

**Rabiner, L. R.** (1989): A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

**Richardson, M.; Domingos, P.** (2006): Markov logic networks. *Machine Learning*, vol. 62, no. 1-2, pp. 107-136.

**Shi, Y.; Tian, Y.; Wang, Y.; Huang, T.** (2017): Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510-1520.

**Song, J.; Gao, L.; Liu, L.; Zhu, X.; Sebe, N.** (2018): Quantization-based hashing: a general framework for scalable image and video retrieval. *Pattern Recognition*, vol. 75, pp. 175-187.

**Song, J.; Gao, L.; Puscas, M. M.; Nie, F.; Shen, F. et al.** (2016): Joint graph learning and video segmentation via multiple cues and topology calibration. *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 831-840.

**Simonyan, K.; Zisserman, A.** (2014): Very deep convolutional networks for large-scale image recognition. *arXiv*: 1409.1556.

**Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. et al.** (2015): Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

**Simonyan, K.; Zisserman, A.** (2014): Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, pp. 568-576.

**Sun, L.; Jia, K.; Yeung, D. Y.; Shi, B. E.** (2015): Human action recognition using factorized spatio-temporal convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4597-4605.

**Song, Y. C.; Kautz, H.; Allen, J.; Swift, M.; Li, Y. et al.** (2013): A markov logic framework for recognizing complex events from multimodal data. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 141-148.

**Singla, G.; Cook, D. J.; Schmitter-Edgecombe, M.** (2009): Tracking activities in complex settings using smart environment technologies. *International Journal of Biosciences, Psychiatry, and Technology*, vol. 1, no. 1, pp. 25.

**Stojanovic, N.; Artikis, A.** (2013): On complex event processing for real-time situational awareness. *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pp.114-121.

**Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M.** (2015): Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489-4497.

**Tran, D.; Ray, J.; Shou, Z.; Chang, S. F.; Paluri, M.** (2017): Convnet architecture search for spatiotemporal feature learning. *arXiv*: 1708.05038.

**Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y. et al.** (2018): A closer look at spatiotemporal convolutions for action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450-6459.

**Tran, S. D.; Davis, L. S.** (2008): Event modeling and recognition using markov logic networks. *European Conference on Computer Vision*, pp. 610-623.

**Wang, H.; Schmid, C.** (2013): Action recognition with improved trajectories. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551-3558.

**Wang, X.; Gao, L.; Wang, P.; Sun, X.; Liu, X.** (2017): Two-stream 3D convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634-644.

**Wang, J.; Cherian, A.** (2018): Learning discriminative video representations using adversarial perturbations. *Proceedings of the European Conference on Computer Vision*, pp. 685-701.

**Wang, Y.; Song, J.; Wang, L.; Van Gool, L.; Hilliges, O.** (2016): Two-stream SR-CNNs for action recognition in videos. *Proceedings of the British Machine Vision Conference*, pp. 1-12.

**Wang, Y.; Long, M.; Wang, J.; Yu, P. S.** (2017): Spatiotemporal pyramid network for video action recognition. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, vol. 6, no. 7, pp. 1529-1538.

**Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D. et al.** (2016): Temporal segment networks: towards good practices for deep action recognition. *European Conference on Computer Vision*, pp. 20-36.

**Wu, Z.; Jiang, Y. G.; Wang, X.; Ye, H.; Xue, X.** (2016): Multi-stream multi-class fusion of deep networks for video classification. *Proceedings of the ACM on Multimedia Conference*, pp. 791-800.

**Xu, H.; Das, A.; Saenko, K.** (2017): R-C3D: region convolutional 3d network for temporal activity detection. *IEEE International Conference on Computer Vision*, pp. 5794-5803.

**Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K.** (2017): Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. *arXiv*: 1712.04851.

**Xiong, Z.; Shen, Q.; Wang, Y.; Zhu, C.** (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.

**Yao, G.; Lei, T.; Zhong, J.** (2019): A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, vol. 118, pp. 14-22

**Zach, C.; Pock, T.; Bischof, H.** (2007): A duality based approach for realtime TV-L 1 optical flow. *Joint Pattern Recognition Symposium*, pp. 214-223.

**Zolfaghari, M.; Oliveira, G. L.; Sedaghat, N.; Brox, T.** (2017): Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. *IEEE International Conference on Computer Vision*, pp. 2923-2932.

**Zhang, D.; Meng, D.; Han, J.** (2017): Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 865-878.

**Zhou, S.; Liang, W.; Li, J.; Kim, J.** (2018): Improved VGG model for road traffic sign recognition. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 11-24.