IET Computer Vision

Special Section: Deep Learning in Computer Vision

IET Journals
The Institution of Engineering and Technology

Survey on deep learning methods in human action recognition

ISSN 1751-9632
Received on 2nd December 2016
Revised 2nd August 2017
Accepted on 15th September 2017
E-First on 24th October 2017
doi: 10.1049/iet-cvi.2016.0355
www.ietdl.org

Maryam Koohzadi[1], Nasrollah Moghadam Charkari[1] ✉

[1]*Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran*

✉ *E-mail: charkari@modares.ac.ir*

Abstract: A study on one of the most important issues in a human action recognition task, i.e. how to create proper data representations with a high-level abstraction from large dimensional noisy video data, is carried out. Most of the recent successful studies in this area are mainly focused on deep learning. Deep learning methods have gained superiority to other approaches in the field of image recognition. In this survey, the authors first investigate the role of deep learning in both image and video processing and recognition. Owing to the variety and plenty of deep learning methods, the authors discuss them in a comparative form. For this purpose, the authors present an analytical framework to classify and to evaluate these methods based on some important functional measures. Furthermore, a categorisation of the state-of-the-art approaches in deep learning for human action recognition is presented. The authors summarise the significantly related works in each approach and discuss their performance.

1 Introduction

Human action recognition is an active field of research in computer vision with numerous applications. The result of various research studies indicates that the success of action recognition problems highly depends on an appropriate feature extraction process. The reason confirms the fact that the appropriate feature extraction would map raw frames into space where the corresponding samples are well distinguishable and reveal the main factors of variations in the frames, related to the problem domain. The hand-craft feature extraction procedures are usually incapable of extracting high-level discriminative information from raw frames due to different issues like high complexity, and noise. Learning the appropriate features from raw data [1–6] is another approach which has been increasingly attracted interest during the last decade. Accordingly, deep learning has been found as one of the successful approaches for feature learning in complex data. Deep learning discovers some effective and valuable patterns in large and complex datasets by building distributed representations. A deep learning is able to increase the capacity of modelling complex data through several representation layers. This leads to remarkable results in the challenging human action recognition task [7–10].

The first achievement of deep learning in computer vision was introduced for image classification problem in 2012. The objective was to employ an efficient method to train a model with 1.2 million high-resolution images to classify new images in 1000 different classes. This approach was based on a deep convolutional neural network (CNN) [11]. After that, other research studies were carried out [12–17] with supervised, unsupervised, reinforcement learning, and other learning paradigms to improve the result. However, the majority of these efforts employed supervised solutions and obtained significant results. On the other hand, unsupervised deep learning methods which have unlimited access to unlabelled data is another interesting solution particularly for the problem of limited training data. Unsupervised learning has been one of the open research problems in deep learning in recent years [6, 18–20].

Although a variety of deep learning techniques in the field of computer vision have been introduced in last decade, there is still a lack of particular structure for evaluating the main approaches in deep learning methods for computer vision. Some interesting approaches in deep learning have been studied in [21, 22]. Furthermore, they were particularly investigated in the area of time-series data [19] and human action recognition [23]. Wu *et al*. [23] introduced recent human action recognition datasets and reviewed important works that applied deep learning methods to these datasets. In this regard, we attempt to provide an analytical framework based on learning models that are wildly used in this area. So, we firstly have studied deep learning approaches proposed in last few years for computer vision both from theoretical and practical points of view, and then evaluated them based on some important functional measures. Then, we categorise best performing deep methods for human action recognition and discuss the most important recent approaches in this area. This categorisation is done, due to a variety of deep learning techniques for modelling temporal dynamics.

The rest of this paper is organised as follows: some background issues are described in Section 2. Also, the importance of using deep learning in computer vision is studied. In Section 3, we explain the main approaches to deep learning in computer vision. The categorised deep learning methods used for human recognition are studied in Section 4. Finally, in Section 5, concluding remarks and future directions for research and development are presented.

2 Deep learning for human action recognition, opportunities and challenges

Automatic human action/activity recognition has been one of the challenging issues in computer vision in recent years. It is of great importance in various applications in artificial intelligence like video surveillance, computer games, robotic, and human computer interactions. In Fig. 1 some examples of five types of actions such as body-motion only, human–object interaction, playing musical instruments, human–human interaction, and sports are depicted.

As shown in Fig. 2, an action recognition system involves three main steps: feature extraction, action representation, and classification. Hence, each step plays an important role to proper recognition rate. Feature extraction and action representation model input raw data to a feature vector. The proper extraction and selection of features would highly affect the classification result.

On the other hand, it is a difficult task to design comprehensive features which provide a desirable performance in the large-scale complex datasets. Over the last decade, action recognition data have been represented with hand-craft feature extraction procedure like HOG, HOF, and iDT [27, 28]. More recently, deep learning methods are introduced for learning high-level representation directly from raw video data. These methods attain a remarkable

Fig. 1 *Action types in UCF101 [24] dataset [25]*

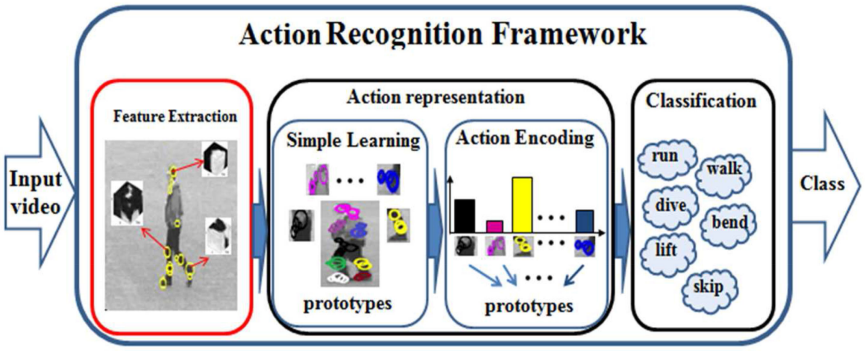

Fig. 2 *General framework for a vision-based action recognition problem. Feature extraction step refers to the extraction primary spatial and temporal features from raw frames. Action representation composes of simple learning and action encoding. In simple learning some pre-processing works are done to combine spatial and temporal features. Action encoding step represents spatial-temporal features into a feature vector. The obtained features are fed into a classifier to distinguish different actions from each other [26]*

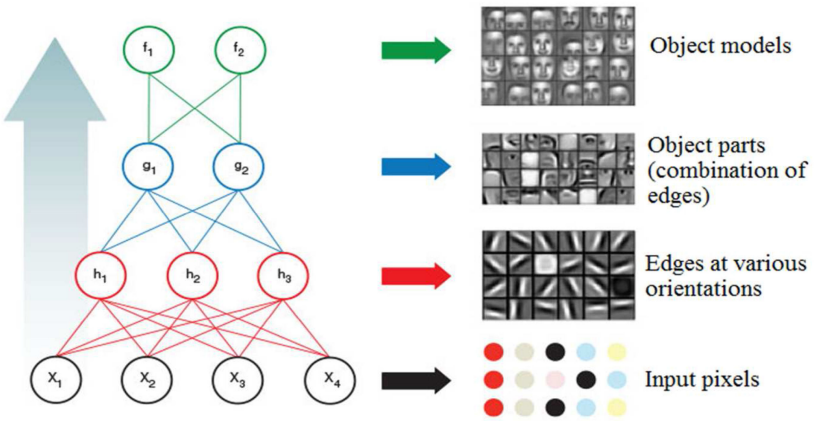

Fig. 3 *Learning several representation layers in deep learning method. Representations learned at each layer is more abstract compared with its previous layer [33]*

success rate on different conventional datasets like UCF101 [24] (94.6%) and HMDB51 [29] (70.3%) [30].

2.1 Challenges of deep learning in computer vision

As a branch of machine learning, deep learning is based on learning several layers of representation for modelling the complex relationships among data. Its hierarchical architecture makes it possible to create high-level concepts from related low-level ones where many models of this group are based on unsupervised learning [12, 31, 32]. Fig. 3 shows several representation layers of a deep model that non-linear learning is done at each layer. As is shown, the output of each layer becomes the input of the next layer where more complex functions are learned in each layer compared with the previous one. The goal of this model is to learn functions through hierarchal layers that are more invariant to changes of worthless information. So, the representation of each layer becomes more abstract compared with its lower layer ones.

In general, building blocks are defined as stacks to shape a deep architecture. These building blocks include different methods of representation learning and data dimension reduction [34] which are capable of building stacks such as deep belief network (DBN), Boltzmann machine (BM), restricted BM (RBM), deep neural network (DNN), auto-encoder (AE), gated AE, CNN, recurrent neural network (RNN), long short-term memory (LSTM), independent subspace analysis (ISA) [18], independent component estimation [35], and slow feature analysis (SFA) [36]. Some of

these methods can significantly benefit action recognition like CNN, RNN, and LSTM.

CNN: this model is inspired by a visual cortex structure that is composed of simple cells and complex cells. Complex cells provide more spatially invariant than simple cells. The simple and complex cells are modelled with convolution operators and pooling operators, respectively. In the convolution layer, the input values are convolved by the receptive field equipped with a given weighted vector as a filter. This process is repeated for all available filters. It reduces the number of required free parameters because of wide weight repetition called as weight sharing. Accordingly, one of the crucial issues is the problem of high dimensionality in the convolved features. Thus, the pooling operator is used to replace each specific region in the convolved feature space with its statistics. The result of applying this operator is invariant to some variations. This model has been found to exhibit excellent performance in various image recognition datasets. The ability to fine-tune end-to-end directly from raw data is the main advantages of CNN [37].

RNN: this neural network models temporal dynamics of the input sequence to the sequence of hidden states, and then hidden states to the outputs. This ability is a result of feedback connections and internal memory. It can be described by the following recurrence equations:

$$\begin{aligned} h_t &= f_{\mathrm{i}}(W_{\mathrm{ih}}x_t + W_{\mathrm{hh}}h_{t-1} + b_{\mathrm{h}}), \\ y_{\mathrm{t}} &= f_{\mathrm{o}}(W_{\mathrm{ho}}h_t + b_{\mathrm{o}}), \end{aligned} \tag{1}$$

where f_{o} and f_{i} are activation functions, x_t is the input, h_t is the hidden state, and y_t is the output at a time. The purpose of this model is to carry out weight sharing over time [38].

LSTM: RNNs perform poorly in learning long-term temporal dynamics. They are not appropriate to deal with the problem of vanishing gradient. LSTM overcomes this problem by considering recurrent forget gates in its architecture. Forget gates are memory units that allow the LSTM to learn from which time step of previously hidden states must be considered in network updating. So, LSTMs have proven successful on very deep learning models while it can perfectly remember important events in long time intervals. Some advantages of LSTM are the ability to fine-tune end-to-end like CNN, modelling varying length sequential data in input and output, and learning in a very deep model without any unsupervised pre-training. These properties make it appropriate for automatic action recognition task [38, 39].

Some of the important benefits of using deep learning are: learning representations with high-level abstraction of complex data [39], creating distributed representation [40], and exploiting unlabelled data [34].

Some of the main challenges of deep learning are: scalability of calculations, non-convex optimisation, disentangling underlying factors of variation, and lack of appropriate representation learning measure [1, 2, 41, 42]. Moreover, there are some other challenges such as the need for an expert to design an appropriate network due to multiple hyper parameters, modelling several factors of variations and the interactions between them [3], large number of local minima [43], performance of Stochastic Gradient Descent (SGD) [44, 45] in many layers, over-fitting problem [46], and the need for a great deal of data.

2.2 Role of deep learning in human action recognition

The history of computer vision goes back to over 50 years. However, there is no satisfactory solution which can be effectively used in different applications such as object recognition in scenes, human action recognition, and behaviour recognition in complex, large-scale data. According to the studies, real environments contributed to a small range of studies due to inefficient feature extraction methods [47–50], while a wide range of applications for human action recognition occurs in real out door environments.

One of the main strategies to cope with action recognition challenges such as learning complex data manifold, providing a data-driven and high-level representation, overcoming noise, and high intra-class diversity is to make use of feature learning methods in deep networks. They obtain impressive results in learning powerful models with high capabilities of representation [19].

Deep learning methods learn a hierarchical, abstract representation directly from the raw data. The learning features obtain remarkable success in many well-known challenging datasets [30]. One of the important advantages of these features is building representations from some simple and local features toward abstract and global ones through layers of the network [41, 51, 52]. The continuity property in deep representation indicates the proximity of the extracted features to the semantic space. Therefore, the gap between observation, representation, and semantic spaces would be decreased.

Consequently, deep learning can resolve the challenge of the constraints on hand-craft features by extracting high-level valuable features [53]. Another challenge of human action recognition is the high variation in gender, size, speed, and gestures of people. Therefore, it is hard to define a unified model which can be employed for all actions compatible with different diversities. As mentioned above, one of the distinct characteristics of the deep learning model is its ability to build representations with a high-level of abstraction from raw input data. These abstract representations are invariant to useless diversity.

Another problem towards reducing the semantic gap is that the manifold of artificial intelligence data like image and video are very complex with many variations. As deep learning models benefit from distributed representation with high reusability, they have high modelling capacity. Motivated by these observations, applying deep learning models in action recognition task can be very efficient [6, 9]. Therefore, the deep learning model has been considered as an appropriate solution to the above-mentioned challenges stated in human action recognition.

3 Main approaches in deep learning for computer vision

In this section, we present some guidelines for choosing and using models for deep learning in computer vision based on machine learning paradigms.

Supervised-deep generative model: if there is a small number of labelled data while sufficient unlabelled data are available, then a high-level of data presentation could be obtained through unsupervised pre-training. This representation can be fine-tuned for a particular task using available labelled data. A generative model that often captures the generation process of y by modelling $p(x|y = +1)$ and $p(x|y = -1)$ can benefit the representation learning. Generative models attempt to understand the basic formation of the individual classes, and thus, carry more information than discriminative models [54–57].

Supervised-deep discriminative model: if the amount of available labelled data is high, the discriminative methods that only focus on the optimisation of boundaries between different classes would be highly efficient. Given an input data point x, a discriminative model computes $p(y|x)$ where the probability of x could be either positive or negative [58–61].

Unsupervised deep model: this approach focuses on the automatic learning from unlabelled data. It can reflect a great deal of information about data structure itself, like data dimension reduction approaches. They are capable of learning new relations and structures in accordance with the data. There has been considerable attention drawn to this approach because very limited labelled data is available [4, 62–64].

Semi-supervised deep model: it refers to the methods in which the objective function includes both reconstruction cost and classification error. For instance, the objective function of the AE model can be composed of discriminative and generative terms at the same time. Semi-supervised learning methods use unlabelled data to modify hypotheses obtained from labelled data [14, 65–67].

Hybrid deep model: this approach includes the methods that are combinations of different deep learning methods or combinations

of deep learning and other machine learning methods. Hence, the representation power of the model increases through the advantages of both models [68–70].

Five different types of the aforementioned approaches in deep learning methods are depicted in Table 1a–c. A brief review of each approach in terms of advantages, challenges, key points, and samples is presented.

So far, the approaches to deep learning used in computer vision have not been compared with each other in a specific structure. The presence of different approaches to deep learning makes the selection of an appropriate approach difficult. However, there are some strategies for evaluating and making a decision on the appropriate approach. These strategies aim to answer the following questions:

i. To what degree the processing time in learning the desired representation is crucial?
ii. How important is the accuracy of the learned representation?
iii. Is learning the representation done on a large dataset?
iv. Is the learned representation only for a particular application or for multiple tasks?
v. Is it important to benefit from the background knowledge for accelerating the learning?
vi. How important is the capability of representation in separating the discriminative factors?

According to the above measures, the aforementioned approaches are compared with each other in Table 2. The hybrid approach was not compared with other approaches because of the dependency of its property on a model which was combined with deep learning methods.

Regarding the cost and accuracy of the mentioned approaches in Table 2, the following issues could be mentioned. In the supervised approaches, the convergence speed increases through pre-training with unlabelled data. The reason is that a good optimisation is achieved earlier [39, 92]. Moreover, in this group, Deep Boltzmann Machine (DBM) method uses an approximate inference based on the mean field which is slower than DBN [93]. Furthermore, accuracy increases in this group with pre-training the unlabelled data since a good regularisation can be achieved [22, 92]. DBM benefits from two-way links. Thus, it is capable of creating a better representation of complex and vague data. The accuracy of AE is better than that of RBM [94]. In the supervised discriminative approach, high accuracy can be achieved if many labelled data exist [7, 8, 11, 79, 95]. Among the available methods in this group, the performance of DSN is better than that of conventional DBN in the case of purely supervised learning [96].

The unsupervised approach has less accuracy than the methods using labelled data. The scaling, multi-task, and generative columns indicate the development features in large datasets, the capability of applying representation to different tasks, and the productive nature of an approach, respectively. The knowledge column is related to the requirements of the approach for exploiting distribution of data, labelled data, and primary knowledge. The discriminatively disentangling column emphasises on the separation of the factors relating to a particular class. From a global perspective, because of the main challenges in this domain like limited dataset and computational cost, a semi-supervised deep model that requires a few labelled samples with a medium

Table 1a Supervised-deep generative and discriminative deep approaches in computer vision

Approach	Benefits	Challenges	Tips and tricks	Samples
supervised-deep generative model	• usability in multitask learning • the generative models with the reconstruction ability makes it easy to find out the captured and lost information at each level of representation [71] • computationally efficient way to use generative modelling to regularise a discriminative system [72, 73] • capture more interesting and wide range of correlation that leads to high-level abstractions and better generalisation [72, 74] • unsupervised pre-training helps prevent over fitting [68]	• learning unrelated factors to the discriminating task • BMs, while having very simple learning algorithm, are very slow to compute in learning [31] • fine-tuning phase of a stochastic gradient descent learning algorithm, which makes it difficult to parallelise across-machines and deploys at large scale [75, 76] • the presence of the partition function causes precise maximum likelihood learning in RBM's intractable. Model selection and complexity control are difficult [34]	• unsupervised pre-training with large amount of unlabelled data. Fine tune the representation with a limited number of labelled data [71, 77] • to identify the joint probability distribution of visible data and relevant class • characterise the high order correlation properties of data [21] • initialise deep models generatively causes better optimisation and reduces generalisation error [68]	DBN[a], DBN (RBM[b]), generative RNN[c], TDNN[d], deep coding networks
supervised-deep discriminative model	• scalable and parallelisable in DSN[e] [78] • powerful in recognition and classification tasks • convex optimisation in DSN [78] • CNN is highly effective in computer vision and image recognition [7, 8, 11] • easy to teach units that do not require the back-propagation and have good speed in ELM[f] [79]	• Due to loss of capacity only models some aspects of the data that are relevant to the discriminative goal [72] • fine-tuning phase of a stochastic gradient descent learning algorithm, which is extremely difficult to parallelise across-machines and impossible at large scale [75, 76]	• requiring large amount of available labelled data • using discriminative criteria to train parameters • use of supervised pre-training • integrating some object functions such as EM-like criteria in SGD to mitigate weak supervision of video data [61]	DSN (tensor, kernel), CNN, ELM, discriminative RNN, DBN-DNN, convolutional DBN, DBN-DNN

[a]Deep belief networks.
[b]Restricted Boltzmann machine.
[c]Recurrent neural network.
[d]Tensor deep neural network.
[e]Deep stacked networks.
[f]Extreme learning machine.

Table 1b Unsupervised and semi-supervised deep approaches in computer vision

Approach	Benefits	Challenges	Tips and tricks	Samples
unsupervised deep model	• robust feature detectors [12] • appropriate to multitask learning since a lot of structures in input data, which is useful for specific tasks, can be discovered [34] • prone to learn common data structures • effective use of plentiful unlabelled data • discover interesting structures, by imposing constraints on the network [80]. • adaptability to learn new structure • effective discriminant features can be learned without any prior assumption [64]	• too slow for large-scale data [81] • lower accuracy than methods that labelled data are used • how to learn useful high level features? Lack of suitable criteria? • if data was independent then this learning task would be very difficult. In the case of structured data, this algorithm can discover some correlations [82, 83]. • include noisy factor and relationship	• structures in unlabelled data can be useful in machine learning tasks • demonstrated with success on dimensionality reduction (coding) tasks [31, 84, 85] • it is a non-linear feature extraction method involving no class labels • most generative model is used • efficiency of unsupervised training of CNN as AEs without [64] • unsupervised learning of temporal dynamics as temporal order verification [86]	SFA, ISA, AE, FA[a], sparse coding, CSAE[b]
semi-supervised deep model	• a natural choice in hard AI tasks with limited labelled data • the highly expressive models can be trained on a small number of labelled data with regularisation effect that provides useful information in data distribution [48]	• how to apply model on larger datasets • generalised and scalable probabilistic approach for semi-supervised learning is still unresolved [87] • determining the semi-supervised strategy for appropriate training	• most generative model • effective generalisation from small labelled data sets to large unlabelled ones • ability to adapt to manifold assumptions through graph neighbourhood • can be partially supervised training where the RBM or auto-encoder gradient is added to the global gradient [14] • simultaneously as a supervised and unsupervised learning	• add a semi-supervised loss to regularise the supervised loss [88] • AE with the objective function of discriminative and generative at the same time can be considered

[a]Factor analysis.

[b]Convolutional sparse auto encoder.

Table 1c Hybrid deep approach in computer vision

Approach	Benefits	Challenges	Tips and tricks	Samples
hybrid deep model	• combine the representational power of deep learning and other learning algorithm [68] • hierarchical Bayesian (HB) models, provide learning from few examples. Compound hierarchical deep architectures try to integrate both characteristics of HB and deep network, and is a full generative model [89] • the dimensionality of both inputs and outputs are variables [22]	• how to integrate deep learning methods with other models to achieve a coherent dynamic model • need for more dynamic models than HMM • joint training of several components is more complex, it is feasible but existing works do not demonstrate significant results [70]	• a combination of deep learning and other methods of machine learning techniques or included of several deep learning methods • separate training of components is more flexible [70]	CD-DNN-HMM[a] [90] convolutional gated restricted Boltzmann machines [91] compound hierarchical-deep models [89] a hybrid of spatial, temporal and fusion deep network for video classification [70]

[a]Context-dependent deep neural network hidden Markov model.

computational cost, is the most attractive paradigm for deep learning.

4 Deep learning approaches to spatio-temporal representation for human action recognition

Despite the advantages and the superiority of deep CNN approaches over the traditional methods in image classification area, there are many unresolved issues in the successful use of deep learning models in human action recognition. When time is added as the third dimension to the 2D images, some problems like excessive computational cost and how to learn temporal dimension emerge. Moreover, limited training samples in the standard dataset for action recognition, for instance, HMDB51 and UCF101, become a major obstacle which leads to over fitting in deep architectures [30, 97, 98]. In following, we introduce three categories of human action recognition using deep learning methods. The overall scheme of these categories is shown in Fig. 4.

Considering how to process temporal dimension, the approaches in action recognition with deep learning methods can be generally divided into three groups; methods that extend spatial

Table 2 Comparison of the most important approaches to training deep models in computer vision

Approach	Computational cost	Accuracy	Scaling	Multitask	Generative	Knowledge	Discriminatively disentangling
supervised-deep generative model	medium	medium	✗	✓	✓	a few labelled samples	✓
supervised-deep discriminative model	low	high	✓ (only for DSN, Tensor Deep Stacking Network (TDSN), Expectation Maximization (EM))	✗	✗	a large number of labelled samples	✓
unsupervised deep model	high	low	✗	✓	✓	a large number of unlabelled samples	✗
semi-supervised deep model	medium	medium	✗	✓	✓	a large number of unlabelled samples and a few labelled samples – background knowledge	✓

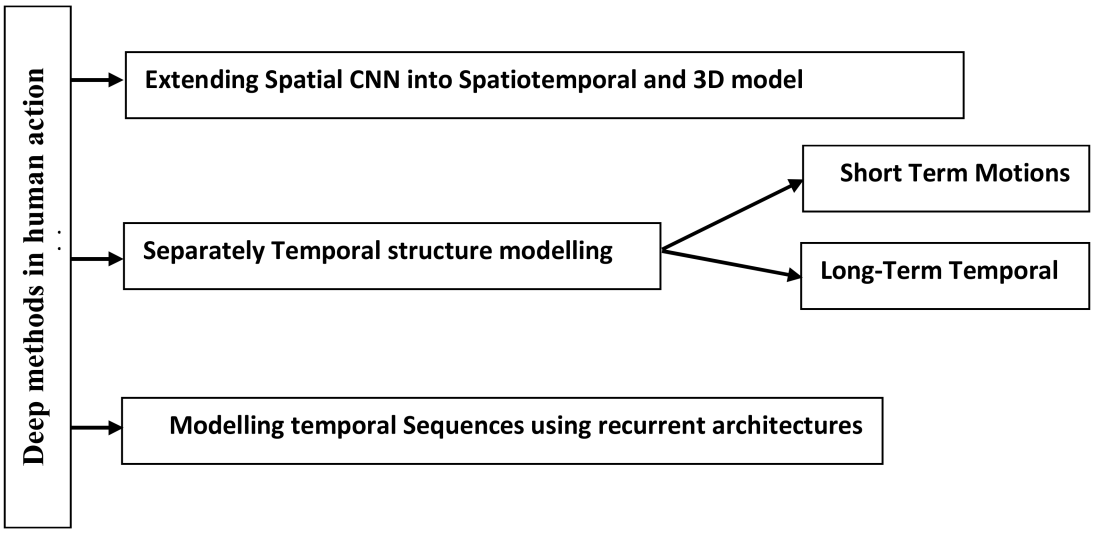

Fig. 4 *Categories in deep human action recognition methods*

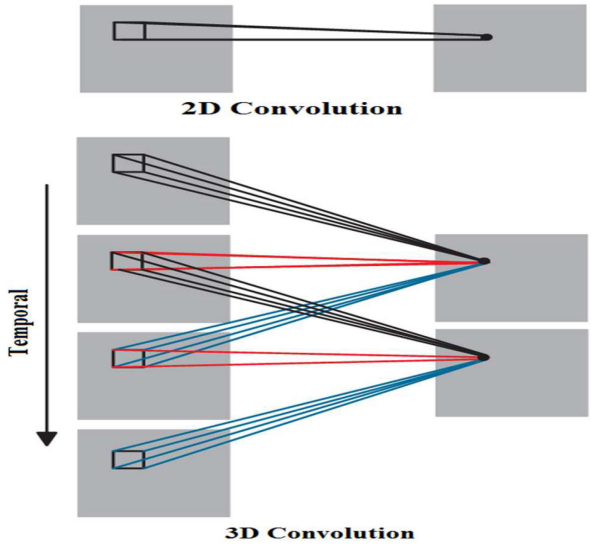

Fig. 5 *Comparison of 2D and 3D convolution [99], similar colours indicate the same weight*

CNNs into the spatiotemporal domain using 3D models. Two other methods consider the intrinsic differences between temporal and spatial domains. There are methods that separately model temporal dimension and methods that make deep sequence model by recurrent architectures.

4.1 Extending spatial CNNs into spatiotemporal and 3D model

This approach takes the human action recognition problem into account as image classification. The extended method makes it possible to solve the problem with the 3D model. For instance, 3D convolutional operations are introduced instead of 2D convolutional ones. 3D CNN is obtained by applying convolution with a 3D kernel. In this structure, feature mapping of a convolution layer is connected to several continuous frames in the previous layer. Therefore, it is possible to capture motion information. In Fig. 5, 2D and 3D convolutions are compared. By applying separate kernels to similar frame locations in previous layers and using several 3D convolutional actions, different weights would be obtained. 3D convolution is invariant to spatial translations in time. The approach has achieved limited success due to some issues. Some major problems in using 3D convolution kernel for human action recognition are high computational cost and the need for a large amount of training data. In the following, some related works are introduced. It is important to point out that these methods focus on a short window length in video frames for the feature extraction phase.

A convolutional gated RBM model for learning latent representations for a sequence of images using pairs of successive images was proposed in [91]. Motion-sensitive features like optical flow are extracted in this model. Human action recognition based on the 3D CNN model for the scenes of noisy and cluttered backgrounds, occlusion, and different view spots was presented in [99]. Their common pyramidal architecture is supervised and requires a large number of labelled data. The experimental results indicated the efficiency and validity of CNN in a real environment.

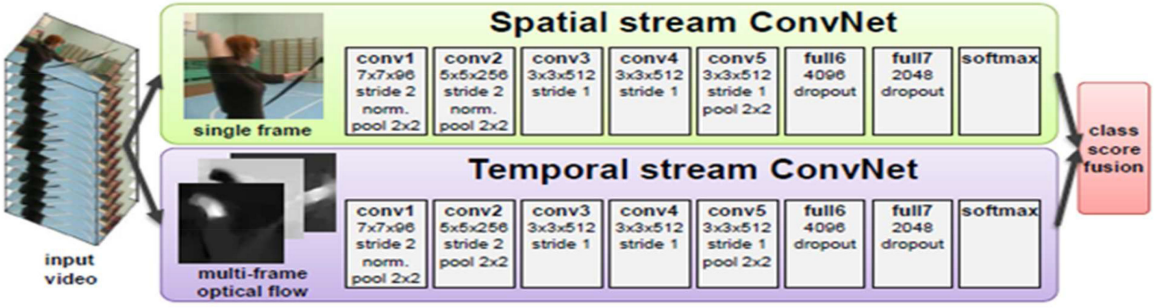

Fig. 6 *General schema of the proposed method in human action recognition [8]*

Karpathy *et al*. [7] presented one of the interesting works in extended CNN-based approaches on large-scale datasets. The different models of information combinations across the temporal domain have been carried out in this work. As a result, a slow fusion model has been found to perform better than early and late fusion models. The key difference between slow fusion models with other models is the use of 3D convolution kernel and averaging pooling in the first three layers. This approach performs convolution on 12-frame clips. Also, the process of consecutive frames in two different resolutions of context stream and fovea stream has been done for speeding up CNN. Baccouche *et al*. [47] proposed a fully-automated deep model. Spatial-temporal features are initially learned automatically using 3D CNN; then a RNN is learned to classify feature sequences at each time step.

Like [7, 100] employed a 3D CNN on large-scale video datasets. It demonstrates that only 3D convolution preserves the temporal cues. They suggested some basis in 3D CNN design to improve performance that is using a small $3 \times 3 \times 3$ convolution kernels which could be the best size of the kernel, and a homogeneous architecture as a suitable model for learning spatio-temporal features. They also proposed the C3D architecture that learns spatio-temporal features on a limited interval of 16 frames. It could learn generic features. So, it achieved proper performance on various tasks without fine tuning for each task. These features attained remarkable success on six different benchmarks. The model conceptually is very simple, easy to train and deploy. Sun *et al*. [101] proposed a factorised spatio-temporal CNNs in order to decompose 3D convolutional kernels to some 2D spatial kernel followed by 1D temporal kernels. This model considers the spatial domain in lower layers, then combines spatial information across temporal dimension in upper layers. One of the advantages of this method is reducing the number of network parameters from $n_x n_y n_t$ to $n_x n_y + n_t$. Therefore, the problem of high kernel complexity would be reduced in order $O(n_t)$. Subsequently, it mitigates the problem of insufficient training data. In [102], a manifold learning concept with a C3D model was used to mitigate the problems of intra-class variations as well as over-fitting. The spatio-temporal manifold constrain is included in the loss function of C3D as a regularisation term like image recognition, and fine-tuned it. Contrary to other methods that leverage from a few video frames, [10] extends local representation of the 3D CNN to global action representation with considering multiple resolutions in the temporal domain. Long-term temporal convolutions (LTC) explore 3D CNNs with significantly longer temporal convolutions to capture action representation at their full temporal scale. Zhu *et al*. [61] proposed a model to confront a weak supervision problem that assigns a video-level label to a limited number of consecutive frames. For this purpose, it integrated EM-like loops in SGD training and optimised both simultaneously.

4.2 Temporal structure modelling

The approaches in this category operate on temporal dimension in a separate way from spatial one. It decreases the calculations since only the difference between successive frames in the form of optical flow is sufficient to detect motion. However, all the image pixels are required to further identify the object. Due to the window length of frames being processed, we divided these methods into two categories i.e. short-term motions and long-term motions.

4.2.1 Short-term motions: Methods in this category generally rely on dense temporal sampling with a pre-defined sampling interval. So, these methods usually detect local changes within a small time window. In [48], spatial, temporal, and audio low-level features were extracted, subsequently the DBN method was employed to obtain the compression vector of the features. Two-stream CNNs that are composed of both spatial and temporal deep models were discussed in [8]. The spatial net is pre-trained on ImageNet dataset. Furthermore, the temporal net is trained on optical flow features. The number of layers in this model is relatively small as shown in Fig. 6. Thus, in the case of insufficient training data, it would be a precise deep learning method for action recognition tasks. Wang *et al*. [103] show how the number of layers in a network can benefit action recognition.

In line with [8], some solutions like fusing, spatial-temporal pooling or encoding of CNN based features have been studied. Feichtenhofer *et al*. [104] investigated different approaches to fuse the information of spatial and temporal networks. As a consequence of this work, it was found that a convolution layer could perform efficiently as a fusion layer. Accordingly, a new CNN architecture for fusion of spatial and temporal information at several levels of granularity has been proposed. It improves the accuracy without increasing the number of parameters. The method emphasises on fusing the spatial and temporal features at upper layers of CNN. Wang *et al*. [105] proposed a trajectory-pooled deep convolutional descriptor (TDD). The method exploits from hand-crafted trajectories for pooling convolutional feature maps obtained from two-stream CNNs. It initially detects some improved trajectory points, and then local CNN features are passed through the spatio-temporal tubes based on the trajectories to build TDDs. The results of experiments indicate that TDDs outperform other hand-crafted feature approaches. Misra *et al*. [86] proposed an unsupervised sequential verification method that can be used in human action recognition. In contrast to [8] that captures spatial and temporal information concurrently through two CNNs. In this model, capturing temporal information is performed using the verification of temporal order after that underlying spatial representation is extracted by CNN. A limited number of spatial features are selected and fine-tuned for sequence verification task. Since the final features must be ground-trothed both spatially and temporally, they capture object transformations and local motions.

4.2.2 Long-term motions: Generally, the methods in this group focus on capturing long-term motion patterns associated with certain actions, using the CNN. Therefore, only a small number of models might successfully act at the video level because of high computational cost. Wang *et al*. [9] proposed a precise CNN architecture for long-range temporal structure modelling using the temporal segment network (TSN). TSN is the first framework for end-to-end temporal structure modelling at the whole video-level. Firstly, sparse sampling from a long video sequence is done to make short snippets. Then, the two-stream network is employed on snippets to capture both spatial and temporal representations. At the second step, a segmental structure is performed to aggregate information from representations of snippets like consensus

function in ensemble methods. Feichtenhofer *et al*. [30] presented a CNN in the video domain that is a stacking of transformed temporal filters. These filters are made by 1×1 convolutional dimensionality mapping filters in ResNets as temporal filters. They exploit from residual connections in the basic two-stream method to facilitate training of very deep models. Consequently, their spatial CNNs have functionality like spatio-temporal models. In [106], dynamic images represent a video with a single RGB image. This image is made by applying rank pooling on the video frames. Although this idea is simple, it creates a great ability to take advantages from the existing CNN models on the spatial domain and large image datasets.

4.3 Modelling temporal sequences using recurrent architectures

Recently, several works have attempted to model long-range temporal structures. These methods do not take advantages of the entire video as the input because of high computational cost. They usually process sequences with pre-defined window length ranging from 64 to 120 frames. This category mainly leverages from RNN or LSTM techniques. In general, some models in this approach train both spatial and temporal deep models concurrently and others perform temporal dimension learning over underlying spatial features obtain from trained CNN. Du *et al*. [49] proposed an end-to-end hierarchical RNN for skeleton-based action recognition. In this study, the action representation is formed by a hierarchical model. At the first level of this architecture, five low-level body parts are modelled by bidirectional RNNs. Then, these representations are concatenated to make high-level body parts at upper layers. They also use LSTM neurons at the last layer to mitigate the vanishing gradient problem. Donahue *et al*. [38] developed a novel long-term recurrent convolutional architecture benefited from the complex concept when a limited number of training data is available. This model is end-to-end trainable architecture. It learns spatial and temporal domains with deep CNN and a deep sequence model consists of a long-term RNN. The model is in contrast to other existing RNN-based 3D action recognition methods that only model temporal domain. In [107], the relationship of each joint of 3D skeleton data in both spatial and temporal domain is considered simultaneously. Contrary to most methods that concatenate features of the joints to present high-level body parts, this model is a 2D LSTM network that applied deep recurrent learning on both spatial and temporal dimension to encode the spatio-temporal context. Sharma *et al*. [97] proposed a spatially and temporally deep model with long- and short-term memory (LSTM) units as well as a soft attention model. The soft attention model leverages some important parts in each frame and attaches different scores to them depending on a particular task. This model concluded that the proposed dynamically pooling performs better than other pooling methods in CNN. Li *et al*. [108] present a multi-stream model to span different granularities of actions from a single frame to the entire video. Each stream is made by 2D or 3D CNNs. LSTMs are used to further model temporal dynamics of streams. The final decision is made by aggregating stream-level predictions. Srivastava *et al*. [109] proposed a deep encoder/decoder using the LSTM building block to learn a representation of video frames. This model encodes input sequence into a fixed length representation. It works in an unsupervised paradigm and has only access to un-labelled video frames.

In the following, those methods which do not model the spatial domain and temporal dimension concurrently will be discussed. Yue-Hei Ng *et al*. [110] employed several convolutional temporal feature pooling architectures and RNN based on the LSTM to combine the temporal domain over longer window of video frames. The performance of the method outperformed previous results on the sports dataset [7]. The max length window reaches 120 frames. As a result, they concluded that max pooling of the last convolutional layers across frames performs better than slow, local, or late pooling, as well as temporal convolution. They also investigated a recurrent network with LSTM for sequence modelling which receives the underlying CNN features as the input. However, the recurrent network did not perform as well as maximum pooling of convolutional features. Lev *et al*. [111] introduced a novel approach for sequence representation using a RNN and Fisher vector. At the first step, the Visual Geometry Group (VGG) [112] is applied to extract spatial features from the frames of the video. Furthermore, the RNN is trained to predict the sequence order of the extracted features. Consequently, the back propagation mechanism in RNN provides the gradients which are required for the computation of fisher vector. Yue-Hei Ng et al. [113] added a differential gating scheme inside the LSTM neural network. This model is capable of discovering salient motion patterns by computing different-orders of derivative of state (DoS). The DoS of the LSTM indicates changes of the input sequence. The differential LSTM receives concatenation of the input features for each frame as the input. Escorcia *et al*. [114] proposed a model that takes advantage of deep learning models and memory cells to retrieve temporal segments from videos, which are likely to compose actions. It encodes the sequence of spatial features extracted with a deep CNN using LSTM, and predicts the location of an action in the input stream.

4.4 Comparison between advanced models

In the previous section, we presented a short review of several approaches in deep learning for an action recognition problem. In general, all the methods in this area confront with some obstacles like limited training samples and high computational cost. Therefore, very deep models in action recognition usually fall into an over fitting problem due to the limited training samples. To cope with these obstacles, some solutions like cross-modality pre-training, regularisation, and enhanced data augmentation have been introduced [9]. In extending spatial CNNs into spatio-temporal and 3D model approaches, 3D CNN needs much more training data due to the extensive complexity of training 3D kernel. So, Ji *et al*. [99] did not attain remarkable results even in the presence of simple action recognition samples on the KTH dataset. In the UCF-101, we observed that two-stream CNN [8] outperforms other reported studies in [7, 101]. Moreover, Karpathy *et al*. [7] gain similar performance by a purely spatial network. Methods in a separately temporal structure modelling approach specifically cope with the problem of excessive computational cost in a long video. Therefore, these methods have to break the video into very small clips and assign video labels to them. Consequently, some new challenges like the risk of missing important information, limited access to the temporal context, and weak supervision remain unresolved. On the other hand, recurrent architectures are used to model temporal dynamics using LSTM units. Although these methods are efficient in very deep learning models, they also focus on the pre-defined window length ranging from 64 to 120 frames. Thus, they face similar challenges as mentioned above. A comparison between the accuracy of different methods on two conventional benchmarks is depicted in Table 3. As seen in this table, the highest accuracy belongs to the long-range temporal approach that is based on the CNN.

The results indicate that purely deep learning methods are inferior to those obtained by a combination of hand-craft and deep learning features in the action recognition domain. If we consider only purely deep methods, the best performance belongs to the approaches that access to the video-level temporal context. These observations confirm the effectiveness of temporal representation. Furthermore, as already mentioned, clip-level information is incomplete. Making a video-level decision based on the aggregation of the clip level and possibly imperfect decision is likely to be suboptimal. Even this aggregation could be done with a perfect method like LSTM, as mentioned in [38]. Moreover, this observation confirms the superiority of homogeneous architectures of CNN.

5 Conclusion

Learning the representation directly from raw video data is an active research area in computer vision. In this study, deep learning as the most powerful solution for creating a high-level conceptual

Table 3 Accuracy of some recent deep learning methods in two popular datasets

Approach	Method	UCF101	HMDB51
extending spatial CNN into spatiotemporal and 3D model	C3D [100]	90.5	—
	FSTCN [101]	88.1	59.1
	LTC [10]	92.7	67.2
short-term motions	TDD [105]	91.5	65.9
	two-stream (Support Vector Machine (SVM) fusion) [104]	88.0	59.4
	two-stream (conv. fusion) [104]	92.5	65.4
long-term motions	TSN [9]	94.2	69.4
	ResNet [30]	93.4	66.4
	dynamic image network [106]	89.1	65.2
	ST-ResNet* + IDT [30]	94.6	70.3
modelling temporal sequences using recurrent architectures	AE/DE LSTM [109]	75.8	44.1
	RNN fisher vectors [111]	94.08	67.71

representation of complex data with appropriate abstraction has been studied.

There are various methods available in deep learning that care should be taken to a proper selection of an appropriate one. In this regard, it is crucial to categorise and evaluate these methods in a structural view. Some related review articles have been presented [19, 23, 37]; however, these papers have not investigated specifically different action recognition approaches in temporal dimension modelling. We have analysed and categorised methods in this area based on modelling temporal dynamics in a comparative manner. In this study, because of the importance of image classification in action recognition, the most interesting approaches to deep learning in computer vision have been analysed in a framework for learning the appropriate representation. They also were compared with proper functional measures. Then, we categorise the state-of-the-art deep learning methods for action recognition, based on processing temporal information. The reason for this type of categorisation is that most recent successful works have been conducted on separation of temporal information.

The proposed structural issues of deep learning methods in action recognition and computer vision tasks are comprehensive and include main benefits and challenges of each approach.

6 References

[1] Bengio, Y.: 'Deep learning of representations: looking forward'. Statistical Language and Speech Processing, 2013, pp. 1–37

[2] Bengio, Y., Courville, A., Vincent, P.: 'Representation learning: a review and new perspectives', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (8), pp. 1798–1828

[3] Reed, S., Sohn, K., Zhang, Y., *et al.*: 'Learning to disentangle factors of variation with manifold interaction'. Proc. 31st Int. Conf. on Machine Learning (ICML-14), 2014

[4] Coates, A., Ng, A.Y., Lee, H.: 'An analysis of single-layer networks in unsupervised feature learning'. Int. Conf. on Artificial Intelligence and Statistics, 2011

[5] Bengio, Y., Courville, A.C., Vincent, P.: 'Unsupervised feature learning and deep learning: a review and new perspectives', *CoRR*, 2012, **1**, abs/1206.5538

[6] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444

[7] Karpathy, A., Toderici, G., Shetty, S., *et al.*: 'Large-scale video classification with convolutional neural networks'. 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014

[8] Simonyan, K., Zisserman, A.: 'Two-stream convolutional networks for action recognition in videos'. Advances in Neural Information Processing Systems, 2014

[9] Wang, L., Xiong, Y., Wang, Z., *et al.*: 'Temporal segment networks: towards good practices for deep action recognition'. European Conf. on Computer Vision, 2016

[10] Varol, G., Laptev, I., Schmid, C.: 'Long-term temporal convolutions for action recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, doi: 10.1109/TPAMI.2017.2712608

[11] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, 2012

[12] Le, Q.V.: 'Building high-level features using large scale unsupervised learning'. 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2013

[13] Peng, X., Zou, C., Qiao, Y., *et al.*: 'Action recognition with stacked fisher vectors'. Computer Vision–ECCV 2014, 2014, pp. 581–595

[14] Rifai, S., Bengio, Y., Courville, , *et al.*: 'Disentangling factors of variation for facial expression recognition'. Computer Vision–ECCV 2012, 2012, pp. 808–822

[15] Ciresan, D., Meier, U., Schmidhuber, J.: 'Multi-column deep neural networks for image classification'. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012

[16] Zeiler, M.D.: '*Hierarchical convolutional deep learning in computer vision*' (New York University, 2013)

[17] Mnih, V., Kavukcuoglu, K., Silver, D., *et al.*: 'Human-level control through deep reinforcement learning', *Nature*, 2015, **518**, (7540), pp. 529–533

[18] Chen, G., Clarke, D., Giuliani, M., *et al.*: 'Combining unsupervised learning and discrimination for 3D action recognition', *Signal Process.*, 2015, **110**, pp. 67–81

[19] Längkvist, M., Karlsson, L., Loutfi, A.: 'A review of unsupervised feature learning and deep learning for time-series modeling', *Pattern Recognit. Lett.*, 2014, **42**, pp. 11–24

[20] Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., *et al.*: 'Discriminative unsupervised feature learning with convolutional neural networks'. Advances in Neural Information Processing Systems, 2014

[21] Deng, L.: 'Three classes of deep learning architectures and their applications: a tutorial survey', *APSIPA Trans. Signal Inf. Process.*, 2012, 10.1017/atsip.2013.9. Available online: https://www.microsoft.com/en-us/research/publication/three-classes-of-deep-learning-architectures-and-their-applications-a-tutorial-survey/ (accessed 7 July 2016)

[22] Deng, L., Yu, D.: 'Deep learning: methods and applications', *Found. Trends Signal Process.*, 2014, **7**, (3-4), pp. 197–387

[23] Wu, D., Sharma, N., Blumenstein, M.: 'Recent advances in video-based human action recognition using deep learning: a review'. 2017 Int. Joint Conf. on Neural Networks (IJCNN), 2017

[24] Soomro, K., Zamir, A.R., Shah, M.: 'UCF101: a dataset of 101 human actions classes from videos in the wild', arXiv preprint arXiv: 1212.0402, 2012

[25] Cho, H., Lee, H., Jiang, Z.: 'Evaluation of LC-KSVD on UCF101 action dataset'. THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes, 2013

[26] Shabani, A.H., Clausi, D.A., Zelek, J.S.: 'Evaluation of local spatio-temporal salient feature detectors for human action recognition'. 2012 Ninth Conf. on Computer and Robot Vision (CRV), 2012

[27] Wang, H., Kläser, A., Schmid, C., *et al.*: 'Dense trajectories and motion boundary descriptors for action recognition', *Int. J. Comput. Vis.*, 2013, **103**, (1), pp. 60–79

[28] Wang, H., Schmid, C.: 'Action recognition with improved trajectories'. Proc. IEEE Int. Conf. on Computer Vision, 2013

[29] Kuehne, H., Jhuang, H., Stiefelhagen, R., *et al.*: 'HMDB51: a large video database for human motion recognition'. High Performance Computing in Science and Engineering '12, 2013, pp. 571–582

[30] Feichtenhofer, C., Pinz, A., Wildes, R.: 'Spatiotemporal residual networks for video action recognition'. Advances in Neural Information Processing Systems, 2016

[31] Deng, L.: 'A tutorial survey of architectures, algorithms, and applications for deep learning', *APSIPA Trans. Signal Inf. Process.*, 2014, **3, p**. e2

[32] Dosovitskiy, A., Fischer, P., Springenberg, J.T., *et al.*: 'Discriminative unsupervised feature learning with exemplar convolutional neural networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (9), pp. 1734–1747

[33] Vogt, R.L., Chinn, K.B., Kotta, P., *et al.*: '*Science & technology review June 2016*' (Lawrence Livermore National Lab. (LLNL), Livermore, CA, 2016)

[34] Salakhutdinov, R.: '*Learning deep generative models*' (University of Toronto, 2009)

[35] Dinh, L., Krueger, D., Bengio, Y.: 'NICE: non-linear independent components estimation', arXiv preprint arXiv: 1410.8516, 2014

[36] Sun, L., Jia, K., Chan, T.H., *et al.*: 'DL-SFA: deeply-learned slow feature analysis for action recognition'. 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014

[37] Srinivas, S., Sarvadevabhatla, R.K., Mopuri, K.R., *et al.*: 'A taxonomy of deep convolutional neural nets for computer vision', arXiv preprint arXiv:1601.06615, 2016

[38] Donahue, J., Anne Hendricks, L., Guadarrama, S., *et al.*: 'Long-term recurrent convolutional networks for visual recognition and description'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015

[39] Schmidhuber, J.: 'Deep learning in neural networks: an overview', *Neural Netw.*, 2015, **61**, pp. 85–117

[40] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., *et al.*: 'Deep learning applications and challenges in big data analytics', *J. Big Data*, 2015, **2**, (1), p. 1

[41] Bengio, Y.: 'Learning deep architectures for AI', *Found. Trends Mach. Learn.*, 2009, **2**, (1), pp. 1–127

[42] Erhan, D., Manzagol, P.A., Bengio, Y., *et al.*: 'The difficulty of training deep architectures and the effect of unsupervised pre-training'. Int. Conf. on Artificial Intelligence and Statistics, 2009

[43] Kawaguchi, K.: 'Deep learning without poor local minima'. Advances in Neural Information Processing Systems, 2016

[44] Dean, J., Corrado, G., Monga, R., *et al.*: 'Large scale distributed deep networks'. Advances in Neural Information Processing Systems, 2012

[45] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016

[46] Srivastava, N., Hinton, G.E., Krizhevsky, A., *et al.*: 'Dropout: a simple way to prevent neural networks from overfitting', *J. Mach. Learn. Res.*, 2014, **15**, (1), pp. 1929–1958

[47] Baccouche, M., Mamalet, F., Wolf, C., *et al.*: 'Sequential deep learning for human action recognition'. Human Behavior Understanding, 2011, pp. 29–39

[48] Yang, Y.: '*Learning hierarchical representations for video analysis using deep learning*' (University of Central Florida, Orlando, FL, 2013)

[49] Du, Y., Wang, W., Wang, L.: 'Hierarchical recurrent neural network for skeleton based action recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015

[50] Rahmani, H., Mian, A., Shah, M.: 'Learning a deep model for human action recognition from novel viewpoints', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **abs/1602.00828**, doi: 10.1103/PhysRevD.94.065007

[51] Higgins, I., Matthey, L., Glorot, X., *et al.*: 'Early visual concept learning with unsupervised deep learning', arXiv preprint arXiv:1606.05579, 2016

[52] Sun, Y., Wang, X., Tang, X.: 'Hybrid deep learning for face verification', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (10), pp. 1997–2009

[53] Schulz, H., Behnke, S.: 'Deep learning', *KI-Künstliche Intelligenz*, 2012, **26**, (4), pp. 357–363

[54] Koohzadi, M., Keyvanpour, M.R.: 'An analytical framework for event mining in video data', *Artif. Intell. Rev.*, 2014, **41**, (3), pp. 401–413

[55] Hinton, G.E., Salakhutdinov, R.R.: 'Reducing the dimensionality of data with neural networks', *Science*, 2006, **313**, (5786), pp. 504–507

[56] Nair, V., Hinton, G.E.: '3D object recognition with deep belief nets'. Advances in Neural Information Processing Systems, 2009

[57] Rezende, D., Danihelka, I., Gregor, K., *et al.*: 'One-shot generalization in deep generative models'. Int. Conf. on Machine Learning, 2016

[58] Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., *et al.*: 'Learning convolutional feature hierarchies for visual recognition'. Advances in Neural Information Processing Systems, 2010

[59] Zeiler, M.D., Fergus, R.: 'Stochastic pooling for regularization of deep convolutional neural networks', arXiv preprint arXiv:1301.3557, 2013

[60] Wen, Y., Zhang, K., Li, Z., *et al.*: 'A discriminative feature learning approach for deep face recognition'. European Conf. on Computer Vision, 2016

[61] Zhu, W., Hu, J., Sun, G., *et al.*: 'A key volume mining deep framework for action recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016

[62] Ngiam, J., Chen, Z., Koh, P.W., *et al.*: 'Learning deep energy models'. Proc. 28th Int. Conf. on Machine Learning (ICML-11), 2011

[63] Le, Q.V., Zou, W.Y., Yeung, S.Y., *et al.*: 'Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis'. 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011

[64] Kallenberg, M., Petersen, K., Nielsen, M., *et al.*: 'Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring', *IEEE Trans. Med. Imaging*, 2016, **35**, (5), pp. 1322–1331

[65] Yang, Y., Shu, G., Shah, M.: 'Semi-supervised learning of feature hierarchies for object detection in a video'. 2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013

[66] Zhu, X.: 'Semi-supervised learning literature survey', *Computer Science, University of Wisconsin-Madison*, 2006, **2**, (3), p. 4

[67] Gan, J., Li, L., Zhai, Y., *et al.*: 'Deep self-taught learning for facial beauty prediction', *Neurocomputing*, 2014, **144**, pp. 295–303

[68] Dahl, G.E., Yu, D., Deng, L., *et al.*: 'Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition', *IEEE Trans. Audio, Speech Lang. Process.*, 2012, **20**, (1), pp. 30–42

[69] Sun, Y., Wang, X., Tang, X.: 'Hybrid deep learning for face verification'. 2013 IEEE Int. Conf. on Computer Vision (ICCV), 2013

[70] Wu, Z., Wang, X., Jiang, Y.G., *et al.*: 'Modeling spatial-temporal clues in a hybrid deep learning framework for video classification'. Proc. 23rd ACM Int. Conf. on Multimedia, 2015

[71] Erhan, D., Bengio, Y., Courville, A., *et al.*: 'Why does unsupervised pre-training help deep learning?', *J. Mach. Learn. Res.*, 2010, **11**, pp. 625–660

[72] Ranzato, M.A., Susskind, J., Mnih, V., *et al.*: 'On deep generative models with applications to recognition'. 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011

[73] Jia, X., Li, K., Li, X., *et al.*: 'A novel semi-supervised deep learning framework for affective state recognition on EEG signals'. 2014 IEEE Int. Conf. on Bioinformatics and Bioengineering (BIBE), 2014

[74] Bornschein, J., Bengio, Y.: 'Reweighted wake-sleep', arXiv preprint arXiv:1406.2751, 2014

[75] Deng, L.: 'An overview of deep-structured learning for information processing'. Proc. Asian-Pacific Signal & Information Processing Annual Summit and Conf. (APSIPA-ASC), 2011

[76] Deng, L., Yu, D.: '*Deep learning for signal and information processing*' (2013), Microsoft Research Monograph Microsoft Research One Microsoft Way Redmond, WA 98052, https://www.microsoft.com/en-us/research/publication/deep-learning-for-signal-and-information-processing/

[77] Hinton, G.E.: 'Learning multiple layers of representation', *Trends Cogn. Sci.*, 2007, **11**, (10), pp. 428–434

[78] Hutchinson, B., Deng, L., Yu, D.: 'Tensor deep stacking networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (8), pp. 1944–1957

[79] Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: 'Extreme learning machine: theory and applications', *Neurocomputing*, 2006, **70**, (1), pp. 489–501

[80] Fukushima, K.: 'Improved generalization ability using constrained neural network architectures'. Proc. 1993 Int. Joint Conf. on Neural Networks, 1993. IJCNN'93-Nagoya, 1993

[81] Raina, R., Madhavan, A., Ng, A.Y.: 'Large-scale deep unsupervised learning using graphics processors'. Proc. 26th Annual International Conf. on Machine Learning, 2009

[82] Ng, A.: 'Sparse autoencoder', CS294A Lecture notes, 72, 2011

[83] Ouyang, Y., Liu, W., Rong, W., *et al.*: 'Autoencoder-based collaborative filtering'. Int. Conf. on Neural Information Processing, 2014

[84] Wang, W., Huang, Y., Wang, Y., *et al.*: 'Generalized autoencoder: a neural network framework for dimensionality reduction'. 2014 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2014

[85] Wang, M., Sha, F., Jordan, M.I.: 'Unsupervised kernel dimension reduction'. Advances in Neural Information Processing Systems, 2010

[86] Misra, I., Zitnick, C.L., Hebert, M.: 'Shuffle and learn: unsupervised learning using temporal order verification'. European Conf. on Computer Vision, 2016

[87] Kingma, D.P., Mohamed, S., Rezende, D.J., *et al.*: 'Semi-supervised learning with deep generative models'. Advances in Neural Information Processing Systems, 2014

[88] Weston, J., Ratle, F., Mobahi, H., *et al.*: 'Deep learning via semi-supervised embedding', in '*Neural networks: tricks of the trade*' (Springer, Berlin Heidelberg, 2012), pp. 639–655

[89] Salakhutdinov, R., Tenenbaum, J.B., Torralba, A.: 'Learning with hierarchical-deep models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (8), pp. 1958–1971

[90] Miao, Y., Metze, F.: 'Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training', 2013

[91] Taylor, G.W., Fergus, R., LeCun, Y., *et al.*: 'Convolutional learning of spatio-temporal features'. Computer Vision–ECCV 2010, 2010, pp. 140–153

[92] Bengio, Y., Lamblin, P., Popovici, D., *et al.*: 'Greedy layer-wise training of deep networks', *Advances in Neural Information Processing Systems*, 2007, **19**, p. 153

[93] Khanna, R., Awad, M.: '*Efficient learning machines: theories, concepts, and applications for engineers and system designers*' (Apress, 2015)

[94] Tan, C.C., Eswaran, C.: 'Performance comparison of three types of autoencoder neural networks'. Second Asia Int. Conf. on Modeling & Simulation, 2008 (AICMS 08), 2008

[95] Charalampous, K., Gasteratos, A.: 'A tensor-based deep learning framework', *Image Vis. Comput.*, 2014, **32**, (11), pp. 916–929

[96] Deng, L., Yu, D., Platt, J.: 'Scalable stacking and learning for building deep architectures'. 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2012

[97] Sharma, S., Kiros, R., Salakhutdinov, R.: 'Action recognition using visual attention', arXiv preprint arXiv:1511.04119, 2015

[98] Park, E., Han, X., Berg, T.L., *et al.*: 'Combining multiple sources of knowledge in deep CNNs for action recognition'. 2016 IEEE Winter Conf. on Applications of Computer Vision (WACV), 2016

[99] Ji, S., Xu, W., Yang, M., *et al.*: '3D convolutional neural networks for human action recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (1), pp. 221–231

[100] Tran, D., Bourdev, L., Fergus, R., *et al.*: 'Learning spatiotemporal features with 3d convolutional networks'. Proc. IEEE Int. Conf. on Computer Vision, 2015

[101] Sun, L., Jia, K., Yeung, D.Y., *et al.*: 'Human action recognition using factorized spatio-temporal convolutional networks'. Proc. IEEE Int. Conf. on Computer Vision, 2015

[102] Li, C., Chen, C., Zhang, B., *et al.*: 'Deep spatio-temporal manifold network for action recognition', arXiv preprint arXiv:1705.03148, 2017

[103] Wang, L., Xiong, Y., Wang, Z., *et al.*: 'Towards good practices for very deep two-stream ConvNets', arXiv preprint arXiv:1507.02159, 2015

[104] Feichtenhofer, C., Pinz, A., Zisserman, A.: 'Convolutional two-stream network fusion for video action recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016

[105] Wang, L., Qiao, Y., Tang, X.: 'Action recognition with trajectory-pooled deep-convolutional descriptors'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015

[106] Bilen, H., Fernando, B., Gavves, E., *et al.*: 'Dynamic image networks for action recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016

[107] Liu, J., Shahroudy, A., Xu, D., *et al.*: 'Spatio-temporal LSTM with trust gates for 3D human action recognition'. European Conf. on Computer Vision, 2016

[108] Li, Q, *et al.*: 'Action recognition by learning deep multi-granular spatio-temporal video representation'. Proc. 2016 ACM Int. Conf. on Multimedia Retrieval, 2016

[109] Srivastava, N., Mansimov, E., Salakhudinov, R.: 'Unsupervised learning of video representations using LSTMS'. Int. Conf. on Machine Learning, 2015

[110] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., *et al.*: 'Beyond short snippets: deep networks for video classification'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015

[111] Lev, G., Sadeh, G., Klein, B., *et al.*: 'RNN fisher vectors for action recognition and image annotation'. European Conf. on Computer Vision, 2016

[112] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv:1409.1556, 2014

[113] Veeriah, V., Zhuang, N., Qi, G.-J.: 'Differential recurrent neural networks for action recognition'. Proc. IEEE Int. Conf. on Computer Vision, 2015

[114] Escorcia, V., Heilbron, F.C., Niebles, J.C., *et al.*: 'DAPS: deep action proposals for action understanding'. European Conf. on Computer Vision, 2016