Multimed Tools Appl (2018) 77:20415–20453
https://doi.org/10.1007/s11042-017-5438-7

Video scene analysis: an overview and challenges on deep learning algorithms

Qaisar Abbas[1] · **Mostafa E. A. Ibrahim**[1,2] · **M. Arfan Jaffar**[1]

Received: 26 April 2017 / Revised: 29 September 2017 / Accepted: 20 November 2017 /
Published online: 9 December 2017

Abstract Video scene analysis is a recent research topic due to its vital importance in many applications such as real-time vehicle activity tracking, pedestrian detection, surveillance, and robotics. Despite its popularity, the video scene analysis is still an open challenging task and require more accurate algorithms. However, the advances in deep learning algorithms for video scene analysis have been emerged in last few years for solving the problem of real-time processing. In this paper, a review of the recent developments in deep learning and video scene analysis problems is presented. In addition, this paper also briefly describes the most recent used datasets along with their limitations. Moreover, this review provides a detailed overview of the particular challenges existed in real-time video scene analysis that has been contributed towards activity recognition, scene interpretation, and video description/captioning. Finally, the paper summarizes the future trends and challenges in video scene analysis tasks and our insights are provided to inspire further research efforts.

Keywords Deep learning · Computer vision · Video processing · Activity classification · Scene interpretation · Video description · Video captioning

1 Introduction

Video scene analysis is an automatic process to recognize humans and objects from live-video sequences. In last several decades, the computer vision and artificial intelligence areas have been considered as an active research domain for the development of automatic application. Particularly, the area of video analysis consists of human action recognition, activity classification, scene interpretation, and video description or captioning. Human activity recognition is

✉ Qaisar Abbas
qaisarabbasphd@gmail.com

[1] Department of Computer Science, Al Imam Muhamad Ibn Saud Islamic University, Riyadh, Kingdom of Saudi Arabia

[2] Benha Faculty of Engineering, Benha University, Qalubia, Benha, Egypt

thoroughly associated with other research areas that evaluate human motion from images and video. The recognition of movement can be executed at numerous levels. So sometimes action can be represented by some activities so activity and actions are closely related to each other. Similarly, what is happening in the videos to determine action or activity, scene interpretation is also important. To describe scenes, video description and captioning are also required. Therefore, in this paper, we will discuss mainly those approaches that can deal with a variety of human action recognition, activity classification, scene interpretation, video description and video captioning.

Human detection and tracking in crowded environments are the most challenging tasks for automatic video scene analysis due to tracking of humans, which is itself unsatisfactory due to analyze complex interactions between humans and other objects. Whatever the system is developed, it must be able to analyze the complex movements of humans with respect to other objects in a dynamic environment. In each frame of video sequence, the system must be capable to identify detailed states of all objects. Such an analysis is particularly essential for many important applications including surveillance and military systems.

For video scene analysis and understanding, the authors utilized the machine learning techniques to obtain robust results. It is due to the fact that the learning-based methods have ability to update their structure consistent with input and respective output data through a training process. Hence, the learning-based methods are provided important solution for video scene analysis. For a dynamic environment, the learning-based methods did not provide an appropriate solution for real-time scene analysis because it is difficult to obtain a prior knowledge about all the objects. Still, the learning-based methods are adopted due to their robust nature.

Numerous computer vision researches were conducted in the beginning of this decade to define superior visual features for classification tasks involve in video scene analysis through conventional learning-based methods such as neural network (NN), support vector machine (SVM) and AdaBoost. These learning-based methods need the hand-crafted features extracted through Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Speeded-Up Robust Features (SURF), and Local Binary Pattern (LBP). During learning and extraction phase of these low-level features, the machine learning algorithms have invested lots of computational time by making them unusable for real-time video analysis tasks. Instead of using these low-level features, the recent trend of machine learning algorithms has introduced the new way of learning visual representations through the deep learning techniques.

Deep learning is a subfield of machine learning and it is an emerging approach in the domain of video scene analysis. In particular, the deep learning algorithms have many variants to represent visual features such as the convolutional neural network (CNN), recurrent neural network (RNN), deep belief networks (DBN), restricted Boltzmann machine (RBM) and AutoEncoder. This removes the requirement of handcrafted feature approaches that is needed for action representation like the regular method. Unlike the traditional handcrafted approach, it uses the concept of a trainable feature extractor followed by a trainable classifier, presenting the idea of end-to-end learning. In practice, this learning-based representation techniques employs computational models with many hidden to show numerous levels of abstraction [78]. The learning in the deep-neural network was built through a group of techniques to execute the data in a raw way and automatically convert it into an appropriate representation. This procedure is performed through multi-layer architecture. In the first layer, the group of pixels is extracted from each image. Afterward, in the second layer, these pixels gathers as subjects by identifying the specific edges in an image. The third layer can corporate the

subjects into small segments. Finally, the next layers could convert it into the recognizable objects. These layers are erudite from the raw data using a common learning process that does not require to be calculated manually by the experts [1].

In this survey article, our primary focus is to provide a useful information about computer vision and multimedia researchers who are interested in the state-of-the-art in deep learning-based methods known as deep vision systems (DVS). This paper provides an overview about various deep learning algorithms and their applications, especially those that can be applied in the computer vision domain. The details about these deep-learning architectures are described in detail in section 2.

Since the initial efforts on artificial neural networks, they have practiced many ups and downs, but have all the time been of distinctive concern for researchers. Neural networks based techniques have been magnificently smeared to classification, clustering, forecasting, approximation and recognition delinquents in medicine, biology, commerce, robotics etc. The modern development in the arena has been initiated by the development of deep learning techniques [21, 95, 104], prompted by the improvement of parallel computing hardware and software. The main constituent of deep learning is the multilayered hierarchical data representation usually in the custom of a neural network with beyond two layers. Such techniques let automatically synthesizing data depictions (features) of an upper level based on the lesser ones. In terms of image analysis hierarchy levels can correspond to "pixels -> edges -> combinations of edges" chain.

However deep learning has been perished by neural networks there are nearly efforts to employee its concepts to further forms of prototypes. Deep Learning, currently a hot research topic, is considered the leading machine learning tool in the imaging and computer vision domains. It can be seen as an improvement to the artificial neural networks as it consists of more layers. In other words, Deep Learning constructs many layers of abstraction that help map inputs to higher level representation. The ultimate goal of applying machine learning to video analysis is to recognize patterns in a better and quicker way than humans can, and thus increasing the productivity of video analysis software outcomes. Deep Learning (a form of machine learning) is able to take that further, especially its ability to provide improved predictions from a large amount of data it is trained on, due to the higher levels of abstractions it provides.

The main objective of deep learning is to extract information from large-scale data (e.g. images and videos) through deep architecture models with numerous hidden layers. By using this type of technique, it is easier to attain good result as compare to use raw pixel values or hand-crafted features. The main thing to achieve this is that deep learning is capable of extricating diverse stages of abstractions surrounded by perceived data. Actually, the concept of deep learning begins from 1980 when Fukushima [21] proposed Neocognitron model. In 1989, LeCun et al. [45] suggested a solution solve the problem of handwritten ZIP code recognition by employing the concept of backpropagation onto a deep neural network (DNN). Although, it was difficult to use practically due to extensive time for training on the network. Even, DNN has been used for the recognition of speech for several years, but difficult to be used as generative models. The main reason due to which deep learning architectures need huge training data that was very difficult to be available in those initial days. Hinton et al. [31] analysis these problems and proposed suggestion to solve this problem by using DNN for recognition of speech problem. Since Hinton [30] attained discoveries on training multilayer NN by pre-training single layer at a time as an unsupervised restricted Boltzmann machine and later employed supervised backpropagation for refinement. After that, it has been used in many different research areas to solve different problems.

Video analysis using deep learning technique is the recent research area. Research using deep learning techniques could make better representations and create innovative models to learn these representations from large-scale unlabeled data. Some of the deep learning techniques like Convolutional Deep Neural Networks, Deep Boltzmann Machine, Deep Belief Networks, recurrent neural networks, deep neural networks and stacked autoencoders are applied to practical applications like pattern analysis, audio recognition, computer vision, natural language processing, automatic speech recognition, bioinformatics, vehicle, pedestrian and landmark identification for driver assistance, image recognition, customer relationship management, speech recognition and translation and life sciences where they produce challenging results on various tasks Some of the advantages of deep learning technique are its ability to detect complex interactions among features, capability to learn low-level features from minimally processed raw data, easy to work with high cardinality class memberships and its competence to work with unlabeled data. Recently, DNN approaches have been shown to achieve superior performance over traditional methods. One advantage of DNNs is their capability to jointly learn feature representations and appropriate classifiers.

This paper gives an introduction to the deep learning techniques used in scene analysis applications. Those systems are known as deep vision systems (DVS). This paper summarizes the different deep learning approaches adopted by researchers in the previous years with their advantages, drawbacks and future works. The reminder of this paper is organized as follows: Section 2 demonstrates the deep learning and feature representation methods mostly utilized in Deep Vision Systems (DVS). Section 3 surveys the most recent developments in the fields of activity recognition and classification, scene interpretation, and video description or captioning using variant deep learning approaches. Section 4 depicts the latest data sources employed for appraising the numerous developed algorithms for variant deep scene analysis objectives. Section 5 discuss the challenges of deep scene analysis systems and the future remarks for overcoming the current restrictions. Finally, Section 6 sums up the review, and gives some future directions for research in the area of deep scene analysis.

2 Deep learning architectures

Deep learning is a fast-growing area which models high-level patterns in data as complex multilayered networks. It is mainly used in machine learning and artificial intelligence. Leading companies like Microsoft and Google use deep learning techniques to solve challenging problems in crucial areas like speech recognition, image recognition, object detection, object recognition, and natural language processing. In recent years, deep learning methods have been widely considered in the field of computer vision and as a result, there are many related techniques have developed. Mostly, these deep learning methods can be divided into four types according to the basic technique that they are derived from such as Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Networks (CNN) and Stacked Auto-Encoders (SAD). Some of the deep learning algorithms discussed in this section are described in the subsequent paragraphs. This section also illustrates some feature representation techniques.

The primarily aim of our review article is to emphasis the application of vision system using recent trends about deep learning algorithms. Therefore, we follow in this paper an application-oriented taxonomy for deep vision system. Figure 1 shows a scene analysis application-based taxonomy of deep learning methods recently used for video deep vision systems (DVS).

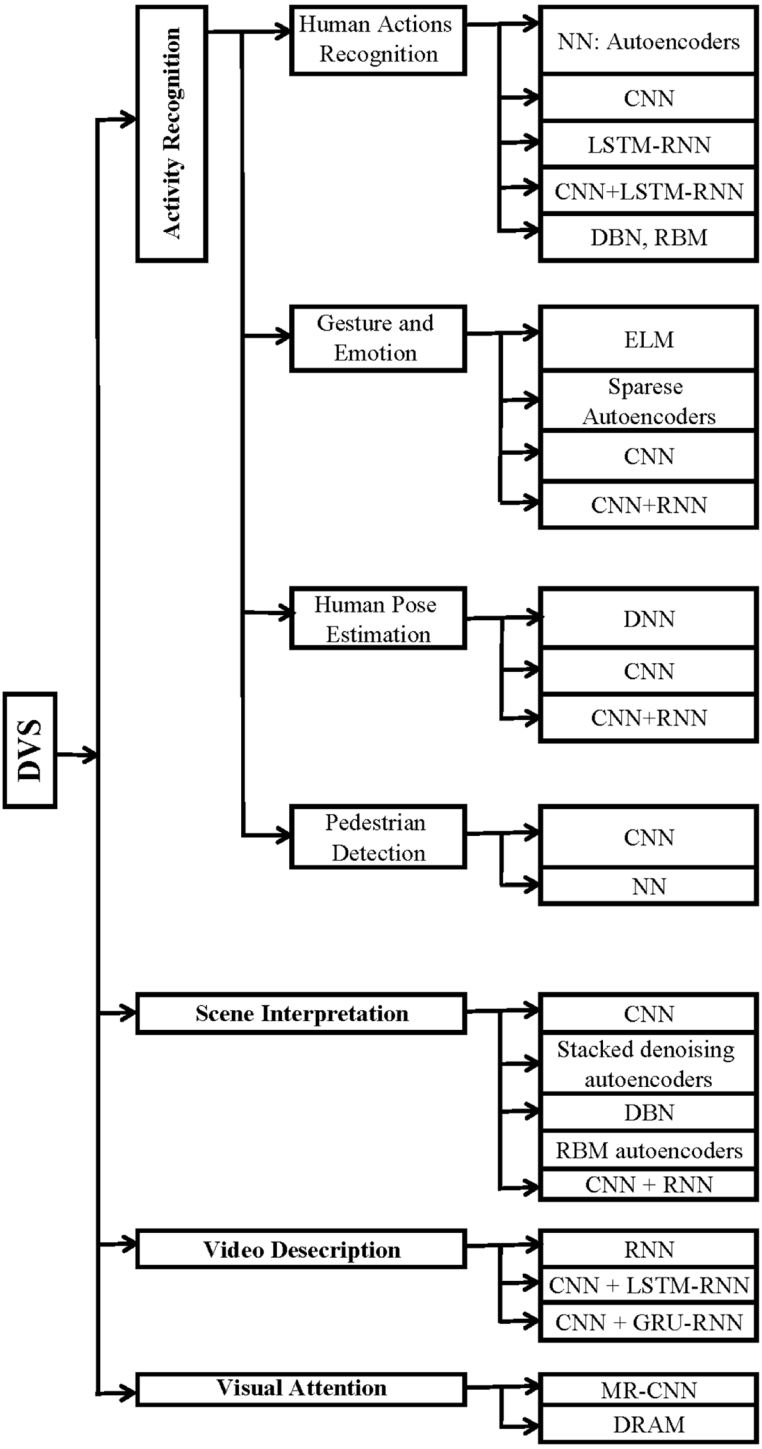


Fig. 1 Deep learning methods used for different DVS applications

2.1 Deep-learning types

2.1.1 Convolutional neural networks (CNNs)

A Convolutional Neural Network is a type of Feed forward neural network architecture in machine learning [104]. CNNs have a collection of small neurons in multiple layers that process the input image in portions called as the receptive fields. The output of these collections is lined in such a way that there is an overlapping of the input regions that gives a clear representation of the original input image. The process is repeated for all the layers (Fig. 2). CNNs are mostly used in video and image recognition, natural language processing and recommender systems.

Compared to still images, real-time videos deliver another important information in the form of motion. However, the researchers attempted to extend the CNNs for video, fed the networks with raw frames. Overall, this step is a much difficult learning problem. The main reason for the achievement of the video approach through CNN is the natural ability of the videos to be separated into spatial and temporal components. The spatial component in the form of frames captures the appearance information like the objects present in the video. The temporal component in the form of motion (optical ow) across the frames captures the movement of the objects. These optical flow estimates can be obtained either from classical approaches or correspond to the hidden units.

The advantages of the convolutional neural network are as follows. First, the usage of shared weights in convolutional layers paves the way to use the same filter for each pixel in the layer. Next, CNNs use relatively little pre-processing which means that the CNN network is responsible for learning the filters where the traditional algorithms are hand-engineered. Thirdly, CNNs are easy to train and are less dependent on the human understanding and effort and also on the previous knowledge in designing the features of the model. CNNS can also design 2D structure of the input image by using local connections and weights followed by pooling technique which in turn results in translation invariant features. The main advantage of CNNs is that it has only fewer parameters when compared to the fully connected networks with the same number of hidden units. The most distinguishing feature of the CNN is that it has the 3D volume of neurons in which the neurons are arranged in three dimensions namely weight, height, and depth. The disadvantage of CNNs is that CNNs require a huge amount of memory requirement to hold all the intermediate results of the convolutional layer for giving as the input to the back propagation layer.

A convolution net layer correlates a bank of K filters with C channels and size $R \times S$ against a small set of N images with C channels and size $H \times W$. We represent filter elements as $G_{k,\,c,\,u,\,v}$ and image elements as $D_{i,\,c,\,x+u,\,y+v}$. The computation of a single covnnet layer output $Y_{i,\,k,\,x,\,y}$ at time-frame t is given by Eq. 1:

$$Y_{i,k,x,y} = \left(\sum_{c=1}^{C} \sum_{v=1}^{R} \sum_{u=1}^{S} \ldots D_{i,c,x+u,y+v}\, G_{k,c,u,v} \right)^{t} \tag{1}$$

and we can write the output of an entire image/filter pair as follows in Eq. 2:

$$Y_{i,k} = \sum_{c=1}^{C} D_{i,c}{}^{*}\, G_{k,c} \tag{2}$$

Where $*$ represents 2D correlation. Figure 3 illustrates the utilization of CNNs in human object identification from video streams.

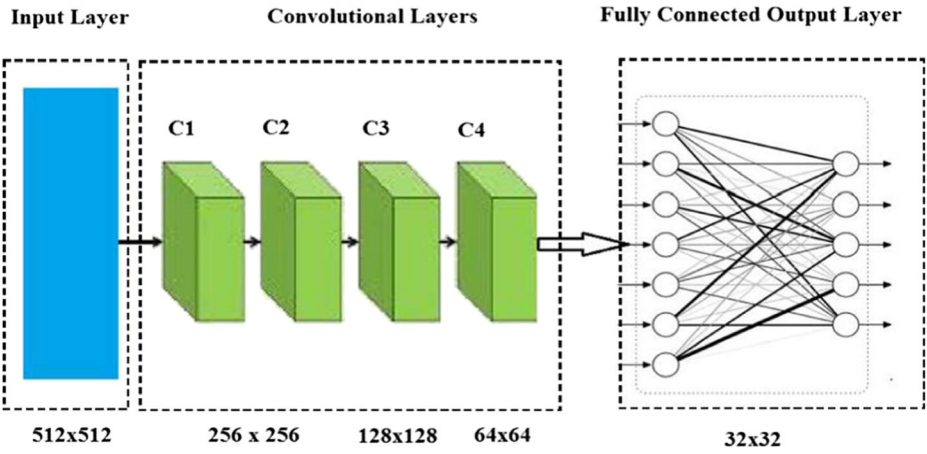


Fig. 2 Example of CNN architecture

Among all these variants of deep learning algorithms, the CNNs is very famous and an entirely supervised learning model. Initially, millions of instances are categorized to produce tens of millions of training samples. After that, the CNNs are trained by using gradient descent and back-propagation methods to make it faster convergent. However, the main challenge that faces the CNN is that, unlike humans, it require millions of semantically labeled imageries to learn a decent representation [71].

2.1.2 Recurrent neural networks (RNNs)

The convolutional neural network (CNNs) model is used widely in the past studies for video scene analysis but they are unsuitable for learning sequences. Learning such patterns requires memory of previous states and feedback mechanisms which are not present in CNNs models. Therefore, the recurrent neural networks (RNNs) are neural nets with at least one feedback connection. This looping structure enables the RNN to have an internal memory and to learn temporal patterns in data.

The recurrent neural network (RNN) is a stochastic multilayer model that is utilized in the past studies to recognize objects in video scene, music, text and motion capture [25]. The RNNs have capability to process real data sequences one step at a time and then it can generate new sequence by prediction based on training dataset. In practice, the RNN use fuzzy rules to

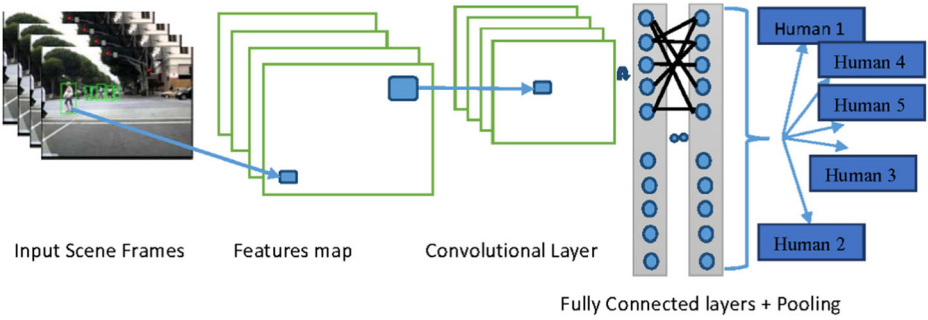


Fig. 3 Human object recognition system from live video sequence using CNN Model

generate new sequence based on hidden layers in order to avoid long sequences. Therefore, these RNN models are best candidate of optimization techniques. It is due to the fact that the standard RNNs are unable to store past inputs for very long.

As with other model problems, the predictions of network are based on few inputs, and these inputs were themselves predicted by the network then it has little opportunity to recover from past mistakes. If we increase the network size or sample size then the network cannot look further back in the past to formulate its predictions. However if we add the noise then the RNN model can go and force to learn more effective solution during training process. The architecture of RNN model is a long short-term memory (LSTM) for better storing and representing information [25].

An example of RNN model with LSTM memory training dataset is shown in Fig. 4 for human activity recognition system from live video sequence. In this figure, the CNN is combined with RNN model to predict human activity such as reading a book or listening a mobile phone from live video sequences.

2.1.3 Deep Boltzmann machine (DBM)

A deep Boltzmann machine (DBM) is one of the famous variant of deep learning algorithm that contains many layers of hidden variables. The DBM model is developed by Salakhutdinov and Hinton [75] as a network of equally-joined binary stochastic units. In DBM learning algorithm, the network has undirected connections between its top two layers and downward directed connections between all its lower layers. The DBM model is separated into two categories: (1) the observable units $v \in \{0, 1\}^D$ that show the data, and (2) the hidden units $h \in \{0, 1\}^N$ that facilitate cravings among observable units by using their common collaborations.

Though, with some sensible selections in the pattern of interactions between the visible and hidden units, more tractable subsets of the model family are conceivable. The advantages of the Deep Boltzmann Machine are their capability to learn efficient representations of complex data, with efficient pre-training technique layer by layer. The most benefit of DBMs is that it could be trained even with unlabeled data and fine-tuned with the possible limit data for a specific application. DBMs could also

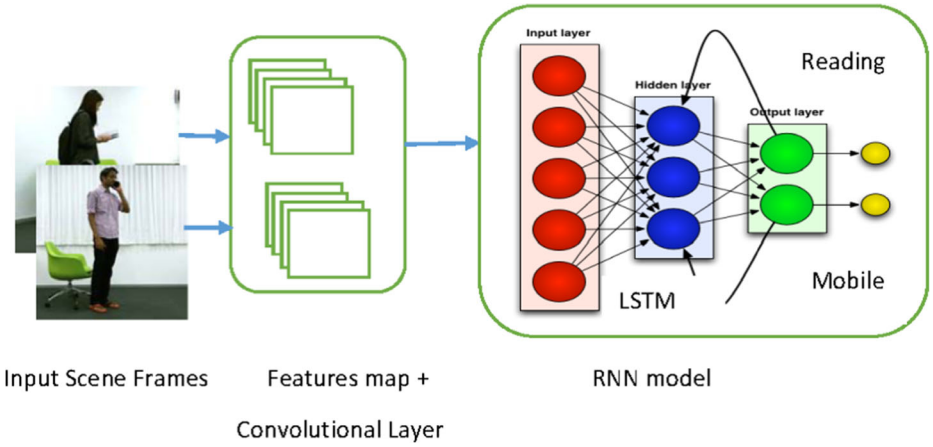


Fig. 4 Human activity recognition system from live video sequence using CNN and RNN Models

predict the uncertainty of the ambiguous input by the way of analyzing the approximate inference procedure found in DBMs. By applying the approximate gradient procedure to all the layers, the parameters in it could be optimized. Which in turn facilitates for the learning of better generation of models. While the disadvantages of the Deep Boltzmann Machine are works well for theoretical purpose rather than a general computational medium and it stop learning correctly when the machine is scaled up to anything larger than a minor machine. The approximate inference procedure followed in DBMs is nearly 50 times slower than which is followed in DBNs. Hence DBMs is not suitable for larger databases.

2.1.4 Deep belief networks (DBNs)

Deep belief networks are highly complex directed acyclic graph, which are formed by a sequence of restricted Boltzmann Machine (RBM) architectures. The principal alteration between DBN and RBM models is that the higher two layers are connected without directions while the lower layers are connected with directions (Fig. 5). DBN could be trained by training RBMs layer by layer from bottom to top. Since RBM could be trained rapidly through layered contrast divergence algorithm, the training avoids a high degree of complexity of training DBNs which in turn simplifies the process to train each RBM. Studies on DBN illustrated that it can solve low convergence speed and local optimum problems in traditional back propagation algorithm in training multilayer neural network.

The advantages of the Deep Belief Network model include the ability to learn an optimum set of parameters quickly even for the models which contain many large number of parameters and the layers with nonlinearity by way of the greedy layer-by-layer algorithm. DBNs use an unsupervised pre training method even for very large unlabeled databases. DBNs could also compute the output values of the variables in the lowest layer using approximate inference procedure. The disadvantages of DBNs include the limitation of the approximate inference procedure to a single bottom-up pass. The greedy procedure learns only the features of one layer at a time and it never readjusts with the other layers or parameters of the network. The wake-sleep algorithm proposed by Hinton for DBNs is very slow and inefficient though it fine tunes globally.

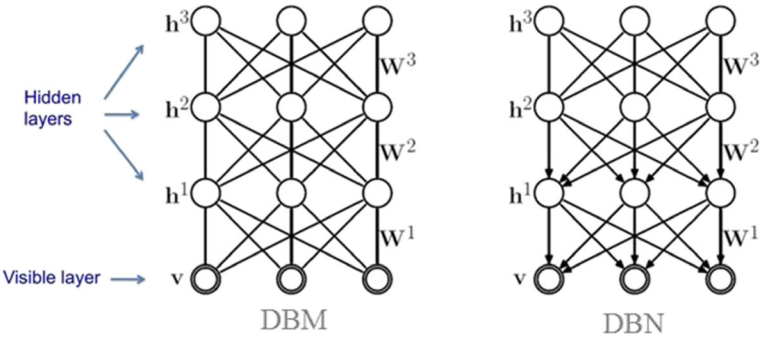


Fig. 5 DBM vs. DBN architectures [75]

They model the joint distribution between observed vector x and the l hidden layers h^k as indicated by Eq. 3:

$$P(x, h^1, \ldots, h^l) = \left(\prod_{k=0}^{l-2} P\left(\left(h^k | h^{k+1}\right)\right)\right) P\left(h^{l-1}, h^l\right) \quad (3)$$

Where $x = (h^0)$, $P(h^{l-1} | h^l)$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{l-1}, h^l)$ is the visible-hidden joint distribution in the top-level RBM. Figure 6 shows an example of using the DBNs in object tracking from video streams.

2.1.5 Stacked Denoising auto encoders (SDA)

The stacked denoising autoencoder was first introduced by Vincent et al. [93] as the protraction of the stacked autoencoder [8]. The idea behind the auto encoder is based on the concept of a well-built representation of any model. An encoder is a mapping f(x) that transforms an input vector i into a hidden layer representation h, where $h = Wt$, c, Wt is the matrix weight and c is the bias / the offset vector. The decoder maps the hidden representation h to the reconstructed input d through $e0a$. The auto encoder process is to compare the reconstructed input with the original by minimizing the error so as to make the reconstructed value as close as possible to the original value. In this technique, the output which is partially corrupted is cleaned. After the encoding function f(x) of the first denoising auto encoder is trained, it is used to uncorrupt the corrupted input where the second level can be trained. After all the layers in the stacked auto encoder is trained, the output can be used as an input to any supervised learning algorithms like Support Vector Machine, Multiclass logistic regression etc. The advantages of stacked encoders are as follows: First, it is a layer wise training method. SDA works very much compatible with Artificial Neural Networks. SDA considers all the real number inputs,

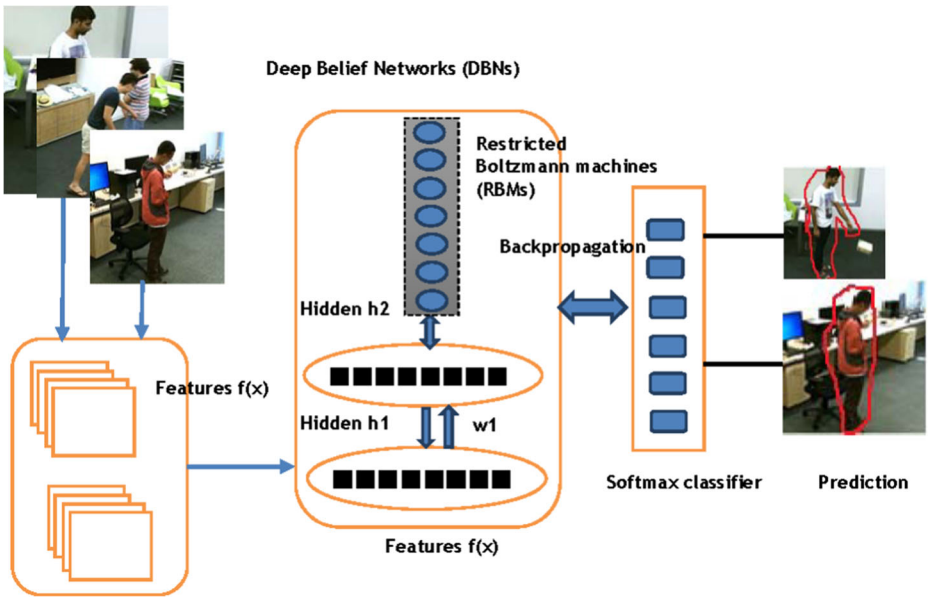


Fig. 6 Tracking Objects from live video sequence using DBN model and softmax linear classifier

binary inputs, probabilistic distribution only by simply changing the loss function and the activation function. SDA is also a self-fine tunable technique by the back-propagation algorithm which reduces the total reconstruction loss.

2.2 Feature representations

We also present some methods that are available to extract the features from the videos. Some existing approaches take into account the temporal variations during features extraction from videos while some approaches extract features for each frame of the videos individually. In such type of approaches, temporal variations can be handled during the classification process. Mainly, we divide image representations into two groups: global representations and local representations. Global representations are attained in a top-down way like the first person is localized in the frame using background subtraction and then the region of interest is segmented that can be used for feature representations. Such type of approaches is good but these approaches heavily rely on accurate localization of the object and background subtraction. These approaches have a problem of sensitivity towards viewpoint, noise, and occlusions. But if this problem handled then global representations generally good perform [24].

There is another local representation approach that combines the different local segments into a final representation known as Bag-of-feature. However, the problem of Bag-of-feature algorithms are less sensitive to noise and partial occlusion and do not firmly need background subtraction [36].

3 Recent advances in scene analysis systems

Recently, Deep Learning (DL) have been aggressively utilized for high-level computer vision tasks such as activity recognition and classification, scene interpretation, and video description or captioning. In this paper, the authors refer to the computer vision systems that employ DL as Deep Vision Systems (DVS). This section demonstrates the most recent tendencies in the aforementioned mentioned classes of DVS related to application-oriented main tasks.

3.1 Activity recognition

Recognizing human actions from videos has long been a pivotal problem in the tasks of video understanding and surveillance. Currently, activity recognition or classification is considered a hot research topic. Activity classification includes but not limited to the following topics: Human action recognition; Emotions and moods recognition; Gesture Recognition. Many deep learning algorithms were developed [27] in the past studies to perform different tasks required in the field of activity classification. As studied from the literature, many deep algorithms were proposed to solve the diverse problems of computer vision tasks, such as image classification, object detection, image retrieval, semantic segmentation and human pose estimation. The literature obviously shows that the CNN is the main utilized deep learning in such vision analysis tasks. Table 1 introduces a summary of the most recent studies in terms of employed deep learning architectures for visual activity recognition. It illustrates the utilized data sources, the achieved results and finally if the study compares its achievements to others.

In computer vision domain, the human action recognition is an important field and many deep learning algorithms have been developed so far for image categorization tasks [7]. In the

Table 1 State-of-the-art deep-learning methods for activity classification

Cited	Scope	Method	Data	Results
Xia et al. [99]	Human Lying-pose detection system	R-CNN	XMULP	Avg. Precision AP 53.8%
Lin et al. [54]	3D Human Activity Recognition	CNN	CAD-120, SBU OA	Acc.: Min 37% and Max. 94%
Lin and Yuan [52]	Human Activity Recognition	CNN + LSTM RNN	UCF-101	Acc.: (1/2 1/2) 85.8% (1/3 2/3) 87.2%
Husain et al. [35]	Content/activity recognition	CNN	UCF-101, HMDB	Acc. HMDB: VGG-3D 46.9% VGG-3D + C3D-FC6–1 net 53.9%
Pei et al. [67]	Action recognition	Independent Subspace Analysis (ISA) NN	Hollywood-2, KTH, UCF sport	Acc.: KTH: 90.0% UCF sports: 85.6% Hoolywood2: 39.4%
Pigou et al. [70]	Gesture recognition	CNN + LSTM RNN	Montalbano gesture	Acc.: Max 97.23% (Conv + RNN, LSTM). Min. 79.92% (Single-frame CNN)
Ho et al. [32]	Emotion prediction	CNN	UGVEmotion	Acc.: 54.2%
Acar et al. [2]	Analyzing the emotional content of user generated video sequences	CNN	DEAP VideoEmotions Sentibank	Acc. 49.19% for VideoEmotions 81.08% for DEAP
Kaya and Salah [39]	In the wild emotions recognition	Extreme Learning Machines (ELM)	EmotiW 2014	Acc.: 50.12%
Liu et al. [58]	Classifying human emotions in videos that were captured under uncontrolled environment	CNN	EmotiW 2013, CK+ MMI	Acc.: 45.28%
Sun et al. [86]	Multimodal Emotion Recognition system	LRCN (CNN + LSTM RNN)	CHEAVD, FER	Acc.: Best 51.85% Worst 38.27%
Tome et al. [88]	Pedestrian detection	CNN	Caltech	Miss Rate 0.248 at 0.1 False Positive Per Image
Ouyang et al. [65]	A pedestrian detection system that considers the overlapping pedestrians.	NN	ETHZ, Caltech Pedestrian, PETS	Miss Rate 37% for Caltech & ETHZ & 50% for PETS
Zhao et al. [108]	Facial expression recognition (FER) is another way to recognize human activities during job activities.	CNN	SFEW	ACC: 57.3%
Liu et al. [57]	3D human action identification derived from 3D skeleton data.	RNN-LSTM	NTU RGB + D, SBU Interaction, UT-Kinect, Berkeley MHAD	ACC: For NTU RGB + D, 69.2% (Cross subject) 77.7% (Cross view)

form of sequential frames, the computational power is significantly increased when used algorithms such as deep learning techniques. However, the authors are still using this latest trend. However, this deep learning technique is better than the handcrafted representation of features that are usually ad-hoc and overfitting to specific data. Still, the usage of deep learning in human action recognition is a natural opinion. In recent years, the human activity recognition has gained a lot of attention and it is demanded in various applications.

3.1.1 Human actions recognition

Many researchers have addressed the problem of recognizing variant human actions. Le et al. [44] presented a framework that employed unsupervised feature learning to learn features directly from video data for activity recognition. They have proposed a technique that acquires features from spatiotemporal data by using the concept of independent subspace analysis. They boosted the technique to huge interested fields by convolution and stacking and acquired hierarchical representations. They have shown outcome very exciting by using the similar parameters through 4 datasets. It also recommends that learning features straight from data is a very significant research way. Baccouche et al. [4] have proposed a neural framework to recognize human action by using a fully automated feature building procedure. In this paper, they have used a convolutional auto-encoder that is trained to construct a sparse shift-invariant demonstration of 2D shapes available in every frame of the video, and the temporal growth of these features is employed to categorize the actions. The development of these mid-level features is constructed by a Recurrent Neural Network trained to categorize every sequence. Ballan et al. [5] have proposed a DBN based technique to categorize human actions. It represents a novel descriptor for spatiotemporal interest points. It performed an ensemble of appearance, motion, effective codebook creation and consignment of feature represents to code-words. They employed image gradient and optic flow to represent the appearance and motion of human actions at portions in the surroundings of local interest points and cogitate numerous spatial and temporal scales. Finally, they have used radius-based clustering with the soft assignment. It has been reported that there is a reduction of computation time due to codebook size decrease with DBNs. While Zuniga et al. [113] suggested an event learning method for video, based on concept formation framework. This technique incrementally learns on-line a hierarchy of states and events by accumulating the attribute values of tracked objects in the scene. The framework can cumulative together numerical and symbolic values. The application of symbolic attributes provides high tractability to the method. The method also suggests the incorporation of attributes as a doublet value-reliability, for seeing the consequence of the event learning procedure of the indecision hereditary from preceding phases of the video analysis procedure. Simultaneously, the method distinguishes the states and events of the traced objects, giving a multi- level explanation of the object condition. The method has been tested for an old care application and a rat performance examination application. Charalampous and Gasteratos [10] also developed an unsupervised on-line deep learning algorithm. They used spatiotemporal data and obtained significant results by developing this deep learning algorithm. Also, compared this deep learning approach with other methods, in many cases, it is shown to outperform in terms of classification accuracy.

One of the early attempts to employ CNN for action recognition was introduced by Ji et al. [36]. They proposed a framework that created features from both spatial and temporal dimensions by using 3D convolutions. Proposed framework produce numerous ways of information from contiguous input frames and accomplish convolution and sub-sampling

independently in every channel. The absolute feature depiction is attained by joining data from all channels. They proposed a framework for regularization and grouping methods to more enhancement the framework performance. Lin et al. [54] introduce a deep CNN model for 3D Human Activity Recognition. Their proposed model differed from tradition deep learning models by considering the latent temporal structure and by defining a radius margin bound. They make use of several datasets for performance evaluation including the CAD-120 and the Office-Activity datasets. They achieved an average accuracy of 93.4% fo the SBU dataset while achieved an average accuracy of 37.3% and 88.9% for Merged-50 and Merged-4 datasets respectively. However, their proposed model requires very heavy computations at a large number of human activities. Which obviously indicates that their model is not scalable. Moreover, their model can only deal with simple human actions, not complex ones. Lin and Yuan [52] developed a human action recognition approach that relied on CNN and RNN for extracting spatial and temporal features from raw video sequences. Their deep learning approach does not include an optical flow CNN which leads to a considerable learning time reduction. However, they do not get the best accuracy but they achieved a comparable accuracy to other related work in case of using the UCF-101 dataset.

Husain et al. [35] presented a model for content/activity recognition that extended a 2D CNN to a 3D CNN. Their model extracted spatiotemporal features from raw video sequences. They utilized two datasets for evaluating their model namely; UCF-10 and HMDB. They achieved an accuracy of 86.7% for the UCF-101 dataset which is the second best accuracy in the literature. While they got an accuracy of 53.9% for the HMDB which is a moderate accuracy compared to the related work for the same dataset. But it worth to mention that their model does not require preprocessing stage of data such as the optical flow stage. Hence, the computational time of their model is relatively low. Ronao and Cho [74] proposed a deep convolutional neural network (CNN) and achieved an almost perfect classification on moving activities, especially very similar ones which were previously perceived to be very difficult to classify. Lastly, ConvNets outperform other state-of-the-art data mining techniques in HAR for the benchmark dataset collected from 30 volunteer subjects. An overall performance of 94.79% was achieved on the raw sensor data set. This high accuracy is due to perfect classification of moving activities. The human hand activities are recognized in [9] by developing a deep-learning-based method to classify hand states such as free vs. active hands, hand gestures, object categories), and discover object categories. The authors privately collected a new dataset with 20 videos captured in three scenes for hand states recognition. The authors fine-tuned the convolutional neural network (CNN) and obtained significant results in four out of five tasks.

Pei et al. [67] presented an NN-based approach for action recognition. Their proposed approach depended on learning spatiotemporal features using NN and Temporal pooling from raw video data. Moreover, their model integrates the denoising Independent Subspace Analysis (ISA). Three datasets were used to evaluate their approach namely; KTH, Hoolywood-2, and UCF sports. They attained an average accuracy of 85.5% and 90% for the UCF sports and the KTH datasets. Hasan and Roy-Chowdhury [28] have suggested a new method for constant learning of activity prototypes from streaming videos by knottily tying organized deep hybrid feature framework and active learning. They considered a deep hybrid feature framework for human activity recognition. It will be trained in an unsupervised way and can use both local and the deep feature framework. They apprise the feature and the recognition framework spending the unlabeled cases that constantly reach from the video stream. They used a mixture of semi-supervised and active erudition method so as to decrease the manual labeling of the

arriving cases. They grow a method to make a choice the finest group of agents from the training data.

Mocanu et al. [63] developed a new version of Conditional Restricted Boltzmann Machines (CRBMs) for human activity recognition system. In this CRBMs model, they incorporated a new label layer and four-way interactions among the neurons from the different layers. The additional layer gives the classification nodes a similar strong multiplicative effect compared to the other layers and avoids that the classification neurons are overwhelmed by the (much larger set of) other neurons. Two sets of experiments one on benchmark datasets and one on a robotic platform for smart companions show the effectiveness of FFW-CRBMs. A spatiotemporal feature learning approach was developed by Pei et al. [68] for action recognition. They automatically detect and track the actor, and then determine the action. The authors used some ensemble classifiers to form a two-layer network classifier. In the first layer, they used multiple RBMs. Each RBM is trained by the data vectors that have the same spatial location. The output of the second layer RBM is the learned spatiotemporal feature. They trained a Support Vector Machine classifier for each class to recognize the actions. Experiments on challenging data sets confirm the effectiveness of this approach. Similarly, Xu et al. [100] used also spatiotemporal feature learning approach with the adapted hidden restricted haphazard field.

Liu et al. [57] presented an RNN based framework for 3D human action identification derived from 3D skeleton data. 3D action recognition is characterized by its invariance to the viewpoint. Unlike old methods that employed RNN with only temporal features. They considered both spatial and temporal features. In order to overcome the noise and occlusion issues, they presented a novel gating technique for LSTM. They evaluated their algorithm using four datasets including the most recent RGB + D dataset. Their results outperformed other state-of-the-art methods for 3D human action recognition. Kong et al. [41] also presented a framework for 3D human action recognition using both RGB and depth features that is extracted using 3D kernel descriptors.

3.1.2 Gesture and emotion recognition

Apart from ordinary human activity classification, Pigou et al. [70] presented a framework that integrated both Deep CNN and RNN for gesture recognition. Besides, they showed that utilizing simple temporal feature pooling is not adequate for gesture recognition. They figured out that using RNN is essential for gesture recognition especially if it is integrated with CNN to enhance the performance. Montalbano dataset is used for evaluating their approach. They attained an accuracy of 97.23% by the integration of CNN and RNN. While Li et al. [49] developed a human hand gesture recognition system by using feature learning approach that is sparse autoencoder and principle component analysis (PCA) for recognizing human actions, i.e. finger-spelling or sign language, for RGB-D inputs. The authors improved the recognition rate from 75% to 99.05% and outperform the state-of-the-art.

Similarly to gesture recognition, other researchers focus on another human activity which is the emotion recognition and classification. Recognition of emotional state is also important for human-robot interaction [38]. In practice, the robot can identify important variables of human behavior and thereby extend the interaction with human-like fashion. As a matter of fact, it is very difficult to recognize the human emotional state. The authors developed a Multichannel CNN to classify human emotional state and achieved 91.3% accuracy on spontaneous emotion expressions. Ho et al. [32] investigated the validity of employing CNN to predict emotions from user-generated videos. They incorporated wheel guided CNN for feature extractions.

Their results indicated an average accuracy rate of 54.2%. Kaya et al. [39] developed a CNN-based multimodal technique for recognizing face expressions and emotions from video taken under uncontrolled environment. Their method integrates both audio and video features with "least squares regression based classifiers and weighted score level fusion". They evaluated the performance of their method using the EmotiW and ChaLearn-LAP data sources. Acar et al. [2] proposed a framework for analyzing the emotional content of video sequences that edited or user-generated videos. Audio and visual features that were extracted using CNN have been utilized to represent the input video sequences. The Video Emotion and a division of the DEAP dataset have been utilized for performance evaluation. They achieved the accuracy of 49:19% in case of using Video Emotions dataset and 81.08% in case of using DEAP dataset. The results demonstrated that the use of audio learned features outperforms the handcrafted features. Moreover, the results illustrated that deep learning do better than multi-class SVMs. The facial expression recognition (FER) is another way to recognize human activities during job activities [108]. In that paper, the authors developed deep convolutional neural networks (deep CNNs) for static facial expression recognition but tested on the wild (SFEW) dataset. They trained multiple deep CNNs by varying network architectures, input normalization, and weight initialization as well as by adopting several learning strategies and obtained an accuracy of 57.3%.

In the same direction, Sun et al. [86] also implemented a DCNN-based multimodal emotion recognition system. The system was capable of recognizing eight facial emotions. Several classifiers along with a decision fusion algorithm to aggregate the different predictions are employed to realize the best performance. The CHEAVD and the FER datasets are involved in the evaluation process of their proposed system. They accomplished a maximum accuracy of 51.85% and a minimum of 38.27%. However, the authors of that work only explore the bag-of-features pooling strategy.

Unlike other researchers, Kaya and Salah [39] utilized Extreme Learning Machines (ELM) for in-the-wild emotions recognition. ELM is used for representing multimodal features. They stick to the proposed algorithms in EmotiW 2014 challenge. Their proposed approach was able to distinguish seven different emotions. Their approach achieved an accuracy of 50:12% on the seven under test emotions. Liu et al. [58] also developed a framework for classifying the human emotions in videos that were captured under uncontrolled environment. The input video sequences were modeled using three different approaches namely; linear subspace, covariance matrix, and Gaussian distribution. Next Riemannian kernels were utilized for estimating the similarity/distance. Three classifiers were employed which are kernel SVM, logistic regression, and partial least squares. At last, the different classifier outputs were fused for enhancing the performance. Three different datasets were used for performance evaluation. The results showed an accuracy of 45.28%. However, they employed a simple score-level fusion strategy.

3.1.3 Human pose estimation

A basic human activity is to estimate the configuration of a human body also known as human pose estimation. Li et al. [48] proposed a framework for 3D human pose estimation. Their framework utilizes CNN for feature extraction. The output features and the pose inputs are then transformed to a joint pose embeddings. The dot product of the image features and the pose embeddings has been considered as a similarity score function. RNN is next utilized for doing the interference with the learned image-embedding and also to fine-tune the pose

coordinates. The Human3.6 m data set is used for evaluating the performance of their system. Xia et al. [99] introduced a system for detecting human lying-pose that utilized CNN. They used normalized binary gradient features and then feed these features to learn a CNN. At last, the location and direction of the human lying-pose are identified using pyramid mean-shift algorithm. Their system attained an average precision of 53.8% for the XMULP dataset. Trumble et al. [89] determined the pose in multiple view-point videos (MVV) and by using an affine invariant pose descriptor is learned using a convolutional neural network (CNN). They included Gaussian processes for the CNN descriptor and so estimation of human pose from MVV input. The learned descriptor and manifold are shown to generalize over a wide range of human poses, providing an efficient performance capture solution that requires no fiducials or other markers to be worn.

Similarly, Zhu et al. [111] developed a supervised deep ConvNets to segment the body from depth images. This approach can be used for object recognition, human pose estimation, and scene recognition through a decomposition into a collection of parts. The authors trained a classifier with spatial relationships which increases generalization performance when compared to classical training minimizing classification error on the training set. They presented an application to human body part estimation from depth images.

Whereas Hong et al. [33] suggested a big data-driven approach for Human Pose Recovery (HPR). They used deep learning neural network (DNN) architecture that has 4 layers namely: 2D features, 2D hidden representations, 3D hidden representations and 3D poses. They enhanced 3 sets of plotting parameters for the DNN. They considered a novel deep architecture known as Multimodal deep autoencoder (MDA) for HPR. At the initial stage, the process of MDA employed Multi-view Hyper-graph that is capable to ensemble numerous features into an integrated manifold representation. Like this, the difference between features and images is lessened. After that, the MDA used 2 auto-encoders to mine hidden depictions for 2D images and 3D poses. During understanding the mapping function, they combined a two-layer NN to map the internal depictions of 2D visual features to those of 3D poses. Since the MDA is supervised and job-explicit, it can mutually discover the inner 2D/3D depictions and the proper association between them.

3.1.4 Pedestrian detection

One of the imperative claims of the activity recognition/classification is the pedestrian detection. Pedestrian detection is crucial for video surveillance and robot systems. Tome et al. [88] introduced a DCNN based approach for pedestrian detection. Their proposed method is tested using both handcrafted and learned features. They achieved a very comparable accuracy, a miss rate of 0.248 at 0.1 False Positive Per Image (FPPI), while significantly reduced the computational time. Their work is an important step toward real-time pedestrian detection. Ouyang et al. [65] developed a pedestrian detection system that considered the overlapping pedestrians. The utilization of overlapping pedestrians enhanced the detection accuracy for the single pedestrian. Three famous datasets were used to appraise the effectiveness of their model. They achieved lowest miss rate of 37% for Caltech & ETHZ datasets compared to other related systems. While they achieved a miss rate of 50% for the challenging PETS dataset.

3.2 Scene interpretation

The scene interpretation is defined as the process of "representing and recognizing structures consisting of several spatially and temporally related components (e.g. object configurations, situations, occurrences, episodes)" [64]. It helps to comprehend decisive activities or behaviors in a given scene such as dodging of barriers. This section illustrates the most recent trends in the scene interpretation using deep learning methods. The literature confirms that the deep learning algorithms outperformed other state-of-the-art methods as reported in Table 2.

Visual path prediction is an important and challenging task from a static or live video sequences using machine learning algorithms [34]. In fact, the scene interpretation is a complicated task for underlying motion patterns in the video sequence. Many researchers utilized CNN for different scene interpretation applications.

Huang et al. [34] suggested that the deep learning CNN framework performs very well to learn visual features. They used a deep semantic understanding of the scenes and motion patterns to improve the performance of the prediction of the visual path. They did comparisons with this approach to the state-of-the-art literature and achieved higher performance. Whereas Sarkar et al. [76] detected the occlusion edges by training a deep CNN approach from video sequences. They utilized the use of CNN to avoid hand-crafting features for automatically determined the occlusion edges from video sequences. There are also complexities, public security, and other surveillance applications require more efficient and intelligence video processing at runtime [106]. To address these challenges, Zhang et al. [106] proposed a deep-learning framework that can express the knowledge hidden in video sequences. Similarly, Zhu et al. [112] presented an unsupervised framework to learn jointly from both visual and

Table 2 State-of-the-art deep-learning methods for scene interpretation

Cited	Scope	Method	Data	Results
Wang et al. [96]	Multimodal representation Learning	DNN	MIR Flickr 25 K, Wikipedia	mAP: 0.6395
Gao and Zhang [22]	Loop closure detection	DNN	RGB-D, and FAB-Map	similarity score 0.79 recall rate 0.45 for FabMap at precision 100%
Revathi and Kumar [72]	Detection of abnormal events	DBN RBM	*UCSD*	EER of 0.75% Precision Rate 85%
Zhang et al. [107]	Events recognition	RBM	*MED'11*, and CCV	mAP 87.53% MED'11 mAP 70.83% CCV
Huang et al. [34]	Visual path prediction	CNN	*Video Tracking*	–
Mathieu et al. [61]	Multi-scale video prediction	CNN	*UCF101*	PSNR 30.0 SSIM 0.90
Ballas et al. [6]	Learning video representations	CNN+ GRU-RNN	*YouTube2Text UCF101*	Acc.: 85.7%
Zhang et al. [105]	Moving object detection based on binary scene modeling	Stacked autoencoder	*Highway PETS2006 Pedestrians*	F-measure 0.7595
Perez et al. [69]	CNN learning algorithm that automatically detect pornographic in advance	CNN	Pornography-800, Pornography-2 k	classification accuracy 97.9%

independently-drawn non-visual data sources for discovering the meaningful latent structure of surveillance video data. Mathieu et al. [61] have proposed a standard of numerous approaches for next frame prediction, by assessing the superiority of the estimate in terms of PSNR, SSIM and image sharpness. They showed outputs on small UCF video clips. The displayed architectures and losses may be employed as main masses for more erudite estimate framework, including memory and recurrence. Contrasting maximum optical flow methods, their framework is fully differentiable, so it can be adjusted for other jobs if required.

In contrast with these approaches, Lee et al. [47] presented a deep-learning approach to detect shadow from moving objects in a video sequence. They concluded that this deep-learning approach is far better than hand-crafted features and learned the visual features of shadow regions from an input source. They compared this approach with the state-of-the-art approaches and found better performance. This algorithm is applied to five different datasets of moving shadow detection for comprehensive experiments. Etezadifar and Farsi [17] developed a new method to summarize a large amount of video compared to other methods of using training and selection sparse dictionary problem simultaneously. This approach was also compared with other methods and achieved the best performance.

Gao and Zhang [22] presented the employment of stacked autoencoders in a new scene interpretation application. They developed an unsupervised DNN based approach to detect loops closure for "simultaneous localization and mapping systems". Their approach employed also a stacked denoising autoencoder (SDA) for detecting the loops. Two datasets RGB-D and Fab-map were used to evaluate the performance of their approach. Their results showed a similarity score of 0.79 and a recall rate 0.45 for Fab-map at a precision of 100%. However, they use the dataset for both learning and evaluation. Xu et al. [101] also presented a deep learning based framework for detecting unusual events in a video surveillance. Their developed framework is based on stacked denoising autoencoders to learn both visual and action features. The UCSD pedestrian anomaly dataset is used for assessing the performance of their proposed approach. The results show that their approach outperforms other state-of-the-art methods. Wang et al. [96] presented a DNN based framework for scene analysis. They developed a new technique for combining multi-modal features in order to learn the DNN. They also adopted a joint learning model. Two benchmarks Wikipedia and MIR Flickr 25 K datasets were used to appraise their proposed framework. They achieved a mean average precision of 0.6395. However, their learning protocol still needs more optimization.

Other researchers utilized other DL methods such as DBN and RBM. Revathi and Kumar [72] introduced a new DBN based approach for abnormal events detection in video sequences. Their system estimated the background, then separated the objects from the background, next extracted features from those objects. Finally, these features were fed to the DBN for classifying the activity as normal or abnormal. They utilized the USCD dataset to assess the performance of their system. They achieved an equal error rate of 0.75% and a precision rate of 85%. However, their system suffers from the large dimensionality of the features which in turn affected the computational time. While Zhang et al. [107] developed an RBM based system for recognizing complicated scene events. Their system provided RBM based autoencoders to combine the different features of different modalities (Human actions, foreground targets, and the scene).

The RBM were unsupervised trained with unlabeled data. The MED11 and CCV datasets were used for assessing the effectiveness of their proposed system. They attained a mean average precision of 87.53% and 70.83% for MED'11 and CCV datasets respectively. Ballas et al. [6] presented another solution to learn visual features from live video sequences. The

author suggested that the low-level percepts reserve a greater spatial resolution by utilizing a recurrent convolutional network (RCN) on the ImageNet dataset from diverse spatial resolutions.

Zhang et al. [105] have suggested an efficient technique for moving object detection. The technique gets a stacked autoencoder that consists of deep learning approach to adaptively form a strong feature representation that was accomplished of fine apprehending the basic structural characteristics of a scene and adaptively finding a group of filter patterns that are strong to complex issues like noise and illumination changes. To enhance the computational efficacy, they have proposed a meek but efficient hash technique to make the features binary. Moreover, authors have proposed a block-wise binary scene framework that proficiently transforms the spatiotemporal dispersal in the development on the scene blocks.

Zhou et al. [109] suggested a Combination framework of Dynamic Pedestrian-Agent to learn the collective dynamics from video sequences in jam-packed sights. The combined dynamics of pedestrians were transformed as the linear dynamic method to get extended variety moving patterns. By using modeling the principles of pedestrians and the omitted situations of annotations, MDA can be fine learned from greatly patchy trajectories instigated by regular tracking batches. Hence, it is appropriate for behavior analysis in crowded situations. By modeling the procedure of pedestrians creating assessments on actions, it can not only categorize collective behaviors but also envisage mutual crowd behaviors.

Recently few researchers addressed a new scene interpretation application: automatic detection of pornographic scenes in video sequences. Unluckily, video scenes that contain high skin exposure, such as in porno movies or even in scenes of people who are taking the sun bath and wrestling, a lot of false alarms are expected. Perez et al. [69] proposed an approach for detecting pornography in video sequences that was based on deep CNN. They combined static and dynamic features to achieve higher accuracy. They have tested their system on MPEG motion vectors and achieved 97.9% classification accuracy and an error lessening of 64.4% compared to the state-of-the-art research work on a collection of 800 test cases.

3.3 Video description and captioning

Recently, video captioning gets higher importance because of the increasing demand for online video search [80]. A significant research is also devoted in the domain of video description. Wu et al. [98] presented a deep video hashing technique to enhance the process of video search. Cho et al. [12] illustrated the recent utilized encoder-decoder architectures for describing video scenes. They demonstrated the role of CNN and RNN in the encoder-decoder model. This kind of architectures is commonly used for divers of missions such as "machine translation, image caption generation, video clip description, and speech recognition". This section demonstrates the up to date achievements in video description or captioning using deep learning methods. Table 3 shows the used deep learning methods, datasets, results and indicates if the results were compared to state-of-the-art algorithms.

The literature showed some attempts to engage RNN in video description systems. Pan et al. [66] described the video content along with natural language by using RNNs deep learning algorithms. The authors suggested that the sentence semantic and visual contents are not discussed in the past studies to describe the relationship between them. Therefore, they named this model as a Long Short-Term Memory with visual-semantic Embedding (LSTM-E), which can simultaneously explore the learning

Table 3 State-of-the-art deep-learning methods for video captioning

Cited	Scope	Method	Data	Results
Yao et al. [102]	Image description based on video sequences.	3D–CNNs + LSTM-RNN	Youtube2Text	NA
Pan et al. [66]	the authors described the video content along with natural language.	RNN	YouTube2Text	Performance: 45.3% and 31.0%
Venugopalan et al. [91]	a solution for end-to-end sequence-to-sequence model that generate captions for videos.	RNN	M-VAD and MPII-MD	Prediction: 49.5%
Rohrbach et al. [73]	descriptions from live video sequences to assist the blind people.	CNN + LSTM RNN	MPII-MD	NA
Donahue et al. [16]	Investigated the use of recurrent models in video description.	CNN + LSTM RNN	ImageNet, UCF101, COCO 2014	Class. Acc.: 60.2% and 57.4% for hybrid and CaffeNet models CIDEr-D metric 0.895 for caption generation for COCO dataset
Venugopalan et al. [92]	Directly interpret video scenes to text.	CNN + LSTM RNN	MRVDC, Flickr30k and COCO2014	SVO Acc.: 87.27%, 42.79%, 24.23% for both COCO and Flickr30k
Zhu et al. [110]	Straightforwardly interpret video scenes to text	CNN + GRU RNN	MovieBook	outperforms the SVM baseline by 30% in recall, and doubles the AP.

of LSTM and visual-semantic embedding. The obtained results indicate that this model is performed outstanding results compared to state-of-the-art techniques. Venugopalan et al. [91] developed a technique to propose a solution for the end-to-end sequence-to-sequence model that generate captions for videos. They used RNN model to on video-sentence pairs and learns the sequence of the event in the video clip. They tested their proposed approach in the standard video sequence dataset that came from YouTube videos such as M-VAD and MPII-MD.

Other researchers integrated both CNN and LTSM RNN models for variety purposes of video description. Yao et al. [102] utilized (3-D CNN) and RNNs deep-learning models to do image description based on video sequences. In this paper, the authors used static images taking from real-time video sequences to integrate the information in the proper manner. In the first step, the authors proposed 3D CNN to represent the short temporal dynamics and then trained on video action recognition tasks. As a result, this study can be used to detect the human motion and behaviors. Afterward, the authors used RNN to model temporal attention to select most relevant segments. This approach is better than the BLEU and METEOR metrics as achieved by these authors. Moreover, Rohrbach et al. [73] generated image descriptions from live video sequences to assist the blind people and human-robotic interactions. The authors utilized data set of MPII-MD and M-VAD video sequences. In that paper, the author made a clear difference among verbs, objects, and places in the setting of movie description. They used robust visual classifiers that are generated from weak annotations of the sentence description in order to complete this research study.

Donahue et al. [16] investigated a variety of hybrid deep learning architectures (recurrent and convolutional) for several scenes analysis tasks such as activity classification, and video captioning or description. Their model captures both spatial and temporal features. The recurrent models can straight forward generate erratic-span outputs such as “natural language text” from the input video sequences. It also can mock-up intricate dynamic temporal features with backpropagation optimization. They developed a hybrid deep learning approach using CNN as well as LSTM layers (which they call LRCN) to recognize activity in video scenes. The UCF101 and Flickr30k datasets have been utilized for evaluating the performance of their proposed approach. Their results illustrated that their proposed LRCN model outperformed other state-of-the-art models. Venugopalan et al. [92] presented a novel deep NN-based approach to straightforwardly interpret video scenes to text. Their approach employed CNN as well as RNN layers. The main challenge they faced is the rarity of data sources with appropriate description. Thus, they collect a huge dataset of over 100,000 images annotated with captions. Their approach shows acceptable performance when applied to video sequences and succeeded to generate text that describes these videos. Another sophisticated application to video description is to match book text to the video movies based on these books [110]. They proposed a deep neural network framework that was unsupervised learned from a variety of books. Their system also employs another NN for calculating the matching between movie scenes and book text. Then a CNN was employed to integrate information from different manifolds. They quantitatively and qualitatively evaluated their framework and the results showed good alignment between movie and book text.

4 Data sources

In this section, the authors argue and depict the most recent data sources used for appraising the numerous developed algorithms for variant scene analysis objectives. The authors in this paper spotlight the data sources utilized mostly since 2015 and hardly ever since 2014.

4.1 Data sources for activity recognition

Kth However, it is a relatively old dataset but it still used for training many DL algorithms for human activity recognition. It consisted of six human actions classes namely: "walking, jogging, running, boxing, hand waving and hand clapping" as shown in Fig. 7. These actions were captured from only four different scenes taken out from exactly 2391 video clips of poor resolution (160 × 120) at 25 fps [77].

Hollywood-2 Composed of 12 different human action patterns. The covered actions in this dataset include AnswerPhone, DriveCar, Eat, GetOutCar, Run, SitDown, SitUp, StandUp, HandShake, Kiss, FlightPerson and HugPerson as shown in Fig. 8. These actions were taken from 10 variant scenes such as in car, in kitchen, out of house, and on the road. These distinct classes are dispersed over more than 3500 video snips in excess of a 20 h video captivated from 69 movies [60].

UCF-sports Comprised of ten dissimilar sport actions. These actions were gathered from a diverse of sports such as diving, golf, and swimming as shown in Fig. 9. The dataset embraces

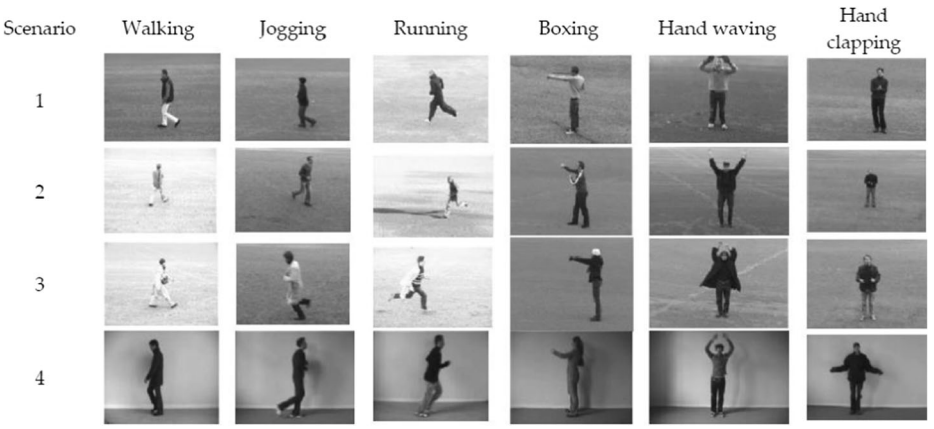


Fig. 7 Variant actions and scenes of the KTH dataset

150 video clip with a spatial resolution of (720 × 480) at 10 fps of a total of 16 min. This data source was frequently utilized in various action classification systems [85].

Caltech pedestrian A dataset composed of almost 250,000 video frames with (640 × 480) spatial resolution at 30 fps of a total of 10 h. These video sequences have been captured for a vehicle driving during normal traffic environment. It incorporates exactly 2300 distinctive annotated pedestrians. The annotation also gives a detailed occlusion labels [15]. Figure 10 shows samples of the Caltech pedestrian dataset.

MuHAVi Multicamera Human Action Video Dataset. This dataset is one of very few datasets that uses multiple cameras for recording the different human actions. In MuHAVi dataset, eight cameras were employed. It consists of 17 variant human actions acted by 14 persons. The video sequences of this dataset are available in both AVI and MPEG-2 formats with a good resolution of (720 × 576) at 25 fps. Some selected action video sequences are annotated manually Silhouette Data [83]. Figure 11 shows samples of the MuHAVi dataset.

NTU-RGB + D A very recent dataset for human action recognition. It consists of more than 56,000 actions including four variant patterns of data for each action namely: "RGB video, depth maps, 3D skeletal, and infrared videos" [79]. This modern dataset comprised large number of human actions, typically 60 different actions include many daily human action categories like moving (walk, jump, sit,...etc.), wearing(shoe, glass, hat,...etc.), and reading/ writing as shown in Fig. 12. Three different Microsoft Kinect sensor based cameras were used

Fig. 8 Samples of the Hollywood-2 dataset actions

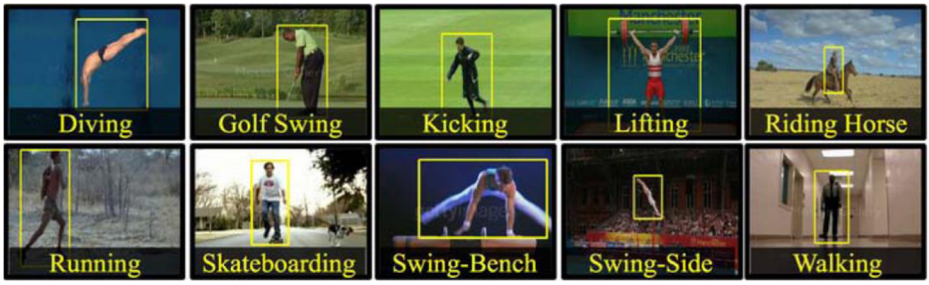


Fig. 9 Samples of the UCF-Sports dataset actions

to capture the RGB video sequences with a high resolution of (1920 × 1080). While the spatial resolution of both the depth maps and infrared videos is (512 × 424) [79].

4.2 Data sources for scene interpretation

MED'11 Multimedia Event Detection (MED) is a dataset that is released 2011 and is dedicated for evaluating algorithms for searching video scenes for pre-defined events. Events are multifarious actions taking place at a definite position and time and engages individual or group of human persons react with further persons or objects. The video sequences of this dataset are supplied in MPEG-4 format. It consists of 15 events of complex actions such as feeding an animal, landing a fish, Wedding ceremony, changing a vehicle tire, Parade, Parkour, and repairing an appliance [59].

CCV Columbia Consumer Video dataset is an important data source for variant scene analysis applications. It comprises more than 9000 YouTube videos snips of a total of 210 h. It includes 20 different semantic classes [37].

4.3 Data sources for video description

YouTube2Text Consists of 1970 video sequences from YouTube. It is mainly used to evaluate algorithms that generate short sentence descriptions of video clips [26]. Figure 13 shows an example of the YouTube2Text dataset.

Microsoft Research video description corpus Consists of slightly more than 120 K sentences collected during the summer of 2010. Workers were partially employed to

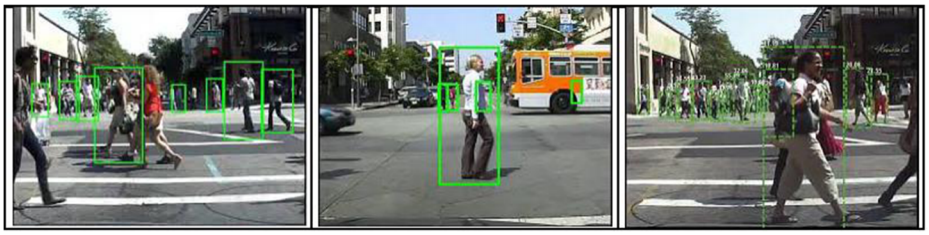


Fig. 10 Samples of the Caltech pedestrian dataset

Fig. 11 Samples of the MuHAVi multicamera human actions dataset

watch more than 2000 very short video sequence. The video snips commonly represent a single, explicit act or event. Then they sum up the action in these video clips in only one sentence. The data set include video descriptions for more than 16 languages. The English language descriptions constitute the majority (70%+) [11]. Figure 14 shows an example of the this dataset.

Flickr30k Has become a standard dataset for paraphrase image/video clip description. It composed of 158 k captions for 30 K+ images. The captions center of attention is on people implicated in daily actions and behaviors (Fig. 15) [103].

Fig. 12 Samples of the RGB + D 3D human actions dataset

A woman is cooking onions.
Someone is cooking in a pan.
someone preparing something
a person coking.
racipe for katsu curry

A girl is ballet dancing.
A girl is dancing on a stage.
A girl is performing as a
ballerina.
A woman dances.

A man is sitting and playing
a guitar
A man is playing guitar
Street artists play guitar.
A man is playing a guitar.
a lady is playing the guitar.

A train is rolling by.
A train passes by Mount Fuji.
A bullet train zooms through
the countryside.
A train is coming down the
tracks.

Fig. 13 Samples of the YouTube2Text dataset

5 Discussion

This section recapitulates the cutting-edge research in deep scene analysis, signalizes the challenges and finally sets apart auspicious remarks for the upcoming research to address these restrictions.

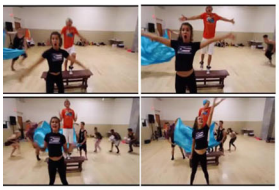
In this paper, the authors surveyed the most recent advances in deep learning methodologies employed for the various scene understanding fields. The literature review showed that variant deep learning approaches have been utilized such as stack autoencoders, RBMs, CNN, and RNN. Figure 16 represents the contribution percentage of each deep learning architecture surveyed in this paper.

It is obvious from Fig. 16 that the CNN is the most widely utilized DL method in video processing applications. Therefore, Table 4 demonstrates the architectural information and results of well-known CNN models.

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.



1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

Fig. 14 Samples of the microsoft research video description corpus dataset

Fig. 15 Samples of the Flicker30K dataset

No matter what variant DVS have been utilized in the past studies for video processing, this paper spotlights three categories of scene analysis or understanding namely: activity classification, scene interpretation, and video description or captioning.

5.1 Results of deep vision systems

In this section, the authors sum up the modern deep scene understanding such as activity recognition, scene interpretation and video captioning and description.

5.1.1 Activity classification

Deep learning methods are widely utilized for a wide spectrum of activity classification claims. Many researchers have proposed deep learning algorithms for human activity recognition and classification. These algorithms differed in the deep learning model used, the way of representing features, as well as the way of training input data. Le et al. [44]; Baccouche et al. [4]; Ballan et al. [5]; Lee et al. [46]; Pei et al. [67]; Hasan and Roy-Chowdhury [28]; Mocanu et al. [63]; Pei et al. [68]; Xu et al. [100], presented deep learning algorithms for human actions classification based on NN, RBMs. They extracted spatiotemporal features that are next used to train the deep learning architectures in a supervised manner. While Zuniga et al. [113], and Charalampous and Gasteratos [10] attempted to do unsupervised on-line learning. Other researchers utilized the CNN for human action classification such as Ji et al. [36]; Lin et al. [54]; Lin and Yuan [52]; Husain et al. [35]; Ronao and Cho [74]; Chan et al. [9].

Another important activity category is the pedestrian detection which was addressed by Tome et al. [88] and Ouyang et al. [65]. Besides, several CNN deep learning approaches are

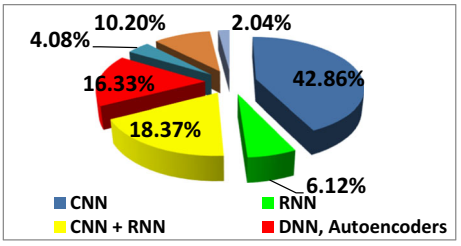


Fig. 16 Overall deep learning architectures utilized in recent scene analysis studies

Table 4 Recently utilized CNN based architectures

CNN	Year	# of Layers	Architecture	Description
AlexNet Krizhevsky et al. (2012)	2012	8	5 convolutional +3 fully connected layers	An early base CNN architecture for variant of current architectures
VGG-16 Simonyan et al. (2014)	2014	16	13 convolutional +3 fully connected layers	CNN deep architecture with much more conv. Layers by using smaller conv. Filters of 3 × 3
GoogLeNet Szegedy et al. (2015)	2014	22	21 convolutional +1 fully connected layers	CNN deeper architecture with much more conv. Layers with variant conv. Filter sizes of 1 × 1, 3 × 3 and 5 × 5
SPP He et al. (2014)	2014	8	5 convolutional +3 fully connected layers	CNN architecture equipped with spatial pyramid pooling layers to allow for input images with variant sizes

used for human pose estimation and recovery. Li et al. [48] and Xia et al. [99] proposed human pose estimation methods. Similarly, Trumble et al. [89] determined the human pose but in multiple viewpoint videos. Zhu et al. [111] presented an approach for segmenting the human body from depth images which can be employed in object recognition, human pose estimation. In the same direction, Hong et al. [33] developed a DNN architecture for human pose recovery.

Apart from ordinary human activity classification Pigou et al. [70] and Li et al. [49] developed frameworks for human gesture recognition. Similarly to gesture recognition, other researchers focus on another human activity which is the emotion recognition and classification. Ho et al. [32] and Acar et al. [2] proposed CNN based systems to predict emotions from user-generated videos. In the same direction, Sun et al. [86] also utilized CNN but for multimodal emotion recognition system. Moreover, Zhao et al. [108] developed an approach for facial expression recognition (FER). While Kaya and Salah [39] utilized Extreme Learning Machines (ELM) for in-the-wild

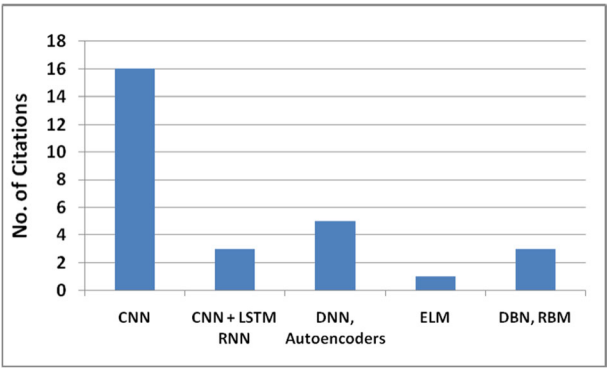


Fig. 17 Deep learning architectures employed in recent studies for activity classification

emotions recognition. Also, Liu et al. [57, 58] developed a framework for classifying the human emotions in videos that were captured under uncontrolled environment using kernel SVM, logistic regression, and partial least squares classifiers. Figure 17 characterizes the different deep learning architectures that are recently employed for activity recognition and classification.

However, activity classification faces some key challenges [41] such as: 1) Intra and Interclass variations, 2) Noisy or dynamic background, and 3) Camera jitter. An example of the intra-class variation is the variation of the individual action, for example, the walk action can be slow or fast. An instance of the inter-class variation is the resemblance of two or more different actions such as walk and run actions.

5.1.2 Scene interpretation

The main goal of scene interpretation process is to describe the video sequence frames to a momentous semantics [23]. Many researchers utilized CNN for different scene interpretation applications. Huang et al. [34] used a deep semantic understanding of the scenes and motion patterns to improve the performance of the prediction of the visual path. Whereas Sarkar et al. [76] developed a CNN based system to detect the occlusion edges in video sequences. Zhang et al. [106], Zhu et al. [112] and Mathieu et al. [61] proposed a CNN deep-learning framework that can express the knowledge hidden in video sequences. While Ballas et al. [6] combined both CNN and RNN for scene interpretation. Lee et al. [47] presented a deep-learning approach to identify shadow from moving objects in a video sequence. They proved that DL methods outperform hand-crafted features. Another application for scene interpretation was introduced by Gao and Zhang [22] to detect loops closure for "simultaneous localization and mapping systems" using deep NN stacked autoencoders. Wang et al. [96] presented a DNN based framework for combining multi-modal features in order to learn the DNN. Other researchers utilized other DL methods such as DBN and RBM. Revathi and Kumar [72] developed a DBN based approach for abnormal events detection in video sequences. Zhang et al. [107] developed an RBM based system for recognizing complicated scene events. Recently few researchers Perez et al. [69] addressed a new application: automatic detection of pornographic scenes in video sequences. Figure 18 illustrates the diverse deep learning methods that are recently engaged in scene interpretation systems.

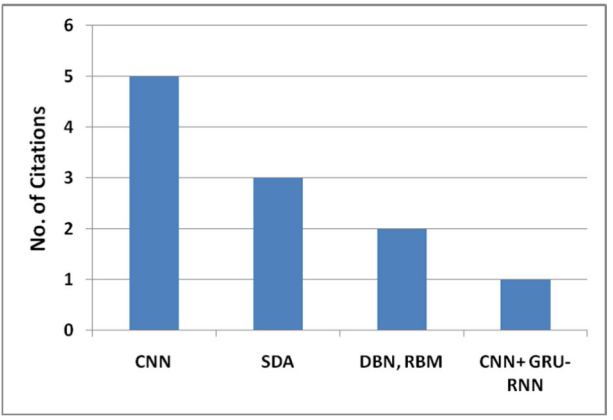


Fig. 18 Deep learning techniques recently utilized for scene interpretation

Although deep learning approaches gain significant importance in the few past years, but they are still far from real time video scene analysis. In robotic applications, the deep learning methods are not performing sophisticated tasks, done by human video scene analyzers, due to the dynamic and varied nature of the environment.

5.1.3 Video captioning and description

Recently, many researchers presented DL frameworks for video description or at list captioning. Pan et al. [66] and Venugopalan et al. [91] described the video content along with natural language by using RNNs deep learning algorithms. While Yao et al. [102]; Rohrbach et al. [73]; Donahue et al. [16]; Venugopalan et al. [92]; Zhu et al. [110] integrated both CNN and LTSM RNN models for variety purposes of video description. However, Yao et al. [102]; Rohrbach et al. [73]; Donahue et al. [16] utilized CNN and RNNs deep-learning models to do image description based on video sequences. Venugopalan et al. [92]; Zhu et al. [110] developed a CNN-RNN based approach for video scenes to text translation.

5.2 Data sources for deep scene analysis applications

A further important factor that significantly affects the evaluation of the developed algorithms for deep scene understanding is the openly accessible data sources. Which permit a fair way of comparing the results of different algorithms based on common datasets. The authors, momentarily describe the most frequently utilized datasets in the past three years. The datasets are first categorized based on the specific scene understanding application into human actions, scene interpretation, and video description data sources. Moreover, the datasets differ in the number of classes, video clips they contain, and spatial resolution. In addition, they differ in the way of capturing these video clips either using single viewpoints such as KTH, UCF-Sport and Caltech pedestrian or multi-view points as MuHAVi, and NTU RGB + D as illustrated in Section 4. Table 5 lists the main features of recent used datasets for various scene analysis tasks such as deep activity recognition and classification, deep scene interpretation, and for video captioning.

5.3 Deep learning challenges

Tuning the current deep learning architectures to extend the use of multimodal method improve recognition/classification of objects in video scene understanding. Although the great advents in both deep learning and video processing approaches, there still a room for further improvements especially in contexts such as camera jitter, dynamic background, intra-class similarity, and video occlusion.

The variants of deep-learning algorithms are widely used in computer vision applications but still, the theory behind, human vision, and 3D recovery, real-time processing complexity are too important issues to ignore for the long-term viability of the trending DVS systems. Hereafter, we discuss in detail these challenges.

5.3.1 Theory behind DL

Although DL is currently used on broad band of computer vision application, the theory behind the DL is still not well understood. The optimal number of layers cannot be easily

Table 5 Scene analysis utilized data sources in recent years

Ref.	Data source	# Classes/ actions	# Scenes	# Subjects	# Video Clips/ frames	Spatial resolution	Frame rate (fps)
Koppula et al. [42]	CAD-120	20	NA	4	120 clips RGB-D	240 × 320	NA
Lin et al. [54]	Office Activity (OA)	20 over two Subsets	NA	10	1180 clips	NA	NA
Soomro and Zamir [85]	UCF-101	101	NA	NA	13,320 clips	320 × 240	25
Marszalek et al. [60]	Hollywood-2	12	10	NA	3500 clips	NA	NA
Schuldt et al. [77]	KTH	6	4	25	2391 clips	160 × 120	25
Soomro and Zamir [85]	UCF sport	10`	NA	NA	150 clips	720 × 480	10
Dollar et al. [15]	Caltech Pedestrian	2300 Pedstrian	NA	NA	250,000 frames	640 × 480	30
Singh et al. [83]	MuHAVi	17		14	952 Clips	720 × 576	25
Shahroudy et al. [79]	NTU RGB + D	60	NA	NA	56,880 clips	1920 × 1080 RGB 512 × 424 D-Map	NA
Lai et al. [43]	RGB-D	300	14	NA	900 clips 22 annotated video sequences	640 × 480	30 HZ
Chen and Dolan [11]	YouTube2Text	241	NA	45	1970 clips	NA	25
Lin et al. [53]	COCO2014	80	NA	NA	300,000 images	NA	NA

decided because there is no clear formula to estimate the number of layers. Even more, the architecture of the DL network i.e. how many layers to be convolutional, recurrent or pooling also is not easy task.

5.3.2 Human vision

In computer vision tasks, the human visual system (HVS) has performed the tasks without having the problems of geometric transformations, background changes and occlusion in an efficient manner. Therefore, the human visual system is more effective and efficient compared to deep-learning algorithms for video processing tasks. This gap should be minimized in the future. More studies of the human brain should be integrated into these deep machine learning architectures for better performance. Moreover, the variants of the DL algorithms should build more layers for selecting middle- or high-level features to simulate the human brain structure. In particular, the performance of the DL algorithms for video processing is not matched with that of the human visual system.

5.3.3 Limited datasets

DL algorithms require a training phase which needs the existence of many different datasets that consist of huge number of images or video frames. The limited datasets even in the number of publically available ones or those with limited number of classes and instances badly influence the training accuracy of DL architecture which in turn harshly decrease the classification or detection

accuracy. This paper reviews the most recent utilized datasets in the visual objects detection, recognition and tracking applications. The issue of limited datasets have been tackled in the surveyed studies in two ways: 1) using dataset mirroring e.g. by flipping the dataset images/ frames horizontally, 2) using weak supervised learning methods.

5.3.4 Processing at real time

In the early version of DL algorithms, lot of computational resources were required and now, we are moving these algorithms toward the real-time video processing. Therefore recently, Li et al. [51] and Wu et al. [97] developed new DL architectures allowing for real-time video processing applications. A series of experiments were conducted to reduce the running cost, filter size and propose dynamic activation function that is not fast enough for real-time applications. Therefore, fixing the time complexity is required in deep-learning algorithms and eliminated all the duplicate computations in the forward and backward propagation. In addition, the GPU-based implementation of the DNN [20] will provide the high efficiency compare to the simple model of this machine learning algorithms.

5.4 Future trends for deep learning

Recently the DL architectures are improved and get deeper in term of number of layers and number of processing units per layer to become in line with the latest progress in image and video processing applications. Still, the DL architectures need to be more influential in terms of their computational power. Henceforth, we explore some schemes to increase the DL architectures computational power.

1) Building deeper DL architectures by adding more layers to improve the performance. Recently GoogleNet DL model reaches 22 layers.
2) Using hierarchical features learning i.e. learning variant layers with variant features. Liu et al. [56] presented the DeepIndex that utilized the hierarchical features learning solution.
3) Using different kinds of features not only the raw images such as the SIFT or SURF features.
4) Design application specific DL models not only relying on current models. Since every computer vision application has its own characteristics. For instant some computer vision application require pixel level annotations while other require object level annotations and other even require scene level annotations.

Integrate different DL approaches such as CNN and RNN to make use of both of these DL models. Few researchers have followed this direction Lin and Yuan [52] combined RNN with CNN for human action recognition application While Yao et al. [102] combined CNN and RNN for Image description based on video sequences. Rohrbach et al. [73] combined CNN and RNN for generating descriptions from live video sequences to assist the blind people. Venugopalan et al. [92] and Zhu et al. [110] combined CNN and RNN for direct interpretation of video scenes to text. In this review paper, we have surveyed the visual application of deep learning algorithms. There is another important and fundamental visual processing task that exists in the literature known as visual attention. It is the process that describes a set of mechanisms that limit some processing to a subset of incoming stimuli [3, 13, 14, 18, 19, 50]. It involves the assortment of spatial as well as a temporal region of interest, restriction of features and features dimensions, prevailing the stream of data through

the DL architecture, and moving between different selected regions of interest [55, 62, 81, 84, 90, 94]. In the future work, we should also explore visual attention related to deep learning architectures

6 Conclusion

This paper discussed the recent challenges present in real-time video scene analysis that have been contributed to activity recognition, scene interpretation, and video description captioning. Firstly, this study described the general framework about deep learning algorithms that have been adopted in the past years. Secondly, the deep learning methods have applied better representation and classification when configured and trained properly. However, there is still a room for improving these learning algorithms specifically for time series data. Thirdly, there is also needing to change the internal structure of the deep learning algorithms to capture both short and long-term time dependencies. For real-time video processing, there is a dire need to extend this deep learning concept in terms of better learning of features and faster to train. Furthermore, we need larger datasets for real-time scene analysis especially in scene interpretation and action recognition algorithms that should direct research efforts to realistic settings. The scene interpretation and activity classification domains have many applications in practice such as human surveillance, HCI, and robotic poses different challenges. The real-time video scene analysis has a broad range of application and it is expected that these challenges will be addressed in the near future in terms of advancement of deep learning algorithms.

References

1. Abdulnabi AH, Wang G, Lu J, Jia K (2015) Multi-task CNN model for attribute prediction. IEEE Trans Multimedia 17(11):1949–1959. https://doi.org/10.1109/TMM.2015.2477680
2. Acar E, Hopfgartner F, Albayrak S (2016) A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material. J Multimedia Tools Appl 76(9):11809–11837. https://doi.org/10.1007/s11042-016-3618-5
3. Ba J, Mnih V, Kavukcuoglu K (2015) Multiple object recognition with visual attention. In: *Proceedings of Int Conf on Learning Representations (ICLR'15)*. San Diego, California, USA
4. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2012) Sparse shift-invariant representation of local 2D patterns and sequence learning for human action recognition. *In:* Proceedings of the 21st Int Conf on pattern recognition (ICPR'12), pp 3823–3826. doi:10.11385.6048
5. Ballan L, Bertini M, Bimbo AD, Seidenari L, Serra G (2012) Effective codebooks for human action representation and classification in unconstrained videos. IEEE Trans Multimedia 14(4):1234–1245. https://doi.org/10.1109/TMM.2012.2191268
6. Ballas N, Yao L, Pal C, Courville AC (2016) Delving deeper into convolutional networks for learning video representations. *In: Proceedings of Int Conf on Learning Representations (ICLR'16)*, San Juan, Puerto Rico
7. Barros P, Jirak D, Weber C, Wermter S (2015) Multimodal emotional state recognition using sequence dependent deep hierarchical features. J Neural Netw 72:140–151. https://doi.org/10.1016/j.neunet.2015.09.009.
8. Bengio Y, Lamblin P, Popovici D, Larochelle H (2006) Greedy layer-wise training of deep networks. In: Proceedings of the 19th Int Conf on neural information processing systems (NIPS'06). MIT Press, Canada, pp 153–160
9. Chan C-S, Chen S-Z, Xie P-X, Chang C-C, Sun M (2016) Recognition from hand cameras: a revisit with deep learning. In: Proceedings part IV of 14th European Conf computer vision (ECCV'16). Springer Int Publishing, Amsterdam, The Netherlands, pp 505–521. https://doi.org/10.1007/978-3-319-46493-0 31
10. Charalampous K, Gasteratos A (2016) On-line deep learning method for action recognition. *J of*. Pattern Anal Applic 19(2):337–354. https://doi.org/10.1007/s10044-014-0404-8.

11. Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. *In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR, USA
12. Cho K, Courville A, Bengio Y (2015) Describing multimedia content using attention-based encoder-decoder networks. IEEE Tran Multimedia 17(11):1875–1886. https://doi.org/10.1109/TMM.2015.2477044
13. Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J (2012) Deep neural networks segment neuronal membranes in electron microscopy images. *In: Proceedings of Conf on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, pp. 2852–2860
14. Couprie C, Farabet C, Najman L, LeCun Y (2013) Indoor semantic segmentation using depth information. *In: Internatinal Conf on Learning Representation (ICLR'13)*, Scottsdale, AZ, USA, pages 8
15. Dollar P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761. https://doi.org/10.1109/TPAMI.2011.155
16. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2016) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of IEEE Conf on computer vision and pattern recognition (CVPR'15). MA, USA, Boston, pp 2625–2634
17. Etezadifar P, Farsi H (2016) Scalable video summarization via sparse dictionary learning and selection simultaneously. J Multimedia Tools Appl 76(6):7947–7971. https://doi.org/10.1007/s11042-016-3433-z
18. Evans KK, Horowitz TS, Howe P, Pedersini R, Reijnen E, Pinto Y, Kuzmova Y, Wolfe JM (2011) Visual Attention. Wiley Interdiscip Rev Cogn Sci 2(5):503–514
19. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929. https://doi.org/10.1109/TPAMI.2012.231
20. Farrajota M, Rodrigues JMF, du Buf, JMH (2016) A Deep Neural Network Video Framework for Monitoring Elderly Persons. *In: Proceedings Part II of 10th International Conference Universal Access in Human-Computer Interaction (UAHCI2016),* pp. 370–381, Toronto, ON, Canada, July 2016
21. Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *J of.* Biol Cybern 36(4):193–202. https://doi.org/10.1007/BF00344251
22. Gao X, Zhang T (2015) Unsupervised learning to detect loops using deep neural networks for visual SLAM system. *J of.* Auton Robot 41(1):1–8. https://doi.org/10.1007/s10514-015-9516-2
23. Gilani SO, Jamil M, Fazal Z, Naveed MS, Sakina R (2016) Automated scene analysis by image feature extraction. *In: Proceedings of IEEE 14th Intl Conf on Dependable, Autonomic and Secure*. Computing: 530–536. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2016.102
24. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. *In: Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR'14)*, IEEE computer society, Columbus, Ohio, USA, pp. 580–587, doi:https://doi.org/10.1109/CVPR.2014.81
25. Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. Proceedings of the IEEE Int Conf on Acoustics, Speech and Signal Processing:6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947
26. Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K (2013) YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. *In:* Proceedings of IEEE Int Conf on computer vision (ICCV'13), pp. 2712–2719, doi: https://doi.org/10.1109/ICCV.2013.337
27. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS (2016) Deep learning for visual understanding. J of Neurocomput 187(C):27–48. https://doi.org/10.1016/j.neucom.2015.09.116
28. Hasan M, Roy-Chowdhury AK (2015) A continuous learning framework for activity recognition using deep hybrid feature models. IEEE Trans Multimedia 17(11):1909–1922. https://doi.org/10.1109/TMM.2015.2477242
29. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9):1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
30. Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11(10):428–434
31. Hinton G, Deng L, Yu D, Dahl GE, Mohamed RA, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Proc. Magaz 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
32. Ho C-T, Lin Y-H, Wu J-L (2016) Emotion prediction from user-generated videos by emotion wheel guided deep learning. *In: Proceedings of 23rd Int Conf on Neural Information Processing (ICONIP'16)*, springer Int publishing, Kyoto, Japan, pp. 3–12, doi:https://doi.org/10.1007/978-3-319-46687-3 1
33. Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoencoder for human pose recovery. IEEE Trans Image Proc 24(12):5659–5670. https://doi.org/10.1109/TIP.2015.2487860

34. Huang S, Li X, Zhang Z, He Z, Wu F, Liu W, Tang J, Zhuang Y (2016) Deep learning driven visual path prediction from a single image. IEEE Trans Image Proc. 25(12):5892–5904. https://doi.org/10.1109/TIP.2016.2613686
35. Husain F, Dellen B, Torras C (2016) Action recognition based on E_cient deep feature learning in the Spatio-temporal domain. IEEE Robo Auto Lett 1(2):984–991. https://doi.org/10.1109/LRA.2016.2529686
36. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59
37. Jiang Y-G, Ye G, Chang S-F, Ellis D, Loui AC (2011) Consumer video understanding: a benchmark database and an evaluation of human and machine performance. *In: Proceedings of ACM Int Conf on Multimedia Retrieval (ICMR'11)*, Trento, Italy
38. Jiu M, Wolf C, Taylor G, Baskurt A (2014) Human body part estimation from depth images via spatially-constrained deep learning. Pattern Recogn Lett 50(C:122–129. https://doi.org/10.1016/j.patrec.2013.09.021
39. Kaya H, Salah AA (2016) Combining modality-specific extreme learning Machines for Emotion Recognition in the wild. *J on Multimodal User*. Interfaces 10(2):139–149. https://doi.org/10.1007/s12193-015-0175-6.
40. Krizhevsky A, Sutskever I, and Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12) vol 1, USA, p 1097–1105
41. Kong Y, Fu Y (2016) Human activity recognition and prediction, springer Int publishing, Switzerland, chapter "action recognition and human interaction", pp. 23-48. doi:https://doi.org/10.1007/978-3-319-27004-3 2
42. Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from RGB-D videos. Int J Rob Res (IJRR) 32(8):951–970
43. Lai K, Bo L, Ren X, Fox D (2011) A large-scale hierarchical multi-view RGB-D object dataset. *In: proceedings of IEEE International Conference on Robotics and Automation (ICRA'11)*, shanghai, China, pp. 1817–1824, doi:https://doi.org/10.1109/ICRA.2011.5980382
44. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant Spatio-temporal features for action recognition with independent subspace analysis. *In: Proceedings of IEEE Conf on Computer Vision and Pattern Recognition (CVPR'11)*, Colorado Springs, USA, pp. 3361-3368, 24 https://doi.org/10.1109/CVPR.2011.5995496
45. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. J of Neural Comput 1(4):541–551. https://doi.org/10.1162/neco.1989.1.4.541
46. Lee K, Su Y, Kim T-K, Demiris Y (2013) A syntactic approach to robot imitation learning using probabilistic activity grammars. J of Robot Auton Syst 61(12):1323–1334. https://doi.org/10.1016/j.robot.2013.08.003
47. Lee JT, Lim K-T, Chung Y, Sugimoto A (2016) Moving shadow detection from background image and deep learning. In: Proceedings of Image and Video Technology (IVT'15), workshops, Auckland, New Zealand, pp. 299–306, doi:https://doi.org/10.1007/978-3-319-30285-0 24
48. Li S, Zhang W, Chan AB (2015a) Maximum-margin structured learning with deep networks for 3D human pose estimation. In: Proceedings of IEEE Int Conf on computer vision(ICCV), pp. 2848–2856, doi: https://doi.org/10.1109/ICCV.2015.326g
49. Li S-Z, Yu B, Wu W, Su S-Z, Ji R (2015b) Feature learning based on SAE-PCA network for human gesture recognition in RGBD images. J Neurocomputing 151:565–573
50. Li T, Chang H, Wang M, Ni B, Hong R, Yan S (2015c) Crowded scene analysis: a survey. IEEE Trans Circuits Syst Video Technol 25(3):367–386. https://doi.org/10.1109/TCSVT.2014.2358029
51. Li H, Li Y, Porikli F (2016) DeepTrack: learning discriminative feature representations online for robust visual tracking. IEEE Trans Image Process 25(4):1834–1848. ISSN 1057-7149. https://doi.org/10.1109/TIP.2015.2510583
52. Lin Z, Yuan C (2016) A very deep sequences learning approach for human action recognition. *In: Proceedings of 22nd Int Conf on MultiMedia Modeling*, springer Int publishing, Miami, FL, USA, pp. 256–267. doi:https://doi.org/10.1007/978-3-319-27674-8 23
53. Lin T et al (2014) Microsoft COCO: common objects in context. In: Proceedings of the 13th European conference on computer vision (ECCV'14), Zurich, Switzerland, pp. 740–755. doi:https://doi.org/10.1007/978-3-319-10602-1_48
54. Lin L, Wang K, Zuo W, Wang M, Luo J, Zhang L (2016) A deep structured model with radius-margin bound for 3D human activity recognition. Int J Comput Vision 118(2):256–273. https://doi.org/10.1007/s11263-015-0876-z.
55. Liu N, Han J, Zhang D, Wen S, Liu T (2015a) Predicting eye fixations using convolutional neural networks. *In:* Proceedings of IEEE Conf on computer vision and pattern recognition (CVPR'15*)*, pp. 362–370. doi:https://doi.org/10.1109/CVPR.2015.7298633

56. Liu Y, Guo Y, Wu S, Lew M (2015b) DeepIndex for accurate and efficient image retrieval. *In: Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR'15)*, shanghai, China, pp. 43–50, doi:https://doi.org/10.1145/2671188.2749300
57. Liu J, Shahroudy A, Xu D, Wang G (2016a) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Proceedings of the 14th European Conf computer vision (ECCV'16). Netherlands, Amsterdam, pp 816–833. https://doi.org/10.1007/978-3-319-46487-9-50
58. Liu M, Wang R, Li S, Huang Z, Shan S, Chen X (2016b) Video modeling and learning on Riemannian manifold for emotion recognition in the wild. *J on Multimodal User*. Interfaces 10(2):113–124. https://doi.org/10.1007/s12193-015-0204-5.
59. Ma Z, Yang Y, Sebe N, Zheng K, Hauptmann AG (2013) Multimedia event detection using a classifier-specific intermediate representation. IEEE Trans on Multimedia 15(7):1628–1637. https://doi.org/10.1109/TMM.2013.2264928
60. Marszalek M, Laptev I, Schmid C (2009) Actions in context. *In:* Proceedings of IEEE Conf on computer vision and pattern recognition (CVPR'09), pp. 2929–2936. doi:https://doi.org/10.1109/CVPR.2009.5206557
61. Mathieu M, Couprie C, LeCun Y (2016) Deep multi-scale video prediction beyond mean square error. *In: Proceedings of Int Conf on Learning Representations (ICLR'16)*, San Juan, Puerto Rico
62. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. *In: Collections of Advances in Neural Information Processing Systems*, No. 27, Curran Associates, Inc., pp. 2204–2212
63. Mocanu DC, Bou Ammar H, Lowet D, Driessens K, Liotta A, Weiss G, Tuyls K (2015) Factored four way conditional restricted Boltzmann Machines for Activity Recognition. Pattern Recogn Lett 66(C:100–108. https://doi.org/10.1016/j.patrec.2015.01.013
64. Neumann B, Möller R (2008) On scene interpretation with description logics. *J of*. Image Vis Comput 26(1):82–101. https://doi.org/10.1016/j.imavis.2007.08.013
65. Ouyang W, Zeng X, Wang X (2016) Learning mutual visibility relationship for pedestrian detection with a deep model. Int J Comput Vision 120(1):14–27. https://doi.org/10.1007/s11263-016-0890-9
66. Pan Y, Mei T, Yao T, Li H, Rui Y (2016) Jointly modeling embedding and translation to bridge video and language. *In:* Proceedings of IEEE Conf on computer vision and pattern recognition (CVPR'16*)*, pp. 4594–4602, doi:https://doi.org/10.1109/CVPR.2016.497
67. Pei L, Ye M, Zhao X, Dou Y, Bao J (2016a) Action recognition by learning temporal slowness invariant features. J Visual Comput 32(11):1395–1404. https://doi.org/10.1007/s00371-015-1090-2
68. Pei L, Ye M, Zhao X, Xiang T, Li T (2016b) Learning Spatio-temporal features for action recognition from the side of the video. J SIViP 10(1):199–206. https://doi.org/10.1007/s11760-014-0726-4.
69. Perez M, Avila S, Moreira D, Moraes D, Testoni V, Valle E, Goldenstein S, Rocha A (2017) Video pornography detection through deep learning techniques and motion information. J Neurocomput 230: 279–293. https://doi.org/10.1016/j.neucom.2016.12.017
70. Pigou L, van den Oord A, Dieleman S, Herreweghe MV, Dambre J (2016) Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video. *Int J of Computer Vision* https://doi.org/10.1007/s11263-016-0957-7
71. Poppe R (2010) A survey on vision-based human action recognition. J Image Vision Comput 28(6):976–990. https://doi.org/10.1016/j.imavis.2009.11.014.
72. Revathi AR, Kumar D (2016) An efficient system for anomaly detection using deep learning classifier. *J of*. SIViP 11(2):1–9. https://doi.org/10.1007/s11760-016-0935-0
73. Rohrbach A, Rohrbach M, Schiele B (2015) The long-short story of movie description. *In: Proceedings of 37th German Conf on Pattern Recognition (GCPR'15)*, springer Int publishing, Aachen, Germany, pp. 209–221, doi:https://doi.org/10.1007/978-3-319-24947-6 17
74. Ronao CA, Cho S-B (2016) Human activity recognition with smartphone sensors using deep learning neural networks. J Expert Syst Appl 59:235–244. https://doi.org/10.1016/j.eswa.2016.04.032
75. Salakhutdinov R, Hinton GE (2009) Deep Boltzmann Machines. *In:* Proceedings of the twelfth Int Conf on artificial intelligence and statistics (AISTATS'09), Clearwater Beach, Florida, USA, pp. 448–455
76. Sarkar S, Venugopalan V, Reddy K, Ryde J, Jaitly N, Giering M (2016) Deep learning for automated occlusion edge detection in RGB-D frames. J Signal Process Syst 88(2):205–217. https://doi.org/10.1007/s11265-016-1209-3
77. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th Int Conf on Pattern Recognition (ICPR'04), vol 3, pp. 32–36
78. Sermanet P, Kavukcuoglu K, Chintala S, Lecun Y (2013) Pedestrian detection with unsupervised multistage feature learning. *In: Proceedings of the 2013 I.E. Conf on Computer Vision and Pattern Recognition (CVPR'13)*, IEEE computer society, Portland, Oregon, pp. 3626–3633, doi:https://doi.org/10.1109/CVPR.2013.465
79. Shahroudy A, Liu J, Ng T-T, Wang G (2016) NTU RGB+D: a large scale dataset for 3D human activity analysis. *In: Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, pp. 1010–1019, doi:https://doi.org/10.1109/CVPR.2016.115

80. Shen J, Wang M, Chua T-S (2016) Accurate online video tagging via probabilistic hybrid modeling, journal of. Multimedia Systems 22(1):99–113
81. Shuai B, Wang G, Zuo Z, Wang B, Zhao L (2015) Integrating parametric and non-parametric models for scene labeling. In: Proceedings of the IEEE Conf on computer vision and pattern recognition (CVPR'15). MA, USA, Boston, pp 4249–4258. https://doi.org/10.1109/CVPR.2015.7299053
82. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computing research repository (CoRR), vol abs/1409.1556
83. Singh S, Velastin SA, Ragheb H (2010) MuHAVi: a multicamera human action video dataset for the evaluation of action recognition methods. *In:* Proceedings of the 7th IEEE Int Conf on advanced video and signal based surveillance, pp. 48–55, doi:https://doi.org/10.1109/AVSS.2010.63
84. Singh S, Hoiem D, Forsyth D (2015) Learning a sequential search for landmarks. *In:* Proceedings of IEEE Conf on computer vision and pattern recognition (CVPR'15*)*, pp. 3422–3430, doi:https://doi.org/10.1109/CVPR.2015.7298964
85. Soomro K, Zamir AR (2014) Computer vision in sports, *Springer Int Publishing*, chapter "action recognition in realistic sports videos", pp. 181-208. doi:https://doi.org/10.1007/978-3-319-09396-3 9
86. Sun B, Xu Q, He J, Yu L, Li L, Wei Q (2016) Audio-video based multimodal emotion recognition using SVMs and deep learning. In: Proceedings of 7th Chinese Conf on pattern recognition (CCPR2016). Springer Singapore, Chengdu, pp 621–631. https://doi.org/10.1007/978-981-10-3005-5 51
87. Szegedy C, Liu W, Jia Y (2015) Going deeper with convolutions. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, 2015, pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
88. Tome D, Monti F, Baroffo L, Bondi L, Tagliasacchi M, Tubaro S (2016) Deep convolutional neural networks for pedestrian detection. J of Signal Processing: Image Communication 47:482–489. https://doi.org/10.1016/j.image.2016.05.007
89. Trumble M, Gilbert A, Hilton A, Collomosse JP (2016) Learning Markerless human pose estimation from multiple viewpoint video. In: Proceedings part III of computer vision (ECCV'16). Workshops, Amsterdam, The Netherlands, pp 871–878. https://doi.org/10.1007/978-3-319-49409-8 70
90. Varior RR, Wang G, Lu J, Liu T (2016) Learning invariant color features for person re-identification. IEEE Trans. on Image Proc. 25(7):3395–3410. https://doi.org/10.1109/TIP.2016.2531280
91. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K (2015a) Sequence to Sequence-Video to Text. *In:* Proceedings of IEEE Int Conf on computer vision (ICCV'15*)*, pp. 4534–4542, doi:https://doi.org/10.1109/ICCV.2015.515
92. Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney RJ, Saenko K (2015b) Translating videos to natural language using deep recurrent neural networks. *In: Proceedings of Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Denver, Colorado, USA, pp. 1494–1504
93. Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with Denoising autoencoders. *In: Proceedings of the 25th Int Conf on Machine Learning (ICML'08)*, ACM, Helsinki, Finland, pp. 1096–1103, doi:https://doi.org/10.1145/1390156.1390294
94. Wang D (2007) Challenges for computational intelligence, springer, berlin, Germany, chapter "computational scene analysis", pp. 163-191
95. Wang L, Sng D (2015) Deep learning algorithms with applications to video analytics for a Smart City: a survey. *CoRR*, https://arxiv.org/abs/1512.03131v1
96. Wang C, Yang H, Meinel C (2016) A deep semantic framework for multimodal representation learning. *J of*. Multimedia Tools Appl 75(15):9255–9276. https://doi.org/10.1007/s11042-016-3380-8.
97. Wu C, Cheng H-P, Li S, Li HH, Chen Y (2016) ApesNet: a pixel-wise efficient segmentation network. In proceedings of the 14th ACM/IEEE symposium on embedded Systems for Real-Time Multimedia (ESTIMedia'16), pp. 2-8, Pittsburgh, PA, USA, October 2016. ACM. ISBN 978-1-4503-4543-9. doi:https://doi.org/10.1145/2993452.2994306
98. Wu G, Liu L, Guo Y, Ding G, Han J, Shen J, Shao L (2017). Unsupervised deep video hashing with balanced rotation. In processing of the twenty-sixth international joint conference on artificial intelligence (IJCAI'17), pp. 3076-3082, Melbourne, Australia, august 2016. 10.24963/ijcai.2017/429
99. Xia D-X, S-Z S, Geng L-C, G-X W, Li S-Z (2016) Learning rich features from Objectness estimation for human lying-pose detection. J Multimedia Syst 23(4):515–526. https://doi.org/10.1007/s00530-016-0518-5
100. Xu W, Miao Z, Zhang J, Tian Y (2015) Learning Spatio-temporal features for action recognition with modified hidden conditional random field. *In: Proceedings, Part I of Computer Vision (ECCV'14)*, workshops, springer Int publishing, Zurich, Switzerland, pp. 786–801, doi:https://doi.org/10.1007/978-3-319-16178-5 55
101. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Elsevier J Comput Vis Image Underst 156:117–127. https://doi.org/10.1016/j.cviu.2016.10.010.

102. Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. *In:* Proceedings of IEEE Int Conf on computer vision (ICCV'15*)*, pp. 4507–4515, doi:https://doi.org/10.1109/ICCV.2015.512
103. Young P, Lai A, Hodosh M, Hockenmaier J (2014), From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans. of the* Association for Computational Linguistics (TACL), 2(Feb.):67–78.
104. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. *In: Proceedings Part I of the 13th European Conf Computer Vision (ECCV'14)*, Zurich, Switzerland, pp. 818–833, https://doi.org/10.1007/978-3-319-10590-1 53
105. Zhang Y, Li X, Zhang ZM, Wu F, Zhao L (2015) Deep learning driven Blockwise moving object detection with binary scene modeling. J Neurocomputing 168:454–463. https://doi.org/10.1016/j.neucom.2015.05.082.
106. Zhang W, Duan P, Gong W, Lu Q, Yang S (2016a) A load-aware pluggable cloud framework for real-time video processing. IEEE Trans Industrial Inf 12(6):2166–2176. https://doi.org/10.1109/TII.2016.2560802
107. Zhang X, Zhang H, Zhang Y, Yang Y, Wang M, Luan H, Li J, Chua TS (2016b) Deep fusion of multiple semantic cues for complex event recognition. IEEE Trans Image Proc. 25(3):1033–1046. https://doi.org/10.1109/TIP.2015.2511585
108. Zhao F, Huang Y, Wang L, Xiang T, Tan T (2016) Learning relevance restricted Boltzmann machine for unstructured group activity and event understanding. Int J Comput Vis 119(3):329–345. https://doi.org/10.1007/s11263-016-0896-3
109. Zhou B, Tang X, Wang X (2015) Learning collective crowd behaviors with dynamic pedestrian-agents. Int J Comput Vis 111(1):50–68. https://doi.org/10.1007/s11263-014-0735-3
110. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: towards story-like visual explanations by watching movies and reading books. *In: Proceedings of IEEE Int Conf on Computer Vision (ICCV'15)*, pp. 19-27, doi:https://doi.org/10.1109/ICCV.2015.11
111. Zhu F, Shao L, Xie J, Fang Y (2016a) From handcrafted to learned representations for human action recognition: a survey. J Image Vis Comput 55:42–52. https://doi.org/10.1016/j.imavis.2016.06.007.
112. Zhu X, Loy CC, Gong S (2016b) Learning from multiple sources for video summarisation. Int J Comput Vis 117(3):247–268. https://doi.org/10.1007/s11263-015-0864-3
113. Zuniga MD, Bremond F, Thonnat M (2013) Hierarchical and incremental event learning approach based on concept formation models. J of Neurocomputing 100:3–18. https://doi.org/10.1016/j.neucom.2012.02.038

Dr. Qaisar Abbas received Doctor of engineering degree (Ph.D) from the University of HUST at (Wuhan, China) in 2011. He has published more than 30 research papers in both reputable journals and conferences. He is currently working as an Assistant Professor in the Department of Computer Science at the Al-Imam Ibn Saud Islamic University. His research interests include: image processing, pattern recognition and computer vision.

Dr. Mostafa E. A. Ibrahim received PhD degree in electronics and communication Engineering in In January 2010 from Cairo University, Egypt in conjunctionwith Vienna University of Technology, Vienna, Austria under a channel supervision grant. He works as an assistant professor at Faculty of Engineering - Benha University, Egypt. Currently, he works as assistant professor in College of Computer and Information Sciences – Al-Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia. He is the author of more than 20 journal and conference research papers. His research interests include Image and Video Processing, Wireless Sensor Networks, Software Defined Radio.

Dr. M. Arfan Jaffar received the PhD degree of computer in 2009 from FAST-NUCES, Islamabad, Pakistan. He is currently an Assistant Professor. He is interested in conducting research in areas related with image processing, machine learning, computer vision, artificial intelligence and medical image processing. I am specially interests in biologically inspired ideas like genetic algorithms and artificial neural networks, and their soft-computing applications. Recently, I have been involved in solving image/video restoration problems using neuro-fuzzy techniques, homogeneous and heterogeneous combination of classifiers using genetic programming, optimization of shaping functions in digital watermarking and image fusion. Currently, I am working in these fields.