Video-based Human Action Recognition using Deep Learning: A Review

Hieu H. Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin

Abstract—Human action recognition is an important application domain in computer vision. Its primary aim is to accurately describe human actions and their interactions from a previously unseen data sequence acquired by sensors. The ability to recognize, understand and predict complex human actions enables the construction of many important applications such as intelligent surveillance systems, human-computer interfaces, health care, security and military applications. In recent years, deep learning has been given particular attention by the computer vision community. This paper presents an overview of the current state-of-the-art in action recognition using video analysis with deep learning techniques. We present the most important deep learning models for recognizing human actions, analyze them to provide the current progress of deep learning algorithms applied to solve human action recognition problems in realistic videos highlighting their advantages and disadvantages. Based on the quantitative analysis using recognition accuracies reported in the literature, our study identifies state-of-the-art deep architectures in action recognition and then provides current trends and open problems for future works in this filed.

Index Terms—Human action recognition, deep learning, CNNs, RNN-LSTMs, DBNs, SDAs.

1 Introduction

In recent years, human action recognition continues to be an increasingly active research in the computer vision community due to the interest in the development of many intelligent systems involving surveillance, control, and analysis. The main goal of this area is to determine, and then predict what humans do in a video or a sequence of images. There are many potential applications such as intelligent surveillance systems [1], [2], [3], human-computer interfaces [4], [5], health care [6], virtual reality [7], or security and military applications [8], [9].

1.1 Motivation

An action can be defined as a spatio-temporal sequence of human body movements. There are many ways to define an action from the literature [15], [16], [17]. Here, we consider "an action" as a single motion or complex sequences of motions performed by a single person or several humans. Actions are understood as episodic examples of human dynamics that have starting and ending temporal points. From the viewpoint of computer vision, given an image sequence that contains one or many actions, human action recognition attempts to label each frame or a sequence of frames with a corresponding name of an action. In general, human action recognition is a hierarchical process, where the lower levels are on human detection and segmentation. The objective of those levels is to identify the regions of interest (ROIs) corresponding to static or moving humans in video. The visual information of actions is extracted at the next level and represented by features. These features are then used for recognizing actions. So, recognizing an action from features can be considered as a classification problem. Early attempts at human action recognition systems used independent frame-by-frame analysis methods, e.g. shape matching techniques [18], while later research has focused on the spatio-temporal analysis of human motions.

A rapid increase in the number of researchers and techniques focusing on human action recognition has significantly improved its accuracy. However, action recognition is still a challenging problem due to many issues including the large intra-class difference, fuzzy boundary between classes, viewpoint, occlusion, appearance, influence of environments and recording settings [17], *in particular from realistic videos*. Moreover, to have a complete human action recognition system, we need a mating of several disciplines including psychology and ontology [20], [21].

1.2 Scope of the review, taxonomy and organization

Human action recognition is a big topic in computer vision. Many different approaches have been published in the last two decades [22]. In recent years, the advances of computer vision algorithms, especially machine learning, has opened up a new direction for researchers. Therefore, it is timely that progress in this field is reviewed. In this paper, we focus on surveying publications that use deep learning, a technique that has won numerous contests in machine learning including the recognition of human actions. Our

- *Hieu H. Pham and Louahdi Khoudour are with the Centre d'Etudes et d'Expertise sur les Risques, l'environnement la mobilité et l'aménagement (CEREMA), 1 Avenue du Colonel Roche, 31400 Toulouse, France. Address all correspondence to: hieu.ph@vinuni.edu.vn*
- *Hieu H. Pham is with the College of Engineering and Computer Science and VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam.*
- *Alain Crouzil is with the Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France.*
- *Pablo Zegers is with the Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Mons. Álvaro del Portillo 12455, Las Condes, Santiago 7620001, Chile.*
- *Sergio A. Velastin is with the Department of Computer Science, Applied Artificial Intelligence Research Group, University Carlos III Madrid, Av. Gregorio Peces-Barba 22, Colmenarejo, 28270 Madrid, Spain.*

main goal is to present a review of the work that has been reported in literature, compare the performance of deep learning based approaches and other existing work in order to identify its advantages and limitations. For instance, we divide deep learning approaches for action recognition based on their architectures. Many of the most important models are covered including Convolutional neural networks (CNNs), Recurrent Neural Network with Long Short-Term memory (RNN-LSTMs), Deep Belief Networks (DBNs) and Stacked Denoising Autoencoders (SDAs). In addition, some combination architectures will also be discussed.

The review is organized as follows: First, we introduce related surveys and publicly available datasets in Section 2. Then, we present the key deep learning architectures for human action recognition in Section 3, including the main ideas and mathematical models behind each architecture. Section 4 reviews the state-of-the-art in using deep models for human action recognition and related tasks. In Section 5, we give a quantitative analysis about the recognition accuracies of deep learning approaches and discuss their pros and cons. In that section, we also provide some promising directions for future research. Finally, we conclude our paper in Section 6.

2 Related surveys and publicly available datasets

2.1 Previous surveys

In this section, we first consider related earlier surveys in human action recognition. Looking at the major conferences and journals [23], [24], [25], [26], [27], several earlier surveys have been published. Aggarwal and Cai [28] reviewed methods for human motion analysis focusing on three major areas including: motion analysis involving human body parts, tracking a moving human from a single view or multiple cameras and recognizing human activities from image sequences. Moeslund and Granum [29] reviewed papers on human motion capture considering a general structure for systems analyzing human body motion as a hierarchical process with four steps: initialization, tracking, pose estimation and recognition. Wang *et al*. [30] presented a survey of work on human motion analysis, in which motion analysis was illustrated as a three-level process including human detection (low-level vision), human tracking (intermediate-level vision), and behavior understanding (high-level vision). Moeslund *et al*. [15] described the work in human capture and analysis based on 280 papers from 2000 to 2006, centered on initialization of human motion, tracking, pose estimation, and recognition.

Turaga *et al*. [16] considered that "actions" are characterized by simple motion patterns typically executed by a single person while "activities" are more complex and involve coordinated actions among a small number of humans and reviewed the major approaches for recognizing human action and activities. Poppe [17] focused on image representation and action classification methods. A similar survey by Weinland *et al*. [31] also concentrated on approaches for action representation and classification. Popoola and Wang [32] presented a survey focusing on contextual abnormal human behavior detection for surveillance applications. Ke *et al*. [33] reviewed human activity recognition methods for both static and moving cameras, covering many problems such as feature extraction, representation techniques, activity detection and classification. Aggarwal and Xia [34] presented a survey of human activity recognition based on 3D data, especially on using RGB and depth information acquired by consumer 3D sensors as the Kinect [12] sensor. Guo and Lai [35] gave a survey of existing approaches on still image-based action recognition.

Recently, Cheng *et al*. [36] reviewed approaches on human action recognition using an approach-based taxonomy, in which all methodologies are classified into two categories: single-layered approaches and hierarchical approaches. In addition, Vrigkas *et al*. [37] categorized human activity recognition methods into two main categories including "unimodal" and "multimodal". Then, they reviewed classification methods for each of these two categories. The survey of Subetha and Chitrakala [38] mainly focused on human activity recognition and human-object interaction methods. Presti *et al.* [39] provided a survey of human action recognition based on 3D skeletons, summarizing the main technologies, including both hardware and software for solving the problem of action classification inferred from time series of 3D skeletons. In addition, another survey was presented by Kang and Wildes [40]. It summarized various action recognition and detection algorithms, focused on encoding and classifying features. The latest survey on human action recognition was published in early 2016 by Herath *et al.* [41], in which the authors reviewed methods based on hand-crafted features and some deep architectures for recognizing actions. Table 1 summarizes previous surveys on human action and activity recognition published from 1997 to 2017 and reviewed in this paper. The surveys in the literature have shown that the common approaches in human action recognition have focused on using hand-designed local features such as HOG/HOF [42], [43], SIFT [44], or SURF [45]. In addition, these approaches are also extended for more robustness in video processing such as Cuboids [46], HOG3D [47]. To the best of our knowledge, there is no review on human action recognition based on deep learning techniques including comparisons of the performance of deep learning based approaches with traditional methods and with each other. Moreover, deep learning is a rapidly growing field, where novel algorithms appear in very short time duration and change the way of understanding and recognizing actions from visual data. That has prompted us to perform this work. Not only to provide a comparative analysis about the current state of human action recognition using deep learning algorithms, but also to point out the new trends in this field. Our survey will add to the latest reviews on human action recognition in the literature.

2.2 Benchmark datasets for human action recognition

With the increase in study of human action recognition algorithms, many datasets have been recorded and published for the research community. Much of the progress in action recognition was demonstrated on standard benchmark datasets. These datasets allow us to develop, evaluate and compare new methods. In this section we summarize the most important public datasets in the area. From the early dataset which contained very simple actions and acquired under controlled environments, to recent benchmark

TABLE 1
Summary of previous surveys and their key points ordered by year of publication.

Authors	Year	Main topics / Area of Interest
Aggarwal *et al.* [28]	1997	Human motion analysis, tracking.
Moeslund *et al.* [29]	2001	Motion initialization, tracking, pose estimation, recognition.
Wang *et al.* [30]	2003	Human detection, tracking, activity understanding.
Moeslund *et al.* [15]	2006	Human motion capture, action, and behavior analysis.
Turaga *et al.* [16]	2008	Recognizing human behavior.
Poppe [17]	2010	Feature extraction and classification of human action.
Weinland [31]	2011	Full-body action segmentation, and recognition.
Popoola *et al.* [32]	2012	Human motion analysis, abnormal behavior recognition.
Ke *et al.* [33]	2013	Human activity recognition from static and moving camera.
Aggarwal *et al.* [34]	2014	Human activity recognition from 3D and depth data.
Guo *et al.* [35]	2014	Human action recognition using still image.
Cheng *et al.* [36]	2015	Single-layered and hierarchical approaches for action recognition.
Vrigkas *et al.* [37]	2015	Human activity classification.
Subetha *et al.* [38]	2016	Human activity recognition and human-object interactions.
Presti *et al.* [39]	2016	Action classification based on skeleton.
Kang *et al.* [40]	2016	Human action recognition and detection.
Herath *et al.* [41]	2016	Human action recognition based on handcrafted features and deep learning approaches

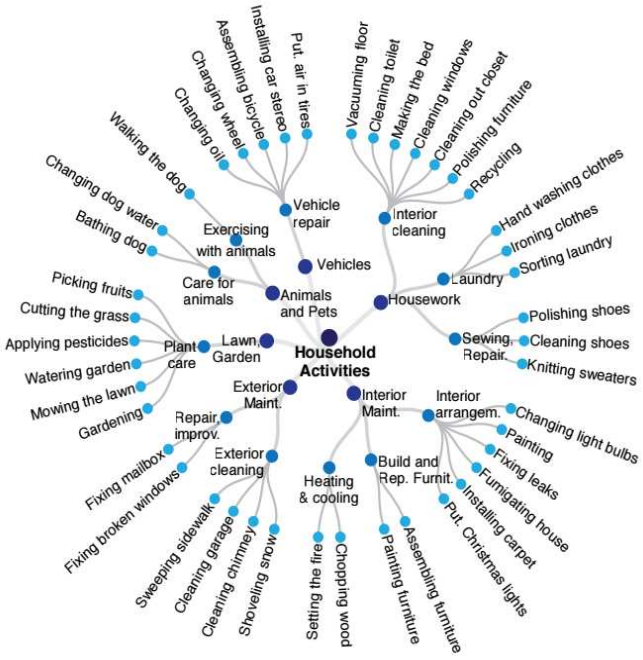


Fig. 1. Household activities from the ActivityNet [69] dataset.

datasets with thousands of video samples and millions of frames providing complex actions and human behaviors from the real world. Table 2 shows the datasets and their descriptions. To guide readers in the selection of the most suitable dataset for evaluating their work, we divide benchmarks into four categories including single action (category I), human-human interaction, human-object interaction and behavior (category II), surveillance (category III) and sport videos and other types (category IV). The complexity of each dataset depends on its recorded setting. For example, early benchmark datasets such such as KTH [48] or Weizmann [49] were made under laboratory conditions for idealized human actions: all of them are composed of simple and unrealistic actions and homogeneous background. Many methods have already achieved very high recognition rates on these datasets. Performances have increased over the years and have reached perfect accuracy, e.g., 100% on the Weizman [49] by Ikizler *et al.* [71] or Brahnam *et al.* [72]. In other words, we can say that the unrealistic datasets have already already solved by our action recondition systems. Another dataset named IXMAS has also been produced under laboratory conditions, but with multiple viewpoints [50].

After the success of the action recognition systems on benchmarks produced "in the lab", more complex benchmarks have been released. For instance, MSR Action3D [56], UT-Interaction [55], Daily-Activity-3D [57], Cornell Activity CAD-60 [73], Cornell Activity CAD-120 [63], VIRAT 2.0 [60], SBU-Kinect Interaction [64]. These datasets aim to provide challenging videos of human action under unconstrained environments with complex background and illumination conditions. However, they are not "real" actions. Then, many researchers have extracted realistic situations from movie or sport videos on social networks such as YouTube to make new realistic benchmark datasets. See for example Hollywood-1 [51], Hollywood-2 [52], YouTube [53], HMDB-51 [61], UCF-50 [66], UCF-101 [67], Sports-1M [68], ActivityNet [69]. The general approach in these datasets is to collect videos from "in-the-wild" sources with many clips and action classes. It is easy to see that several datasets are designed with deep learning algorithms in mind due to their very large scale. For example, in Sports-1M [68] there are around 1 million YouTube videos belonging to a taxonomy of 487 classes of sports, ActivityNet [69] provides more than 200 activity classes with 10,024 training videos, 4,926 validation videos and 5,044 testing videos. Figure 1 shows some actions in a class of the ActivityNet [69] dataset. These large scale datasets are an important premise for the development of deep learning methods because they require a large number of training data and tuning them on small and out-of-date datasets such as KTH [48] or Weizmann [49] leads to low performance. Most recently, Shahroudy *et al.* introduced NTU RGB+D dataset [70], a very large-scale RBD-D dataset for human action recognition. The NTU RGB+D dataset contains more than 56 thousand video samples, 4 million frames with 60 different action classes and performed by 40 different subjects. To our best knowledge, this is the newest dataset for action recognition tasks. Some samples of RGB, depth, joints, and IR image are shown in Figure 2. Experiments on realistic human action datasets have so far given limited results specially when dealing with a large and varied range of actions (e.g., table 3 shows recognition results methods on the HMDB-51 [61] dataset). Therefore, the current problem

TABLE 2
Some popular datasets for human action recognition (ordered by year of publication).

Dataset (category)	Author	Year	# Classes	Description
KTH (I)	Schuldt *et al.* [48]	2004	6	Walking, jogging, running, boxing, hand waving, and hand clapping.
Weizman (I)	Gorelick *et al.* [49]	2005	10	Walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping, jack skip.
IXMAS (I)	Weinland *et al.* [50]	2006	13	Check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, etc.
Hollywood-1 (II)	Laptev *et al.* [51]	2008	8	Answer phone, get out car, hand shake,hug person, kiss, sit down, sit up, stand up.
Hollywood-2 (II)	Marszalek *et al.* [52]	2008	12	Answer phone, drive car, eat, fight person,hug person, kiss, run,etc.
YouTube (II)	Liu *et al.* [53]	2009	8	Basketball shooting, volleyball spiking, soccer juggling, cycling, diving, etc.
MuHAVi (II)	Singh *et al.* [54]	2010	17	Walk turn back, run stop, punch, kick, shot gun collapse, pull heavy object, pick up through object, walk fall.
UT-Interaction (II)	Ryoo *et al.* [55]	2010	6	Shake-hands, point, hug, push, kick, and punch.
MSR Action3D (II)	Li *et al.* [56]	2010	20	High arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, etc.
Daily-Activity-3D (II)	Wang *et al.* [57]	2010	16	Drink, eat, read book, call cellphone, cheer up, sit still, toss paper, play game, etc.
MSR Action3D (II)	Li *et al.* [58]	2010	20	High arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, etc
Olympic Sports (IV)	Niebles *et al.* [59]	2010	16	High jump, long jump, triple jump, pole vault discus throw, hammer throw, etc.
VIRAT 2.0 (III)	Oh *et al.* [60]	2011	12	Loading an object to a vehicle, opening a vehicle trunk, getting into a vehicle, etc.
HMDB-51 (II)	Kuehne *et al.* [61]	2011	51	Smile, laugh, chew, talk, smoke, eat, drink, etc.
Cornell Activity CAD-60 (II)	Sung *et al.* [62]	2011	12	Rinsing mouth, brushing teeth, talking on the phone, drinking water, etc.
Cornell Activity CAD-120 (II)	Koppula *et al.* [63]	2012	20	Making cereal, taking medicine, stacking objects, reaching, moving, pouring, eating, etc.
SBU-Kinect Interaction (II)	Kiwon *et al.* [64]	2012	8	Approach, depart, push, kick, punch, exchange objects, hug, and shake hands.
LIRIS (II)	Wolf *et al.* [65]	2012	10	Discussion between two or more people, give an object to another person, put (take) an object into (from) a box (desk), etc.
UCF-50 (IV)	Reddy *et al.* [66]	2012	50	Diving, drumming, fencing, tennis swing, trampoline jumping, playing piano, etc.
UCF-101 (IV)	Soomro *et al.* [67]	2012	101	Horse riding, hula hoop, ice dancing, skiing, skijet, sky diving, etc.
Sports-1M (IV)	Karpathy *et al. et al.* [68]	2014	487	Juggling club, pole climbing, tricking, foot-bag, skipping rope, slack-lining, etc.
ActivityNet (II)	Heilbron *et al.* [69]	2015	203	Personal care, eating and drinking, household, caring and helping, working, socializing, etc.
NTU RGB+D (II)	Shahroudy *et al.* [70]	2016	60	Drinking, eating, reading, punching, kicking, hugging, etc.

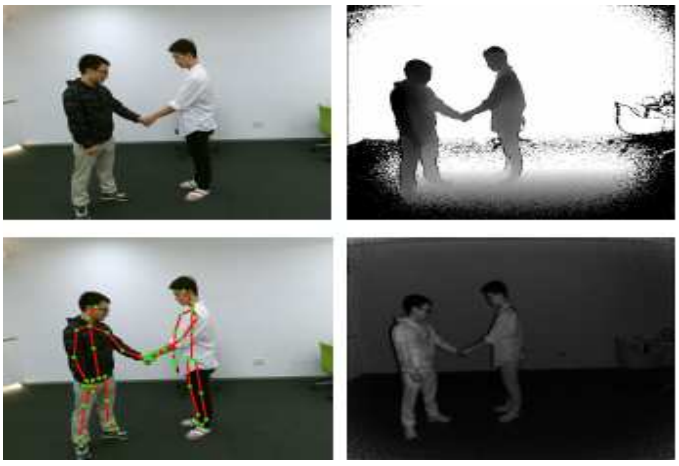


Fig. 2. Some samples of RGB, depth, skeleton and IR image from the NTU RGB+D dataset [70].

in action recognition that needs solving by computer vision community is recognizing *complex actions and behaviors* on *realistic scenarios*. Furthermore, there is also the need to build cost-effective real-world applications. This explains state-of-the-art benchmark datasets such as UCF-101 [67], HMDB-51 [61], Sports-1M [68], ActivityNet [69] and NTU RGB+D [70]. Researchers who want to evaluate their algorithms on state-of-the-art benchmark datasets can participate in the THUMOS challenge [74], a common benchmark for action classification and detection for computer vision community from around the world.

Recent developments in low-cost depth sensor technology have brought many opportunities for solving human action recognition tasks. RGB-D and skeleton data allow better understanding of the 3D structure of human body motion. Related to RGB-D and skeleton datasets, interested readers are referred to the recent work of Zhang *et al.* [75] and Firman [76]. In the next section, we will present deep learning-based approaches, one of the most interesting techniques in recent years in this field to answer the challenges highlighted here.

TABLE 3
Accuracy on the HMDB-51 dataset [61]

Approach	Author	Year	Acc.(%)
RGB + optical flow fusion	Wang *et al.* [77]	2016	**62.0**
F_{ST} + SCI fusion	Sun *et al.* [78]	2015	59.1
Two-stream CNN + SVM	Simonyan *et al.* [79]	2014	59.4
Improved dense trajectory	Wang *et al.* [80]	2013	57.2
W-flow dense trajectories	Jain*et al.* [81]	2013	52.1
Dense trajectory	Wang *et al.* [82]	2013	46.6
TRAJMF	Jiang *et al.* [83]	2012	40.7
Binary ranking models	Can *et al.* [84]	2013	39.0
MIP	Kliper-Gross *et al.* [85]	2012	29.2
GIST 3D	Solmaz *et al.* [86]	2012	29.2
Action bank	Sadanand *et al.* [87]	2012	26.9
C2	Kuehne *et al.* [61]	2011	23.0
HOG/HOF	Kuehne *et al.* [61]	2011	20.0

3 Deep Learning: A Short Presentation

For the sake of completeness, we present this section especially for readers who might not be very familiar with deep learning techniques. A full discussion is clearly outside the scope of this paper. Before discussing deep learning, we would like to briefly summarize the concept of machine learning (ML). ML is the branch of algorithms that allows computers to automatically learn from data. We can use ML systems for identifying objects in images, detecting spam emails, understanding text, finding genes associated with a particular disease and numerous other applications. The primary goal of ML is to develop general-purpose algorithms which are able to make accurate predictions in many different tasks. *In other words, ML algorithms try to match the density function that produced the data*. For example in classification problems, we need to identify a set of categories $\mathcal{C}$ from a space of all possible examples $\mathcal{X}$. Given any set of labeled examples $(\mathbf{x}_1, \mathbf{c}_1), ..., (\mathbf{x}_m, \mathbf{c}_m)$, where $\mathbf{x}_i \in \mathcal{X}$ and $\mathbf{c}_i \in \mathcal{C}$; the goal of ML is to find a concept $\mathcal{F}(\cdot)$ that satisfies $\mathbf{c}_i = \mathcal{F}(\mathbf{x}_i)$ for all i. In general, ML algorithms include two main steps. The first step is to define the representations of the raw data acquired by sensors, called "*feature extraction*". Then, these features are mapped into labels and called "*feature-to-label mapping*". This process produces an ML model that can be applied for new unlabeled data. Depending on the way of learning, (e.g., learn from labeled data or unlabeled data, learn with feedback or non-feedback), ML methods are typically classified into four categories including supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning.

Deep Learning (DL) is a class of techniques in machine learning. In 2012, DL became a major breakthrough in computer vision after the authors of AlexNet [88] achieved record performance on a highly challenging dataset named ImageNet. AlexNet [88] was able to classify 1.2 million high-resolution images from 1000 different classes with the best error rate. In general, DL methods are machine learning methods that consists and operate on multiple (multi-layer) levels of representation.

Various DL architectures have been proposed over the years (see Table 4) and have been shown to produce state-of-the-art results on many tasks, not least within human action recognition. In this section, we describe the most important DL architectures for human action recognition including Convolutional Neural Networks (CNNs or ConvNets) [92], [93], [94], [95], Recurrent Neural Networks with Long Short-Term Memory (RNN-LSTMs) [96], Deep Belief Networks (DBNs) [97], and Stacked Denoising Autoencoders (SDAs) [98].

3.1 Convolutional neural networks (CNNs)

After obtaining breakthrough results in object recognition with AlexNet [88] for the ImageNet project in 2012, CNNs become one of the most important deep learning models and play a dominant role for solving visual-related tasks. A CNN is a type of artificial neural network, designed for processing visual and other two-dimensional data. The main benefit of this model is that it operates directly on the raw data without any hand-crafted feature extraction. The idea of CNNs was firstly presented in 1980 by Fukushima [92] inspired by the structure of the visual nervous system [105]. CNN models continued to be proposed and developed e.g. by Rumelhart *et al*. [93], LeCun *et al*. [94]

TABLE 4
Popular deep learning architectures.

Architecture	Main articles
CNNs	Fukushima (1980) [92]; Rumelhart *et al.* (1986) [93]; LeCun *et al.* (1989) [94]; Krizhevsky *et al.* (2012) [95] Szegedy *et al.* (2015) [99] Simonyan *et al.* (2014) [100] He *et al.* (2015) [101]
RNN-LSTMs	Hochreiter and Schmidhuber [96]
DBNs	Hinton et al. [97]
DBMs	Salakhutdinov et al. (2006) [102]
Sparse Coding	Olshausen and Field (1996) [103]; Lee et al. (2006) [104]
SDAs	Vincent et al. (2008) [98]

and Krizhevsky *et al*. [95]. There are three key ideas behind a CNN architecture including *"local connections"*, *"shared weights"*, and *"pooling"*.

Local connections: In regular neural networks, each hidden layer consists of a set of neurons, where each neuron is fully connected to all neurons in the previous layer (Figure 3a). This model does not work efficiently when the input vector has a hight dimension. To make this more efficient, the idea is to reduce the number of connections between the first hidden layer to the input or each hidden layers to each other. Given an image as an input vector, every input pixel is not connected to every neuron in the first hidden layer. Instead, neurons in the first hidden layer are connected to localized regions of the input image. This sub-region is called the "*local receptive field*". For each local receptive field, we can identify a neuron in the first hidden layer as shown in Figure 3b.

Shared weights: For standard neural networks such as multilayer perceptrons [107] (MLP), the neurons of the first layer are computed by the dot product function of input vector $\vec{\mathbf{x}}$ and its weights $\vec{\mathbf{w}}$ where many different $\mathbf{w}_i$ values are used. In a CNN, we use a technique called "*weight sharing*" which is able to reduce the number of parameters $\mathbf{w}_i$. Specifically, in weight sharing, some of the parameters in the CNN model are constrained to be equal to each other [108]. Mathematically, the weight sharing technique can be performed using a convolution operator. In this process, we apply the filters to many local receptive fields in the input image, a "*feature map*" is generated by sliding a filter over the input matrix and computing the dot product. We can use many different filters and each of them will produce one feature map.

Pooling: "*Pooling*" also called "*subsampling*" is a sample-based discretization process. Its main goal is to reduce the dimensionality of the input representation while retaining the most important information in feature maps. This process reduces the computational cost and at the same time it provides a form of translation invariance. Max-pooling is performed by applying a max filter, it computes the max value of a selected set of output neurons from the feature map in the convolutional layer.

These concepts above can now be put together to form a complete CNN architecture that consists of a series of stages. The first few stages are structured by one convolutional layer and one max-pooling layer. These layers are followed by one or more fully connected layers at the top of the model. In a CNN, the convolution layer plays the role of a local feature extractor while the max-pooling layer merges semantically similar features into one. The last layer is a standard neural network working as a classifier (or a standard classifier such as an SVM). So the network learns a set of good features (c.f. with arbitrarily chosen or hand-crafted features) to use with a classifier. To prevent over-fitting and train the CNNs faster, Rectified Linear Units (ReLUs) and Dropout Layers [109], [110] have also been used. However, we do not discuss these two layers here as it is beyond the scope of this paper.

3.2 Recurrent Neural Networks with Long-Short Term Memory (RNN-LSTMs)

Recurrent Neural Network (RNN) is a good choice to model the complex dynamics of various actions in video because its architecture allows to store and access the long range contextual information of a temporal sequence. The main difference between an RNN and a multilayer perceptron is the presence of cyclical connections (Figure 4). This way, an RNN can learn to map from the entire history of previous inputs to each output [114]. However, they are very difficult to train due to the "*vanishing gradient problem*" [115], [116]. The Long Short-Term Memory (LSTM) approach [96] has been proposed to solve these problems. Figure 5 describes the LSTM's structure and its information flow.

RNNs not only are able to make use of previous context but also able to exploit future context as well. Bidirectional RNNs [117] has been proposed to do this by processing data in both directions with two separate hidden layers. All the information are then sent forwards to the same output layer. By replacing the nonlinear units in the Bidirectional RNNs architecture by LSTM cells, we can obtain Bidirectional-LSTM as shown in Figure 6. In subsection 4.2, we will see how to apply Bidirectional-LSTMs to model and recognize human actions in video.

3.3 Deep belief networks (DBNs)

DBNs [97] have been used successfully for many recognition tasks such as handwritten digits recognition [118], object recognition [119], or modeling human motion [120]. DBNs are probabilistic generative models that are constructed by stacking several restricted Boltzmann machines (RBMs) [121], [122] (Figure 7b). RBMs are shallow networks containing two layers: one layer of "*visible*" units that represents the input data and one layer of "*hidden*" units that learns to represent features. In an RBM architecture, all visible units of the visible layer are connected to all hidden units of the hidden layer, but there are no connections between two units of the same layer (Figure 7a). The standard type of RBM has binary-valued hidden and visible units, meaning that each unit can only be in one of two states, "0" or "1". The probability of setting a unit to "1" is a sigmoid function of its bias, weights on connections, and the state of other units. More detail, given a binary RBM

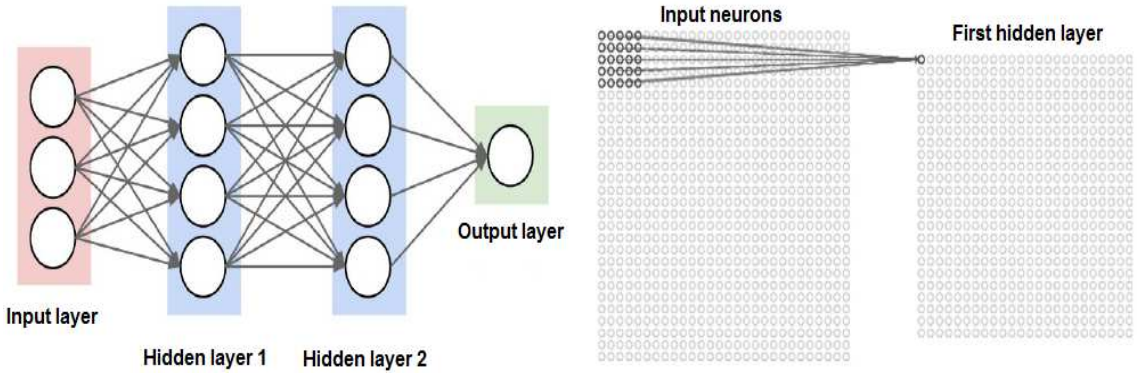


Fig. 3. **(a)** Illustration of a fully-connected model in a regular 3-layer neural network. **(b)** Illustration of the local receptive field in the input neurons [106].

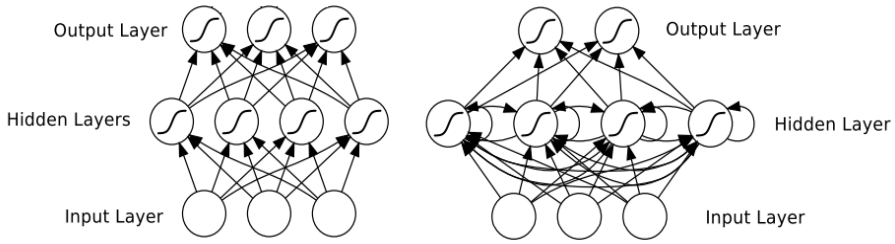
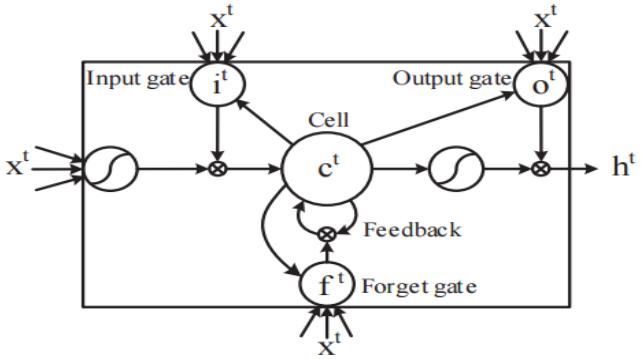


Fig. 4. Illustration of: **(a)** a multilayer perceptron and **(b)** a recurrent neural network.



$$
\begin{aligned}
i^t &= \sigma(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i) \\
f^t &= \sigma(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f) \\
c^t &= f^t c^{t-1} + i^t tanh(W_{xc}x^t + W_{hc}h^{t-1} + b_c) \\
o^t &= \sigma(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^t + b_o) \\
h^t &= o^t tanh(c^t)
\end{aligned}
$$

Fig. 5. Diagram of an LSTM unit [114]. A typical LSTM unit contains an input gate i^t, a forget gate f^t, an output gate o^t, an output state h^t and a memory cell state c^t. The information flow is described by the above equations where σ is the sigmoid activation; x^t is the input to the network at time t; all the matrices W are the connection weights between units. $\odot$ denotes element-wise product; and u^t denotes the modulated input function.

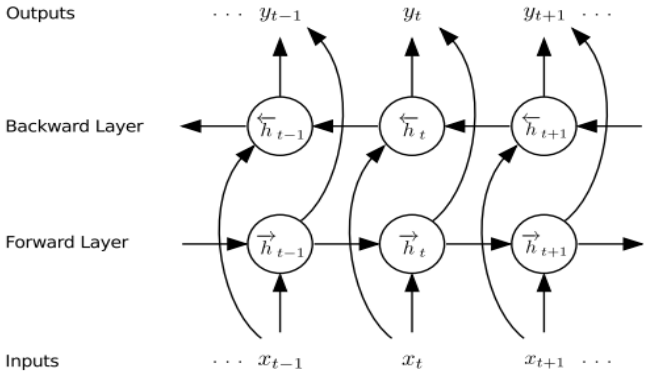


Fig. 6. Architecture of a Bidirectional-LSTM. The circular nodes represent LSTM cells.

with m visible units $\mathcal{V} = \{v_i\}, i \in (1, ..., m)$ and n hidden units $\mathcal{H} = \{h_j\}, j \in (1, ..., n)$, where v_i and h_j are the binary states of visible unit i and hidden unit j or $(v_i, h_j) \in (0, 1)^{m+n}$, the joint probability distribution for visible and hidden units is defined as [123]:

$$P(v_i, h_j) = \frac{1}{Z} e^{-E(v_i, h_j)} \tag{1}$$

where Z is the partition function computed by summing over possible pairs of (v_i, h_j):

$$Z = \sum_{v_i, h_j} e^{-E(v_i, h_j)} \tag{2}$$

and $E(v_i, h_j)$ is the energy function given by:

$$E(v_i, h_j) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i,j} v_i h_j w_{i,j}. \quad (3)$$

In function 3, a_i and b_j are biases, $w_{i,j}$ is the weight between v_i and h_j units. In a binary RBM model, there are no direct connections between visible units nor between hidden units. So, given the input data $\mathbf{v}$ through the visible units, the binary state of each unit h_j is 1 with probability:

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{i,j}). \quad (4)$$

Given a hidden vector $\mathbf{h}$, we can also reconstruct the states of a visible unit by:

$$p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j w_{i,j}) \quad (5)$$

where $\sigma(x)$ is the sigmoid function with form: $\frac{1}{1+e^{-x}}$. For estimating the weights $w_{i,j}$ and biases a_i, b_j, we use:

$$\frac{\partial \log p(\mathbf{v})}{\partial w_{i,j}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (6)$$

$$\frac{\partial \log p(\mathbf{v})}{\partial a_i} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \quad (7)$$

$$\frac{\partial \log p(\mathbf{v})}{\partial b_j} = \langle h_j \rangle_{data} - \langle h_j \rangle_{model} \quad (8)$$

The conditional distribution $p(h_j|\mathbf{v})$ in equation 4 shows that the hidden layer can be constructed by updating the state of units h_j when given a data vector $\mathbf{v}$. In practice, since all units in the hidden layer are conditionally independent given the visible layer, the state of each unit can be computed by using block Gibbs sampling [97]. This technique allows to update the state of all the units in parallel. As shown in Figure 7b, a DBN could be viewed as a stack of several RBMs. Therefore, training a DBN is performed through training each of its RBM. The work of Hinton *et al*. [97] provided an efficient procedure for training DBNs. In this process, the units of the current hidden layer are regarded as visible layer for the next hidden layer and training a DBN starts from the lowest RBM. The procedure is repeated layer-to-layer until the highest RBM is reached and known as the "*greedy layer-wise training strategy*". Each component (an RBM) of the DBNs acts as a feature extractor on inputs. It extracts "*low level*" features at the bottom hidden layer, as well as more "*abstract*" features at the higher hidden layers. To improve the performance of DBNs for classification tasks, the DBN model could be extended by adding a soft-max layer on the top of its architecture.

3.4 Stacked Denoising Autoencoders (SDAs)

SDA is another important technique in DL. It is an extension of a classical autoencoder [124] and was first introduced in 2008 by Vincent *et al*. [98]. The idea of an autoencoder is shortly described here: Given a set of data points $\mathbf{x} = \{x_1, x_2, ..., x_m\}$, map $\mathbf{x}$ to another set of data points $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ where $n < m$. From the compressed set $\mathbf{y}$, we reconstruct a set of $\tilde{\mathbf{x}}$, which approximates the original data $\mathbf{x}$. The mapping $\mathbf{x} \mapsto \mathbf{y}$ is called "*encoding*" and the mapping $\mathbf{y} \mapsto \tilde{\mathbf{x}}$ is called "*decoding*". Formally, the processes of encoding and decoding are performed as follows:

$$\mathbf{y} = W_1 x_i + b_1 \quad (9)$$

$$\tilde{\mathbf{x}} = W_2 y_i + b_2. \quad (10)$$

where $W_1 \in \mathbb{R}^{m \times m}, W_2 \in \mathbb{R}^{n \times n}$. Figure 8 illustrates the network architecture of a typical autoencoder. To achieve the goal of reconstructing $\tilde{\mathbf{x}}$ to approximate the original data $\mathbf{x}$, we minimize the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ by minimizing the function:

$$J(W_1, b_1, W_2, b_2) = \sum_{i=1}^{m} (\tilde{x}_i - x_i)^2. \quad (11)$$

From equations 10 and 11, we have:

$$J(W_1, b_1, W_2, b_2) = \sum_{i=1}^{m} (W_1 W_2 x_i - 1) x_i + b_1 W_2 + b_2)^2. \quad (12)$$

SDAs are constructed by stacking several autoencoders together to create a "*deep*" architecture where the weights are fine-tuned with a back-propagation algorithm [125]. The "*unsupervised pre-training*" of each autoencoder is performed in a greedy layer by layer manner. Once the a SDAs is learnt, its output will then be used as the input representation of a supervised learning algorithm for recognition tasks.

4 Human action recognition approaches based on DL

This section reviews current studies of deep learning on human action recognition. We categorized publications based on the proposed taxonomy, including: human action recognition based on CNNs (subsection 4.1); human action recognition based on DBNs (subsection 4.3); human action recognition based on SDAs (subsection 4.4); human action recognition based on RNN-LSTMs (subsection 4.2), and some other architectures (subsection 4.5).

4.1 Human action recognition based on CNNs

Many works on human action recognition and related tasks based on DL models have been proposed and reported in the literature. Among them, one of the most used deep models is CNNs (see subsection 3.1) and its extensions. Researchers have successfully applied CNN-based architectures for many visual tasks such as people detection and tracking [126], [127], [128], pose estimation [129], [130], [131], [132], [133], [134], action recognition [79], [135], [136], [137], [138], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], event detection and crowded scene understanding [159], [160], [161], [162]. Early work on applying CNNs was made in 1995 by Nowlan *et al*. [129] for hand tracking and recognizing. In their work, a CNN model is proposed to locate the hand and recognize whether it is close or open with accuracies of 99.7% and 99.1% on a dataset of 900 video images from 18 different subjects for each task. However, the complex structured backgrounds of images may have a significant impact on the recognition accuracy. Starting from the work of Fukushima [92], Giese and Poggio

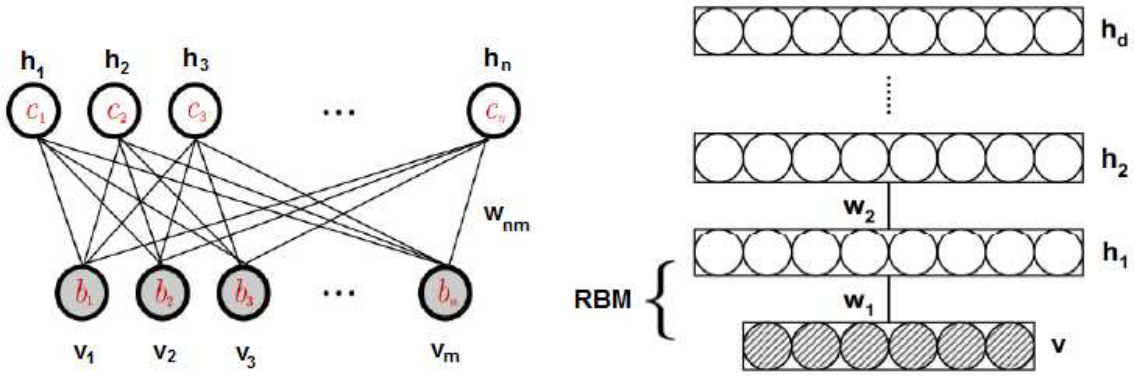


Fig. 7. **(a)** An example of a RBM with m visible units and n hidden units. **(b)** The schematic overview of a deep belief networks composed of d RBMs. W_1,W_2,...,W_h are the weights matrices between the connections.

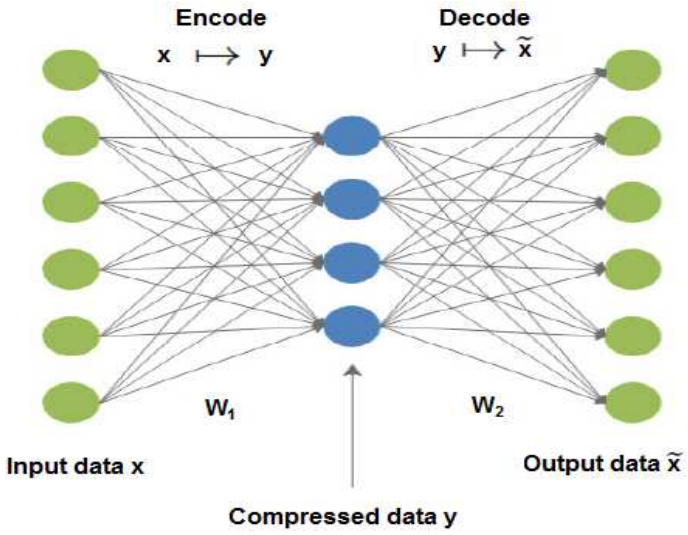


Fig. 8. The typical structure of an autoencoder.

[135] proposed a hierarchical feedforward architecture for the recognition of biological movements such as walking, running or various full-body actions. In a related paper, Sigala *et al*. [136] also developed a hierarchical model for detecting a walker based on the use of the neural detectors that are able to extract motion features with different levels of complexity. Jhuang *et al*. [137] proposed an extension model from the work of Giese and Poggio [135] for the recognition of actions from video sequences (Figure 9).

In 2007, Kim *et al*. [138] used a modified CNN model and a weighted fuzzy min-max neural network (WFMM) [163] for human action recognition. In their paper, the CNN generates a set of feature maps from the pretreated data and a WFMM [163] plays the role of a classifier. Normally, the CNNs have been primarily applied on two-dimensional data (2D-CNN) in which these models compute features from the spatial dimensions only. In order to exploit the temporal information of human motion, Ji *et al*. [139] presented a novel three-dimensional convolutional neural network (3D-CNN) architecture for recognizing human action. This architecture used 3D kernels in the convolution stages to extract motion features from both spatial and temporal dimensions. This improvement can be applied to contiguous frames in video to extract multiple features.

Experiments on TRECVID-2008 [164] datasets have shown that this model outperforms the frame-based 2D-CNN model and two other methods proposed by Lazebnik *et al*. [165] and Yang *et al*. [166] which follow the state-of-the-art bag-of-words (BoW) [167]. Motivated by Ji *et al*. [139], Wang *et al*. [140] has also used 3D-CNN for building a deep architecture for human activity understanding using RGB-D data. In addition, Tran *et al*. [142] investigated in detail the 3D-CNN model and showed that it outperforms the 2D-CNN in modeling human motion information on various recognition tasks. Moreover, Tran *et al*. [142] found that the best kernel length for 3D-CNN is $3 \times 3 \times 3$ size. Varol *et al*. [168] also used 3D-CNN for learning action representation in video but with long-term temporal convolutions at the input layer. This study demonstrated that this solution can significantly improve the performance on state-of-the-art action recognition datasets. A visible disadvantage of 3D-CNN model is the increasing number of parameters of the network. To reduce the complexity of the model, Sun *et al*. [78] proposed a factorized spatio-temporal convolutional network that factorizes the 3D convolution kernels into 2D spatial kernels and followed by 1D temporal kernels.

After finding more efficient ways to train CNNs using GPU computing [169] and the success of AlexNet [88] in the ILSVRC-2012 competition, much work on human action recognition has been published. Ijjina *et al*. [170] recognize human actions in videos by using the standard action bank [171] as a feature detector and a CNN as a classifier. Gkioxari *et al*. [132] gave state-of-the-art performance for predicting actions on the PASCAL VOC 2012 detection and action train set [172] by using the same CNN architecture as AlexNet [88] and extracting region proposals on input image with R-CNN technique [173]. Chéron *et al*. [134] designed a new CNN-based pose descriptor for human action recognition from RGB and optical flow information. Two distinct CNNs with an architecture similar to AlexNet [88] have been used.

The two-stream convolutional network proposed by Simonyan and Zisserman [79] has shown strong performance for human action recognition in videos. This model is a two-stream architecture including the spatial stream and

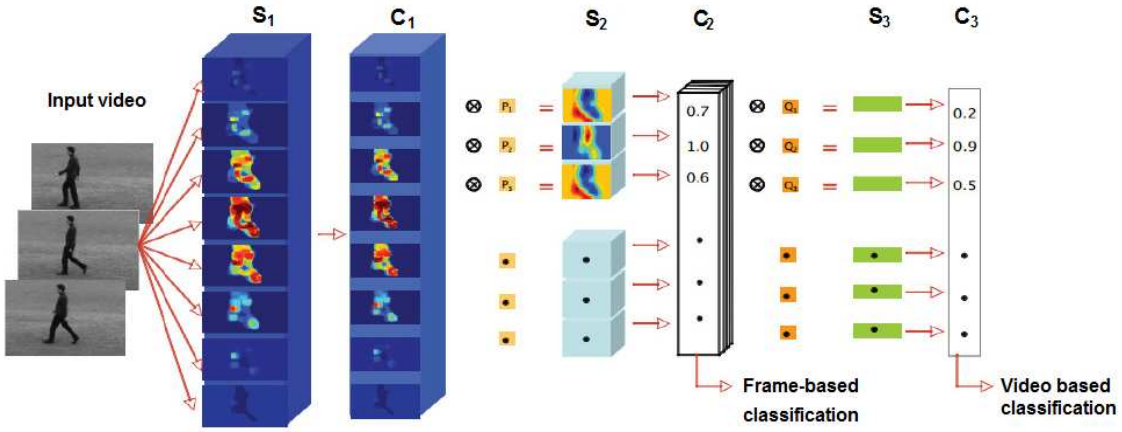


Fig. 9. The framework for recognizing human action proposed by Jhuang *et al.* [137]. Given a gray-value video sequence as input data, the S_1 stage locates the object in image frame by using spatio-temporal filters. Each C_1 unit is computed by applying a local max over for each S_1 unit for down-sampling. From the C_1 stage, we perform a template matching operation for identifying intermediate-level features of the model. The C_2 stage is constructed by computing the global max over each S_2 unit. The high-level features are extracted in S_3 through a template matching and the C_3 features are computed from S_3 using the same way like computing C_2. The last stage is a linear multiclass SVM classifier that is able to recognize the actions using the C_3 features as input.

the temporal stream where each stream is executed by a CNN. The first stream recognizes actions from a single frame, while the second recognizes actions from motion information of multi-frame optical flow. These two streams are then combined for the classification task. The experimental results show that using multi-frame optical flow for training model allows to achieve very good performance with limited training data. This architecture has been seen as the most effective approach of applying DL to action recognition with limited training data. Inspired by the work of Simonyan and Zisserman *et al*. [79], many different authors have developed two-stream convolutional networks for solving action recognition problems, e.g., Wang *et al.* [174], [175], [176], Xiong *et al.* [177]. Unlike the two-stream architecture developed by Simonyan and Zisserman *et al*. [79], Liu *et al*. [145] added a module called stCNN (Spatio-Temporal Convolutional Neural Network) to the standard CNN model for exploiting motion and content-dependent features concurrently. Experiments on KTH [48] and UCF-101 [178] datasets showed that the recognition accuracy for motion-content combined was better when compared with motion alone. Singh *et al*. [148] addressed the problem of understanding egocentric activities by using a three-stream CNN architecture. More specifically, the authors proposed a framework for the recognition of wearer's actions. First, a CNN model called "*Ego Convnet*" is trained for learning features from egocentric cues including hand mask, head motion, and a saliency map. Then, Ego Convnet is extended by adding two more streams corresponding to spatial and temporal streams as the model proposed by Simonyan and Zisserman *et al*. [79]. Experiments showed that the model with the Ego Convnet stream alone achieved state-of-the-art accuracy on different egocentric videos datasets. In addition, the three-stream architecture.

In a recent study, Wang *et al.* [77] divided an input video consisting of t frames $\mathcal{X} = \{x_1, x_2, ..., x_t\}$ into two sets: the precondition state frames $\mathcal{X}_p = \{x_1, ..., x_{z_p}\}$ and effect state frames $\mathcal{X}_e = \{x_{z_e}, ..., x_t\}$. The Siamese network architecture has been designed for learning action features. In fact, this is a two-stream CNN models where the first stream is trained on the precondition state frames and the second is trained on the effect state frames as shown in Figure 10.

Advances of 3D sensors such as Microsoft Kinect [12] brings up new opportunities in computer vision, even though they tend to be limited to small indoor environments. RGB-D data is able to provide additional information about human motion. Take advantage of depth maps provided by Kinect sensors, Wang *et al.* [179] proposed the use of CNNs to learn actions from sequences of depth maps. Given a sequence of depth maps, 3D points are created and three Depth Motion Maps (DMMs) are constructed by projecting the 3D points to the three orthogonal planes. Three CNNs are constructed based on AlexNet architecture [88] to extract motion features from each DMM and then classify them into classes. This study is extended in [180] and [143]. State-of-the-art results have been shown on MSR Action3D Dataset [56], an extension of the MSR Action3D Dataset, UTKinect-Action Dataset [181], and MSR-Daily-Activity3D Dataset [57]. Dobhal *et al*. [144] also used depth information and a CNN for recognizing human activities. Given a sequence of 2D images, background subtraction is performed. All binary frames are then stacked into a single image called Binary Motion Image (BMI) which contains the flow of the action and is used as the input for the CNN in training and testing phases. The CNN's architecture is same the architecture introduced by LeCun and Bengio [182]. Their approach is extended for extracting BMI from 3D depth maps and achieved competitive performance on Weizmann [49] and MSR Action3D Dataset [56]. The key ideas behind CNNs such as "*local connections*" or "*shared weights*" and the improvements on GPU computing technology have enabled CNNs to train on very large scale datasets. Karpathy *et al*. [183] studied the performance of CNNs by trying to predict and classify on Sports-1M [184] dataset which consists of more than one million sport videos. Multiresolution CNN architecture with two separate streams of

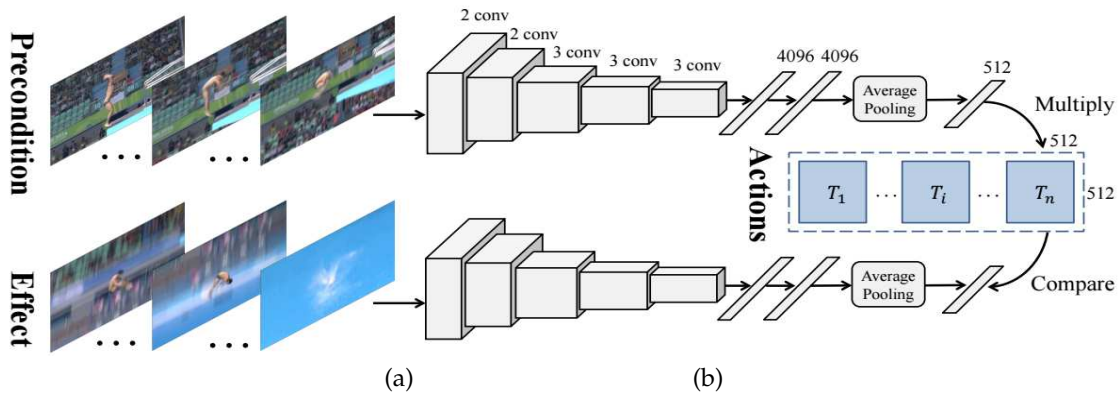


Fig. 10. The Siamese network architecture proposed by Wang *et al.* [77].

processing has been proposed for reducing training time. The results show that CNNs are capable of learning powerful features and significantly outperform the feature-based baseline. Figure 11 shows some examples of predictions on Sports-1M dataset [184].

In addition to RGB-D information, the acquisition of the skeleton data has become easier with the support of RGB-D sensor. Mo *et al*. [147] presented a deep model which combines a CNN with a multilayer perceptron [107] for recognizing the human activities based on skeleton data acquired from a Kinect sensor [12]. The method achieves a recognition accuracy of 81.8% on the CAD-60 dataset [73]. Skeleton data has been used by Wang *et al.* [185]. Firstly, the spatio-temporal information of the joint trajectories is encoded into color images. Then, a CNN based on the AlexNet architecture [88] is used to learn the color distribution and to classify actions. The idea of encoding the spatio-temporal information of a skeleton sequence into color texture images and using a standard CNN architeture such as AlexNet [88] can also be found in the work of Hou *et al.* [186].

Among the local space-time features, trajectories are one of the best ways to describe motion [80], [187], [188]. Wang *et al*. [141] combined the benefits of improved trajectories [80] and two-stream CNN architecture from the work of Simonyan and Zisserman *et al*. [79] for designing an effective representation of video feature called "*Trajectory-Pooled Deepconvolutional Descriptor (TDD)*". The experimental results show that this framework has obtained state-of-the-art performance for recognizing action on the UCF-101 [178] and HMDB51 datasets [189]. Inspired by the work of Wang *et al*. [141], Cao *et al*. [146] proposed a novel 3D deep convolutional descriptor based on joint positions named "*Joints-Pooled 3D Deep Convolutional Descriptors (JDD)*". Promising experimental results on sub-JHMDB [190], Penn Action [191], and Composable Activities [192] have shown that using joint-based descriptor with deep model is an effective and robust way for understanding human action. A new powerful and simple representation of videos for action recognition based on DL, especially CNNs, called "Dynamic Image" has been presented in the work of Bilen *et al.* [193]. The idea of this paper is summarizing the video content in a single standard RGB image, then using a pre-trained CNN model such as AlexNet [88] on a dataset of dynamic images with fine-tuning technique. The authors also proposed to train CNN from scratch by generating more dynamic images from video segments. Experiments on HMDB-51 and UCF-101 datasets shown the effectiveness of the "Dynamic Image" representation.

Very deep convolutional neural networks such as VGGNet [100], GoogLeNet [99] have achieved significant success for object recognition and classification tasks. Several authors started to exploit these architectures for action recognition problems. Wang *et al.* [194] introduced very deep two-stream CNNs for action recognition based on VGG-16 (VGGNet C with 13 convolutional layers and 3 fully-connected layers) and GoogLeNet [99] with 22-layers network. Feichtenhofer *et al.* [195] proposed a CNNs-based novel architecture for spatio-temporal fusion of two stream networks in which the deep CNN model VGG-M-2048 [196] and very deep model VGG-16 [100] have been used. The performance comparison between deep (VGG-M-2048) and very deep (VGG-16) models on UCF-101 and HMDB-51 datasets shown that the use of deeper networks improves performance. In addition, GoogLeNet [99] and VGGNet [100] have also been used to design the two-stream CNNs in the work of Wang *et al.* [197]. Fernando *et al.* [198] trained VGG-16 [100] on HMDB-51 [189], UCF-101 [178] and Hollywood2 [199] datasets for obtaining VGG-16 CNN features. The CNN feature vectors are then encoded by a method called "hierarchical rank pooling". This method allows encoding the temporal dynamics of a video sequence for action recognition. A video sequence is encoded at multiple levels in which the output of the each level is a sequence of vectors which captures higher-order dynamics of its previous level. The final representation can be used to learn an SVM classifier for activity recognition as descriptors.

Very recently, the residual learning framework (ResNets) [101], a state-of-the-art CNN and the deepest CNN model at the moment has been exploited for human action recognition by Feichtenhofer *et al.* [200]. In the main ResNet paper [101], authors have suggested different architectures of ResNet with 18, 34, 50, 101, 152, and 1202 layers. The underlying network with 50 layer ResNet has been used in the work of Feichtenhofer *et al.* [200] to design a two-stream network. Experiments shown a state-of-the-art performance

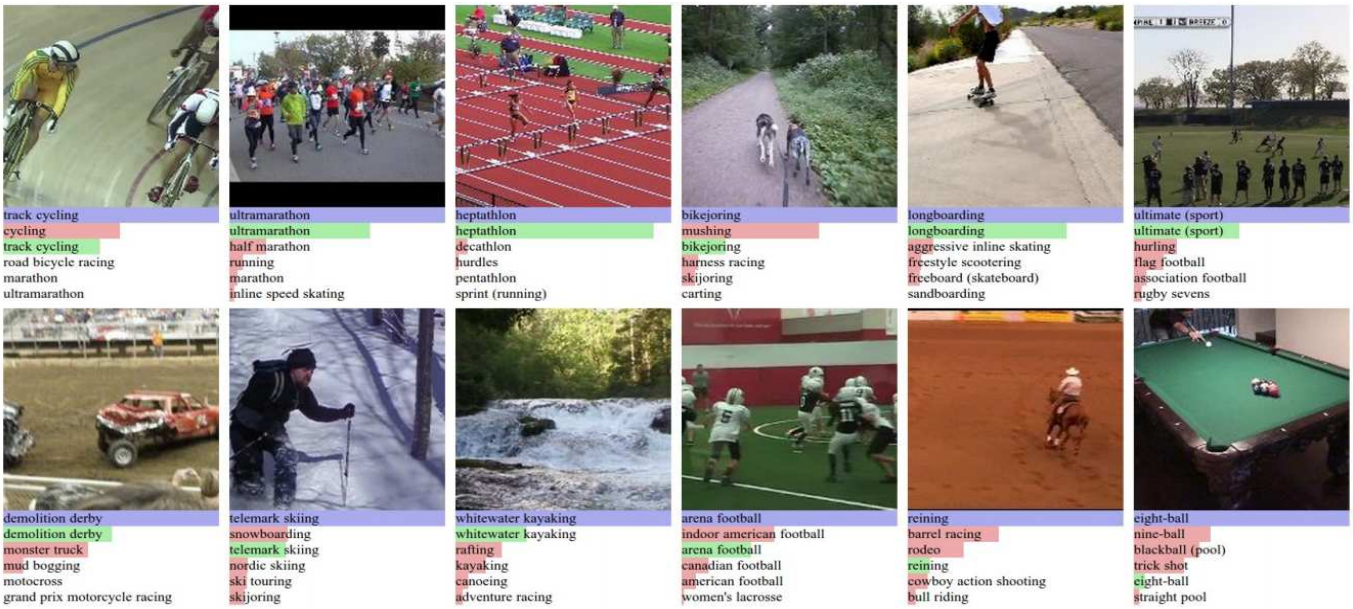


Fig. 11. Action prediction on Sports-1M dataset [184]. The first row indicates ground truth label and the bars below show model predictions. Green and red distinguish correct and incorrect predictions, respectively. [183].

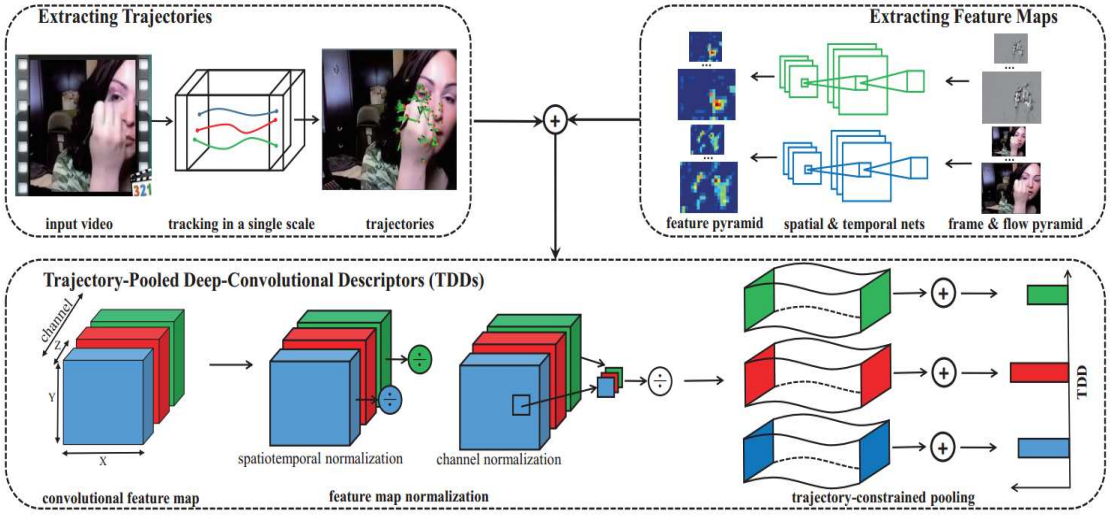


Fig. 12. The framework for action recognition proposed by Wang *et al.* [141]. Given an input video, the model extracts trajectories. Multiscale convolutional feature maps are extracted by a CNN at the same time. Trajectory Pooled deep-Convolutional Descriptors (TDDs) are then estimated from a set of improved trajectories and convolutional feature maps.

TABLE 5
Performance comparison of deep model VGG-M-2048 with very deep model VGG-16 on the UCF-101 [178] and HMDB-51 reported by Feichtenhofer *et al.* [195].

Dataset	**UCF101**		**HMDB51**	
Model	VGG-M-2048	VGG-16	VGG-M-2048	VGG-16
Spatial	74.22%	82.61%	36.77%	47.06%
Temporal	82.34%	86.25%	51.50%	55.23%
Spatio-Temporal	85.94%	90.62%	54.90%	58.17%

on UCF-101 [178] and HMDB51 [189] datasets.

CNNs are also applied for solving more complex tasks related to human action recognition such as event detection, crowd analysis or behavior prediction. Xu *et al.* [201] proposed a CNN-based approach for event detection on the large scale video datasets, i.e., TRECVID MEDTest 13 [**?**] and TRECVID MEDTest 14 [**?**]. The encoding technique is used for improving the performance and the video representation is compressed for reducing the computation costs. Gan *et al.* [159] presented a CNN-based framework called "DevNet" for detecting events in videos. Shao *et al.* [160] built a large-scale crowd dataset called WWW Crowd Dataset and designed a CNN model to learn and recognize attributes prediction in crowd video. A similar study can be found in the work of Castro *et al.* [161]. Xiong *et al.* [162] presented a CNN-based approach which contains two-channels CNN for recognizing complex events from static images. This

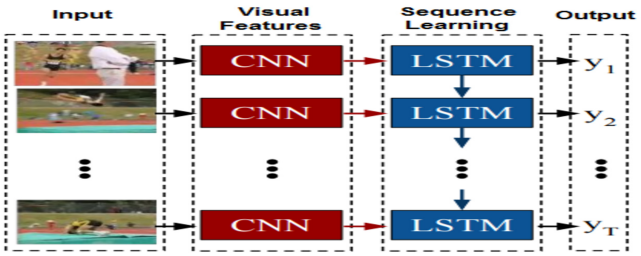


Fig. 13. Deep learning framework combining CNN and RNN-LSTM for action recognition proposed by Donahue *et al.* [205]

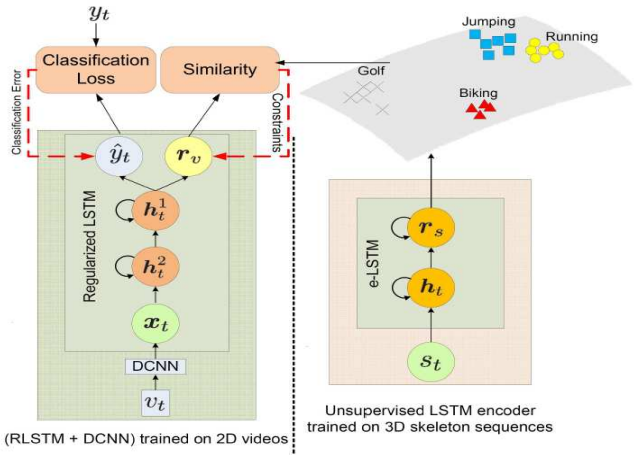


Fig. 14. The parallel deep learning architecture with RNN-LSTM proposed by Mahasseni *et al.* [219].

system is able to detect the objects, predict events, and has given a state-of-the-art result on a very large dataset.

4.2 Human action recognition based on RNN-LSTMs

As pointed out in subsection 3.2, the main advantage of RNN-LSTMs is the capacity to model the long-term contextual information of temporal sequences. This advantage puts RNN-LSTM at one of the best sequence learners for time-series data including visual information of human action. Grushin *et al.* [202] has demonstrated the robustness of the LSTM network's performance on the human action recognition task with the hand-crafted feature HOF [43]. As discussed in subsection 4.1, CNNs has been shown its effectiveness in learning features from raw data. Therefore, the works of Baccouche *et al.* [203], Ng *et al.* [204], Donahue *et al.* [205], Giel *et al.* [206], Sharma *et al.* [207], Ibrahim *et al.* [208], Singh *et al.* [209], Li *et al.* [210], Wu *et al.* [211], Wang *et al.* [212], Chen *et al.* [213] tackle the question of understanding human actions by combining a CNN and an RNN-LSTM network. The general idea of these papers is to use the standard CNN models such as AlexNet [88], VGGNet [100], or GoogLeNet [99] for extracting motion features from input video. Then, RNN-LSTM network is connected to the output of the CNN to classify sequences using learned features. Figure 13 shows an example of using CNN and RNN-LSTM for human action recognition from the work of Donahue *et al.* [205] . While all the work above just uses RNN-LSTMs as a sequence classification, several studies have proposed the use of RNN-LSTMs as an end-to-end learning framework for skeleton based action recognition. E.g., the work of Du *et al.* [214], Song *et al.* [215], Zhu *et al.* [216], Li *et al.* [217], Liu *et al.* [218]. RNN-LSTMs learn directly motion features and classify them into classes from 3D human-skeleton sequences provided by depth sensors. Experiments on the state-of-the-art datasets demonstrate the effectiveness of these methods. In another study of Mahasseni *et al.* [219] used a parallel architecture to recognize actions with multi-source data. A RNN-LSTM is trained in unsupervised manner on 3D human-skeleton sequences. In the same time, another RNN-LSTM with a CNN is trained on 2D videos. The outputs are then compared to improve the ability of the system.

4.3 Human action recognition based on DBNs

DBNs have become popular DL models after the key paper by Hinton *et al.* [97] presented in 2006. A comparative evaluation by Tang [220] showed that DBNs seem ideal for semi-supervised learning, in which we do not need much labeled data. Early work on DBNs was successfully applied for handwritten digits recognition [97] and object recognition [119], [221]. In 2007, Taylor *et al.* [120] extended the RBM model by connecting two more visible layers to the hidden layer for modeling human motion. The new model, called the conditional RBM (cRBM) allows to find a single set of parameters that simultaneously capture several different kinds of motion after training on skeleton data. Then, the authors successfully constructed a DBN from cRBMs. Experiments on two motion datasets have demonstrated that this model is able to effectively learn different kinds of motion, as well as the transitions between these kinds.

In another research, Zhang *et al.* [222] used a modified DBN model for recognizing human actions in real-time from skeleton data. To achieve this goal, the authors used cRBMs as proposed by Taylor *et al.* [120] to create the new DBN architecture with two hidden layers as shown in Figure 15. The proposed model is trained and tested by using the skeletal representation of MSR Action3D [56] and MIT datasets [223]. Results show that the recognition accuracy depends on the number of frames. For example, on the MIT datasets [223], the accuracy when using one frame is 98.34%. Meanwhile, when the number of frames is more than 30, accuracy can reach 100%. Foggia *et al.* [224] proposed a DBN-based method for recognizing human actions with depth images. A DBN model is constructed as shown in Figure 16. Three types of well-known feature including the Average Depth Image (ADI), the Motion History Image (MHI), and the Depth Difference Image (DDI) are computed and encoded as low-level data representation in the first layer. The high level representation is then extracted by the proposed model for recognition task. The achieved results on MIVIA [225] and MHAD [226] datasets are very promising. Ali and Wang [227] presented a framework based on DBN to recognize and identify human actions. To speed up learning time, the Fast Fourier Transform (FFT) [228] technique is used for

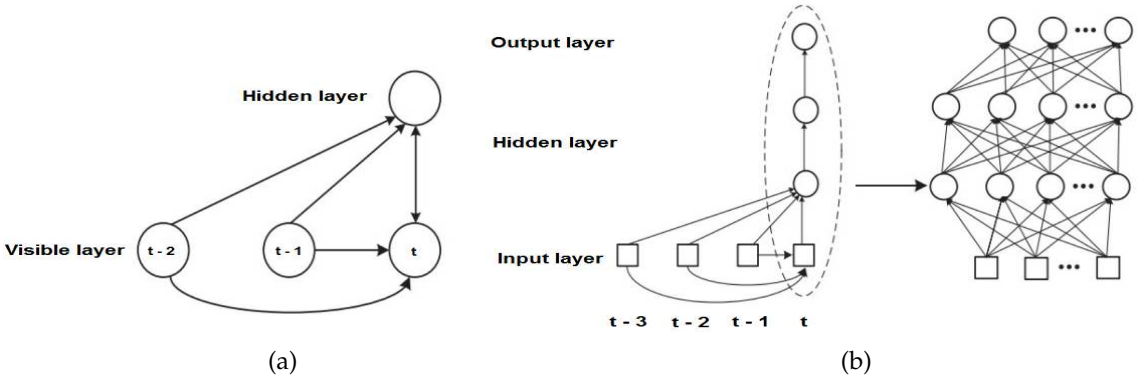


Fig. 15. **(a)** The **cRBM** model proposed by Taylor et al. [120]. **(b)** A modified DBN model designed by Zhang et al. [222].

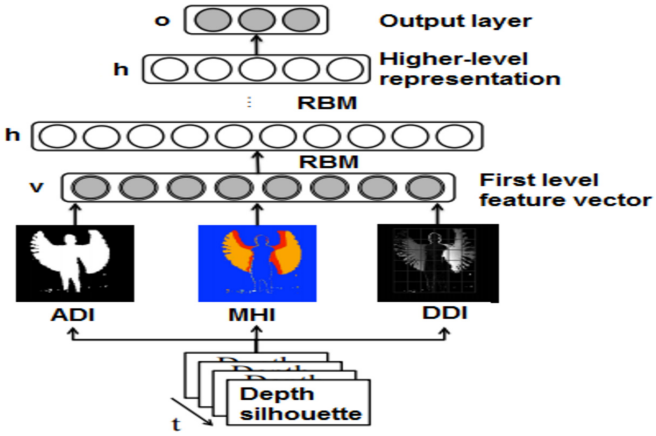


Fig. 16. An overview of the DBN architecture for human action recognition proposed by Foggia et al. [224]. Three derived images (ADI, MHI,DDI) are computed from depth images and feed into the first level of the network. A more abstract representation is obtained at higher level. Finally, the classification is done using a feed-forward neural network.

converting images to the frequency domain. The model is first pre-trained with KTH dataset [48] and then is used for predicting actions. Experiments showed that the proposed model is better than all published approaches in the literature. More details about this comparison are shown in Table 6. We can also find in the literature some other human action recognition applications based on DBNs. For example, Nam *et al*. [235] employed a DBN for developing a real-time human activity recognition using 3D joint positions from RGB-D sensor. The achieved results from these studies confirmed that DBN-based approaches are a good choice for many human action recognition problems.

4.4 Human action recognition based on SDAs

As pointed out in subsection 3.4 SDAs can be trained to reconstruct the input from a corrupted version of it. The first

TABLE 6
Average recognition accuracy of human action on KTH [48] dataset

Approach	Accuracy
DBN by Ali et Wang [227]	**94.3%**
ISA + Norm-thresholding by Le et al. [229]	93.9%
Harris3D [230] + HOF [43] by Wang *et al.* [231]	92.1%
Harris3D [230] + HOG/HOF [43] by Wang *et al.* [231]	91.8%
HMAX [137]	91.7%
3D CNN [139]	90.2%
Cuboids [46] + ISA [230]	90.0%
GRBM [232]	90.0%
Dense + HOF [43] by Wang *et al.* [231]	88.0%
pLSA [233]	83.3%
Volumetric [234]	62.7%

Here, accuracy (ACC) is computed as: $\text{ACC} = \frac{TP + TN}{N}$.

successful application based on the encoder-decoder model is presented in 2007 by Huang *et al*. [236] for object recognition tasks. A few years later, based on the principle of the model of Huang *et al*. [236], Baccouche *et al*. [237] proposed a solution for learning of sparse spatio-temporal features based on autoencoder scheme. Experiments on KTH [48] and GEMEP-FERA datasets [238] showed the best results when compared to methods using hand-crafted features. Some other autoencoder-based approaches have also been proposed in the works of Wu *et al*. [239], Xie *et al*. [240], Hasan *et al*. [241], and Budiman *et al*. [242]. For instance, Wu *et al*. [239] constructed a 3-layer SDA architecture for human action recognition using skeleton information captured by Kinect [12] sensor. Budiman *et al*. [242] have also performed a similar study when using a SDA model to learn skeleton feature for human body pose classification. To recognize human action, Xie *et al*. [240] used a SDA architecture with 3-hidden layers to learn contour features from a single depth frame. Hasan *et al*. [241] presented an autoencoder-based framework for learning human activity models continuously from streaming videos. This method is executed through two phases: "*initial learning*" phase and "*incremental learning*" phase. Given a streaming video with a few labeled activities, the first phase will extract space-time interest points (STIP) [230] of the motion then encode these feature vectors by a sparse autoencoder. A softmax function

is used as a classification model that provides action label. To recognize human activities in unlabeled frames, the incremental learning phase uses the sparse autoencoder and the parameters of activity classification model in initial learning phase, but in an unsupervised manner. In this phase, the active learning technique [243] has also been used to reduce the amount of manual labeling of classes.

The long training time is a disadvantage of SDAs when working with large-scale datasets. To overcome this limitation, Chen *et al*. [244] proposed a novel variant of SDAs named "mSDA". Experiments on the same dataset showed that mSDA matched the performance of SDA but reducing the training time down to 450 times. Taking advantage of the mSDA, Gu *et al*. [245] trained an mSDA network for multi-view action recognition. An mSDA is trained over all the camera views and the trained network is then used to generate features for each camera view respectively. These obtained features from all the camera views are then combined to create a single integrated representation, which can then be used as the input of a classifier. The evaluation on three benchmark multi-view action datasets provided that this model achieved the state-of-the-art recognition performance.

4.5 Other deep architectures for human action recognition

Some other deep architectures have also been used for human action recognition and related recognition tasks such as group activity analysis, or prediction of physical interactions. Sparse coding [103], [104], [246] is also another potential deep model for recognizing human action. The success of the sparse representation in various fields including pattern recognition [247], [248] or image classification [249] have shown that it could flexibly adapt to diverse low level natural signals. The sparse representations of the signals are then used as image features which are sent directly into the classifiers. Therefore, many authors [250], [251], [252], [253], [254] have exploited the advantages of sparse coding for solving human action recognition problems. Recently, some novel deep architectures for recognizing human action have been published in the literature [255], [256], [257]. For instance, Ullah and Petrosino [255] employed a CNN and a pyramidal neural network (PyraNet) [258] to recognize human action. A strict 3D pyramidal neural network (3DPyraNet) was constructed which allows to learn spatio-temporal features of human motion. These works continued to be expanded by the same authors [256] and achieved competitive results on some action datasets. Rahmani *et al*. [257] presented the "*Robust Non-Linear Knowledge Transfer Model*" (R-NKTM), a deep fully-connected neural network which is capable of understanding human action from cross-view by learning features from dense trajectories of synthetic 3D human models and real motion capture data. Figure 17 illustrates the procedure to train this network. Experiments on cross-view human action datasets including IXMAS [50], UWA3DII [259], N-UCLA Multiview Action3D [260], and UCF Sports [261] have shown that this method outperforms existing state-of-the-art.

The paper published by Le *et al*. [229] reports that we can combine the different network models to build a single deep architecture for improving its performance. Based on two key ideas, "*convolution*" and "*stacking*" in CNN architecture (subsection 3.1), the authors constructed a deep model by using the Independent Subspace Analysis (ISA) [262] (see Figure **??**a) and Principal Component Analysis (PCA) [263]. The ISA is trained on small input patches for learning feature directly from unlabeled video data. It is then convolved with a larger region of the input image. The PCA algorithm is applied on the top of ISA for reducing dimensions. The responses are then used as the input layer for another ISA.

The method is evaluated on KTH [48] , Hollywood2 [199], UCF sports [261] and YouTube datasets [53]. Table 7 shows that this deep architecture advanced the state-of-the-art in human action recognition when the paper was published.

TABLE 7
Comparison of Le's method and the best methods before

Method	**KTH**	**Hollyhood2**	**UCF**	**YouTube**
Measure	AA	Mean AP	AA	AA
Le et al. [229]	**93.9**%	**53.3**%	**86.5**%	**75.8**%
Previous best result	92.1%	50.9%	85.6%	71.2%
Improvements	1.8%	2.4%	0.9%	**4.6%**

Here, the average accuracy is noted by AA.

Srivastava *et al*. [264] constructed a model which consists of two LSTMs - the encoder LSTM and the decoder LSTM to learn representations of sequences of images. The state of the LSTM encoder is the representation of the input video. Then, the LSTM decoder will reconstruct the input sequence from this representation. It can be used for reconstructing the input sequence as well as predicting the future sequence.

Very recently, Luo *et al.* [265] combined many different models to build a deep learning framework for recognition human motion in Videos. The idea is designing a network which is able to predict the future 3D motions in videos (see Figure 18). Given input frames, the model will predict 3D flows in future frames, then use the features to recognize activities. To do that, a Recurrent Neural Network based Encoder-Decoder framework has proposed. During the encoding process, CNNs (the standard VGG-16 networks) are used for extracting a low-dimensionality feature from the input frames. Then, the LTSMs have been used to learn the temporal representation of motion. The learned representation is then decoded in the decoding process to generate the atomic 3D flows. This approach achieved the state-of-the-art result on NTU-RGB+D dataset [70] and MSR Daily Activity3D [58]. To the best of our knowledge, this model is the best learning framework at the moment for action recognition using different input modalities (RGB, Depth, RGB-D).

A new unsupervised learning approach called Generative Adversarial Networks (GANs) was proposed by Ian *et al.* [266]. In 2016, Radford *et al.* [267] introduced a set of architectures called Deep Convolutional GANs (DCGANs) in order to train GANs in a better way. This study showed that GANs can learn good representations of images for supervised learning and generative modeling. After that, GANs have started to show their real potential. E.g. Vondrick *et al.* [268] capitalized on recent advances in GANs

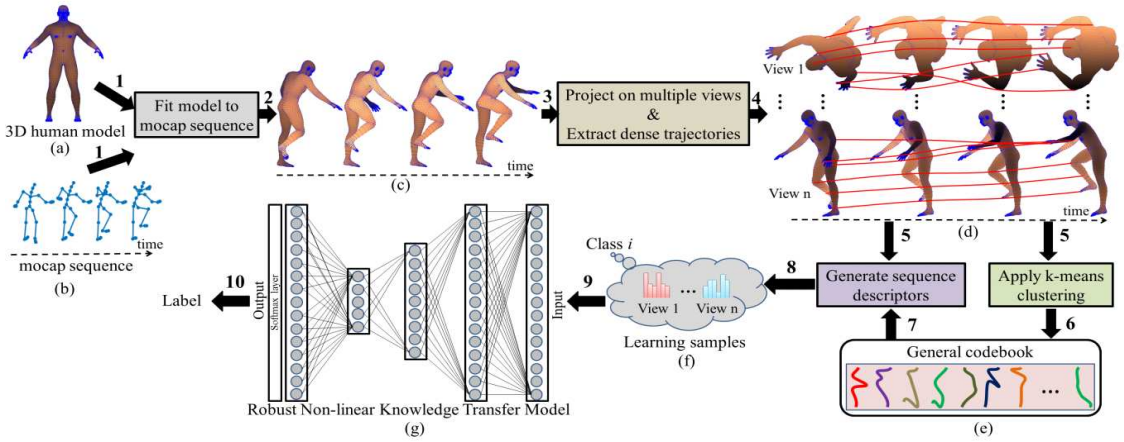


Fig. 17. Architeture of R-NKTM and its learning process [257]. Firstly, 3D human models are fitted to real motion capture data for generating realistic 3D videos. These 3D videos are then projected on 2D planes for calculating dense trajectories. A general codebook is learned from trajectories which is then used as the input of R-NKTM. By this way, the R-NKTM can learn features of human action videos and use it for testing process.

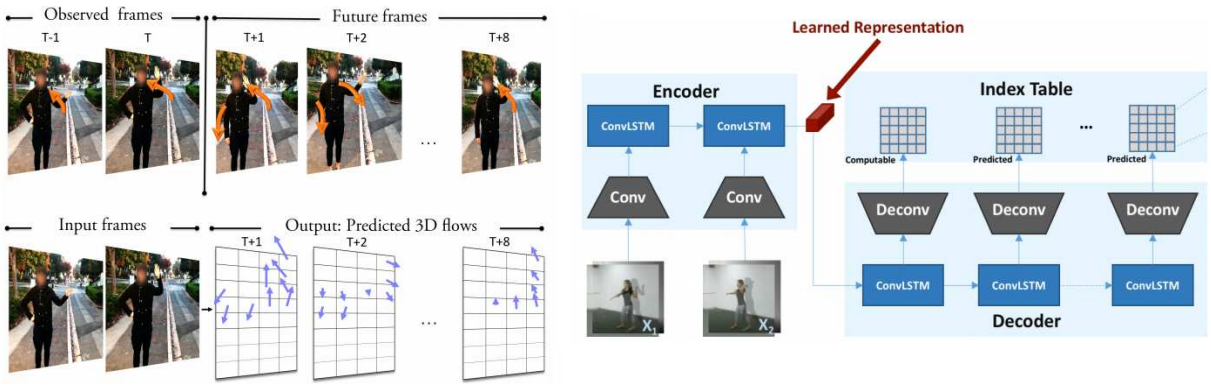


Fig. 18. **(a)** Illustration of the idea of learning a video representation by predicting a sequence of basic motions described as atomic 3D flows. The learned representation is then used for action recognition. **(b)** The learning framework architecture based on the Recurrent Neural Network based Encoder-Decoder proposed in the work of Luo *et al.* [265]

for both action classification and prediction in video. A two-stream generative model has built for learning scene dynamics. This study is an open research opportunity for designing of predictive models for understanding human actions.

5 Discussion

Human action recognition has become one of the most active research topics in computer vision during the last two decades. In particular, the appearance of the DL models as well as the advances of parallel computing techniques, e.g. GPU computing, opened up more new opportunities for this field. Many DL based approaches have developed and applied for various applications related to human action recognition. Their studies indicate various methods to learn motion features from videos and use them to recognize and classify actions. In this section, we provide a detailed analysis of the mentioned classes of architectures. The pros and cons of each class and the link between them will be discussed. Based on these analyses, we point out challenges, current trends and potential directions future research in this field.

After reviewing more than two hundred papers, our study shows that human action recognition has advanced rapidly from recognition in controlled environment with small size benchmark datasets to recognition of actions in realistic videos with very large scale benchmarks. DL techniques play an important role in this progress. In the literature of human action recognition based on DL, CNNs seem to be the most important model for learning spatio-temporal features of human action directly from RGB and RGB-D videos without pre-processing. Almost outstanding architectures, such as networks proposed by Ji *et al*. [139], Tran *et al*. [142], Simonyan *et al*. [79], Wang *et al*. [176], Feichtenhofer *et al*. [200], Luo *et al*. [265], etc. have used 3D convolutional filters to extract motion features. The key ideas behind CNNs allow them to work directly on image structure and obtaining high-level features by composing lower-level ones. CNNs are not only working as an end-to-end solution, they were also used as a feature extractor and were a part in another frameworks. However, CNNs achieve very goof performance when they were trained on very large datasets. If not, overfit will happen. Some techniques have been developed to prevent overfitting in convolutional layers such as dropout, data augmentation (e.g. random

cropping, flipping, color effect, etc). When training a very deep CNN architecture, millions of connections between neurons will be involved. Therefore, another limiting factor of CNNs is the high energy consumption due to its high computational complexity. Normally, GPU computing is required to work with this type algorithm.

Recurrent neural networks with long short-term memory (LSTM-RNNs) have been designed for solving time series problems. LSTM-RNNs have been used successfully in modeling the long-term context information of motion sequences, specifically with skeleton data as the work of Du *et al.* [214], Song *et al.* [215], Zhu *et al.* [216], Li *et al.* [217], Liu *et al.* [218]. The success of LSTM-RNNs for human action recognition comes from their ability to take advantage the entire history motion frames. Even so, most of LSTM-RNN based models can not work directly on raw data. For example, skeleton data need to be preprocessed before feeding into LSTM-RNNs. It is difficult to build an LSTM-RNN based end-to-end learning framework with RGB-D data. Consequently, many authors used CNN to extract color features and then fed into the LSTM for sequences learning and prediction.

Deep belief network (DBNs) and Stacked Denoising Autoencoders (SDAs) are also very promising choice for action recognition tasks. For DBNs, these networks can be trained in an semi-supervised way with less labeled data from a set of examples to classify its inputs. The limitation of DBNs is that they require hand-crafted features [224] or converting input data to appropriate form [227]. SDAs can learn motion features in unsupervised manner and are capable of generating robust features. However, it has several drawbacks related to its optimization process.

5.1 A quantitative analysis

- *Hand-crafted approaches and deep learning approaches: A small comparison*

In order to have a general view on recognition accuracies reported by hand-crafted approaches and deep learning approaches, we have carried out a small performance comparison on KTH [48] dataset. This dataset has been used to evaluate many action recognition solutions, both the traditional approaches based on hand-crafted features and deep learning based approaches over many years.

- *A performance comparison between deep learning models*

We provide a quantitative analysis of the deep learning approaches on a state-of-the-art benchmark for human action recognition in realistic and challenging settings. Figure 19 shows our comparison based on the performance of many deep learning solution on UCF-101 dataset that have been reviewed in this paper. This comparison helps us to see clearly the current state of this field and also provide the best architectures proposed in the literature. The accuracies are reported directly from the original papers and all of these work use the same measure. We found that the networks proposed by Varol *et al.* [168], Feichtenhofer *et al.* [195], Tran *et al*. [142].

5.2 The future of DL for human action recognition

- *Developing unsupervised learning models*

As labeling of data is very costly in terms of money and manpower, we expect that learning features directly from videos in an unsupervised manner is a very important research direction [89]. Unsupervised learning procedures such as DBNs or deep autoencoders will continue to be developed strongly because they could learn features without requiring labeled data or requiring very limited labeled data in pre-training process.

- *Deeper CNNs*

The success of some very deep learning models such as VGGNet [100], GoogLeNet [99], and ResNets [101] provided that deeper CNN models can boost the recognition accuracy. It appears that the new algorithms allow us to train deeper network easier. For example, He *et al.* [101] released ResNets in which it has fewer filters and lower complexity than VGGNet [100]. Therefore, we expect deeper CNNs will be more fully exploited in this field.

- *Combining different deep learning models*

Taking full advantage of the different deep learning models and combining them into a single learning framework is a trend in action recognition. Specifically, the use of CNNs with LSTM-RNNs has improved the stare-of-the-art in many benchmark datasets [203], [204], [205], [206], [207], [208], [209], [210], [211], [212], [213]. We believe that this trend will be continued in the future.

- *Fusion of hand-crated and deep learning solutions*

We found that hand-crafted features such as the trajectory descriptors or optical flow frames have been used in most of state-of-the-art DL models as reported in the work of Varol *et al.* [168], Feichtenhofer *et al.* [195], Tran *et al*. [142], and Wang *et al.* [77]. We expect much of future progress in human action recognition to come from systems that use both hand-crafted and DL solutions to solve challenges in this field.

- *Transfer learning*

One of the main difficulty in training deep networks comes from the scarcity of data. To solve this problem, many authors explored a technique called "*transfer learning*". Instead of training an entire deep network from scratch, we pretrain the network on a very large dataset, and then use the network either as an initialization for the task of interest. We believe that this trend will be continued in computer vision, including the human action recognition in video.

6 Conclusion

Our goal in carrying out this research is to bring readers a detailed view of the development process and especially of current progress of deep learning models applied to recognize human action in video. A comprehensive review of various DL architectures and their applications in action recognition and related tasks has been provided over more than two hundred related publications. Our analysis and comparisons about the recognition accuracy between DL based approaches and other techniques shown that deep learning is at the moment the best choice for recognizing and classifying human action as well as predicting human

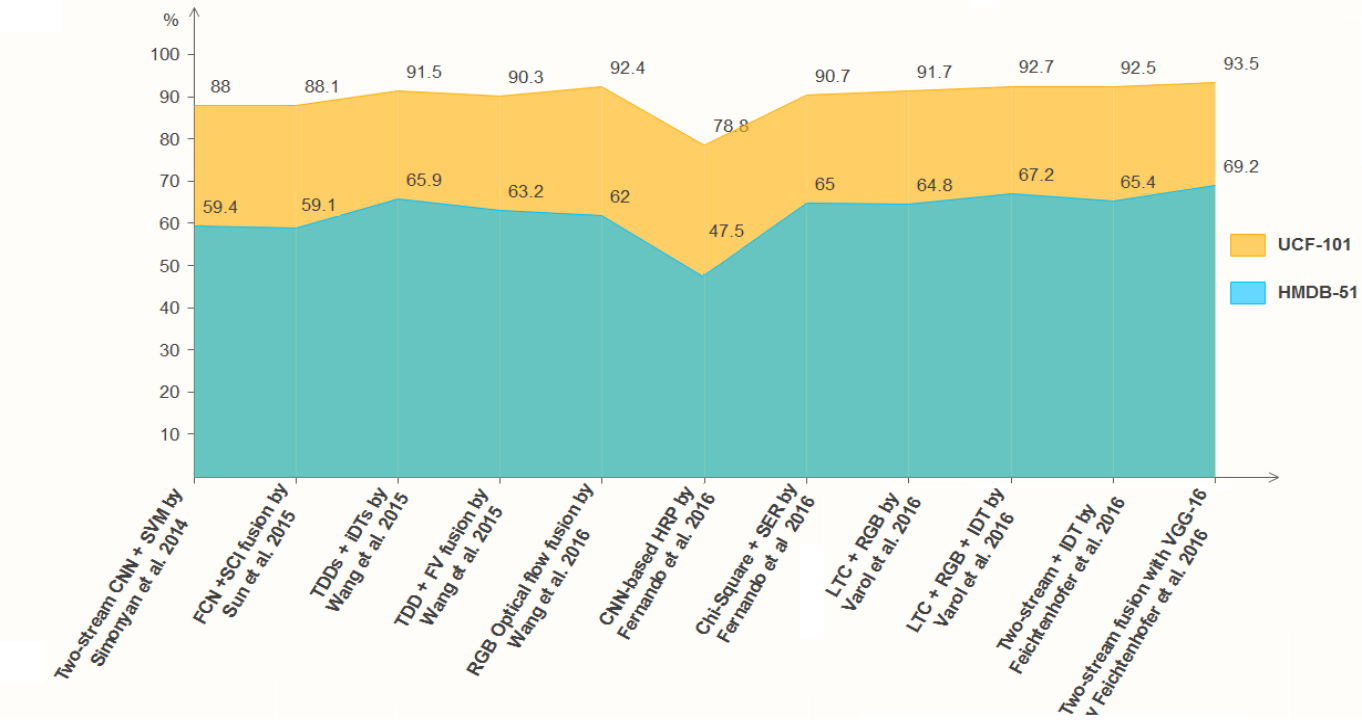


Fig. 19. The recognition performance of different deep learning based solutions on HMDB-51 and UCF-101 datasets.

behavior. In addition, the characteristics of the most important DL architectures for action recognition have been also analyzed to provide current trends and open problems for future works in this field. With a list of datasets in different complexity levels, this paper will help interested readers in choosing approximate algorithms and datasets to develop new solutions. Although there has been significant progress over the last years, there are still many challenges in applying DL models to build vision-based action recognition systems and to bring their benefits to our life. We are still looking forward to new DL based approaches to improve the performance of recognition systems while decreasing computational cost and requiring less labeled data. We hope this survey is helpful for researchers in this field.

Acknowledgments

This work was supported by the Centre d'Etudes et d'Expertise sur les Risques, l'environnement la mobilité et l'aménagement (CEREMA). The authors would like to express our thanks to all the people who have made helpful comments and suggestions on a previous draft.

References

[1] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, vol. 1, June 2004, pp. 719–722 Vol.1.

[2] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, April 2005.

[3] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *2008 IEEE International Symposium on Circuits and Systems*, May 2008, pp. 2737–2740.

[4] C. A. Pickering, K. J. Burnham, and M. J. Richardson, "A research study of hand gesture recognition technologies and applications for human vehicle interaction," in *2007 3rd Institution of Engineering and Technology Conference on Automotive Electronics*, June 2007, pp. 1–15.

[5] P. Sonwalkar, T. Sakhare, A. Patil, and S. Kale, "Hand gesture recognition for real time human machine interaction system," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 19, no. 5, pp. 262–264, 2015.

[6] N. Zouba, F. Bremond, M. Thonnat, A. Anfosso, É. Pascual, P. Mallea, V. Mailland, and O. Guerin, "Assessing computer systems for monitoring elderly people living at home," in *The 19th IAGG World Congress of Gerontology and Geriatrics, Paris*, 2009.

[7] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, and N. García, "Human-action recognition module for the new generation of augmented reality applications," in *2015 International Symposium on Consumer Electronics (ISCE)*, June 2015, pp. 1–2.

[8] M. S. Ryoo and J. K. Aggarwal, "Hierarchical recognition of human activities interacting with objects," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[9] T. McKenna, "Video surveillance and human activity recognition for anti-terrorism and force protection," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, July 2003, p. 2.

[10] M. S. Ryoo and J. K. Aggarwal, "Observe-and-explain: A new approach for multiple hypotheses tracking of humans and objects," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[11] A. B. Albu, B. Widsten, T. Wang, J. Lan, and J. Mah, "A computer vision-based system for real-time detection of sleep onset in fatigued drivers," in *2008 IEEE Intelligent Vehicles Symposium*, June 2008, pp. 25–30.

[12] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[13] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2642–2649.

[14] J. F. Kooij, N. Schneider, and D. M. Gavrila, "Analysis of pedestrian dynamics from a vehicle perspective," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, 2014, pp. 1445–1450.

[15] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis,"

Computer Vision and Image Understanding, vol. 104, no. 2–3, pp. 90 – 126, 2006, special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[16] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, Nov 2008.

[17] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010.

[18] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames," 2001.

[19] "Southampton Human ID at a Distance Database," http://www.gait.ecs.soton.ac.uk/, accessed: 2016-10-12.

[20] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Computing Surveys*, vol. 46, no. 4, pp. 43:1–43:33, Mar. 2014. [Online]. Available: http://doi.acm.org/10.1145/2523819

[21] U. Akdemir, P. Turaga, and R. Chellappa, "An ontology based approach for activity recognition from video," in *Proceedings of the 16th ACM International Conference on Multimedia*, ser. MM '08. New York, NY, USA: ACM, 2008, pp. 709–712. [Online]. Available: http://doi.acm.org/10.1145/1459359.1459466

[22] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, pp. 428–440, 1999.

[23] "IEEE Conference on Computer Vision and Pattern Recognition (CVPR)," http://ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000147, accessed: 2016-10-12.

[24] "Image and Vision Computing," http://www.journals.elsevier.com/image-and-vision-computing/, accessed: 2016-10-12.

[25] "Computer Vision and Image Understanding," http://www.journals.elsevier.com/computer-vision-and-image-understanding/, accessed: 2016-10-12.

[26] "Machine Vision and Applications," http://link.springer.com/journal/138, accessed: 2016-10-12.

[27] "Special Issue on Recognition and Action for Scene Understanding," http://www.journals.elsevier.com/neurocomputing/call-for-papers/recognition-and-action-for-scene-understanding/, accessed: 2016-10-12.

[28] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," in *Proceedings IEEE Nonrigid and Articulated Motion Workshop*, Jun 1997, pp. 90–102.

[29] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer vision and image understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[30] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[31] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224 – 241, 2011.

[32] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, Nov 2012.

[33] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[34] J. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70 – 80, 2014, celebrating the life and work of Maria Petrou.

[35] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[36] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: a survey," *arXiv preprint arXiv:1501.05964*, 2015.

[37] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015. [Online]. Available: http://journal.frontiersin.org/article/10.3389/frobt.2015.00028

[38] T. Subetha and S. Chitrakala, "A survey on human activity recognition from videos," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, Feb 2016, pp. 1–7.

[39] L. Lo Presti and M. La Cascia, "3d skeleton-based human action classification," *Pattern Recogn.*, vol. 53, no. C, pp. 130–147, May 2016. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2015.11.019

[40] S.-M. Kang and R. P. Wildes, "Review of action recognition and detection methods," *CoRR*, vol. abs/1610.06906, 2016.

[41] S. Herath, M. T. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *CoRR*, vol. abs/1605.04988, 2016. [Online]. Available: http://arxiv.org/abs/1605.04988

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[43] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[45] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[46] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[47] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[48] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, Aug 2004, pp. 32–36 Vol.3.

[49] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, Dec 2007.

[50] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[51] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[52] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2929–2936.

[53] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1996–2003.

[54] S. Singh, S. A. Velastin, and H. Ragheb, "MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug 2010, pp. 48–55.

[55] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010, accessed: 2016-10-18.

[56] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14.

[57] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1290–1297.

[58] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010.

[59] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the 11th European Conference on Computer Vision: Part II*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 392–405. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888028.1888059

[60] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "Avss 2011 demo session: A large-scale benchmark dataset for event recognition in surveillance video," in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2011, pp. 527–528.

[61] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2556–2563.

[62] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images," *CoRR*, vol. abs/1107.0169, 2011.

[63] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *CoRR*, vol. abs/1210.1207, 2012. [Online]. Available: http://arxiv.org/abs/1210.1207

[64] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 28–35.

[65] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot *et al.*, "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding*, vol. 127, pp. 14–30, 2014.

[66] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0450-4

[67] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012. [Online]. Available: http://arxiv.org/abs/1212.0402

[68] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[69] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[70] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," *CoRR*, vol. abs/1604.02808, 2016. [Online]. Available: http://arxiv.org/abs/1604.02808

[71] N. Ikizler and P. D. Sahin, "Human action recognition using distribution of oriented rectangular patches," in *Workshop on Human Motion*, 2007.

[72] S. Brahnam and L. Nanni, "High performance set of features for human action classification," in *IPCV*, 2009.

[73] "Cornell Activity Datasets: CAD-60," http://pr.cs.cornell.edu/humanactivities/data.php, accessed: 2016-11-17.

[74] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *CoRR*, vol. abs/1604.06182, 2017.

[75] J. Zhang, W. Li, P. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[76] M. Firman, "RGBD datasets: Past, present and future," *CoRR*, vol. abs/1604.00999, 2016. [Online]. Available: http://arxiv.org/abs/1604.00999

[77] X. Wang, A. Farhadi, and A. Gupta, "Actions transformations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[78] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," *CoRR*, vol. abs/1510.00562, 2015.

[79] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[80] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[81] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2555–2562.

[82] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[83] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *European Conference on Computer Vision*. Springer, 2012, pp. 425–438.

[84] E. F. Can and R. Manmatha, "Formulating action recognition as a ranking problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 251–256.

[85] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *ECCV*, 2012.

[86] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Mach. Vis. Appl.*, vol. 24, pp. 1473–1485, 2012.

[87] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *CVPR*, 2012.

[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[89] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[90] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer International Publishing, 2014, pp. 818–833.

[91] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[92] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980. [Online]. Available: http://dx.doi.org/10.1007/BF00344251

[93] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[94] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[95] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[96] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[97] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[98] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[99] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2015.

[100] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[101] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[102] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines." in *Artificial Intelligence and Statistics Conference (AISTATS)*, vol. 1, 2009, p. 3.

[103] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature*, vol. 381 6583, pp. 607–9, 1996.

[104] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, 2006, pp. 801–808.

[105] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[106] "Neural Networks and Deep Learning", Determination Press, 2015," http://neuralnetworksanddeeplearning.com/chap6.html, accessed: 2016-11-01.

[107] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature selection using a multilayer perceptron," *Journal of Neural Network Computing*, vol. 2, no. 2, pp. 40–48, 1990.

[108] n. . A. Quoc V. Le, howpublished = http://cs.stanford.edu/~quocle/tutorial2.pdf, "A Tutorial on Deep Learning."

[109] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[110] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[111] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *The 2011 International Joint Conference on Neural Networks*, July 2011, pp. 2809–2813.

[112] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[113] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[114] A. Graves, "Supervised sequence labelling with recurrent neural networks," in *Studies in Computational Intelligence*, 2008.

[115] Y. Bengio, P. Y. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, pp. 157–166, 1994.

[116] J. F. Kolen and S. C. Kremer, *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*. Wiley-IEEE Press, 2001, pp. 237–243. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5264952

[117] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[118] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[119] V. Nair and G. E. Hinton, "3D object recognition with deep belief nets," in *Advances in Neural Information Processing Systems*, 2009, pp. 1339–1347.

[120] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Modeling human motion using binary latent variables," in *Advances in neural information processing systems*, 2006, pp. 1345–1352.

[121] P. Smolensky, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, ch. Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281. [Online]. Available: http://dl.acm.org/citation.cfm?id=104279.104290

[122] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, "Boltzmann machines: Constraint satisfaction networks that learn," Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, Tech. Rep., 1984.

[123] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[124] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, ch. Learning Internal Representations by Error Propagation, pp. 318–362. [Online]. Available: http://dl.acm.org/citation.cfm?id=104279.104293

[125] M. Cilimkovic, "Neural networks and back propagation algorithm," *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, vol. 15, 2015.

[126] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Networks*, vol. 21, pp. 1610–1623, 2010.

[127] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[128] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3119–3127.

[129] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," *Advances in Neural Information Processing Systems*, pp. 901–908, 1995.

[130] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," *arXiv preprint arXiv:1312.7302*, 2013.

[131] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 302–315.

[132] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-CNNs for pose estimation and action detection," *arXiv preprint arXiv:1406.5212*, 2014.

[133] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 5, p. 169, 2014.

[134] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3218–3226.

[135] M. A. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 179–192, 2003.

[136] R. Sigala, T. Serre, T. Poggio, and M. Giese, "Learning features of intermediate complexity for the recognition of biological motion," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 241–246.

[137] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[138] H.-J. Kim, J. S. Lee, and H.-S. Yang, "Human action recognition using a modified convolutional neural network," in *International Symposium on Neural Networks*. Springer, 2007, pp. 715–723.

[139] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[140] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3D human activity recognition with reconfigurable convolutional neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 97–106.

[141] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[142] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 4489–4497.

[143] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," *arXiv preprint arXiv:1501.04686*, 2015.

[144] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using binary motion image and deep learning," *Procedia Computer Science*, vol. 58, pp. 178 – 185, 2015.

[145] C. Liu, W. Xu, Q. Wu, and G. Yang, "Learning motion and content-dependent features with convolutions for action recognition," *Multimedia Tools and Applications*, pp. 1–17, 2015.

[146] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3D deep convolutional descriptors," in *25th International Joint Conference on Artificial Intelligence*, New York, NY, USA, July, 2016.

[147] L. Mo, F. Li, Y. Zhu, and A. Huang, "Human physical activity recognition based on computer vision with deep learning model," in *Instrumentation and Measurement Technology Conference Proceedings (I2MTC), 2016 IEEE International*. IEEE, 2016, pp. 1–6.

[148] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *in the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016.

[149] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Exploiting deep residual networks for human action

recognition from skeletal data," *Computer Vision and Image Understanding*, vol. 170, pp. 51–66, 2018.

[150] P. Huy Hieu, L. Khoudor, A. Crouzil, P. Zegers, and S. A. Velastin Carroza, "Exploiting deep residual networks for human action recognition from skeletal data," 2018.

[151] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, S. A. Velastin, and P. Zegers, "A unified deep framework for joint 3d pose estimation and action recognition from a single rgb camera," *Sensors*, vol. 20, no. 7, p. 1825, 2020.

[152] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Learning to recognise 3d human action from a new skeleton-based representation using deep convolutional neural networks," *IET Computer Vision*, vol. 13, no. 3, pp. 319–328, 2019.

[153] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Spatio–temporal image representation of 3d skeletal movements for view-invariant action recognition with deep convolutional neural networks," *Sensors*, vol. 19, no. 8, p. 1932, 2019.

[154] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Skeletal movement to color map: A novel representation for 3d action recognition with inception residual networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3483–3487.

[155] ——, "Learning and recognizing human action from skeleton movement with deep residual neural networks," 2017.

[156] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "A deep learning approach for real-time 3d human action recognition from skeletal data," in *International Conference on Image Analysis and Recognition*. Springer, 2019, pp. 18–32.

[157] H.-H. Pham, "Architectures d'apprentissage profond pour la reconnaissance d'actions humaines dans des séquences vidéo rgb-d monoculaires: application à la surveillance dans les transports publics," Ph.D. dissertation, Université Paul Sabatier-Toulouse III, 2019.

[158] H.-H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Skeletal movement to enhanced color map: A novel representation for rgb-d based 3d human action recognition with densely connected convolutional networks," in *16th International Conference on Image Analysis and Recognition (ICIAR 2019), august 27-29, 2019, Waterloo, Canada*, 2019.

[159] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 2568–2577.

[160] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE, 2015, pp. 4657–4666.

[161] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa, "Predicting daily activities from egocentric images using deep learning," in *proceedings of the 2015 ACM International symposium on Wearable Computers*. ACM, 2015, pp. 75–82.

[162] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1600–1609.

[163] H.-J. Kim, J. Lee, and H.-S. Yang, "A weighted fmm neural network and its application to face detection," in *International Conference on Neural Information Processing*. Springer, 2006, pp. 177–186.

[164] "TREC Video Retrieval Evaluation 2008 (TRECVID)," http://www-nlpir.nist.gov/projects/tv2008/tv2008.html, accessed: 2016-10-11.

[165] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[166] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human action detection by boosting efficient motion features," in *2006 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2009, pp. 522–529.

[167] M. T. Law, N. Thome, and M. Cord, "Bag-of-words image representation: Key ideas and further insight," in *chapter 2, Fusion in Computer Vision - Understanding Complex Visual Content, Advances in Computer Vision and Pattern Recognition*. Springer, 2014, pp. 29–52.

[168] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[169] D. Steinkraus, I. Buck, and P. Y. Simard, "Using GPUs for machine learning algorithms," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, Aug 2005, pp. 1115–1120 Vol. 2.

[170] E. P. Ijjina and C. K. Mohan, "Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks," in *2014 13th International Conference on Machine Learning and Applications*, Dec 2014, pp. 178–182.

[171] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1234–1241.

[172] "The PASCAL Visual Object Classes Homepage," http://host.robots.ox.ac.uk/pascal/VOC/, accessed: 2016-11-13.

[173] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[174] Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Two-stream sr-cnns for action recognition in videos." The British Machine Vision Conference (BMVC), 2016.

[175] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[176] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 4325–4334.

[177] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang, "Cuhk & ethz & siat submission to activitynet challenge 2016," *arXiv preprint arXiv:1608.00797*, 2016.

[178] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[179] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1119–1122. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806296

[180] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, Aug 2016.

[181] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 20–27.

[182] Y. LeCun and Y. Bengio, "Convolutional networks for images speech and time series," 1995.

[183] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[184] "Sports-1M Dataset," http://cs.stanford.edu/people/karpathy/, accessed: 2016-11-17.

[185] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *CoRR*, vol. abs/1612.09401, 2016. [Online]. Available: http://arxiv.org/abs/1612.09401

[186] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[187] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 3169–3176.

[188] C. Beaudry, R. Péteri, and L. Mascarilla, "An efficient and sparse approach for large scale human action recognition in videos," *Machine Vision and Applications*, vol. 27, no. 4, pp. 529–543, 2016.

[189] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[190] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3192–3199.

[191] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.

[192] I. Lillo, A. Soto, and J. Carlos Niebles, "Discriminative hierarchical modeling of spatio-temporally composable human activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 812–819.

[193] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[194] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *CoRR*, vol. abs/1507.02159, 2015.

[195] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[196] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014.

[197] L. Wang, Z. Wang, Y. Xiong, and Y. Qiao, "Cuhk&siat submission for thumos'15 action recognition challenge," *THUMOS'15 Action Recognition Challenge*, vol. 3, no. 4, 2015.

[198] B. Fernando, P. Anderson, M. Hutter, and S. Gould, "Discriminative hierarchical rank pooling for activity recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[199] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2929–2936.

[200] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Advances In Neural Information Processing Systems*, 2016, pp. 3468–3476.

[201] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1798–1807.

[202] A. Grushin, D. Monner, J. A. Reggia, and A. Mishra, "Robust human action recognition via long short-term memory," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013.

[203] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *HBU*, 2011.

[204] J. Y.-H. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[205] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *CoRR*, vol. abs/1411.4389, 2015.

[206] A. Giel and R. Diaz, "Recurrent neural networks and transfer learning for action recognition," 2015.

[207] S. Sharma, J. R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *CoRR*, vol. abs/1511.04119, 2015.

[208] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[209] B. Singh, T. K. Marks, M. J. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[210] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," in *ICMR*, 2016.

[211] J. Wu, G. Wang, W. Yang, and X. Ji, "Action recognition with joint attention on multi-level deep features," *CoRR*, vol. abs/1607.02556, 2016.

[212] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li, "Hierarchical attention network for action recognition in videos," *CoRR*, vol. abs/1607.06416, 2016.

[213] H. Chen, J. Chen, R. Hu, C. Chen, and Z. Wang, "Action recognition with temporal scale-invariant deep learning framework," *China Communications*, vol. 14, no. 2, pp. 163–172, February 2017.

[214] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[215] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," *CoRR*, vol. abs/1611.06067, 2016.

[216] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," *CoRR*, vol. abs/1603.07772, 2016.

[217] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *ECCV*, 2016.

[218] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," *CoRR*, vol. abs/1607.07043, 2016.

[219] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[220] H. Tang, "A comparative evaluation of deep belief nets in semi-supervised learning." Report for CSC2515,Department of Computer Science University of Toronto, 2008.

[221] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 609–616.

[222] H. Zhang, F. Zhou, W. Zhang, X. Yuan, and Z. Chen, "Real-time action recognition based on a modified deep belief network model," in *2014 IEEE International Conference on Information and Automation (ICIA)*, July 2014, pp. 225–228.

[223] E. Hsu, K. Pulli, and J. Popović, "Style translation for human motion," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 1082–1089.

[224] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Exploiting the deep learning paradigm for recognizing human actions," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2014, pp. 93–98.

[225] P. Foggia, G. Percannella, A. Saggese, and M. Vento, "Recognizing human actions by a bag of visual words," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 2910–2915.

[226] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, Jan 2013, pp. 53–60.

[227] K. H. Ali and T. Wang, "Learning features for action recognition and identity with deep belief networks," in *2014 International Conference on Audio, Language and Image Processing*, July 2014, pp. 129–132.

[228] P. Heckbert, "Fourier Transforms and the Fast Fourier Transform (FFT) Algorithm," *Computer Graphics*, vol. 2, pp. 15–463, 1995.

[229] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 3361–3368. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995496

[230] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[231] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[232] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*. Springer, 2010, pp. 140–153.

[233] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[234] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 166–173.

[235] S. Nam, S. Park, J. Park, and T. Kim, "A single depth sensor based human activity recognition via deep belief network," in *Proceedings of the 4th World Conference on Applied Sciences, Engineering and Technology 24-26 October 2015*, 2015, pp. 015–019.

[236] F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[237] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification." in *British Machine Vision Conference 2012 (BMVC'12)*, 2012, pp. 1–12.

[238] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*. IEEE, 2011, pp. 921–926.

[239] D. Wu, W. Pan, L. Xie, and C. Huang, "An adaptive stacked denoising auto-encoder architecture for human action recognition." *Applied Mechanics & Materials*, 2014.

[240] L. Xie, W. Pan, C. Tang, and H. Hu, "A pyramidal deep learning architecture for human action recognition," *International Journal of Modelling, Identification and Control*, vol. 21, no. 2, pp. 139–146, 2014.

[241] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *European Conference on Computer Vision*. Springer, 2014, pp. 705–720.

[242] A. Budiman, M. I. Fanany, and C. Basaruddin, "Stacked denoising autoencoder for feature representation learning in pose-based action recognition," in *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*. IEEE, 2014, pp. 684–688.

[243] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[244] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized stacked denoising autoencoders," in *Proceedings of the Learning Workshop, Utah, UT, USA*, vol. 36, 2012.

[245] F. Gu, F. Flórez-Revuelta, D. Monekosso, and P. Remagnino, "Marginalised stacked denoising autoencoders for robust representation of real-time multi-view action recognition," *Sensors*, vol. 15, no. 7, pp. 17 209–17 231, 2015.

[246] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," in *CVPR 2011*, June 2011, pp. 1713–1720.

[247] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[248] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3517–3524.

[249] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1794–1801.

[250] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 660–671.

[251] Z. Lu and Y. Peng, "Latent semantic learning by efficient sparse coding with hypergraph regularization," in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, ser. AAAI'11. AAAI Press, 2011, pp. 411–416. [Online]. Available: http://dl.acm.org/citation.cfm?id=2900423.2900488

[252] Z. Lu and Y. Peng, "Latent semantic learning with structured sparse representation for human action recognition," *Pattern Recognition*, vol. 46, no. 7, pp. 1799–1809, 2013.

[253] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[254] A. Alfaro, D. Mery, and A. Soto, "Action recognition in video using sparse coding and relative features," *arXiv preprint arXiv:1605.03222*, 2016.

[255] I. Ullah and A. Petrosino, "A strict pyramidal deep neural network for action recognition," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 236–245.

[256] I. Ullah and A. Petrosino, "Spatiotemporal features learning with 3DPyraNet," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2016, pp. 638–647.

[257] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *arXiv preprint arXiv:1602.00828*, 2016.

[258] S. L. Phung and A. Bouzerdoum, "A pyramidal neural network for visual pattern recognition," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 329–343, 2007.

[259] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," 2016.

[260] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2649–2656.

[261] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[262] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," in *Natural Image Statistics*. Springer, 2009, pp. 151–175.

[263] M. Mudrova and A. Prochazka, "Principal component analysis in image processing," in *Proceedings of the MATLAB Technical Computing Conference, Prague*, 2005.

[264] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMS," *CoRR, abs/1502.04681*, vol. 2, 2015.

[265] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," *CoRR*, vol. abs/1701.01821, 2017.

[266] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[267] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.

[268] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *NIPS*, 2016.

Hieu H. Pham is a Teaching Fellow at the College of Engineering and Computer Science (CECS), VinUniversity, and a Research Fellow at VinUni-Illinois Smart Health Center. He received his Ph.D. in Computer Science from the Toulouse Computer Science Research Institute (IRIT), University of Toulouse, France, in 2019. Previously, he earned the Degree of Engineer in Industrial Informatics from Hanoi University of Science and Technology (HUST), Vietnam, in 2016. His research interests include Computer Vision, Machine Learning, Medical Image Analysis, and their applications in Smart Healthcare. He is the author, co-author of 30 scientific articles appeared in about 20 conferences and journals such as Computer Vision and Image Understanding, Neurocomputing, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Medical Imaging with Deep Learning (MIDL), IEEE International Conference on Image Processing (ICIP), and IEEE International Conference on Computer Vision (ICCV). He is also currently serving as Reviewers for MICCAI, ICCV, CVPR, IET Computer Vision Journal (IET-CVI), IEEE Journal of Biomedical and Health Informatics (JBHI), and Nature Scientific Reports. Before joining VinUniversity, Dr. Hieu worked at Vingroup Big Data Institute (VinBigData) as a Research Scientist and Head of the Fundamental Research Team. With this position, he led several research projects on Medical AI, including collecting various types of medical data, managing and annotating data, and developing new AI solutions for medical analysis.

Louahdi Khoudour, Director of Research, received his Ph.D. in Computer Vision from the University of Lille in 1997 and Habilitation à

Diriger des Recherches (HDR) degree from the University of Paris in 2006. He worked at Ifsttar (formerly INRETS: French National Institute on traffic and safety research) for many years. He moved to CEREMA in 2011, where he is head of a research group working on safety and security in transport.

Alain Crouzil received his Ph.D. degree in Computer science from the Paul Sabatier University of Toulouse in 1997. He is currently an associate professor and a member of the Traitement et Comprehension d'Images group of Institut de Recherche en Informatique de Toulouse. His research interests concern stereo vision, shape from shading, camera calibration, image segmentation, change detection, and motion analysis.

Pablo Zegers received his B.S. and P.E. degrees in Engineering from the Pontificia Universidad Catolica, Chile, in 1992, his M.Sc. from The University of Arizona, USA, in 1998, and his Ph.D., also from The University of Arizona, in 2002. He is currently an Associate Professor of the College of Engineering and Applied Sciences of the Universidad de los Andes, Chile. His interests are artificial intelligence, machine learning, neural networks, and information theory. From 2006 to 2010 he was the Academic Director of this College, and for a brief period at the end of 2010, the Interim Dean. He is a Senior Member of the IEEE, and currently the Secretary of the Chilean IEEE section.

Sergio A. Velastin obtained his Ph.D. in 1982 from the University of Manchester, UK. He is a professor of applied computer vision and was the director of the Digital Imaging Research Centre at Kingston University until 2012. He then worked as a research professor at the University of Santiago (Chile) and is currently UC3M-Marie Curie Research Professor at the University Carlos III de Madrid, Spain, working on human action recognition. He is a fellow of the IET and a Senior Member of the IEEE.