Deep learning for action and gesture recognition in image sequences: a survey*

Maryam Asadi-Aghbolaghi MASADIA@CE.SHARIF.EDU
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain
Department of Mathematics and Informatics, University of Barcelona, Barcelona, Spain

Albert Clapés ACLAPES@CVC.UAB.CAT
Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain
Department of Mathematics and Informatics, University of Barcelona, Barcelona, Spain

Marco Bellantonio MARCO.BELLANTONIO@EST.FIB.UPC.EDU
Facultat d'Informatica, Polytechnic University of Barcelona, Barcelona, Spain

Hugo Jair Escalante HUGOJAIR@INAOEP.MX
Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico, 72840

Víctor Ponce-López VICTOR.PONCE@EURECAT.ORG
Eurecat, Barcelona, Catalonia, Spain

Xavier Baró XBARO@UOC.EDU
EIMT, Open University of Catalonia, Barcelona, Spain

Isabelle Guyon GUYON@CHALEARN.ORG
UPSud and INRIA, Université Paris-Saclay, Paris, France
ChaLearn, Berkeley, California

Shohreh Kasaei SKASAEI@SHARIF.EDU
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Sergio Escalera SERGIO@MAIA.UB.ES
Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain
Department of Mathematics and Informatics, University of Barcelona, Barcelona, Spain

Editor: Leslie Pack Kaelbling

Abstract

Interest in automatic action and gesture recognition has grown considerably in the last few years. This is due in part to the large number of application domains for this type of technology. As in many other computer vision areas, deep learning based methods have quickly become a reference methodology for obtaining state-of-the-art performance in both tasks. This chapter is a survey of current deep learning based methodologies for action and gesture recognition in sequences of images. The survey reviews both fundamental and cutting edge methodologies reported in the last few years. We introduce a taxonomy that summarizes important aspects of deep learning for approaching both tasks. Details of the proposed architectures, fusion strategies, main datasets, and competitions are reviewed.

*. A reduced version of this appeared appeared as: M. Asadi-Aghbolaghi et al. **A survey on deep learning based approaches for action and gesture recognition in image sequences.** *In Proceedings of 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), 2017.*

Also, we summarize and discuss the main works proposed so far with particular interest on how they treat the temporal dimension of data, their highlighting features, and opportunities and challenges for future research. To the best of our knowledge this is the first survey in the topic. We foresee this survey will become a reference in this ever dynamic field of research.

Keywords: Action Recognition; Gesture Recognition; Deep Learning Architectures; Fusion Strategies;

1. Introduction

Automatic human behavior analysis has grown in interest in the last few years. This is due in part to the large number of application domains for this technology, from any kind of human-computer interaction scenario (e.g. affective robotics (Wilson and Lewandowska-Tomaszczyk, 2014)), to security (e.g. video surveillance (Vishwakarma and Agrawal, 2013)), e-Health (e.g. therapy (Mousavi Hondori and Khademi, 2014) or automatic diagnosis (Scharcanski and Celebi, 2014)), language/communication (e.g. sign language recognition (Pigou et al., 2015a)), or entertainment (e.g. interactive gaming (Marks, 2011)). Because of this, we can find, in the specialized literature, research works dealing with different aspects of human behavior analysis: action/gesture recognition (Feichtenhofer et al., 2016b; Simonyan and Zisserman, 2014), social interaction modeling (Deng et al., 2016; Ibrahim et al., 2016), facial emotion analysis (Araujo and Kamel, 2014), and personality traits identification (Joo et al., 2014), just to mention some of them.

Two key tasks for human behavior understanding that have an impact in many application scenarios are action and gesture recognition. The former is focused on recognizing generic human actions (e.g. "walking", "eating", "answering phone", etc) performed by one or more subjects, whereas the latter is focused on recognizing more fine-grained upper body movements performed by a user that have a meaning within a particular context (e.g. "come", "hi", "thumbs up", etc). While both tasks present different complications, they are interrelated in that both are based on analyzing the posture and movement of body across video sequences.

Action and gesture recognition have been studied for a while within the fields of computer vision and pattern recognition. Since the earliest works two decades ago (Kuniyoshi et al., 1990; Yamato et al., 1992), researchers have reported substantial progress for both tasks. As in the case of several computer vision tasks (e.g. object or face recognition), deep learning has also recently irrupted in action/gesture recognition, achieving outstanding results and outperforming "non-deep" state-of-the-art methods (Simonyan and Zisserman, 2014; Wang et al., 2015b; Feichtenhofer et al., 2016a).

The extra (temporal) dimension in sequences typically turned action/gesture recognition into a challenging problem in terms of both amounts of data to be processed and model complexity – which in particular are crucial aspects for training large parametric deep learning networks. In this context, authors proposed several strategies, such as frame subsampling, aggregation of local frame-level features into mid-level video representations, or temporal sequence modeling, just to name a few. For the latter, researchers tried to exploit recurrent neural networks (RNN) in the past Waibel et al. (1990). However, these models typically faced some major mathematical difficulties identified by Hochreiter Hochreiter (1991) and Bengio et al Bengio et al. (1994). In 1997, authors' effort led to the development

of the long short-term memory (LSTM) Hochreiter and Schmidhuber (1997) cells for RNNs. Today, LSTMs are an important part of deep models for image sequence modeling for human action/gesture recognition Singh et al. (2016a); Liu et al. (2016a). These, along with implicit modeling of spatiotemporal features using 3D convolutional nets Ji et al. (2010); Tran et al. (2015), pre-computed motion-based features Simonyan and Zisserman (2014); Feichtenhofer et al. (2016a), or the combination of multiple visual Singh et al. (2016b), resulted in fast and reliable state-of-the-art methods for action/gesture recognition.

Although the application of deep learning to action and gesture recognition is relatively new, the amount of research that has been generated in these topics within the last few years is astounding. Because of this overwhelming amount of work and because of the race for getting the best model/performance in these tasks for which the use of deep learning is still in its infancy, we think it is critical to compile the recent advances and, in general, the historical state of the art on action and gesture recognition with deep learning solutions. In this direction, this chapter aims to collect and review all of the existent work on deep learning for action and gesture recognition. To the best of our knowledge, there is no previous survey that collects and reviews all of the existent work on deep learning for those tasks. This chapter aims at capturing a snapshot of current trends in this direction, including an in depth analysis of different deep models, with special interest on how they treat the temporal dimension of the data.

The remainder of this chapter is organized as follows. Section 2 presents a taxonomy in this field of research. Next, Section 3 reviews the literature on human action/activity recognition with deep learning models. Section 4 summarizes the state-of-the-art on deep learning for gesture recognition. Finally, Section 5 discusses the main features of the reviewed deep learning for the both studied problems.

2. Taxonomy

We present a taxonomy that summarizes the main concepts related to deep learning in action and gesture recognition. The taxonomy is shown in Figure 1. The reader should note that with *recognition* we refer to either classification of pre-segmented video segments or localization of actions in long untrimmed videos.

The rest of this section elaborates on the main aspects and findings derived from the taxonomy. We first explain the categorized architectures, and then explore the fusion strategies used in deep learning-based models for action/gesture recognition. We also include a summary of datasets used for such tasks. Finally, we report main challenges have been held for human action and gesture recognition.

2.1 Architectures

The most crucial challenge in deep-based human action and gesture recognition is how to deal with the temporal dimension. Based on the way it is dealt with, we categorize approaches into four non-mutually exclusive groups. The first group consists in 2D CNNs, which are basically able to exploit appearance (spatial) information. These approaches (Sun et al. (2015); Wang et al. (2016g)) sample one or more frames from the whole video and then apply a pre-trained 2D models on each of these frames, separately. They finally label the actions by averaging the result of the sampled frames. The main advantage of

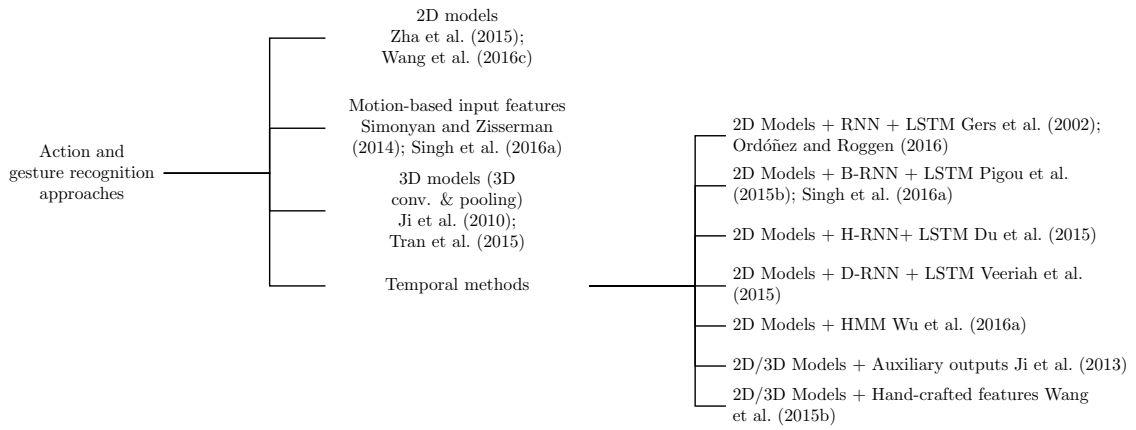


Figure 1: Taxonomy of deep learning approaches for gesture and action recognition

this kind of models is possibility to use pre-trained models on larger image datasets, such as ImageNet Krizhevsky et al. (2012). Gesture recognition methods mainly fall into this category Jain et al. (2014a); Li et al. (2015b); Liang et al. (2016).

Methods in the second group, first extract 2D motion features like optical flow and then utilize these features as a different input channel of 2D convolutional networks Simonyan and Zisserman (2014); Wang et al. (2015b); Gkioxari and Malik (2015); Sun et al. (2015); Weinzaepfel et al. (2015). In other words, these methods take into account the temporal information from the pre-computed motion features. Third group uses 3D filters in the convolutional layers Baccouche et al. (2011); Ji et al. (2013); Liu et al. (2016b); Varol et al. (2016). The 3D convolution and 3D pooling allow to capture discriminative features along both spatial and temporal dimensions while maintaining the temporal structure in contrast to 2D convolutional layers. The spatiotemoral features extracted by this kind of models proven to surpass 2D models trained on the same video frames. Figure 2a-2b illustrate these first three groups.

Finally, the fourth group combines 2D (or 3D) convolutional nets, which are applied at individual (or stacks of) frames, with a temporal sequence modeling. Recurrent Neural Network (RNN) Elman (1990) is one of the most used networks for this task, which can take into account the temporal data using recurrent connections in hidden layers. The drawback of this network is its short memory which is insufficient for real world actions. To solve this problem Long Short-Term Memory (LSTM) networks Gers et al. (2002) were proposed, and they are usually used as a hidden layer of RNN. Bidirectional RNN (B-RRN) Pigou et al. (2015b), Hierarchical RNN (H-RNN) Du et al. (2015), and Differential RNN (D-RNN) Veeriah et al. (2015) are some successful extensions of RNN in recognizing human actions. Other temporal modeling tools like HMM are also applied Wu et al. (2016a) in this context. We show an example of this fourth approach on Figure 2c.

For all methods in the four groups, their performance can be boosted by combining its output with auxiliary hand-crafted features Ji et al. (2013), e.g. improved dense trajectories (iDT) Wang et al. (2015b).

2.2 Fusion strategies

Information fusion is common in deep learning methods for action and gesture recognition. The goal of the fusion is, in most cases, to exploit information complementariness and redundancy for improving the recognition performance. At times, fusion is used to combine the information from different parts in a segmented video sequence (i.e., temporal dimension) (Wang et al., 2016c). Although, it is more common to fuse information from multiple modalities (e.g. RGB, depth, and/or audio cues), where often, information from the same modality, but processed differently is combined as well. Another variant of information fusion widely used in action and gesture recognition consist of combining models trained with different data samples and learning parameters (Neverova et al., 2014).

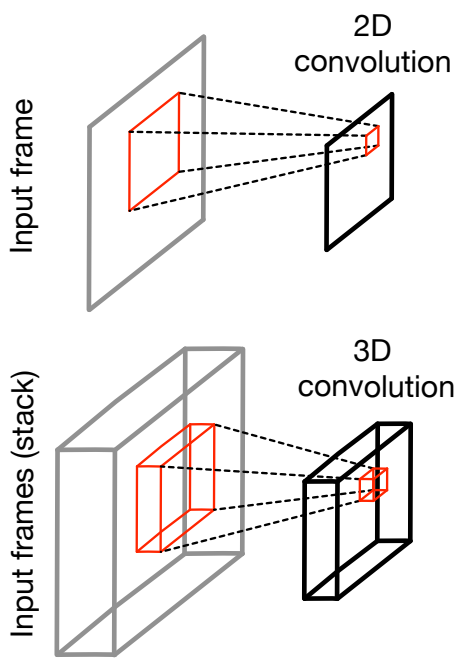
In general terms, there are several variants in which information can be fused (see e.g. (Escalante et al., 2008)). Most notably, early (fusing information before the data is fed into the model, or the model is used to fuse information directly from multiple sources), late (where the outputs of deep learning models are combined, with another layer of a deep network, a classifier or even by majority voting), and middle (in which intermediate layers fuse information, not directly form the different modalities) fusion. An excellent illustration of the effective use of these three traditional fusion schemes is described by (Neverova et al., 2015b). Modifications and variants of these schemes have been proposed as well, for instance, see the variants introduced in (Karpathy et al., 2014) for fusing information in the temporal dimension. Ensembles or stacked networks are also common strategies for information fusion in deep learning based approaches for action and gesture recognition (Wang et al., 2016c; Varol et al., 2016; Neverova et al., 2014). In Figure 2d, we illustrate an example of middle fusion of temporal information into a spatiotemporal stream.

2.3 Datasets

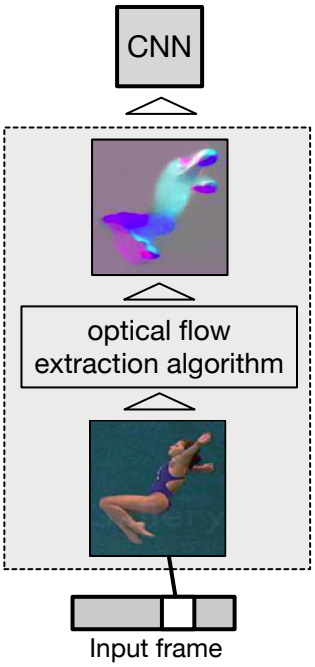
We list the most relevant datasets according to action/activity and gesture recognition in Table 1 and Table 2, respectively. For each dataset, we specify year of creation; problems for which the dataset was defined action classification (AC), temporal localization (TL), spatio-temporal localization (STL), and gesture recognition (GR); involved body parts (U for upper body, L for lower body, F for full body, and H for hands); data modalities available; number of classes and the state-of-the-art result. The last column provides a hint of how difficult the dataset is.

Figure 3 and 4 show some frames for each of the aforementioned datasets. From these few examples it is possible to understand the main differences: constrained/controlled environment (IXMAS, KTH, MPII Cooling, Berkeley MHAD, etc), unconstrained condition of the scene (ActivityNet, CollectiveActivity, Highfive, HMDB51, etc). Some frames also reveal the high complexity of the dataset, with regard to scene diversity (ActivityNet), low image quality (KTH), to mention few.

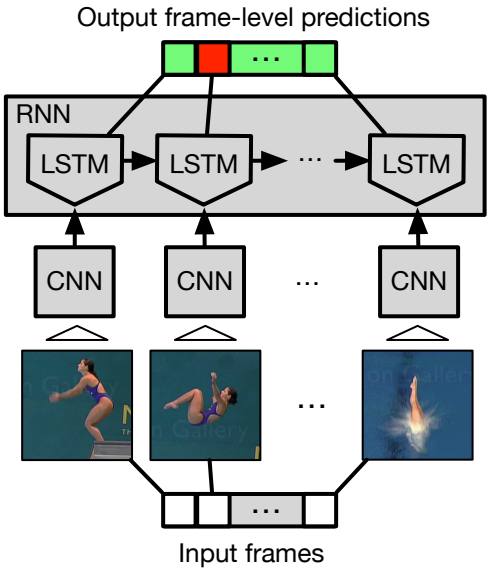
Table 4 and Table 5 summarize the most recent approaches that obtained remarkable results against two of the most well-known and challenging datasets in action recognition, UCF-101 and THUMOS-14. Reviewing top ranked methods at UCF-101 dataset, we find that the most significant difference among them is the strategy for splitting video data and combine sub-sequence results. Wang et al. (2016g) encodes the changes in the environment



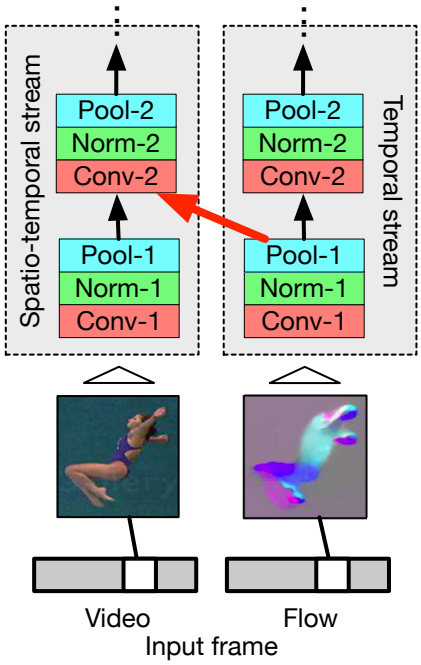
(a) 2D vs. 3D convolutions



(b) Motion-based features (pre-computed optical flow maps as inputs)



(c) Temporal sequence modeling (via LSTM)



(d) Fusion strategies (temporal fused into the spatial stream

Figure 2: Illustrative examples of the different architectures and fusion strategies

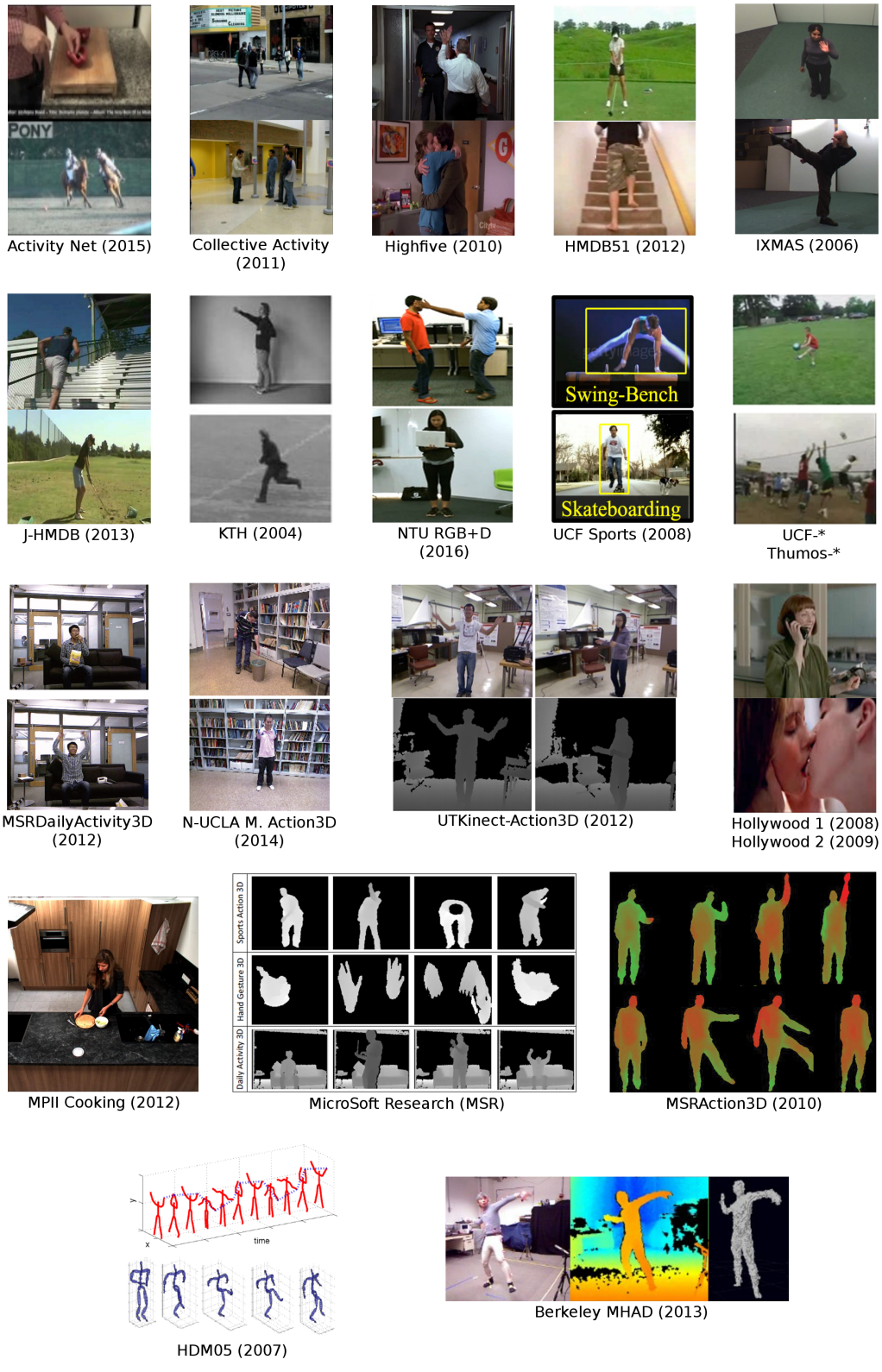


Figure 3: Action datasets: sample images

Table 1: Action datasets.
Notation:
In the Modality column: Depth, Skeleton, Audio, grayscale Intensity, InfraRed.
In Performance column: Accuracy, mean Average Precision, Intersection over Union

Year	Database	Problem	Body Parts	Modality	No.Classes	Performance
2004	KTH	AC	F	I	6	98.67% Acc Zhou et al. (2016)
2006	IXMAS	AC	F	RGB, A	13	98.79% Acc Turaga et al. (2008)
2007	HDM05	AC	F	S	100	98.17% Acc Chaudhry et al. (2013)
2008	HOHA (Hollywood 1)	AC, TL	F, U, L	RGB	8	71.90% Acc Saha et al. (2016), 0.787@0.5 mAP Mettes et al. (2016)
2008	UCF Sports	AC, STL	F	RGB	10	95.80% Acc Shao et al. (2016), 0.789@0.5 mAP Mettes et al. (2016)
2009	Hollywood 2	AC	F, U, L	RGB	12	78.50 mAP Liu et al. (2017)
2009	UCF11 (YouTube Action)	AC, STL	F	RGB	11	93.77% Acc Peng et al. (2014), -
2010	Highfive	AC, STL	F,U	RGB	4	69.40 mAP Wang et al. (2015a), 0.466 IoU Avgerinakis et al. (2015)
2010	MSRAction3D	AC	F	D, S	20	97.30% Acc Luo et al. (2013)
2010	MSRAction II	STL	F	RGB	3	85.00@0.125% mAP Chen and Corso (2015)
2010	Olympic Sports	AC	F	RGB	16	96.60% Acc Li et al. (2016a)
2011	Collective Activity (Extended)	AC	F	RGB	6	90.23% Acc Amer et al. (2013)
2011	HMDB51	AC	F, U, L	RGB	51	73.60% Acc Wang et al. (2016a)
2012	MPII Cooking	AC, TL	F, U	RGB	65	72.40 mAP Zhou et al. (2015), -
2012	MSRDailyActivity3D	AC	F,U	RGB, D, S	16	97.50% Acc Shahroudy et al. (2016b)
2012	UCF101	AC,TL	F, U, L	RGB	101	94.20% Acc Wang et al. (2016c), 46.77@0.2 mAP (split 1) phi
2012	UCF50	AC	F, U, L	RGB	50	97.90% Acc Duta et al. (2017)
2012	UTKinect-Action3D	AC	F	RGB, D, S	10	98.80% Acc Kerola et al. (2017)
2013	J-HMDB	AC, STL	F, U, L	RGB, S	21	71.08 Acc Peng and Schmid (2016), 73.1@0.5 mAP Saha et al. (2016)
2013	Berkeley MHAD	AC	F	RGB, D, S, A	11	100.00% Acc Chaudhry et al. (2013)
2014	N-UCLA Multiview Action3D	AC	F	RGB, D, S	10	90.80% Acc Kerola et al. (2017)
2014	Sports 1-Million	AC	F, U, L	RGB	487	73.10% Acc Yue-Hei Ng et al. (2015)
2014	THUMOS-14	AC, TL	F, U, L	RGB	101, 20 *	71.60 mAP Jain et al. (2015c), 0.190@0.5 mAP Shou et al. (2016a)
2015	THUMOS-15	AC, TL	F, U, L	RGB	101, 20 *	80.80 mAP Li et al. (2016a), 0.183@0.5 mAP (a)
2015	ActivityNet	AC, TL	F, U, L	RGB	200	93.23 mAP (b), 0.594@0.5 mAP Montes et al. (2016)
2016	NTU RGB+D	AC	F	RGB, D, S, IR	60	{69.20, 77.70}[1] Acc Liu et al. (2016a)

* A different number of classes is used for different problems. For TL/STL, "@" indicates amount overlap with groundtruth considered for positive localization. For instance, @0.5 indicates a 50% of overlap.
(a) Winner method from (`http://activity-net.org/challenges/2016/program.html#leaderboard`).
(b) Winner method from `http://www.thumos.info/results.html`.
[1] {cross-subject accuracy, cross-view accuracy}.

by dividing the input sequence into two parts, precondition and effect states, and then look for a matrix transformation between these two states. Li et al. (2016a) processes the input video as a hierarchical structure over the time in 3 levels, i.e. short-term, medium-range and long-range. Varol et al. (2016) achieve the best performance by using different temporal resolutions of RGB and optical flow.

Looking at the top ranked deep models on the THUMOS 2014 challenge, almost all the winners in 2015 use different combinations of appearance and motion features. For

Table 2: Gesture datasets.
Notation:
In the Modality column: Depth, Skeleton.
In the Performance column: Accuracy, Intersection over Union

Year	Database	Problem	Body Parts	Modality	No.Class	Performance
2011	ChaLearn Gesture	GC	F, U	RGB, D	15	-
2012	MSR-Gesture3D	GC	F, H	RGB, D	12	98.50% Acc Chen et al. (2016)
2014	ChaLearn (Track 3)	GC, TL	U	RGB, D, S	20	98.20 Acc Molchanov et al. (2016) 0.870 IoU Neverova et al. (2015b)
2015	VIVA Hand Gesture	GC	H	RGB	19	77.50% Acc Molchanov et al. (2015)
2016	ChaLearn conGD	TL	U	RGB, D	249	0.315 IoU Camgoz et al. (2016)
	ChaLearn isoGD	GC				67.19% Acc Duan et al. (2016)

Figure 4: Gesture datasets: sample images

the appearance ones, most of the methods extract frame-level CNN descriptors, and video representation is generated using a pooling method over the sequence. The motion-based features used by the top ranked methods can be divided into three groups, FlowNet, 3D CNN, and iDTs. In Qiu et al. (2015), we provide a comparison of those showing 3D CNN achieves the best result.

2.4 Challenges

Every year computer vision organizations arrange competitions providing useful datasets with annotations carefully designed according to the problem to face. Table 3 shows 5 main challenges in computer vision. For each challenge we report the year in which it took place, the dataset provided to the participant along with the task to be faced, the associated event, the winner of the challenge, and a list of top results obtained against the competition dataset.

Table 3: Challenges

Year	Challenge	Database	Task	Event	Winner	Results
2011	Opportunity	Opportunity	AR	-	CSTAR	Sagha et al. (2011b) Chavarriaga et al. (2011) Sagha et al. (2011a)
2012	HARL	LIRIS	AR	ICPR	Ni et al. (2013)	Wolf et al. (2014) Gu et al. (2016)*
2012	VIRAT	VIRAT DB	AR	CVPR	-	Vondrick and Ramanan (2011) Oh (2011)
2012	ChaLearn	CGD	GR	-	Alfnie	Konecny and Hagara (2014)* Escalante et al. (2015)
2013		Montalbano	GR	-	Wu et al. (2013)	Bayer and Silbermann (2013)
2014		HuPBA 8K+	AR	ECCV	Peng et al. (2015)	-
		Montalbano	GR		Neverova et al. (2014)	Pigou et al. (2015b) Neverova et al. (2015b) Shu et al. (2015)
2015		HuPBA 8K+	AR	CVPR	Wang et al. (2015e)	-
2016		isoGD, conGD	GR	ICPR	Chai et al. (2016)	Karpathy et al. (2014), Wang et al. (2017)
2013	Thumos	UCF101	AR	ICCV	Jiang et al. (2013)	Sultani and Shah (2016) Soomro et al. (2015) Peng et al. (2013) Karaman et al. (2013)
2014		Thumos-14	AR	ECCV	Jain et al. (2014b)	Jain et al. (2015c) Shou et al. (2016a) Richard and Gall (2016)
2015		Thumos-15	AR	CVPR	Xu et al. (2015a)	Wang et al. (2015c) Yuan et al. (2016)
2015	VIVA	VIVA	GR	CVPR	Molchanov et al. (2015)	Ohn-Bar and Trivedi (2014)
2016	ROSE	NTU RGB+D	AR	ACCV	SEARCH	Shahroudy et al. (2016a)

* Non-deep learning method.

Table 4: UCF-101 dataset results

Ref.	Year	Features	Architecture	Score
Feichtenhofer et al. (2016a)	2016	ST-ResNet + iDT	2-stream ConvNet and ResNet	94.6%
Lev et al. (2016)	2016	RNN Fisher Vector	C3D + VGG-CCA + iDT	94.1%
Varol et al. (2016)	2016	Opt. Flow, RGB, iDT	LTC-CNN	92.7%
Wang et al. (2016h)	2016	conv5	2-Stream SR-CNN	92.6%
Feichtenhofer et al. (2016b)	2016	conv5, 3D pool	VGG-16, VGG-M, 3D CNN	*92.5%*
Wang et al. (2016g)	2016	CNN	Siamese VGG-16	92.4%
Li et al. (2016a)	2016	CNN fc7	2 CNNs (spatial + temporal)	92.2%
Wang et al. (2016b)	2016	3D CNN + RNN hierarchical local	Volumetric R-CNN (DANN)	91.6%
Wang et al. (2015b)	2015	CNN, Hog/Hof/Mbh	2-stream CNN	91.5%
Mansimov et al. (2015)	2015	CNN feat	3D CNN	89.7%
Bilen et al. (2016)	2016	Dynamic feat maps	BVLC CaffeNet	89.1%
Jain et al. (2015c)	2015	H/H/M, iDT, FV+PCA+GMM	8-layer CNN	88.5%
Sun et al. (2015)	2015	CNN	F_{ST}CN: 2 CNNs (spat + temp)	88.1%
Simonyan and Zisserman (2014)	2014	CNN	Two-stream CNN (CNN-M-2048)	88.0%
Mahasseni and Todorovic (2016)	2016	eLSTM, DCNN fc7	eLSTM, DCNN+LSTM	86.9%
Zhang et al. (2016)	2016	CNN	2 CNNs (spatial + temporal)	86.4%
Ye and Tian (2016)	2016	dense trajectory, C3D	RNN, LSTM, 3DCNN	85.4%
Peng and Schmid (2015)	2015	CNN fc6, HOG/HOF/MBH	VGG19 Conv5	79.52%±1.1% (tr2) 66.64% (tr1)
Karpathy et al. (2014)	2014	CNN features	2 CNN converge to 2 fc layers	65.4%, 68% mAP
Jain et al. (2015b)	2015	ImageNet CNN, word2vec GMM	CNN	63.9%
phi	2015	CNN	Spatial + motion CNN	54.28% mAP

Table 5: THUMOS-14 dataset results

Ref.	Year	Features	Architecture	Score
Jain et al. (2015c)	2015	H/H/M, IDT, FV+PCA+GMM.	8-layer CNN	71.6%
Zhang et al. (2016)	2016	CNN	2 CNNs (spatial + temporal)	61.5%
Jain et al. (2015b)	2015	ImageNet CNN, word2vec GMM	CNN	56.3%
Shou et al. (2016b)	2016	CNN fc6, fc7, fc8	3D CNN, Segment-CNN	19% mAP
Yeung et al. (2016)	2015	CNN fc7	VGG-16, 3-layer LSTM	17.1% mAP
Escorcia et al. (2016)	2016	fc7 3D CNN	C3D CNN net	.084% mAP@50 .121% mAP@100 .139% mAP@200 .125% mAP@500

3. Action Recognition

This section reviews deep methods for action (or activity) recognition according to the way they treat the temporal dimension: using 3D convolutions, pre-computed motion-based features, and temporal sequence models.

3.1 2D Convolutional Neural Networks

In these kind of approaches, action recognition is often performed at frame-level and then somehow aggregated (averaging the class score predictions on individual frames). Some works further explore the possibility of using several frames as input. In particular, Karpathy et al. (2014) studied the different alternatives for considering multiple frames in a 2D model; however they concluded there was not a gain in performance using multiple video frames over averaging single frame predictions. Instead, Wang et al. (2016c) randomly sample video frames from K equal width temporal segments, obtain K class score predictions,

compute the consensus scores, and use these in the loss function to learn from video representations directly, instead from one frame or one stack of frames. Zha et al. (2015) convolve each frame of the video sequence to obtain frame-level CNN features. They then perform spatio-temporal pooling on pre-defined spatial regions over the set of randomly sampled frames (50-120 depending on the sequence) in order to construct a video-level representation, which is later l2-normalized and classified using SVM. Wu et al. (2016d) model scene, object, and more generic feature representations using separate convolutional streams. For each frame, the three obtained representations are averaged and input to a three-layer fully connected network which provides the final output. Bilen et al. (2016) collapse the videos into dynamic images, that can be fed into CNNs for image classification, by using *rank pooling* Fernando et al. (2016). Dynamic images represent are simply the parameters of a ranking function that learned to order the video frames. In Rahmani and Mian (2016), the authors propose a CNN, not to classify actions in depth data directly, but to model poses in a view-invariant high-dimensional space. For this purpose, they generate a synthetic dataset of 3D poses from motion capture data that are later fit with a puppet model and projected to depth maps. The network is first trained to differentiate among hundreds of poses to, then, use the features of the penultimate fully-connected layer for action classification in a non-deep action recognition approach. Ni et al. (2016) exploit the combination of CNNs and LSTM for interactional object parsing on individual frames. Note LSTMs are not used for temporal sequence modeling but for refining object detections. For the action detection task, they then use object detections for pooling improved dense trajectories extracted on temporal segments.

Note that, independently from the discussed method, 2D convolutional filters in 2D CNNs only consider spatial inter-relations of pixels, ignoring their temporal neighborhood. Next we explore the more effective ways of exploiting spatiotemporal information in image sequences, which consist in either using pre-computed motion-based to include implicit temporal information in 2D CNNs or explicitly modeling temporal information with 3D CNNs or temporal sequence modeling methods.

3.2 Motion-based features

Researchers found that motion based features, such as optical flow, were a rich cue that could be fed directly as a network input. There are accurate and efficient methods to compute these kind of features, some of them by exploiting GPU capabilities (Fortun et al., 2015). The use of optical flow demonstrated to boost the performance of CNNs on action recognition-related tasks (Simonyan and Zisserman, 2014; Park et al., 2016; Zhang et al., 2016; Gkioxari and Malik, 2015).

Simonyan and Zisserman (2014) presented a two-stream CNN which incorporated both spatial (video frames) and temporal networks (pre-computed optical flow), and showed that the temporal networks trained on dense optical flow are able to obtain very good performance in spite of having limited training data. Along the same lines,Wang and Hoai (2016) propose a two-stream (spatial and temporal) net for non-action classification in temporal action localization. Similarly, Zhu et al. (2016b) use the same architecture for key-volume mining and classification in this case for spatio-temporal localization of actions. Chéron et al. (2015) extract both appearance and motion deep features from body

part detections instead of whole video frames. They then compute for each body part the min/max aggregation their descriptors over time. The final representation consists of the concatenation of pooled body part descriptors on both appearance and motion cues, which is comparable to the size of a Fisher vector. Park et al. (2016) used the magnitude of optical flow vectors as a multiplicative factor for the features from the last convolutional layer. This reinforces the attention of the network on the moving objects when fine-tuning the fully connected layers. Zhang et al. (2016) explored motion vectors (obtained from video compression) to replace dense optical flow. They adopted a knowledge transfer strategy from optical flow CNN to the motion vector CNN to compensate the lack of detail and noisiness of motion vectors.

Singh et al. (2016a) use a multi-stream network to obtain frame-level features. To the full-frame spatial and motion streams from Simonyan and Zisserman (2014), they add two other actor-centered (spatial and motion) streams that compute the features in the actor's surrounding bounding box obtained by a human detector algorithm. Moreover, motion features are not stacks of optical flow maps between pairs of consecutive frames, but among a central frame and neighboring ones (avoiding object's displacement along the stacked flow maps). Gkioxari and Malik (2015) and Weinzaepfel et al. (2015) propose a similar an approach for action localization. They first generate action region proposals from RGB frames using, respectively, selective search Uijlings et al. (2013) on and EdgeBoxes Zitnick and Dollár (2014). Regions are then linked and described with static and motion CNN features. However, high quality proposals can be obtained from motion. Peng and Schmid (2016) show a region proposals generated by a region proposal network (RPN) Ren et al. (2015) from motion (optical flow) were complementary to the ones generated by an appearance RPN. Note some of the works in Section 3.3 were using pre-computed motion features, which is not mutually exclusive with using motion features approaches. Varol et al. (2016) uses stacks of 60 pre-computed optical flow maps as inputs for the 3D convolutions, largely improving results obtained using raw video frames. Wang et al. (2016d) compute motion-like image representations from depth data by accumulating absolute depth differences of contiguous frames, namely hierarchical depth motion maps (HDMM).

In the literature there exist several methods which extend the deep-based methods with the popular dense trajectory features. Wang et al. (2015b) introduce a video representation called Trajectory-pooled Deep-convolutional Descriptor (TDD), which consists on extending the state-of-the-art descriptors along the trajectories with deep descriptors pooled from normalized CNN feature maps. Peng and Schmid (2015) propose a method based on a concatenation of iDT feature (HOG, HOF, MBHx, MBHy descriptors) and Fisher vector encoding and CNN features (VGG19). For CNN features they use VGG19 CNN to capture appearance features and VLAD encoding to encore/pool convolutional feature maps. Rahmani et al. (2016) utilize dense trajectories, and hence motion-based features, in order to learn view-invariant representations of actions. In order to model this variance, they generate a synthetic dataset of actions with 3D puppets from MoCap data that are projected to multiple 2D viewpoints from which fisher vectors of dense trajectories are used for learning a CNN model. During its training, an output layer is placed with as many neurons as training sequences so fisher vectors from different 2D viewpoints give same response. Afterwards, the concatenation of responses in intermediate layers (except for last one) provide the view-invariant representation for actions.

Differently from other works, Ng et al. (2016) jointly estimate optical flow and recognize actions in a multi-task learning setup. Their models consists in a residual network based on FlowNet He et al. (2016a) with extra additional classification layers, which learns to do both estimate optical flow and perform the classification task.

3.3 3D Convolutional Neural Networks

The early work of Ji et al. (2010) introduced the novelty of inferring temporal information from raw RGB data directly by performing 3D convolutions on stacks of multiple adjacent video frames, namely *3D ConvNets*. Since then, many authors tried to either further improve this kind of models (Tran et al., 2015; Mansimov et al., 2015; Sun et al., 2015; Shou et al., 2016b; Poleg et al., 2016; Liu et al., 2016b) or used them in combination with other hybrid deep-oriented models (Escorcia et al., 2016; Baccouche et al., 2011; Ye and Tian, 2016; Feichtenhofer et al., 2016b; Wu et al., 2016c; Li et al., 2016a).

In particular, Tran et al. (2015) proposed 3D convolutions with more modern deep architectures and fixed 3x3x3 convolution kernel size for all layers, that made 3D convnets more suitable for large-scale video classification. In general, 3D ConvNets can be expensive to train because of the large number of parameters, especially when training with bigger datasets such as 1-M sports dataset Karpathy et al. (2014) (which can take up to one month). Sun et al. (2015) factorized the 3D convolutional kernel learning into a sequential process of learning 2D spatial convolutions in lower convolutional layers followed by learning 1D temporal convolutions in upper layers. Mansimov et al. (2015) proposed initializing 3D convolutional weights using 2D convolutional weights from spatial CNN trained on ImageNET. This not only speeds up the training but also alleviates the overfitting problem on small datasets. Varol et al. (2016) extended the length of input clips from 16 to 60 frames in order model more long-term temporal information during 3D convolutions, but reduced the input's spatial resolution to maintain the model complexity. Poleg et al. (2016) introduced a more compact 3D ConvNet for egocentric action recognition by applying 3D convolutions and 3D pooling only at the first layer. However, they do not use raw RGB frames, but stacked optical flow. In the context of depth data, Liu et al. (2016b) propose re-scaling depth image sequences to a 3D cuboid and the use of 3D convolutions to extract spatio-temporal features. The network consists of two pairs of convolutional and 3D max-pooling followed by a two-layer fully-connected layer net.

3D convolutions are often used in more cumbersome hybrid deep-based approaches. Shou et al. (2016b) propose a multi-stage CNN, in this case for temporal action localization, consisting of three 3D convnets (Tran et al., 2015): a proposal generation network that learns to differentiate background from action segments, a classification network that aims at discriminating among actions and serves as initialization for a third network, the localization network with a loss function that considers temporal overlap with the ground truth annotations. Wang et al. (2016d) applied 3D ConvNets to action recognition from depth data. The authors train a separate 3D ConvNet for each Cartesian plane each of which fed with a stack of depth images constructed from different 3D rotations and temporal scales. Singh et al. (2016b) prove the combination of both 2D and 3D ConvNet can leverage the performance when performing egocentric action recognition. Li et al. (2016a) uses 3D convolutions from Tran et al. (2015) to model short-term action features on a hier-

archical framework in which linear dynamic systems (LDS) and VLAD descriptors are used to, respectively, model/represent medium- and long-range dynamics.

3.4 Temporal deep learning models: RNN and LSTM

The application of temporal sequence modeling techniques, such as LSTM, to action recognition showed promising results in the past (Baccouche et al., 2010; Grushin et al., 2013). Earlier works did not try to explicitly model the temporal information, but aggregated the class predictions got from individual frame predictions. For instance, in Simonyan and Zisserman (2014), sample 25 equally spaced frames (and their crops and flips) from each video and then average their predicted scores.

Today, we find the combination of recurrent networks, mostly LSTM, with CNN models for the task of action recognition. Veeriah et al. (2015) propose a new gating scheme for LSTM that takes into account abrupt changes in the internal cell states, namely *differential RNN*. They use different order derivatives to model the potential saliency of observed motion patterns in actions sequences. Singh et al. (2016a) presented a bi-directional LSTM, which demonstrated to improve the simpler uni-directional LSTMs. Yeung et al. (2016) introduce a fully end-to-end approach on a RNN agent which interacts with a video over time. The agent observe a frame and provides a detection decision (confidence and begin-end), to whether or not emit a prediction, and where to look next. While back-propagation is used to train the detection decision outputs, REINFORCE is required to train the other two (non-differentiable) agent policies. Mahasseni and Todorovic (2016) propose a deep architecture which uses 3D skeleton sequences to regularize an LSTM network (LSTM+CNN) on the video. The regularization process is done by using the output of the encoder LSTM (grounded on 3D human-skeleton training data) and by modifying the standard BPTT algorithm in order to address the constraint optimization in the joint learning of LSTM+CNN. In their most recent work, Wang et al. (2016b) explore contexts as early as possible and leverage evolution of hierarchical local features. For this, they introduce a novel architecture called deep alternative neural network (DANN) stacking alternative layers, where each alternative layer consists of a volumetric convolutional layer followed by a recurrent layer. Lev et al. (2016)introduce a novel Fisher Vector representation for sequences derived from RNNs. Features are extracted from input data via VGG/C3D CNN. Then a PCA/CCA dimension reduction and L_2 normalization are applied and sequential feature are extracted via RNN. Finally, another PCA+L_2-norm step is applied before the final classification.

Liu et al. (2016a) extend the traditional LSTM into two concurrent domains, i.e, spatio-temporal long short-term memory (ST-LSTM). In this tree structure each joint of the network receive contextual information from both neighboring joints and previous frame. Shahroudy et al. (2016a) propose a part aware extension of LSTM for action recognition by splitting the memory cell of the LSTM into part-based sub-cells. These sub-cells can yield the models learn the long-term patterns specifically for each part. Finally, the output of each unit is the combination of all sub-cells.

3.5 Deep learning with fusion strategies

Some methods have used diverse fusion schemes to improve recognition performance of action recognition. In Simonyan and Zisserman (2014), in order to fuse the class-level

predictions of two streams (spatial and temporal), the authors train a multi-class linear SVM on stacked L_2-normalized softmax scores, which showed to improve the fusion by simply averaging scores. Wang et al. (2015d), which improves the former work by making the networks deeper and improved data augmentation techniques, simply perform a linear combination of the prediction scores (2 for temporal net and 1 for the spatial net). Similarly, Wang et al. (2016c) combine RGB, RGB difference, flow, and warped flow assigning equal weight to each channel. Feichtenhofer et al. (2016b) fuse a spatial and temporal convnets at the last convolutional layer (after ReLU) to turn it into a spatio-temporal stream by using 3D Conv fusion followed by 3D pooling. The temporal stream is kept and both loss functions are used for training and testing.

Deng et al. (2015) present a deep neural-network-based hierarchical graphical model that recognizes individual and group activity in surveillance scenes. Different CNNs produce action, pose, and scene scores. Then, the model refines the predicted labels for each activity via multi-step Message Passing Neural Network which captures the dependencies between action, poses, and scene predicted labels. Du et al. (2015) propose an end-to-end hierarchical RNN for skeleton based action recognition. The skeleton is divided into five parts, each of which is feed into a different RNN network, the output of which are fused into higher-layer RNNs. The highest level representations are feed into a single-layer perceptron for the final decision. Singh et al. (2016b) face the problem of first person action recognition using a multi-stream CNN (ego-CNN, temporal, and spatial), which are fused by combining weighted classifier scores. The proposed ego-CNN captures hand-crafted cues such as hand poses, head motion, and saliency map.Wang et al. (2016h) incorporate a region-of-interest pooling layer after the standard convolutional and pooling layers that separates CNN features for three semantic cues (scene, person, and objects) into parallel fully connected layers. They propose four different cue fusion schemes at class prediction level (max, sum, and two weighted fusions).

He et al. (2016b) attempt to investigate human action recognition without the human presence in input video frames. They consider whether a background sequence alone can classify human actions.

Peng and Schmid (2016) perform action localization in space and time by linking via dynamic time warping the action bounding box detections on single frames. For bounding box classification, they concatenate the representations of multiple regions derived from the original detection bounding box. Feichtenhofer et al. (2016a) propose a two stream architecture (appearance and motion) based on residual networks. In order to model spatiotemporal information, they inject 4 residual connections (namely "skip-streams") from motion to the appearance stream (i.e., middle fusion) and also transform the dimensionality reduction layers from ResNet's original model to temporal convolution layers. Wang et al. (2016g) train two Siamese networks modeling, respectively, action's precondition and effect on the observed environment. Each net learns a high-dimensional representation of either precondition or effect frames along with the linear transformation per class that transforms precondition to effect. The nets are connected via their outputs and not sharing weights; i.e., late fusion.

4. Gesture Recognition

In this section we review recent deep-learning based approaches for gesture recognition in videos, mainly driven by the areas of human computer, machine, and robot interaction.

4.1 2D Convolutional Neural Networks

The first method that comes to mind for recognizing a sequence of images, is applying 2D CNNs on individual frames and then averaging the result for classification. Jain et al. (2014a) present a CNN deep learning architecture for human pose estimation and develop a spatial-contextual model that aims at making joint predictions by considering related joints positions. They train multiple convnets to perform independent binary body-part classification (i.e., presence or absence of that body part). These networks are applied as sliding windows to overlapping regions of the input which results in smaller networks and better performance. For human pose estimation, Li et al. (2015a) propose a CNN-based multi-tasking model. The authors use a CNN to extract features from the input image. These features are then used as the input of both joint point regression tasks and body-part detection tasks. Kang et al. (2015) exploit a CNN to extract features from the fully connected layer for sign language gesture recognition (finger spelling of ASL) from depth images.

Neverova et al. (2015a) propose a deep learning model for hand pose estimation that leverages both unlabeled and synthetically generated data for training. The key of the proposed model is that the authors encode structural information into the training objective by segmenting hands into parts, as opposed to including structure in the model architecture. Oyedotun and Khashman (2016) use CNN and *stacked denoising autoencoder* (SDAE) for recognizing 24 American Sign Language (ASL) hand gestures. Liang et al. (2016) propose a multi-view framework for hand pose recognition from point cloud. They form the view image by projecting hand point cloud to different view planes, and then using CNN to extract features from these views. Lin et al. (2015) propose a CNN that first detect hands using a GMM-skin detector and align them to the main axes. Then they apply a CNN comprising pooling and sampling layers, and on top a standard feed-forward NN that acted as classifier (heuristic rules on top of the output of the NN were defined).

In terms of hand pose estimation, Tompson et al. (2014) propose a CNN that recovers 3D joints based on synthetic training data. On top of the last layer a neural network transforms the outputs of the conv layers into heat maps (one per joint), indicating the probability-position for each joint. Poses are recovered from the set of heatmaps by solving an optimization problem.

4.2 Motion-based features

Neural networks and CNNs based on hand and body pose estimation as well as motion features have been widely applied for gesture recognition. If one wants to obtain better performance, temporal information rather than spatial data must be included in the models. For gesture *style* recognition in biometrics, Wu et al. (2016b) proposes a two-stream (spatio-temporal) CNN which learns from a set of training gestures. The authors use raw depth data as the input of spatial network and optical flow as the input of temporal one. For

articulated human pose estimation in videos Jain et al. (2015a) exploit both color and motion features. The authors propose a Convolutional Network (ConvNet) architecture for estimating the 2D location of human joints in video, with an RGB image and a set of motion features as the input data of this network. The motion features used in this methods are the perspective projection of the 3D velocity-field of moving surfaces.

Wang et al. (2017) use three representations of *dynamic depth image* (DDI), *dynamic depth normal image* (DDNI) and *dynamic depth motion normal image* (DDMNI) as the input data of 2D networks for gesture recognition from depth data. The authors construct these dynamic images by using bidirectional rank pooling from a sequence of depth images. These representations can effectively capture the spatio-temporal information. Wang et al. (2016e) propose a similar formulation for gesture recognition in continuous depth video. They first identify the start and end frames of each gesture based on *quantity of movement* (QOM), and then they construct *Improved Depth Motion Map* (IDMM) by calculating the absolute depth difference between current frame and the start frame for each gesture segment which is a kind of motion features as the input data of deep learning network.

4.3 3D Convolutional Neural Networks

Several 3D CNNs have been proposed for gesture recognition, most notably Molchanov et al. (2016); Huang et al. (2015); Molchanov et al. (2015). Molchanov et al. (2015) proposes a 3D CNN for driver hand gesture recognition from depth and intensity data. The authors combine information from multiple spatial scales for final prediction. It also employs spatio-temporal data augmentation for more effective training and to reduce potential overfitting. Molchanov et al. (2016) extend the 3D CNN with a recurrent mechanism for detection and classification of dynamic hand gestures. The architecture consists of a 3D-CNN for spatio-temporal feature extraction, a recurrent layer for global temporal modeling and a softmax layer for predicting class-conditional gesture probabilities.

Huang et al. (2015) proposes 3D CNN for sign language recognition which extracts discriminative spatio-temporal features from raw video stream. To boost the performances, multi-channels (RGB-D and Skeleton data) of video streams, including color information, depth clue and body joint positions are used as input to the 3D CNN. Li et al. (2016b) proposes a 3D CNN model for large scale gesture recognition by combining depth and RGB video. The proposed architecture is based on the model proposed by Tran et al. (2015). In a similar way, Zhu et al. (2016a) adopted the same architecture, but this time under a pyramidal for the same problem. In the same line, the work by Camgoz et al. (2016) builds an end to end 3D CNN using as basis the model of Tran et al. (2015) and applies it to large scale gesture spotting.

4.4 Temporal deep learning models: RNN and LSTM

Interestingly, temporal deep learning models have not been widely used for gesture recognition, despite this is a promising venue for research. We are aware of Neverova et al. (2013), where they propose a multimodal (depth, skeleton, and speech) human gesture recognition system based on RNN. Each modality is first processed separately in short spatio-temporal blocks, where discriminative data-specific features are either manually extracted or learned. Then, RNN is employed for modeling large-scale temporal dependencies, data fusion and

ultimately gesture classification. A multi stream RNN is also proposed by Chai et al. (2016) for large scale gesture spotting.

Eleni (2015) propose a Convolutional Long Short-Term Memory Recurrent Neural Network (CNNLSTM) able to successfully learn gesture varying in duration and complexity. Facing the same problem, Nishida and Nakayama (2016) propose a multi-stream model, called MRNN, which extends RNN capabilities with LSTM cells in order to facilitate the handling of variable-length gestures.

Wang et al. (2016f) propose *sequentially supervised long short-term Memory* (SS-LSTM), in which instead of assigning class label to the output layer of RNNs, auxiliary knowledge is used at every time step as sequential supervision. John et al. (2016) uses a deep learning framework to extract the representative frames from the video sequence and classify the gesture. They utilize a tiled image, created by sampling the whole video, as the input of a deconvenet to generates the tiled binary pattern. Then, These representative frames are given as input to the trained long-term recurrent convolution network. Koller et al. (2016) propose an EM-based algorithm integrating CNNs with Hidden-Markov-Models (HMMs) for weak supervision.

4.5 Deep Learning with fusion strategies

Multimodality in deep learning models has been widely exploited for gesture recognition. Wu et al. (2016a) propose a semi-supervised hierarchical dynamic framework by integrating deep neural networks within an HMM temporal framework, for simultaneous gesture segmentation and recognition using skeleton joint information, depth and RGB images. The authors utilize a Gaussian-Bernoulli Deep Belief Network to extract high-level skeletal joint features by, and 3D CNN to extract features from depth and RGB data. Finally, they applied intermediate (middle) and late fusion to get the final result. Neverova et al. (2015b) propose a multimodal multi-stream CNN for gesture spotting. The whole system operates at three temporal scales. Separate CNNs are considered for each modality at the beginning of the model structure with increasingly shared layers and a final prediction layer. Then, they fuse the result of each network by a meta-classifier independently at each scale; i.e., late fusion.

Pigou et al. (2015b) demonstrate that simple temporal feature pooling strategy (to take into account the temporal aspect of video) is not sufficient for gesture recognition, where temporal information is more discriminative compared to general video classification tasks. They explore deep architectures for gesture recognition in video and propose a new end-to-end trainable neural network architecture incorporating temporal convolutions and bidirectional recurrence. The authors test late and different kinds of middle fusions, to combine the result of CNN applied on each frame. Ouyang et al. (2014) present a deep learning model to fuse multiple information sources (i.e., appearance score, deformation and appearance mixture type) for human pose estimation. Three deep models take as input the output the information source from a state-of-the-art human pose estimator. The authors exploited early and middle fusion methods to integrate the models.

Li et al. (2015b) propose a CNN that learns to score pairs of input images and human poses (joints). The model is formed by two sub-networks: a CNN learns a feature embedding for the input images, and a two layer sub-network learns an embedding for the human

pose. These two kinds of features are separately fed through fully-connected layers, and then mapped into two embedding spaces. The authors then calculate score function by dot-product between the two embeddings; i.e. late fusion. Similarly, Jain et al. (2015a) propose a CNN for estimating 2D joints location. The CNN incorporates RGB image and motion features. The authors utilize early fusion to integrate these two kinds of features. For gesture recognition from RGB-D data Duan et al. (2016) use two general deep-based network; i.e., convolutional two stream consensus voting network (2SCVN) for modeling the RGB and optical flow and 3d depth-saliency ConvNet stream for processing saliency and depth data. Then, they use late fusion to fuse the result of these networks.

5. Discussion

In recent years deep learning methods have continued to be a thriving area of research in computer vision. These methods are end-to-end approaches for automatically learning semantic and discriminative feature representations directly from raw observations in many computer vision tasks. Thanks to the massive ImageNet dataset, CNN models overcome other hand-crafted features and achieve the best results on many recognition tasks. These achievements encourage researchers to design deep based models for learning an appropriate representation of image sequences.

In the following sections, the state of the art methods and deep-based platforms are summarized and then compared. We point out some tricks used for improving the result, and also address some limitations for future work.

5.1 Summary

As the recent success of deep learning models, many researchers have extended deep-based models representation of the sequences of images for human action recognition. Table 6 and 7 list a summary of all methods on human action and gesture recognition respectively. A very simple extension consists in applying the existing 2D networks on individual video frames and then aggregating the predictions over the entire sequence for video classification (hereinafter referred as 2D convolutional models). Since they do not model temporal information of any kind, some methods (the second category) propose utilizing pre-computed motion features as input data for those pre-trained 2D networks. In the third group, different 3D extensions of 2D deep models have been proposed. Methods in the fourth group exploited temporal models (e.g. RNN and LSTM) for processing the temporal dimension.

Table 6: Summary of all deep-based action recognition methods.
Notations:
In the Modality column: Depth, Skeleton.
In the Fusion column: Late, Early, Slow, and Middle

Year	**Reference**	**Model**				**Modality**	**Fusion**
		2D	Motion	3D	Temporal		
2010	Ji et al. (2010)	-	-	✓	-	RGB	-
2011	Baccouche et al. (2011)	-	-	✓	✓	RGB	-
2014	Karpathy et al. (2014)	✓	-	-	-	RGB	E-L-S

Continued on next page

Continued from previous page

2014	Simonyan and Zisserman (2014)	✓	✓	-	-	RGB	L
2015	Chéron et al. (2015)	✓	✓	-	-	RGB	L
2015	Deng et al. (2015)	✓	-	-	-	RGB	L-S
2015	Du et al. (2015)	-	-	-	✓	S	S
2015	Gkioxari and Malik (2015)	✓	✓	-	-	RGB	L
2015	Mansimov et al. (2015)	-	-	✓	-	RGB	-
2015	Peng and Schmid (2015)	✓	-	-	-	RGB	-
2015	Sun et al. (2015)	✓	-	-	-	RGB	-
2015	Tran et al. (2015)	-	-	✓	-	RGB	-
2015	Wang et al. (2015b)	-	✓	-	-	RGB	L
2015	Wang et al. (2015d)	-	✓	-	-	RGB	L
2015	Weinzaepfel et al. (2015)	-	✓	-	-	RGB	L
2015	Zha et al. (2015)	✓	-	-	-	RGB	L
2016	Bilen et al. (2016)	✓	-	-	-	RGB	-
2016	Feichtenhofer et al. (2016b)	✓	✓	-	-	RGB	S
2016	He et al. (2016b)	✓	✓	-	-	RGB	L
2016	Lev et al. (2016)	-	✓	✓	✓	RGB	-
2016	Li et al. (2016a)	✓	-	-	-	RGB	-
2016	Liu et al. (2016b)	-	-	✓	-	D, S	L
2016	Mahasseni and Todorovic (2016)	✓	-	-	-	RGB	-
2016	Ng et al. (2016)	-	✓	-	-	RGB	-
2016	Ni et al. (2016)	✓	-	-	✓	RGB	-
2016	Park et al. (2016)	✓	✓	-	-	RGB	S-L
2016	Peng and Schmid (2016)	✓	✓	-	-	RGB	L
2016	Poleg et al. (2016)	✓	✓	✓	-	RGB	-
2016	Rahmani and Mian (2016)	✓	-	-	-	D	-
2016	Rahmani et al. (2016)	✓	✓	-	-	RGB	E
2016	Shou et al. (2016b)	-	-	✓	-	RGB	-
2016	Singh et al. (2016b)	✓	✓	✓	-	RGB	L
2016	Singh et al. (2016a)	✓	✓	-	✓	RGB	L
2016	Varol et al. (2016)	-	✓	✓	-	RGB	-
2016	Escorcia et al. (2016)	-	-	✓	-	RGB	-
2016	Wang et al. (2016d)	-	-	✓	-	D	L
2016	Wang et al. (2016g)	✓	✓	-	-	RGB	L
2016	Wang et al. (2016b)	✓	-	-	✓	RGB	-
2016	Wang and Hoai (2016)	✓	✓	-	-	RGB	L
2016	Wang et al. (2016c)	✓	✓	-	-	RGB	L
2016	Wang et al. (2016h)	✓	✓	-	-	RGB	L
2016	Wu et al. (2016c)	-	-	✓	✓	RGB	-
2016	Wu et al. (2016d)	✓	-	-	-	RGB	L
2016	Yeung et al. (2016)	✓	-	-	✓	RGB	-
2016	Ye and Tian (2016)	-	✓	✓	✓	RGB	-
2016	Zhang et al. (2016)	✓	✓	-	-	RGB	L
2016	Zhu et al. (2016b)	-	✓	-	-	RGB	L

Table 7: Summary of all deep-based gesture recognition methods.
Notations:
In the Modality column: Depth, Skeleton, Audio, InfraRed.
In the Fusion column: Early, Middle, Late, Slow

Year	Reference	Model				Modality	Fusion
		2D	Motion	3D	Temporal		
2013	Neverova et al. (2013)	-	-	✓	✓	D-S-A	L
2014	Tompson et al. (2014)	✓	-	-	-	RGB-D	-
2014	Jain et al. (2014a)	✓	-	-	-	RGB	-
2014	Ouyang et al. (2014)	✓	-	-	-	RGB	E-M
2015	Molchanov et al. (2015)	-	-	✓	-	RGB-D	L
2015	Huang et al. (2015)	-	-	✓	-	RGB-D-S	L
2015	Lin et al. (2015)	✓	-	-	-	RGB	-
2015	Li et al. (2015a)	✓	-	-	-	RGB	-
2015	Eleni (2015)	✓	-	-	✓	RGB	-
2015	Kang et al. (2015)	✓	-	-	-	D	-
2015	Li et al. (2015b)	✓	-	-	-	RGB-S	L
2015	Jain et al. (2015a)	-	✓	-	-	RGB	E
2015	Neverova et al. (2015a)	✓	-	-	-	D	-
2015	Neverova et al. (2015b)	-	-	✓	-	RGB-S-A	L
2015	Pigou et al. (2015b)	✓	-	-	✓	RGB-D	L-S
2016	Molchanov et al. (2016)	-	✓	✓	✓	RGB-D-IR	L
2016	Wu et al. (2016b)	-	✓	-	-	D	L
2016	Nishida and Nakayama (2016)	✓	-	-	✓	RGB-D	L
2016	Wu et al. (2016a)	-	✓	✓	✓	RGB-D	M-L
2016	Wang et al. (2016f)	✓	-	-	✓	RGB	-
2016	Duan et al. (2016)	✓	✓	✓	-	RGB-D	L
2016	John et al. (2016)	✓	-	-	✓	RGB	-
2016	Oyedotun and Khashman (2016)	✓	-	-	-	RGB	-
2016	Liang et al. (2016)	✓	-	-	-	D	L
2016	Wang et al. (2016e)	-	✓	-	-	D	-
2016	Li et al. (2016b)	-	-	✓	-	RGB-D	L
2016	Zhu et al. (2016a)	-	-	✓	-	RGB-D	E
2016	Camgoz et al. (2016)	-	-	✓	-	RGB	L
2016	Chai et al. (2016)	-	-	-	✓	RGB-D	M
2016	Koller et al. (2016)	✓	-	-	✓	RGB	-
2017	Wang et al. (2017)	-	✓	-	-	D	L

5.2 Comparison

The most crucial challenge in deep-based human action and gesture recognition is temporal analysis, for which many architectures have been proposed. These approaches have been classified into four groups; i.e. 2D models, motion-based input model, 3D models, and temporal models. Generally, there are two main issues for comparing the methods; i.e., *how*

does the method deal with the temporal information? and *how can such a large network be trained with small datasets?*

As discussed, methods in the first category only use the appearance (spatial) information to extract features. In other words, there is no temporal processing for these methods. However, because of the availability of large annotated datasets (e.g. ImageNet), it is easier for these methods to be fine tuned on pre-trained models. In the second group, motion features such as optical flow, computed from data before their usage, are fed to the deep models. It has been shown that using training networks on pre-computed motion features is an effective way to save them from implicit learning of motion features. Moreover, fine-tuning motion-based networks with spatial data (ImageNet) proved to be effective. Allowing networks which are fine-tuned on stacked optical flow frames to achieve good performance in spite of having limited training data. However, these models can only exploit limited (local) temporal information.

Methods in the third category, learn spatio-temporal features by 3D filters in their 3D convolutional and pooling layers. It has been shown 3D networks over a long sequence are able to learn more complex temporal patterns Varol et al. (2016). Because of the amount of parameters to learn, training these networks is a challenging task, specially compared to motion-based methods (Simonyan and Zisserman, 2014). Because of the required amount of data, the problem of weights initialization has been investigated. The transformation of 2D Convolutional Weights into 3D ones yield models to achieve better accuracy than training scratch Mansimov et al. (2015). The most crucial advantage of approaches in the fourth group (i.e. temporal models like RNN and LSTM) is that they are able to cope with longer-range temporal relations. These models are preferred when dealing with skeletal data. Since skeleton features are low-dimensional, these networks have fewer weights, and thus, can be trained with fewer data.

We find from Table 4-5, the methods that achieved the best results on two of the most well-known datasets, still using hand-crafted features alongside deep-based features. In other words, action and gesture recognition has not gained a high performance from deep networks compared with other research areas (like image classification). These fields of research still needs to be grown.

Based on the influence of millions of network parameters, in addition to the different strategies for data augmentation, and the current allowed procedure of the usage of pre-trained models, current comparison among method performances for action and gesture recognition is a difficult task. In this sense, we expect in a near future the definition of protocols that will allow for a more accurate comparison of deep-based action and gesture recognition models. More precisely, we refer to Xu et al. (2015b) as the winner of THUMOS 2015 with the best result. This approach used VGG16 to extract frame-level features from the fully connected layers such as fc6 and fc7. Then, using Fisher vector and VLAD, they aggregated all the frames into single video-level representation. They also extracted *latent concept descriptors* (LCD) extracted by a GoogLeNet with Batch Normalization. An enhanced version of improved dense trajectories (iDT), acoustic features MFCC and ASR were also used in this work.

Recently, new deep architectures have started to be used for action/gesture recognition, such as gate-recurrent-unit RNNs (Ballas et al., 2016) (sparse GRU-RNNs that reduce the number of parameters of the network) and siamese architectures (Wang et al., 2016g) (that

allow multi-task learning). More insights into these architectures, and, of course, the use of more recent ones (like Radford et al. (2016)) are promising venues for research.

5.3 Tricks

Regardless of the model, performance is dependent on a large number of parameters that have to be learned from limited data. Strategies for data augmentation and pre-training are common. Likewise, training mechanisms to avoid overfitting (e.g. dropout) and to control the learning rate (e.g. extensions to SGD and Nesterov momentum) have been proposed. Improvements on those strategies are expected in the next few years. The community is nowadays putting efforts on building larger data sets that can cope with huge-parametric deep models (Abu-El-Haija et al. (2016); Heilbron et al. (2015)) and on challenge organization (with novel data sets and well defined evaluation protocols) that can advance the state-of-the-art in the field and make easier the comparison among deep learning architectures (Shahroudy et al., 2016a; Escalante et al., 2016b).

Taking into account the full temporal scale, results in a huge amount of weights for learning. To address this problem and decrease the number of weights, a good trick is to decrease the spatial resolution while increasing the temporal length.

Another trick to improve the result of deep-based models is data fusion. There could be separated networks, trained on different kinds of input data, different kinds of primary features, different portions of input data, and so on. It is well-known that ensemble learning is a powerful way to boost the performance of any machine learning approach. It proved to reduce the bias and variance errors of the learning algorithm (Neverova et al., 2014). We find new methodologies that ensemble several deep models for action and gesture recognition, not necessarily combining different data modalities, but with different sampling of the data and learning parameters (Wang et al., 2016c; Varol et al., 2016). This provides complementary information learned by the different deep models, being able to recover from uncorrelated errors of individual models (Neverova et al., 2014). Recently it is common to see this kind of strategies in action/gesture recognition competitions, where a minor improvement of the model can make the difference to achieve the best performance Varol et al. (2016).

It has been proved that the result of the temporal models (e.g. RNN) on skeletal data can be improved by extending these models to learn two domains, i.e., spatial and temporal, simultaneously Liu et al. (2016a). In other words, each state of the network receives contextual information from neighboring joints in human skeleton (spatial information) and also from previous frames (temporal information).

Finally, a common way to improve the performance of action or gesture recognition is the combination of deep learning-based features and hand-crafted ones. This combination could be performed in different layers of the deep models.

5.4 Platforms

One of the reasons that supports the applicability of deep learning in several areas is code sharing. In fact, there are many open source libraries implementing standard deep learning models. Many authors have published deep-based toolkits that make the research progress easier for the community. Among the most popular ones are Caffe (Jia et al.,

2014), CNTK (Yu et al., 2014), Matlab (Rahmani et al., 2016), TensorFlow (Abadi et al., 2015b), Theano (Al-Rfou et al., 2016), and Torch (Liu et al., 2016b).

Caffe (Jia et al., 2014), is the first deep learning toolkit developed by the Berkeley Vision and Learning Center. It is a Python Library primary focused on CNN, with a poor support of RNN. Caffe is useful for performing image analysis and benefits from having a large repository of pre-trained neural network models. It includes state-of-the-art models (mostly 2D networks) that achieve world class results on standard computer vision datasets. Caffe has been also used to implement 3D-CNN for action recognition (Tran et al., 2015; Poleg et al., 2016; Shou et al., 2016b; Wang et al., 2016d; Singh et al., 2016b), and motion-based approaches for both action (Simonyan and Zisserman, 2014; Zhang et al., 2016; Singh et al., 2016a; Gkioxari and Malik, 2015) and gesture recognition (Wu et al., 2016b; Wang et al., 2017, 2016e). Caffe is preferred to other frameworks for its speed and efficiency, especially in "fused" architectures for action recognition (Singh et al., 2016b; Deng et al., 2015; Diba et al., 2016; Peng and Schmid, 2016). Popular network types like FNN, CNN, LSTM, and RNN are fully supported by CNTK (Yu et al., 2014), which was started by speech processing researchers. On the other hand, TensorFlow (Abadi et al., 2015a) is an C++ toolkit in deep learning under an open source Apache 2.0 License by Google. It fully supports 2D CNNs and RNNs implementations, but not 3D CNNs.

Torch (Collobert et al., 2002) is a script language based on the Lua programming language that provides a rich set of RNN functions. For this reason it has been efficiently used for temporal models in action recognition (Liu et al., 2016a; Shahroudy et al., 2016a). Moreover, most of the 3D CNN-based methods utilized Torch to implement their networks. CUDA is a parallel computing platform and application programming interface (API) model created by Nvidia in order to use GPU. Cuda-convnet and CuDNN support all the mainstream softwares such as Caffe, Torch, Theano. Few methods also use MATLAB, e.g. Rahmani et al. (2016); one of the easiest and most productive software environment for engineers and scientists, widely used also in machine learning, signal and image processing, and computer vision.

5.5 Future work

Deep learning methods emerged not so long ago in the fields of human action and gesture recognition. Even when there is already too much work on deep learning in these topics, there are still several directions in which we foresee deep learning can have a broad impact in the forthcoming years. We briefly review these possible line of research that will be fruitful in the short term future.

Regarding applications, deep learning techniques have been successfully used in surveillance Ahmed et al. (2015), health care Liang et al. (2014), robotics Yu et al. (2013), human-computer interaction Mnih et al. (2015), and so on. We anticipate deep learning will prevail in emerging applications/areas like fine grained action recognition, action description generation, social signal processing, affective computing, and personality analysis, among others.

Another important trend of current deep-based models for action and gesture recognition is the inclusion of contextual cues. While it has been partially considered for gesture recognition (e.g. part-based human-models and scene understanding in combination with depth maps), until recent years very few works considered robust contextual cues for action

recognition. We anticipate context information will be critical for developing explanatory deep learning models for action and gesture recognition. Classical action recognition tasks were mainly addressed by the description of spatio-temporal local patches. Nowadays we can find strategies that incorporate environment recognition, and articulated human body Wang et al. (2016g), places Zhou et al. (2014), and objects Jain et al. (2015c). Moreover, we expect novel architectures and fusion schemes to exploit context and enhanced articulated human body pose estimation to keep progressing in the next few years. It is also expected that there will be advances in hybrid models combining handcrafted and learned descriptors Neverova et al. (2014); Wang et al. (2015b); Ji et al. (2013). Similarly, we think the community will pay attention to deep learning solutions for large scale and real time action and gesture recognition (Han et al., 2016; Zhang et al., 2016). Finally, it is important to mention that most of the surveyed methods targeted merely recognition/classification on already pre-segmented action/gesture clips. Additional effort is expected to advance in the research of methods able to simultaneously perform both detection and recognition tasks in long, realistic videos (Gkioxari and Malik, 2015; Shou et al., 2016b). As such, we envision other related problems like early recognition Escalante et al. (2016a), multi task learning Xu et al. (2016), captioning, recognition from low resolution sequences Nasrollahi et al. (2015) and from lifelog devices Rhinehart and Kitani (2016) will receive special attention within the next few years.

These days, we need to solve the problem of action recognition in more realistic long untrimmed videos. There are some other challenges in human action recognition with deep-based models that have been addressed by few researchers so far, like simultaneous detection and localization Gkioxari and Malik (2015). Another venue for research is early recognition of actions and gestures Escalante et al. (2016a). We need to know if the input video contains an action or not and then localizing temporally and spatially the action by finding the frames and regions in those frames, in which action is performed. Then after detection and localization, the action will be classified. It is anticipated that in the near future research will expand on both action detection and localization.

Acknowledgments

This work has been partially supported by the Spanish projects TIN2015-66951-C2-2-R and TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya. Hugo Jair Escalante was supported by CONACyT under grants CB2014-241306 and PN-215546.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke,

Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015a. URL `http://tensorflow.org/`. Software available from tensorflow.org.

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org*, 1, 2015b.

S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv:1605.02688*, 2016.

M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu. Monte carlo tree search for scheduling activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1353–1360, 2013.

R. Araujo and M. S. Kamel. A semi-supervised temporal clustering method for facial emotion analysis. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

K. Avgerinakis, K. Adam, A. Briassouli, and Y. Kompatsiaris. Moving camera human activity localization and recognition with motionplanes and multiple homographies. In *ICIP*, pages 2085–2089. IEEE, 2015.

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action classification in soccer videos with long short-term memory recurrent neural networks. In *International Conference on Artificial Neural Networks*, pages 154–159. Springer, 2010.

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

N. Ballas, L. Yao, and A. Courville. Delving deeper into convolutional networks for learning video representations. In *Proc. International Conference on Learning Representations*, 2016.

I. Bayer and T. Silbermann. A multi modal approach to gesture recognition from audio and video data. In *ICMI*, pages 461–466, 2013. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2532592. URL `http://doi.acm.org/10.1145/2522848.2532592`.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *TNN*, 5(2):157–166, 1994.

H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016.

N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *Proceedings IEEE International Conference of Pattern Recognition (International Conference on Pattern Recognition), ChaLearn Workshop*, 2016.

X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen. Two streams recurrent neural networks for large-scale continuous gesture recognition. In *Proc. of International Conference on Pattern RecognitionW*, 2016.

R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 471–478, 2013.

R. Chavarriaga, H. Sagha, and J. del R. Milln. Ensemble creation and reconfiguration for activity recognition: An information theoretic approach. In *SMC*, pages 2761–2766, 2011. ISBN 978-1-4577-0652-3. URL `http://dblp.uni-trier.de/db/conf/smc/smc2011.html#ChavarriagaSM11`.

C. Chen, B. Zhang, Z. Hou, J. Jiang, M. Liu, and Y. Yang. Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features. *Multimedia Tools and Applications*, pages 1–19, 2016.

W. Chen and J. J. Corso. Action detection by implicit intentional motion clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3298–3306, 2015.

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*, pages 3218–3226, 2015.

R. Collobert, S. Bengio, and J. Marithoz. Torch: A modular machine learning software library, 2002.

Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori. Deep structured models for group activity recognition. In M. W. J. Xianghua Xie and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 179.1–179.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.179. URL `https://dx.doi.org/10.5244/C.29.179`.

Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

A. Diba, A. Mohammad Pazandeh, H. Pirsiavash, and L. Van Gool. Deepcamp: Deep convolutional action and attribute mid-level patterns. In *IEEE CVPR*, 2016.

Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, June 2015. doi: 10.1109/CVPR.2015.7298714.

J. Duan, S. Zhou, J. Wan, X. Guo, and S. Z. Li. Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition. *arXiv preprint arXiv:1611.06689*, 2016.

I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe. Spatio-temporal vlad encoding for human action recognition in videos. In *International Conference on Multimedia Modeling*, pages 365–378. Springer, 2017.

T. Eleni. Gesture recognition with a convolutional long short term memory recurrent neural network. In *ESANN*, 2015. URL `https://books.google.cl/books?id=E8qMjwEACAAJ`.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 172–179, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460125. URL `http://doi.acm.org/10.1145/1460096.1460125`.

H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan. Principal motion components for gesture recognition using a single example. *PAA*, 2015.

H. J. Escalante, E. F. Morales, and L. E. Sucar. A naïve bayes baseline for early gesture recognition. *PRL*, 73:91–99, 2016a.

H. J. Escalante, V. Ponce, J. Wan, M. Riegler, A. Clapes, S. Escalera, I. Guyon, X. Baro, P. Halvorsen, H. Müller, and M. Larson. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Proc. International Conference on Pattern Recognition*, 2016b.

V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. DAPs: Deep action proposals for action understanding. *European Conference on Computer Vision*, 2016.

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

C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016b.

B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with lstm recurrent networks. *JMLR*, 3(Aug):115–143, 2002.

G. Gkioxari and J. Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015.

A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra. Robust human action recognition via long short-term memory. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–8. IEEE, 2013.

F. Gu, M. Sridhar, A. Cohn, D. Hogg, F. Flrez-Revuelta, D. Monekosso, and P. Remagnino. Weakly supervised activity analysis with spatio-temporal localisation. *Neurocomputing*, 2016. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2016.08.032. URL `http://www.sciencedirect.com/science/article/pii/S0925231216308748`.

S. Han, H. Mao, and W. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *Proc. International Conference on Learning Representations*, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016a.

Y. He, S. Shirakabe, Y. Satoh, and H. Kataoka. Human action recognition without human. In *Proc. European Conference on Computer Vision 2016 Workshops*, pages 11–17. Springer, 2016b.

F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-e video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, page 91, 1991.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

J. Huang, W. Zhou, H. Li, and W. Li. Sign language recognition using 3d convolutional neural networks. In *ICME*, pages 1–6, 2015.

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

A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. In *International Conference on Learning Representations*, pages 1–14. Cornell University, 2014a.

A. Jain, J. Tompson, Y. LeCun, and C. Bregler. *MoDeep: A deep learning framework using motion features for human pose estimation*, volume 9004, pages 302–315. 2015a.

M. Jain, J. van Gemert, and C. G. M. Snoek. University of amsterdam at thumos challenge 2014. In *ECCV THUMOS Challenge 2014*, Zürich, Switzerland, September 2014b.

M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *IEEE ICCV*, 2015b. URL `http://arxiv.org/abs/1510.06939`.

M. Jain, J. C. van Gemert, and C. G. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, pages 46–55, 2015c.

S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 495–502, 2010.

S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.59. URL `http://dx.doi.org/10.1109/TPAMI.2012.59`.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678. ACM, 2014.

Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `/ICCV13-Action-Workshop/`, 2013.

V. John, A. Boyali, S. Mita, M. Imanishi, and N. Sanma. Deep learning-based fast hand gesture recognition using representative frames. In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–8. IEEE, 2016.

J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–223, 2014.

B. Kang, S. Tripathi, and T. Q. Nguyen. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. *ACPR*, abs/1509.03001, 2015.

S. Karaman, L. Seidenari, A. D. Bagdanov, and A. D. Bimbo. L1-regularized logistic regression stacking and transductive crf smoothing for action recognition in video. In *Results of the THUMOS 2013 Action Recognition Challenge with a Large Number of Classes*, 2013.

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*, pages 1725–1732, 2014.

T. Kerola, N. Inoue, and K. Shinoda. Cross-view human action recognition from depth maps using spectral graph sequences. *Computer Vision and Image Understanding*, 154: 108–126, 2017.

O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2016.

J. Konecny and M. Hagara. One-shot-learning gesture recognition using hog-hof features. *JMLR*, 15:2513–2532, 2014. URL `http://jmlr.org/papers/v15/konecny14a.html`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Y. Kuniyoshi, H. Inoue, and M. Inaba. Design and implementation of a system that generates assembly programs from visual recognition of human action sequences. In *Intelligent Robots and Systems' 90.'Towards a New Frontier of Applications', Proceedings. IROS'90. IEEE International Workshop on*, pages 567–574. IEEE, 1990.

G. Lev, G. Sadeh, B. Klein, and L. Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850. Springer, 2016.

S. Li, Z.-Q. Liu, and A. B. Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *IJCV*, 113(1):19–36, May 2015a. ISSN 0920-5691. doi: 10.1007/s11263-014-0767-8. URL `http://dx.doi.org/10.1007/s11263-014-0767-8`.

S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, pages 2848–2856, 2015b.

Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos. Vlad3: Encoding dynamics of deep features for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1951–1960, 2016a.

Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song. Large-scale gesture recognition with a fusion of rgb-d data based on c3d model. In *Proc. of International Conference on Pattern RecognitionW*, 2016b.

C. Liang, Y. Song, and Y. Zhang. Hand gesture recognition using view projection from point cloud. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4413–4417. IEEE, 2016.

Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu. Deep learning for healthcare decision making with emrs. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 556–559. IEEE, 2014.

H.-I. Lin, M.-H. Hsu, and W.-K. Chen. Human hand gesture recognition using a convolution neural network. In *CASE*, pages 1038–1043, 2015.

A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *TPAMI*, 39(1):102–114, 2017.

J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016a.

Z. Liu, C. Zhang, and Y. Tian. 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*, 55:93–100, 2016b.

J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1809–1816, 2013.

B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016.

E. Mansimov, N. Srivastava, and R. Salakhutdinov. Initialization strategies of spatio-temporal convolutional neural networks. *arXiv preprint arXiv:1503.07274*, 2015.

R. Marks. System and method for providing a real-time three-dimensional interactive environment, Dec. 6 2011. US Patent 8,072,470.

P. Mettes, J. C. van Gemert, and C. G. Snoek. Spot on: Action localization from pointly-supervised proposals. In *European Conference on Computer Vision*, pages 437–453. Springer, 2016.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3d convolutional neural networks. In *CVPRW*, pages 1–7, June 2015. doi: 10.1109/CVPRW.2015.7301342.

P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *CVPR*, 2016.

A. Montes, A. Salvador, and X. Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *arXiv preprint arXiv:1608.08128*, 2016.

H. Mousavi Hondori and M. Khademi. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of Medical Engineering*, 2014, 2014.

K. Nasrollahi, S. Escalera, P. Rasti, G. Anbarjafari, X. Bar, H. J. Escalante, and T. B. Moeslund. Deep learning based super-resolution for improved action recognition. In *IPTA*, pages 67–72, 2015. ISBN 978-1-4799-8637-8. URL `http://dblp.uni-trier.de/db/conf/ipta/ipta2015.html#NasrollahiERABE15`.

N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout. A multi-scale approach to gesture detection and recognition. In *ICCVW*, pages 484–491, 2013. URL `http://liris.cnrs.fr/publis/?id=6330`.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multi-scale deep learning for gesture detection and localization. In *ECCVW*, volume 8925 of *LNCS*, pages 474–490, 2014.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. *ACCV*, volume 9005 of *LNCS*, chapter Hand Segmentation with Structured Convolutional Learning, pages 687–702. Cham, 2015a. ISBN 978-3-319-16811-1. doi: 10.1007/978-3-319-16811-1_45. URL `http://dx.doi.org/10.1007/978-3-319-16811-1_45`.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE TPAMI*, 2015b.

J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis. Actionflownet: Learning motion representation for action recognition. *arXiv preprint arXiv:1612.03052*, 2016.

B. Ni, Y. Pei, Z. Liang, L. Lin, and P. Moulin. Integrating multi-stage depth-induced contextual information for human action recognition and localization. In *FG*, pages 1–8, April 2013. doi: 10.1109/FG.2013.6553756.

B. Ni, X. Yang, and S. Gao. Progressively parsing interactional objects for fine grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2016.

N. Nishida and H. Nakayama. Multimodal gesture recognition using multi-stream recurrent neural network. In *PSIVT*, pages 682–694, 2016.

S. Oh. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153–3160, 2011. ISBN 978-1-4577-0394-2. doi: 10.1109/CVPR.2011.5995586. URL `http://dx.doi.org/10.1109/CVPR.2011.5995586`.

E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE-ITS*, 15(6):2368–2377, Dec 2014. ISSN 1524-9050. doi: 10.1109/TITS.2014.2337331.

F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. *CVPR*, pages 2337–2344, 2014.

O. K. Oyedotun and A. Khashman. Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, pages 1–11, 2016.

E. Park, X. Han, T. L. Berg, and A. C. Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

X. Peng and C. Schmid. Encoding feature maps of cnns for action recognition. In *CVPR, THUMOS Challenge 2015 Workshop*, 2015.

X. Peng and C. Schmid. Multi-region two-stream r-cnn for action detection. In *European Conference on Computer Vision*, pages 744–759. Springer, 2016.

X. Peng, L. Wang, Z. Cai, Y. Qiao, and Q. Peng. Hybrid super vector with improved dense trajectories for action recognition. In *ICCV Workshops*, volume 13, 2013.

X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*, pages 581–595. Springer, 2014.

X. Peng, L. Wang, Z. Cai, and Y. Qiao. *Action and Gesture Temporal Spotting with Super Vector Representation*, pages 518–527. 2015. ISBN 978-3-319-16178-5. doi: 10.1007/978-3-319-16178-5_36. URL `http://dx.doi.org/10.1007/978-3-319-16178-5_36`.

L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen. *European Conference on Computer Vision '14*, chapter Sign Language Recognition Using Convolutional Neural Networks, pages 572–578. Cham, 2015a. ISBN 978-3-319-16178-5. doi: 10.1007/978-3-319-16178-5_40. URL `http://dx.doi.org/10.1007/978-3-319-16178-5_40`.

L. Pigou, A. V. D. Oord, S. Dieleman, M. V. Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *CoRR*, abs/1506.01911, 2015b. URL `http://arxiv.org/abs/1506.01911`.

Y. Poleg, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

Z. Qiu, Q. Li, T. Yao, T. Mei, and Y. Rui. Msr asia msm at thumos challenge 2015. In *CVPR workshop*, volume 8, 2015.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. International Conference on Learning Representations*, 2016.

H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2016.

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

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

N. Rhinehart and K. M. Kitani. Learning action maps of large environments via first-person vision. In *Proc. European Conference on Computer Vision*, 2016.

A. Richard and J. Gall. Temporal action detection using a statistical language model. In *CVPR*, 2016.

H. Sagha, J. del R. Milln, and R. Chavarriaga. Detecting anomalies to improve classification performance in opportunistic sensor networks. In *PERCOM Workshops*, pages 154–159, March 2011a. doi: 10.1109/PERCOMW.2011.5766860.

H. Sagha, S. T. Digumarti, J. del R. Millán, R. Chavarriaga, A. Calatroni, D. Roggen, and G. Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *IEEE SMC*, pages 36 –40, Oct. 2011b. doi: 10.1109/ICSMC.2011.6083628.

S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.

J. Scharcanski and M. E. Celebi. *Computer vision techniques for the diagnosis of skin cancer*. Springer, 2014.

A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016a.

A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *arXiv preprint arXiv:1603.07120*, 2016b.

L. Shao, L. Liu, and M. Yu. Kernelized multiview projection for robust action recognition. *International Journal of Computer Vision*, 118(2):115–129, June 2016. URL `http://nrl.northumbria.ac.uk/24276/`.

Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016a.

Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016b.

Z. Shu, K. Yun, and D. Samaras. *Action Detection with Improved Dense Trajectories and Sliding Window*, pages 541–551. Cham, 2015. ISBN 978-3-319-16178-5. doi: 10.1007/978-3-319-16178-5_38. URL `http://dx.doi.org/10.1007/978-3-319-16178-5_38`.

K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576. 2014.

B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016a.

S. Singh, C. Arora, and C. Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016b.

K. Soomro, H. Idrees, and M. Shah. Action localization in videos through context walk. In *ICCV*, 2015.

W. Sultani and M. Shah. Automatic action annotation in weakly labeled videos. *CoRR*, abs/1605.08125, 2016. URL `http://arxiv.org/abs/1605.08125`.

L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4597–4605, 2015.

J. Tompson, Y. L. Murphy Stein, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM-ToG*, 33(5):169:1–169:10, Sept. 2014. ISSN 0730-0301. doi: 10.1145/2629500. URL `http://doi.acm.org/10.1145/2629500`.

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)*, pages 4489–4497. IEEE, 2015.

P. Turaga, A. Veeraraghavan, and R. Chellappa. Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *CVPR*, pages 1–8. IEEE, 2008.

J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *arXiv preprint arXiv:1604.04494*, 2016.

V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4041–4049, 2015.

S. Vishwakarma and A. Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013.

C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *NIPS*, 2011.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *Readings in speech recognition*, pages 393–404, 1990.

H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, pages 1–20, 2015a.

H. Wang, W. Wang, and L. Wang. How scenes imply actions in realistic videos? In *ICIP*, pages 1619–1623. IEEE, 2016a.

J. Wang, W. Wang, R. Wang, W. Gao, et al. Deep alternative neural network: Exploring contexts as early as possible for action recognition. In *Advances in Neural Information Processing Systems*, pages 811–819, 2016b.

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*, pages 4305–4314, 2015b.

L. Wang, Z. Wang, Y. Xiong, and Y. Qiao. CUHK&SIAT submission for thumos15 action recognition challenge. In *THUMOS Action Recognition challenge*, pages 1–3, 2015c.

L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015d.

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*, pages 20–36. Springer, 2016c.

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*, 46(4):498–509, 2016d.

P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona. Large-scale continuous gesture recognition using convolutional neural networks. *Proc. of International Conference on Pattern RecognitionW*, 2016e.

P. Wang, Q. Song, H. Han, and J. Cheng. Sequentially supervised long short-term memory for gesture recognition. *Cognitive Computation*, pages 1–10, 2016f.

P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. *arXiv preprint arXiv:1701.01814*, 2017.

X. Wang, A. Farhadi, and A. Gupta. Actions˜ transformations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016g.

Y. Wang and M. Hoai. Improving human action recognition by non-action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2698–2707, 2016.

Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges. Two-stream sr-cnns for action recognition in videos. BMVC, 2016h.

Z. Wang, L. Wang, W. Du, and Y. Qiao. Exploring fisher vector and deep networks for action spotting. In *CVPRW*, pages 10–14, 2015e. doi: 10.1109/CVPRW.2015.7301330.

P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3164–3172, 2015.

P. A. Wilson and B. Lewandowska-Tomaszczyk. Affective robotics: modelling and testing cultural prototypes. *Cognitive computation*, 6(4):814–840, 2014.

C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *CVIU*, 127:14–30, Oct. 2014. ISSN 1077-3142. doi: 10.1016/j.cviu.2014.06.014. URL `http://dx.doi.org/10.1016/j.cviu.2014.06.014`.

D. Wu, L. Pigou, P. J. Kindermans, N. LE, L. Shao, J. Dambre, and J. M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE TPAMI*, PP(99):1–1, feb 2016a.

J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *ICMI*, pages 453–460, 2013. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2532589. URL `http://doi.acm.org/10.1145/2522848.2532589`.

J. Wu, P. Ishwar, and J. Konrad. Two-stream cnns for gesture-based verification and identification: Learning user style. In *CVPRW*, 2016b.

J. Wu, G. Wang, W. Yang, and X. Ji. Action recognition with joint attention on multi-level deep features. *arXiv preprint arXiv:1607.02556*, 2016c.

Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3112–3121, 2016d.

X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *Proc. European Conference on Computer Vision*, 2016.

Z. Xu, L. Zhu, Y. Yang, and A. G. Hauptmann. Uts-cmu at THUMOS 2015. *CVPR THUMOS Challenge*, 2015, 2015a.

Z. Xu, L. Zhu, Y. Yang, and A. G. Hauptmann. Uts-cmu at thumos 2015. 2015b.

J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

Y. Ye and Y. Tian. Embedding sequential information into spatiotemporal features for action recognition. In *CVPRW*, 2016.

S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016.

D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al. An introduction to computational networks and the computational network toolkit. Technical report, TR MSR, 2014.

J. Yu, K. Weng, G. Liang, and G. Xie. A vision-based robotic grasping system using deep learning for 3d object recognition and pose estimation. In *Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on*, pages 1175–1180. IEEE, 2013.

J. Yuan, B. Ni, X. Yang, and A. Kassim. Temporal action localization with pyramid of score distribution features. In *CVPR*, 2016.

J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, pages 4694–4702, 2015.

S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144*, 2015.

B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2016.

B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495. 2014.

T. Zhou, N. Li, X. Cheng, Q. Xu, L. Zhou, and Z. Wu. Learning semantic context feature-tree for action recognition via nearest neighbor fusion. *Neurocomputing*, 201:1–11, 2016.

Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *CVPR*, pages 3323–3331, 2015.

G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *Proc. of International Conference on Pattern RecognitionW*, 2016a.

W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A key volume mining deep framework for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–1999, 2016b.

C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.