A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences

Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, Sergio Escalera

A survey on deep learning based approaches for action and gesture recognition in image sequences

Maryam Asadi-Aghbolaghi[1,2,3], Albert Clapés[2,3] , Marco Bellantonio[4], Hugo Jair Escalante[5], Víctor Ponce-López[2,3,6], Xavier Baró[6], Isabelle Guyon[7], Shohreh Kasaei[1], Sergio Escalera[2,3]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

[2] Department of Mathematics and Informatics, University of Barcelona, Barcelona, Spain

[3] Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain

[4] Facultat d'Informatica, Polytechnic University of Barcelona, Barcelona, Spain

[5] Instituto Nacional de Astrofísica, Óptica y Electrónica Puebla, Mexico

[6] *EIMT, Open University of Catalonia*, Barcelona, Spain

[7] *Université Paris-Saclay*, Paris, France

masadia@ce.sharif.edu

***Abstract*—The interest in action and gesture recognition has grown considerably in the last years. In this paper, we present a survey on current deep learning methodologies for action and gesture recognition in image sequences. We introduce a taxonomy that summarizes important aspects of deep learning for approaching both tasks. We review the details of the proposed architectures, fusion strategies, main datasets, and competitions. We summarize and discuss the main works proposed so far with particular interest on how they treat the temporal dimension of data, discussing their main features and identify opportunities and challenges for future research.**

I. INTRODUCTION

Action and gesture recognition have been studied for a while within the fields of computer vision and pattern recognition and substantial progress has been reported for both tasks in the last two decades. Recently, deep learning has irrupted in these fields achieving outstanding results and outperforming "non-deep" state-of-the-art methods [97, 112, 31].

The temporal dimension in sequences typically causes action/gesture recognition to be 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 sub-sampling, 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 [108]. However, these models typically faced some major mathematical difficulties identified by Hochreiter [39] and Bengio et al [9]. In 1997, authors' effort led to the development of the long short-term memory (LSTM) [40] cells for RNNs. Today, LSTMs are an important part of deep models for image sequence modeling for human action/gesture recognition [98, 92]. These, along with implicit modeling of spatiotemporal features using 3D convolutional nets [47], pre-computed motion-based features [97], and the combination of multiple visual [98], 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 topic within the last few years is astounding. Even so, to the best of our knowledge, there is no previous survey that collects and reviews all of the existent work on deep learning for action and gesture recognition. This paper 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 paper is organized as follows. Section II presents a taxonomy in this field of research. Next, Section III reviews the literature on human action/activity recognition with deep learning models. Section IV summarizes the state-of-the-art on deep learning for gesture recognition. Finally, Section V discusses the main features of the reviewed deep learning for the both studied problems.

II. TAXONOMY

Fig. 1 illustrates a taxonomy of the main works performing action and gesture recognition using deep learning approaches. Note that with *recognition* we refer to either classification of pre-segmented video segments or localization of actions in long untrimmed videos.

A. Architectures

The most crucial challenge in deep-based human action and gesture recognition is how to deal with the temporal dimension. Based on that, we categorize approaches into different three groups. The first group uses 3D filters in the convolutional layer [7, 47, 58, 105]. The 3D convolution and 3D pooling in CNN layers allow to capture discriminative features along both spatial and temporal dimensions while maintaining a certain temporal structure. In the second group,

An extended version of this paper will be made available as: Asadi-Aghbolaghi et al. **Deep learning for action and gesture recognition in image sequences: a survey.** *Book chapter in Springer Series on Challenges in Machine Learning, forthcoming 2018.*

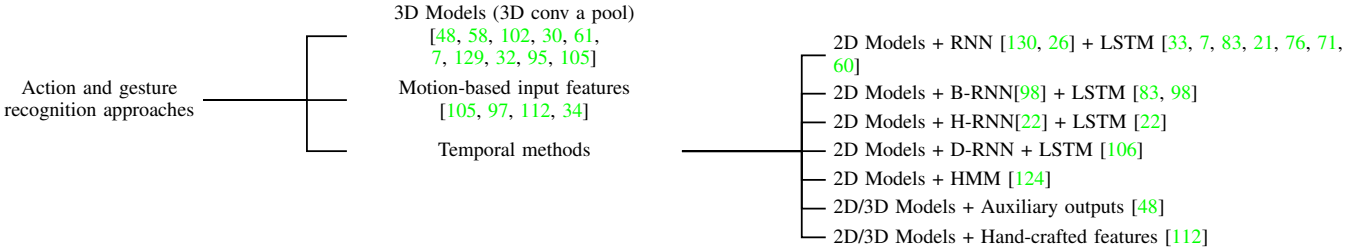

Figure 1: Taxonomy of deep learning approaches for gesture and action recognition.

motion features like 2D dense optical flow maps are pre-computed and input to the networks [105, 97, 112, 34, 102, 122]. Extracted motion features can be fed to the network as additional channels to the appearance ones [105] or input to a secondary network (later combined with former one) [97]. Fig. 2 illustrates these first two groups. The third group combines a 2D (or 3D) CNN applied at individual (or stacks of) frames with a temporal sequence modeling. Recurrent Neural Network (RNN) [26] 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-term memory which is insufficient for real world actions. To solve this problem Long Short-Term Memory (LSTM) [33] was proposed, and it is usually used as a hidden layer of RNN, as seen in Fig. 2. Bidirectional RNN (B-RNN) [83], Hierarchical RNN (H-RNN) [22], and Differential RNN (D-RNN) [106] are some other successful extensions of RNN in recognizing human actions. Other temporal modeling tools like HMM are also applied [124].

For all methods in the three groups, the performance of a deep model can be boosted by combination with hand-crafted features, e.g. improved dense trajectories (iDT) [112].

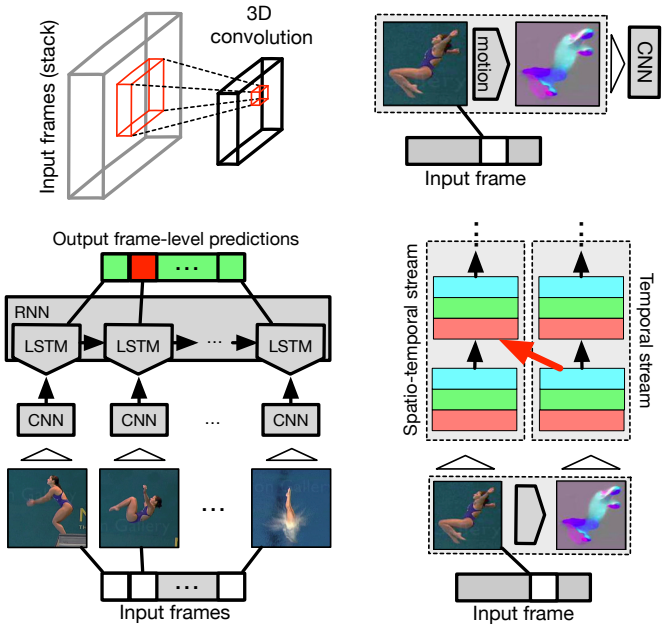

Figure 2: The different architectures and fusion strategies. Top-left: 3D convolution. Top-right: motion pre-computation. Bottom-left: sequential modeling via LSTM. Bottom-right: fusion into a spatio-temporal stream.

B. Fusion strategies

Information fusion is common in deep learning methods for action and gesture recognition. At times, fusion is used to combine the information from parts of a segmented video sequence [51, 115], although, it is more common to fuse information from multiple cues (e.g. RGB and motion, depth, and/or audio) [32], as well combining models trained with different data samples and learning parameters [68].

There are three main variants for information fusion in deep learning models: early (before the data is feed into the model, or the model fuses information directly from multiple sources), late (outputs of deep learning models are combined) and middle (intermediate layers fuse information) fusions [68, 69]. An example of the latter is shown in Fig. 2. Modifications and variants of these schemes have been proposed as well, for instance, see the variants introduced in [51] for fusing information in the temporal dimension. Moreover, ensembles or stacked networks are also considered as fusion strategies [115, 105, 68].

C. Datasets

We list the most relevant datasets according to action (or activity) and gesture recognition in Table I and II respectively. For each dataset we specify year of creation, problems for which the dataset was defined (either classification or temporal/spatiotemporal localization), data modalities available for the task, involved body parts, the number of classes, and state-of-the-art performances to this date (which provide a hint of how difficult the datasets are).

Table III and Table IV summarize the recent approaches which obtained remarkable results against two of the most well-known and challenging datasets in action recognition, respectively, UCF-101 and THUMOS-14. Reviewing top ranked methods on UCF-101 dataset, we find that the most significant difference among them is the strategy for splitting video data and combine sub-sequence results. [119] encodes the changes in the environment by dividing the input sequence into two parts, pre-condition and effect, and model the actions as linear transformation from one to another. [55] processes the input

Table I: Action datasets

Year	Dataset	Problem	Body Parts	Modality	No.classes	Performance
2004	KTH	AC	F	I	6	98.67% Acc [136]
2006	IXMAS	AC	F	RGB, A	13	98.79% Acc [104]
2007	HDM05	AC	F	S	100	98.17% Acc [14]
2008	HOHA (Hollywood 1)	AC, TL	F, U, L	RGB	8	71.90% Acc [91], 0.787@0.5 mAP [62]
2008	UCF Sports	AC, STL	F	RGB	10	95.80% Acc [94], 0.789@0.5 mAP [62]
2009	Hollywood 2	AC	F, U, L	RGB	12	78.50 mAP [56]
2009	UCF11 (YouTube Action)	AC, STL	F	RGB	11	93.77% Acc [82], -
2010	Highfive	AC, STL	F,U	RGB	4	69.40 mAP [109], 0.466 IoU [6]
2010	MSRAction3D	AC	F	D, S	20	97.30% Acc [59]
2010	MSRAction II	STL	F	RGB	3	85.00@0.125% mAP [17]
2010	Olympic Sports	AC	F	RGB	16	96.60% Acc [55]
2011	Collective Activity (Extended)	AC	F	RGB	6	90.23% Acc [5]
2011	HMDB51	AC	F, U, L	RGB	51	73.60% Acc [110]
2012	MPII Cooking	AC, TL	F, U	RGB	65	72.40 mAP [137], -
2012	MSRDailyActivity3D	AC	F,U	RGB, D, S	16	97.50% Acc [93]
2012	UCF101	AC,TL	F, U, L	RGB	101	94.20% Acc [115], 46.77@0.2 mAP (split 1) [122]
2012	UCF50	AC	F, U, L	RGB	50	97.90% Acc [24]
2012	UTKinect-Action3D	AC	F	RGB, D, S	10	98.80% Acc [52]
2013	J-HMDB	AC, STL	F, U, L	RGB, S	21	71.08 Acc [79], 73.1@0.5 mAP [91]
2013	Berkeley MHAD	AC	F	RGB, D, S, A	11	100.00% Acc [14]
2014	N-UCLA Multiview Action3D	AC	F	RGB, D, S	10	90.80% Acc [52]
2014	Sports 1-Million	AC	F, U, L	RGB	487	73.10% Acc [133]
2014	THUMOS-14	AC, TL	F, U, L	RGB	101, 20 *	71.60 mAP [46], 0.190@0.5 mAP [95]
2015	THUMOS-15	AC, TL	F, U, L	RGB	101, 20 *	80.80 mAP [55], 0.183@0.5 mAP (a)
2015	ActivityNet	AC, TL	F, U, L	RGB	200	93.23 mAP (b), 0.594@0.5 mAP [65]
2016	NTU RGB+D	AC	F	RGB, D, S, IR	60	{69.20, 77.70}[1] Acc [57]

Problems: action classification (AC), temporal localization (TL), and spatiotemporal localization (STL). **Body parts**: full body (F), upper body (B), and lower body (L). **Modalities**: audio (A), depth (D), grayscale intensity (I), infrared (IR), skeleton (S), and color (RGB).
Performance: Acc (accuracy), mAP (mean average precision), IoU (intersection-over-union).
* A different no. classes used for different problems. For TL/STL, "@" indicates amount overlap with groundtruth considered for positive localization.
(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}.

Table II: Gesture datasets

Year	Dataset	Problem	Body Parts	Modality	No.classes	Performance
2011	ChaLearn Gesture	GC	F, U	RGB, D	15	-
2012	MSR-Gesture3D	GC	F, H	RGB, D	12	98.50% Acc [16]
2014	ChaLearn (Track 3)	GC, TL	U	RGB, D, S	20	98.20 Acc [64], 0.870 IoU [69]
2015	VIVA Hand Gesture	GC	H	RGB	19	77.50% Acc [63]
2016	ChaLearn conGD	TL	U	RGB, D	249	0.315 IoU [11]
	ChaLearn isoGD	GC				67.19% Acc [23]

Problems: gesture classification (GC). **Body parts**: hands (H).
Check also tablenotes on Table I for additional notation.

video as a hierarchical structure over the time in 3 levels, i.e. short-term, medium-range, and long-range. [105] achieves good performance by using a two-stream network (RGB and motion) with extended temporal resolution respect to previous works (from 16 to 60 frames). [135] could get the best accuracy on UCF101 by using Trajectory pooling to pool the extracted convolutional features from from the optical flow nets of Two-Stream ConvNets and the frame-diff layers of spatial network to get local descriptors.

Looking at the top ranked deep models on THUMOS 2014 challenge, almost all winner methods combined appearance and motion features. For appearance, most of the methods extract frame-level CNN descriptors, and video representation is generated using a pooling method over the sequence. On the other hand, motion-based approaches in the top ranked methods can be divided into three groups, FlowNet, 3D CNN, and iDTs. In [84], we provide a comparison of those showing 3D CNN achieves the best result.

D. Challenges

Every year computer vision organizations arrange competitions providing useful annotated datasets. Table V shows 5 main challenge series in computer vision. For each, we report the year in which it took place, the name of the dataset along with the task to be faced (either action- or gesture-related), the

Table III: UCF-101 dataset results

Ref.	Year	Features	Architecture	Score
[135]	2015	CNN, IDT	2 CNN + iDT pooling	93.78%
[105]	2016	Opt. Flow, 3D CNN, IDT	LTC-CNN	92.7%
[32]	2016	conv5, 3D pool	VGG-16, VGG-M, 3D CNN	92.5%
[119]	2016	CNN	Siamese VGG-16	92.4%
[55]	2016	CNN fc7	2 CNNs (spatial + temporal)	92.2%
[112]	2015	CNN, Hog/Hof/Mbh	2-stream CNN	91.5%
[61]	2015	CNN feat	3D CNN	89.7%
[10]	2016	Dynamic feat maps	BVLC CaffeNet	89.1%
[46]	2015	H/H/M, IDT, FV+PCA+GMM	8-layer CNN	88.5%
[102]	2015	CNN	F_{ST}CN: 2 CNNs (spat + temp)	88.1%
[97]	2014	CNN	Two-stream CNN (CNN-M-2048)	88.0%
[60]	2016	eLSTM, DCNN fc7	eLSTM, DCNN+LSTM	86.9%
[134]	2016	CNN	2 CNNs (spatial + temporal)	86.4%
[129]	2016	dense trajectory, C3D	RNN, LSTM, 3DCNN	85.4%
[78]	2015	CNN fc6, HOG/HOF/MBH	VGG19 Conv5	79.52%±1.1% (tr2)
[78]	2015	CNN fc6, HOG/HOF/MBH	VGG19 Conv5	66.64% (tr1)
[51]	2014	CNN features	2 CNN converge to 2 fc layers	65.4%, 68% mAP
[45]	2015	ImageNet CNN, word2vec GMM	CNN	63.9%
[122]	2015	CNN	Spat + motion CNN	54.28% mAP

Table IV: THUMOS-14 dataset results

Ref.	Year	Features	Architecture	Score
[46]	2015	H/H/M, IDT, FV+PCA+GMM.	8-layer CNN	71.6%
[134]	2016	CNN	2 CNNs (spatial + temporal)	61.5%
[45]	2015	ImageNet CNN, word2vec GMM	CNN	56.3%
[95]	2016	CNN fc6, fc7, fc8	3D CNN, Segment-CNN	19% mAP
[130]	2015	CNN fc7	VGG-16, 3-layer LSTM	17.1% mAP
[30]	2016	fc7 3D CNN	C3D CNN net	.084% mAP@50 .121% mAP@100 .139% mAP@200 .125% mAP@500

associated event's name, the winner participant, and the more recent results on the challenge's associated dataset.

III. Action/Activity Recognition

This section reviews deep methods to address action recognition divided on how they treat the temporal dimension: 3D convolutions, pre-computed motion features, or temporal modeling.

A. 3D Convolutional Neural Networks

In order to capture temporal information, one approach consists in extending the convolution along the temporal axis, in what is known as 3D CNN [47, 7, 103, 61, 58, 102, 30, 129, 32, 95]. The larger number of parameters w.r.t. 2D models, make them harder to train. To alleviate this problem, [61] initializes the weights of a 3D CNN by using 2D weights learned from ImageNET, while [102] proposes a 3D CNN ($F_{st}CN$) that factorizes the 3D convolutional kernel learning as a sequential process of learning 2D spatial and 1D temporal kernels in different layers. Other authors focused on further improving accuracy of 3D CNNs. [32] performs 3D convolutions over stacks of optical flow maps. [95] uses multiple 3D CNNs in a multi-stage (proposal generation, classification, and fine-grained localization) framework for temporal action localization in long untrimmed videos. We also find 3D CNN models being combined with sequence modeling methods [7] or hand-crafted feature descriptors (VLAD [30] or iDTs [129]).

B. Motion-based features

In recent years, many approaches have focused on incorporating pre-computed temporal features within the deep model,

Table V: Challenges

Challenge	Year	Dataset	Task	Event	Winner	Results
ChaLearn	2012	CGD	G	-	Alfnie	[53]* [27]
	2013	Montalbano	G	-	[125]	[8]
	2014	HuPBA 8K+	A	ECCV	[80]	-
		Montalbano	G		[68]	[83] [69] [96]
	2015	HuPBA 8K+	A	CVPR	[121]	-
	2016	isoGD, conGD	G	ICPR	[13]	[51], [117]
HAL	2012	LIRIS	A	ICPR	[70]	[123] [35]*
Opportunity	2011	Opportunity	A	-	CSTA	[90] [15] [89]
ROSE	2016	NTU RGB+D	A	ACCV	SEARCH	[92]
THUMOS	2013	UCF101	A	ICCV	[75]	[101] [100] [81] [50]
	2014	THUMOS-14	A	ECCV	[44]	[46] [95] [111] [88]
	2015	THUMOS-15	A	CVPR	[128]	[114] [132]
VIVA	2015	VIVA	G	CVPR	[63]	[63] [74]
VIRAT	2012	VIRAT DB	A	CVPR	-	[107] [73]

* Non-deep learning method

e.g. dense optical flow maps or iDTs. [122] detects frame proposals and scores them with a combination of static and motion CNN features for action localization. [97] presents a two-stream CNN which incorporates both spatial (still image) and temporal networks (multi-frame dense optical flow). [134] exploit a motion vector from video compression (instead of optical flow). [34] localizes actions in space and time using a (spatial-temporal) two-stream CNNs whose predictions are late-fused with SVM. [105] extends the convolutions in time, aiming at capturing long-term temporal convolutions, at expenses of spatial resolution. [116] uses view-invariant multi-scale depth maps as a input motion descriptor for CNN. [51] proposes a multi-scale foveated CNN for large-scale video classification. Differently, [85] uses CNNs to obtain canonical human poses for action recognition. [113] simply estimates *actionness* maps from appearance and motion cues. In the same line, [138] introduces a deep-learning method to identify key volumes and classify them simultaneously.

In the literature there exist several methods which extend the CNN capabilities using trajectory features. [112] pools and normalizes CNN feature maps along improved dense trajectories. [78] concatenates iDTs (HOG, HOF, MBHx, MBHy descriptors with fisher vector encoding) and CNN feature (VGG19) descriptors. [86] presents a Robust Non-linear Knowledge Transfer Model (R-NKTM) based on a deep fully-connected network that transfers human actions from any view to a canonical one. R-NKTM is learned using bag-of-features from dense trajectories of synthetic 3D human models and generalizes to real videos of human actions. [120] bases on iDT Descriptors and two-Stream CNN features, using a non-action classifier to down-weight irrelevant video segments. [18] presents a new Pose-based CNN descriptor which aggregates motion and appearance information along tracks of human body parts. [55] proposes *VLAD*^{3} to model long-range dynamic information. It captures short-term dynamics with deep CNN features, relying on linear dynamic systems (LDS) to model medium-range dynamics.

C. Temporal deep learning models: RNN and LSTM

We also find approaches which combine CNN with temporal sequence modeling techniques, such as RNNs or LSTMs. [106] introduces a differential gating scheme for LSTM to emphasize on the change in information gain caused by the salient motions between successive frames. [71] proposes a RNN to perform interactional parsing of objects. The object parsings are used to form object specific action representations for fine grained action detection. [98] presents a multi-stream bi-directional RNN. A tracking algorithm locates a bounding box around a person and two streams (motion and appearance) cropped to the tracked bounding box are trained along with full-frame streams. The CNN is followed by a bidirectional LSTM layer. [130] introduces a fully end-to-end approach based on a RNN agent. The agent observes video frames and decides both where to look next and when to emit a prediction. [60] proposes a deep architecture which uses 3D skeleton sequences to regularize an LSTM network (LSTM+CNN) on the video frames.

D. Deep learning with fusion strategies

Some methods have used diverse fusion schemes to improve recognition performance of action recognition. [37] proposes a novel Subdivision-Fusion Model (SFM), where features extracted with CNN are clustered and grouped into subcategories. [22] learns an end-to-end hierarchical RNN using skeleton data divided into five parts, each of which is feed into a different network. The final decision is taken by single-layer network. [99] faces the problem of first person action recognition using a multi-stream CNN (ego-CNN, temporal, and spatial). [119] focuses on the changes that an action brings into the environment and propose a siamese CNN architecture to fuse precondition and effect information from the environment. [20] proposes a CNN which uses mid-level discriminative visual elements. The method, called DeepPattern, is able to learn discriminative patches by exploring human body parts as well as scene context. [76] proposes DeepConvLSTM, based on convolutional and LSTM recurrent units, which is suitable for multimodal wearable sensors.

IV. Gesture Recognition

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

A. 3D Convolutional Neural Networks

Several 3D CNNs have been proposed for gesture recognition, most notably [64, 41, 63]. [41] proposes a 3D CNN for sign language recognition. The CNN automatically learns a representation from raw video, and processes multimodal information (RGB-D+Skeleton data). Similar in spirit, [63] introduces a 3D CNN for driver hand gesture recognition from depth and intensity data. [64] extends a 3D CNN with a recurrent mechanism for detection and classification of dynamic hand gestures. It consists of a 3D-CNN for spatio-temporal feature extraction, followed by a recurrent layer for global temporal modeling and a softmax layer for predicting class-conditional gesture probabilities.

B. 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. For gesture *style* recognition in biometrics, [126] proposes a two-stream (spatio-temporal) CNN. 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 the authors of [43] 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 authors of [117] use three representations of *dynamic depth image* (DDI), *dynamic depth normal image* (DDNI) and *dynamic depth motion normal image* (DDMNI) for gesture recognition.

[118] first identifies 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, as the input data of deep learning network.

C. Temporal deep learning models: RNN and LSTM

This kind of models have not been widely used for gesture recognition, despite being a promising venue for research. We are aware of [67], where the authors propose a multimodal (depth video, skeleton, and speech) human gesture recognition system based on RNN. [25] presents a Convolutional Long Short-Term Memory Recurrent Neural Network (CNNLSTM) able to successfully learn gesture varying in duration and complexity. [72] proposes a multi-stream model, called MRNN, which extends RNN capabilities with LSTM cells in order to facilitate the handling of variable-length gestures.

D. Deep Learning with fusion strategies

Multimodality has been widely exploited for gesture recognition. [124] proposes a semi-supervised hierarchical dynamic framework based on a HMM for simultaneous gesture segmentation and recognition using skeleton joint information, depth and RGB images. The authors applied intermediate (middle) and late fusion to get the final result. [69] proposes a multimodal multi-stream CNN for gesture spotting. Separate CNNs are considered for each modality at the beginning of the model structure with increasingly shared layers and a final prediction layer. The authors fuse the result of each network by a meta-classifier independently at each scale; i.e., late fusion. [77] presents a deep learning model to fuse multiple information sources for human pose estimation. The deep model takes as input the output of a state-of-the-art human pose estimator. The authors exploited early and middle fusion methods to integrate the models.

[54] proposes a CNN that learns to score pairs of input images and human poses (joints). The model is formed by two subnetworks: a CNN learns a feature embedding for the input images, and a two layer subnetwork learns an embedding for the human pose. The authors then calculate score function by dot-product between the two embeddings; i.e. late fusion. Similarly, [43] proposes 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.

V. Discussion

We presented a comprehensive overview of deep-based models for action and gesture recognition. We defined a taxonomy covering most of basic and crucial information about human action and gesture analysis and then we reviewed recent methods. Key topics identified include architectures, fusion strategies, datasets, and challenges.

Generally, there are two main issues when comparing the methods; i.e. *how does a method deal with temporal information?* and *how can such a large network be trained with small datasets?* As discussed, methods can learn motion features by 3D filters in their 3D convolutional and pooling layers. It has been shown that 3D networks over a long sequence are able to learn complex temporal patterns [105]. 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 [61].

It has been also 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.

Still, both groups can only exploit limited (local) temporal information. Hence, the most crucial advantage of approaches in the third group (i.e. temporal models like RNN, LSTM) is that they are able to cope with longer-range temporal relations. These models are mostly used with the skeletal data. 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.

Regardless of the model, performance is dependent on the amount of data. The community is nowadays putting efforts on building larger data sets that can cope with huge parametric deep models (e.g. [2, 38]) and on challenge organization (with novel data sets and well defined evaluation protocols) that can advance the state-of-the-art of the field and make easier the comparison among deep learning architectures (e.g. [92, 29]). 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. 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.

Yet another trick to improve the result of deep-based models is data fusion. Individual networks can be trained on different

portions of input data, primary features, information cues, etc. and then fused. Ensemble learning proved to reduce the bias and variance errors of the learning algorithm [68]. In this sense, we find new methodologies that ensemble deep models for action and gesture recognition [115, 105, 68]. 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 [105].

We also find works that exploit other cues aside from motion. One valuable cue is spatial structure of actions/gestures. [112] takes advantage of iDTs to pool relevant CNN features along trajectories in video frames. [12], takes advantage of human body spatial constraints, by aggregating convolutional activations of a 3D CNN into descriptors based on joint positions. Another cue is interaction among people in a group. [42] uses an LSTM to model each individual's actions and a second-level LSTM aggregates the outputs of individual LSTMs. Along the same lines, [19] fuse individual CNNs using a probabilistic graphical model.

Contextual cues have also been considered for action/gesture recognition. [4] proposes novel multi-stage recurrent architecture consisting of two stages: in a first stage, the model focuses on global context-aware features, and then combines the resulting representation with the localized, action-aware. [46] enriches their motion representation by encoding a set of 15,000 objects from ImageNet and computing their likelihood in frames.

Pose estimation in RGB videos can be computationally expensive and error-inducing. Nonetheless, in depth imaging – in which pose estimation is fast and reliable – inferred joint localization and pose are effectively exploited [139, 22].

One of the reasons that supports the applicability of deep learning is code sharing. There are many open source libraries implementing standard deep models. Among the popular ones are Caffe [49], CNTK [131], TensorFlow [1], and Theano [3]. Regarding applications, deep learning techniques have been successfully used in traditional ones (e.g. surveillance, health care, robotics), improving performance in action and gesture recognition for human computer-robot or -machine interaction. We anticipate deep learning will prevail in emerging applications/areas like social signal processing, affective computing, and personality analysis, among others.

Advances in hybrid models combining handcrafted and new descriptors are expected [68, 112]. Similarly, we think the community will pay attention to deep learning solutions for large scale and real time action and gesture recognition [36, 134]. Immediate effort is also expected in action/gesture localization in long, untrimmed, and realistic videos [128, 34, 95]. As such, we envision newer problems like early recognition [28], multi-task learning [127], captioning, recognition from low resolution sequences [66] and lifelog devices [87] will receive attention in the next years.

VI. 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. H. J. Escalante was supported by CONACyT under grants CB2014-241306 and PN-215546.

REFERENCES

[1] M. e. a. Abadi. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] 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.

[3] 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.

[4] M. S. Aliakbarian, F. Saleh, B. Fernando, M. Salzmann, L. Petersson, and L. Andersson. Deep action-and context-aware sequence learning for activity recognition and anticipation. *arXiv preprint arXiv:1611.05520*, 2016.

[5] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu. Monte carlo tree search for scheduling activity recognition. In *ICCV*, pages 1353–1360, 2013.

[6] 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.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *HBU*, pages 29–39, 2011.

[8] I. Bayer and T. Silbermann. A multi modal approach to gesture recognition from audio and video data. In *ICMI*, pages 461–466, 2013.

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

[10] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *CVPR*, 2016.

[11] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *ICPR W*, 2016.

[12] C. Cao, Y. Zhang, C. Zhang, and H. Lu. Action recognition with joints-pooled 3d deep convolutional descriptors.

[13] 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 ICPRW*, 2016.

[14] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *CVPRW*, pages 471–478, 2013.

[15] 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.

[16] 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.

[17] 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.

[18] G. Chéron, I. Laptev, and C. Schmid. P-CNN: pose-based CNN features for action recognition. *ICCV*, 2015.

[19] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori. Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191*, 2015.

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

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.

[22] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015.

[23] 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.

[24] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe. Spatio-temporal vlad encoding for human action recognition in videos. In *MMM*, pages 365–378. Springer, 2017.

[25] T. Eleni. Gesture recognition with a convolutional long short term memory recurrent neural network. In *ESANN*, 2015.

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

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

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

[29] H. J. e. a. Escalante. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Proc. ICPR*, 2016.

[30] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. DAPs: Deep action proposals for action understanding. *ECCV*, 2016.

[31] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, pages 3468–3476, 2016.

[32] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

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

[34] G. Gkioxari and J. Malik. Finding action tubes. *CoRR*, 2014.

[35] 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.

[36] S. Han, H. Mao, and W. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *Proc. ICLR*, 2016.

[37] Z. B. Hao, L. Lu, Q. Zhang, J. Wu, E. Izquierdo, J. Yang, and J. Zhao. Action recognition based on subdivision-fusion model. *CoRR*, abs/1508.04190, 2015.

[38] 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.

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

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

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

[42] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. *arXiv preprint arXiv:1511.06040*, 2015.

[43] 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. 2015.

[44] 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 2014.

[45] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015.

[46] 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, 2015.

[47] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. In *ICML*, pages 495–502, Haifa, Israel, June 2010. Omnipress.

[48] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013.

[49] 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.

[50] 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 *ICCV Workshops*, 2013.

[51] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[52] 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.

[53] J. Konecny and M. Hagara. One-shot-learning gesture recognition using hog-hof features. *JMLR*, 15:2513–2532, 2014.

[54] 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, 2015.

[55] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos. Vlad3: Encoding dynamics of deep features for action recognition. In *CVPR*, pages 1951–1960, 2016.

[56] 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.

[57] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, pages 816–833. Springer, 2016.

[58] Z. Liu, C. Zhang, and Y. Tian. 3d-based deep convolutional neural network for action recognition with depth sequences. *IVC*, 2016.

[59] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*, pages 1809–1816, 2013.

[60] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *CVPR*, 2016.

[61] E. Mansimov, N. Srivastava, and R. Salakhutdinov. Initialization strategies of spatio-temporal convolutional neural networks. *CoRR*, abs/1503.07274, 2015.

[62] 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.

[63] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3d convolutional neural networks. In *CVPRW*, pages 1–7, June 2015.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.

[71] B. Ni, X. Yang, and S. Gao. Progressively parsing interactional objects for fine grained action detection. In *CVPR*, 2016.

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

[73] S. Oh. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153–3160, 2011.

[74] 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.

[75] D. Oneata, J. Verbeek, and C. Schmid. The LEAR submission at Thumos 2014, 2014.

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

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

[78] X. Peng and C. Schmid. Encoding feature maps of cnns for action recognition. *CVPR*, 2015.

[79] X. Peng and C. Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, pages 744–759. Springer, 2016.

[80] X. Peng, L. Wang, Z. Cai, and Y. Qiao. *Action and Gesture Temporal Spotting with Super Vector Representation*, pages 518–527. 2015.

[81] 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.

[82] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked

fisher vectors. In *ECCV*, pages 581–595. Springer, 2014.

[83] 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, 2015.

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

[85] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *CVPR*, 2016.

[86] H. Rahmani and A. S. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *CVPR*, pages 2458–2466, 2015.

[87] N. Rhinehart and K. M. Kitani. Learning action maps of large environments via first-person vision. In *Proc. ECCV*, 2016.

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

[89] 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 2011.

[90] 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. 2011.

[91] 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.

[92] A. Shahroudy, J. Liu, T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. *CVPR*, pages 1010–1019, 2016.

[93] 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*, 2016.

[94] L. Shao, L. Liu, and M. Yu. Kernelized multiview projection for robust action recognition. *IJCV*, 118(2):115–129, 2016.

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

[96] Z. Shu, K. Yun, and D. Samaras. *Action Detection with Improved Dense Trajectories and Sliding Window*, pages 541–551. Cham, 2015.

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

[98] 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 *CVPR*, 2016.

[99] S. Singh, C. Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016.

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

[101] W. Sultani and M. Shah. Automatic action annotation in weakly labeled videos. *CoRR*, abs/1605.08125, 2016.

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

[103] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE, 2015.

[104] 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.

[105] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *CoRR*, abs/1604.04494, Apr. 2016.

[106] V. Veeriah, N. Zhuang, and G. Qi. Differential recurrent neural networks for action recognition. *CoRR*, abs/1504.06678, 2015.

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

[108] 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.

[109] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *IJCV*, pages 1–20, 2015.

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

[111] L. Wang, Y. Qiao, and X. Tang. Action recognition and detection by combining motion and appearance features. In *THUMOS Action Recognition challenge*, pages 1–6, 2014.

[112] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015.

[113] L. Wang, Y. Qiao, X. Tang, and L. V. Gool. Actionness estimation using hybrid fully convolutional networks. *CoRR*, abs/1604.07279, 2016.

[114] 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, 2015.

[115] 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. *ECCV*, Aug. 2016.

[116] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona. Deep convolutional neural networks for action recognition using depth map sequences. *CoRR*, abs/1501.04686, 2015.

[117] 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.

[118] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona. Large-scale continuous gesture recognition using convolutional neural networks. *Proc. of ICPRW*, 2016.

[119] X. Wang, A. Farhadi, and A. Gupta. Actions ˜ transformations. *CoRR*, abs/1512.00795, 2015.

[120] Y. Wang and M. Hoai. Improving human action recognition by non-action classification. *CoRR*, abs/1604.06397, 2016.

[121] Z. Wang, L. Wang, W. Du, and Y. Qiao. Exploring fisher vector and deep networks for action spotting. In *CVPRW*, pages 10–14, 2015.

[122] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. abs/1506.01929, Dec 2015.

[123] 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.

[124] 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 2016.

[125] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *ICMI*, pages 453–460, 2013.

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

[127] X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *Proc. ECCV*, 2016.

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

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

[130] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. *CoRR*, abs/1511.06984, 2015.

[131] 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.

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

[133] 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.

[134] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. *CoRR*, abs/1604.07669, 2016.

[135] S. Zhao, Y. Liu, Y. Han, and R. Hong. Pooling the convolutional layers in deep convnets for action recognition. *arXiv preprint arXiv:1511.02126*, 2015.

[136] 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.

[137] 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.

[138] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A key volume mining deep framework for action recognition. In *CVPR*, 2016.

[139] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. *reprint arXiv:1603.07772*, 2016.