Computer Vision and Image Understanding 188 (2019) 102799

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Analyzing human–human interactions: A survey

Alexandros Stergiou *, Ronald Poppe

Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, Netherlands

ARTICLE INFO

Communicated by: Nikos Paragios

MSC:
68T45
68U10

Keywords:
Human-human interaction
Human interaction recognition
Human activity

ABSTRACT

Many videos depict people, and it is their interactions that inform us of their activities, relation to one another and the cultural and social setting. With advances in human action recognition, researchers have begun to address the automated recognition of these human–human interactions from video. The main challenges stem from dealing with the considerable variation in recording setting, the appearance of the people depicted and the coordinated performance of their interaction. This survey provides a summary of these challenges and datasets to address these, followed by an in-depth discussion of relevant vision-based recognition and detection methods. We focus on recent, promising work based on deep learning and convolutional neural networks (CNNs). Finally, we outline directions to overcome the limitations of the current state-of-the-art to analyze and, eventually, understand social human actions.

1. Introduction

Despite significant research progress in the automated analysis of humans and their activities (Cheng et al., 2015; Herath et al., 2017; Koohzadi and Charkari, 2017; Poppe, 2010), the recognition of human interactions from video remains a challenging topic. Integral part of the difficulty is that understanding interactions between people requires more than analyzing the actions of each person in isolation. Rather, it is the coordination, in both space and time, between people that reveals the true nature of their collective behavior. In addition, the context in terms of who is interacting why and where determines to a large extent how the interaction unfolds.

There is a long history of the manual and automatic description of human interactions, see Birdwhistell (1952), Poppe (2017) and Vinciarelli et al. (2009) for overviews. Still, the relation between the observable form of the bodily interaction and the more subjective *interpretation* thereof is relatively understudied. For example, putting a hand on someone's shoulder can be objectively identified, whereas more information is required to know that one person is comforting the other, or trying to get the other's attention. The scarcity of a more social, contextual perspective in the automated analysis of human–human interactions is also reflected in computer vision literature, where interactions are typically reduced to visually and temporally well-defined events. Despite this somewhat artificial view on human behavior, current advances pave the way for a more social perspective. In this paper, we survey the research in the recognition of human–human interactions in videos, with a focus on methods based on convolutional neural networks (CNNs). We then discuss promising directions to leverage the current state-of-the-art to a more social analysis.

1.1. Scope and motivation

In this survey, we focus on *dyadic* interactions between two people. We consider joint actions of both people that can be characterized by the positions, movements and coordination of their bodies (see Fig. 1). For example, we consider a handshake as an interaction that can be part of an *activity* such as an agreement or a greeting. Interactions can be made up of several motions in sequence, such as extending the right arm, grasping the right hand of the other and moving the hands up-and-down. The duration of the interactions that we consider can be anywhere between half of a second and several seconds. There can be considerable variation in the performance of an interaction, most notably in the duration but also in the coordination. This variation can also lead to ambiguities in how they are perceived. For example, the hug interaction in Fig. 1(center) could also be considered a lift interaction. The works discussed in this survey exclusively treat the interaction recognition task as deterministic, which does not fully reflect the more ambiguous nature in the perception of social behavior. We discuss alternative representations and methods in the Discussion section.

The automated recognition of bodily interactions from video mainly benefits content-based video retrieval (Liu et al., 2015; Sempena et al., 2011), security (Aran and Gatica-Perez, 2013) and surveillance (Cristani et al., 2011; Tian et al., 2012; Yi et al., 2014) and interactive human–computer interfaces (Rehg et al., 2013; Sheerman-Chase et al., 2011). The vast majority of the research has considered a functional perspective by labeling the visual aspect of videos. This leaves room for a more contextual interpretation of the joint behavior. Opportunities

* Corresponding author.
E-mail address: a.g.stergiou@uu.nl (A. Stergiou).

https://doi.org/10.1016/j.cviu.2019.102799
Received 13 June 2018; Received in revised form 16 August 2019; Accepted 16 August 2019
Available online 19 August 2019

Fig. 1. Three interactions: handshake, hug or lift, and object passing. These examples show non-standard body poses (left), ambiguous class labeling (center) and the need for temporal information (right).

for a broader use of automated measures arise when computers can *understand* the interactions in terms of communicative and affective intent. In this survey, we present the current basis and potential directions to take the important step from interaction recognition to understanding. We discuss the evolution of the current state-of-the-art in interaction recognition towards this *social* perspective in the Discussion section.

1.2. Main challenges in the field

We identify challenges when dealing with the visual and structural aspects of interaction videos. Additionally, we outline practical challenges in the development of methods of automated human–human action recognition.

1.2.1. Variation in visual appearance

Interactions between people can be observed in many different environments, and under vastly different recording settings. Most notably, a change of viewpoint has a large effect on how the interaction is observed. Especially when people are interacting physically, it is likely that their body parts partially occlude each other. This presents challenges in the recognition of interactions from a single viewpoint, as characteristic movements or the poses of key body parts are not visible. Typically, we do not have access to other viewpoints to deal with potential ambiguities.

Variation in clothing and lighting conditions further adds to the challenge of robustly observing the smaller movements. Especially in low-resolution videos, the level of detail might be insufficient to distinguish between subtly different interaction classes such as handshake and fist bump greetings.

1.2.2. Intra-class variation in interaction performance

The performance of an interaction in terms of body movements and coordination can differ significantly, see Fig. 1(left). Ronchi and Perona (2015) has analyzed the variation for single images. Additionally, there is significant variation in the temporal execution of the movement. While such deviations can be used to differentiate between classes (Anderson and Perona, 2014), the dissimilarity of performance within an interaction class is typically too large to derive general rules.

Interactions, like individual actions, often present an intrinsic sequential nature of movements. For example, an extension of the hand of one person is normally followed with the extension of the other actor's hand. Results from works that aim at the prediction of future actions have immediate impact on the improvement of scene understanding (e.g., Vondrick et al. (2016)). Other works build on the key idea that future actions can be predicted by classifying an action or interaction solely on its start (Ziaeefard et al., 2015). Such an approach might work well for goal-directed interactions (Cao et al., 2013; Ryoo, 2011), but is less successful when the variation in the performance increases (see Fig. 1(right)). This is especially true when the interactions are more social and reactive in a communicative or affective way, such as jokingly stomping someone.

Some works have addressed the estimation of a skeletal representation in order to circumvent having to learn interaction patterns directly from video (Cavazza et al., 2016; Pham et al., 2018; Yub Jung et al., 2015; Yun et al., 2012). Recent methods rely on CNN-based approaches (e.g., Cao et al. (2017), Carreira et al. (2016), Güler et al. (2018), Insafutdinov et al. (2017) and Li et al. (2015), Yang et al. (2017)) and allow to investigate both pose and movement of a person. Skeleton representations are informative for actions and interactions and present an attractive alternative or complement for image features. However, errors and inaccuracies in the pose estimation process might be propagated to the classification task. In addition, there is a need for quantitative units that capture the characteristic information of an interaction in terms of pose, movement and coordination in space and time.

1.2.3. Challenges in data collection and labeling

The study of interactions is further complicated by a relative lack of large datasets. In Section 2, we discuss the most popular resources, but most of them focus on a relatively limited domain (e.g. sports or surveillance). In addition, there is no common labeling of the interaction classes. For example, a handshake might be a category of its own, or might be part of a greeting class. This lack of standardization hinders cross-dataset studies and consequently limits the generalization of methods developed in one particular scenario to address another. While human–human interactions are increasingly part of large datasets containing web videos, the interactions considered are often relatively dissimilar and well-defined (e.g. a handshake and a hug). This puts the focus on dealing with the variations in the visual input, rather than subtle variations in the physical performance of the interactions. Also, this practice neglects issues with potentially ambiguous labeling such as in Fig. 1(center). We deem an increased consideration of the coordination of body movements as a key requirement for successful application in more social settings, in which a multitude of subtly varying interactions may be encountered.

1.3. Survey overview

The survey structure is as follows. Section 2 summarizes publicly available datasets. We then continue with an in-depth discussion of human–human interaction recognition literature. We distinguish between the more traditional methods based on hand-crafted features (Section 3) and those based on deep learning (Section 4). Finally, we discuss the limitations of the state-of-the-art and present promising avenues for further research.

2. Datasets

The availability of labeled datasets and the direct comparisons between methods generally lead to better understanding of the relative algorithmic advantages and limitations and, consequently, progression in performance. Compared to datasets available for individual action recognition (e.g., Heilbron et al. (2015), Kuehne et al. (2011), Rodriguez et al. (2008) and Soomro et al. (2012)), resources for human–human interactions are scarce. Most notably, the limited variation in viewpoint, application context and movement performance has hindered remarkable breakthroughs in the recognition of subtly different interactions such as those encountered in social settings. This section provides an overview of the most common datasets. Example frames appear in Fig. 2. A summary of the datasets appears in Table 1.

2.1. UT-Interaction

UT-Interaction (Ryoo and Aggarwal, 2010) contains 20 sequences and six interaction classes. With almost static background, limited occlusions and a fixed viewpoint, the classification difficulty is low. UT-Interaction is used as benchmark for many methodologies, ranging from bounding boxes techniques (Motiian et al., 2017; Shu et al., 2017) to bags-of-visual-words (Shariat and Pavlovic, 2013; Slimani et al., 2014). Some works have also addressed the detection of interactions in both space and time (Van Gemeren et al., 2018).

Fig. 2. Example video frames from different datasets depicting different interaction categories.

Table 1
Summary of datasets with footage type and quantity, number of action/interaction classes and actors.

Dataset	Footage type	Scripted	Sequences	Duration	Classes	Actors
UT-Interaction	Outside recordings	Yes	60	10–25 s	6	8
TV human interaction	TV shows	Yes	300	1–5 s	4	100+
Hollywood2	Films	Yes	3669	10–15 s	12	100+
ShakeFive2	Lab recordings	Yes	153 with pose data	5–10 s	5	33
SBU Kinect	Lab recordings	Yes	300 with pose data	1–5 s	21	9
AVA	Films	Yes	~57.6k	15 min	80	100+
CMU Panoptic	Lab recordings	Partially	65 multi-view with pose data	10–15 min	N/A	16
SALSA	Inside recordings	No	8 multi-view with sensor data	30 min	N/A	18
Kinetics	YouTube videos	No	~500k	10–15 s	700	100+
Moments in time	YouTube videos	No	~800k	1–5 s	340	100+
HACS	YouTube videos	No	~1.5M clips (~490k positive)	2 s	201	100+

2.2. TV Human Interaction

The *TV Human Interaction* dataset is composed of short video segments of four classes (*handshake*, *hug*, *kiss* and *high-five*), taken from popular TV series (Patron-Perez et al., 2010, 2012). The dataset includes annotations of the upper bodies, head orientations and interaction labels for each person in the scene. Compared to *UT-Interaction*, the video quality is higher, more different viewpoints and scenes are included and there is more variation in the number of people in the scene. All interactions are acted and the recording setting is highly controlled.

2.3. Hollywood2

Hollywood2 (Marszalek et al., 2009) also consists of clips from movies. Subtitles were used to align script data with the corresponding movie scenes. Despite the significant variation in the videos, the controlled nature of the movie domain limits generalization to more realistic domains. The four interaction classes are *fight*, *handshake*, *hug* and *kiss*.

2.4. ShakeFive2

A collection of human interaction clips with complementary skeletal data was introduced by Van Gemeren et al. (2016). The videos are captured with fixed viewpoint and static background. The challenge of the dataset is in the similarity of the interaction classes (*fist bump*, *handshake*, *pass object*, *high-five* and *hug*).

2.5. SBU Kinect Interaction

Additional depth data (RGB-D images), obtained from a Kinect sensor, is available in the *SBU Kinect Interaction* dataset (Yun et al., 2012). It features eight two-person interactions: *approach*, *depart*, *kick*, *punch*, *hand shake*, *hug* and *pass object*. The clips are segmented in time, with the interactions fully occupying the frame.

2.6. CMU Panoptic

The *CMU Panoptic* dataset (Joo et al., 2015) is recorded in a large geometric dome with RGB and Kinect cameras distributed across the surface. The data are comprised of 480 synchronized video streams with additional pose information. Each clip depicts 3–8 people participating in social engagements: *ultimatum*, *prisoner's dilemma*, *mafia*, *haggling* and *007-bang*. The activities are scripted but the interactions are genuine. No action classes have been defined but the participants closely interact.

2.7. Kinetics

The *Kinetics* dataset (Carreira et al., 2019; Kay et al., 2017) contains 700 video classes with approximately 600 videos per class. There are 11 interaction classes, including *handshake*, *hug* and *massage feet*. The dataset is a collection of clips from YouTube videos. The video material is not professionally edited and features a large variety of background clutter, illumination settings and motion blur.

2.8. Atomic Visual Actions (AVA)

The AVA dataset (Gu et al., 2018) is composed of 15-min segments from 432 movies. In addition to the labeling of clips for recognition, the interactions and actions of the actors within scenes are localized for tracking and detection tasks. The dataset contains 80 classes, including 13 interaction categories. The videos contain limited camera blur and most of the scenes have been shot with a still camera.

2.9. Synergetic sociAL Scene Analysis (SALSA)

The SALSA dataset (Alameda-Pineda et al., 2016) contains 30 min of a poster presentation event, and 30 min of a cocktail party. In addition to camera views, the events have been recorded with various other sensors, including microphones and accelerometers. The data is richly annotated in terms of body and head orientation, and group membership. SALSA allows for the analysis of more social (group) interactions.

2.10. Moments in Time

The Moments in Time dataset (Monfort et al., 2018) is composed of three-second clips of events and activities. The dataset contains significant intra-class variation. Apart from common activity and interaction classes such as hugging and handshaking, some classes focus on group events such as dinning, baptizing or autographing.

2.11. Human Action Clips and Segments (HACS)

The HACS dataset (Zhao et al., 2019) contains annotations of roughly 50k YouTube videos that correspond to 1.5M clips in total. The extracted two-second clips from the videos cover 201 classes, and also include negative samples that do not contain any action or interaction of interest, but are shot under the same image conditions. The dataset contains 23 interaction classes, mostly relating to sport activities.

3. Recognition from handcrafted features

Traditionally, the recognition of interactions from video starts with the representation of the scene and events as image features, and the subsequent classification of these features into an interaction class. Image features should be invariant to image conditions and interaction performance, while being sufficiently rich to deal with subtle differences between interaction classes.

We distinguish between local feature approaches that rely on salient points in the video, and template-based approaches that take into account regions in the video that roughly correspond to a person's body or body parts.

3.1. Local features approach

In general, local feature algorithms take a bottom-up approach by first detecting interesting points in a video, and then to aggregate detections over time and space to understand which behavior is being performed. These interesting points are selected locally, typically at edges or motion boundaries. Popular descriptors are based on Harris corners (Marín-Jiménez et al., 2013; Zhang et al., 2013), SIFT descriptors (Delaitre et al., 2010; Lowe, 1999) or optical flow (Yu et al., 2012). There is typically no direct correspondence between a point and a person or body part. As a consequence, factors such as camera motion, dynamic backgrounds and occlusions affect the presence of local features.

To increase the robustness of local descriptors, a distribution of points is usually described as a bag-of-words (BoW) or Fisher vector (FV) (Gao et al., 2016; Oneata et al., 2013). Instances of the same interaction class are assumed to have similar descriptors. To allow for a more complex distribution of the features, Niebles et al. (2008) construct a vocabulary using latent topics models.

Instead of modeling the trajectories of individual points, researchers have addressed the sequential nature of interactions by modeling the changes in the distribution of interest points over time. Zhang et al. (2012) use spatio-temporal phases to create a histogram of bag-of-phases. Each phase is composed of local words with specific ordering and spatial position. Instead of jointly mapping both dimensions, authors have addressed separation as well (Shariat and Pavlovic, 2013; Tran et al., 2014). The computed histograms represent similar features in single or multiple frames. Histograms of visual words have also been utilized by Kong et al. (2012). Here, the words derived from the quantization of the spatial–temporal descriptors were clustered to form a high-level representation of dyadic interactions, termed interactive phases. These phases include motion relationships such as the shaking of two hands. This idea has been extended to localize interactions by spatially clustering the phrases (Tran et al., 2013). To allow for variation in the temporal domain, Prabhakar and Rehg (2012) model the causality of the occurrence of visual words.

Not all motions and attributes are informative, such as the positioning of the feet when performing certain greetings. Kong et al. (2014) consider only body parts that characterize the interaction. Their method pools BoW responses in a coarse grid. This allows them to identify specific motion patterns relative to a person's location. The level of detail of the analysis is limited by the granularity of the patches and the accuracy of the person detector. Additionally, they take into account the temporal nature of interactions by linking subsequent detections into trajectories. Mohammadi et al. (2015) extend this approach by grouping the motion patterns as BoW vectors. Similarly, Turchini et al. (2016) introduce an approach to localize interactions from the trajectories of multiple local feature types. Wang and Schmid (2013) have introduced Improved Dense Trajectories (DT), a widely adopted way of finding and describing trajectories of points. In DT, a point is encoded as a combination of Histograms of Oriented Gradients (HOG), Histograms of Oriented Flow (HOF) and Motion Boundary Histograms (MBH). Points are linked over time.

Local features can be used to isolate a person in video first. Extensive work has been done on the detection of humans from local features, encoded with HOG and HOF descriptors (Caba Heilbron et al., 2016). Once a person has been localized, the context of motions and actions of other people in the scene can provide useful cues for the recognition of their interactions. Reddy and Shah (2013) exploit the information obtained through a scene context descriptor which combines the location and surroundings extracted with optical flow and 3D-SIFT, based on the moving and stationary pixels. Cho et al. (2017) introduced the compositional interaction descriptor that takes into account the local, global and individual movement in video sequences. By linking local features to persons, we can describe their surroundings. Lan et al. (2012) presented an Action Context (AC) descriptor that is based on connected action probability vectors of several people. Similarly, Choi and Savarese (2014) perform joint tracking, classification of the actions of an individual and the recognition of collective activities by considering bounding boxes of extracted local features.

3.2. Template-based approaches

When applied to a single frame, a HOG descriptor can represent a characteristic pose. For example, a high-five interaction can be described as two people facing each other with outstretched hands that meet above their heads. This notion was adopted by Bourdev et al. (2010) to detect people engaged in specific actions, and was applied to human–human interactions by Raptis and Sigal (2013). Sefidgar et al. (2015) have formulated an implementation with discriminative key frames and their relative distance and timing within the interaction. Alternatively, Sener and Ikizler-Cinbis (2015) formulate interaction detection as a multiple-instance learning problem to focus on relevant frames, because not all frames in an interaction are considered informative.

The motion around a characteristic pose can provide complementary information. Van Gemeren et al. (2014) combine HOG and HOF descriptors to encode the characteristic frame of a two-person interaction. Yu and Yuan (2015) concatenate HOG and HOF descriptors and applied FV to make the detection linearly separable, thus allowing the model to concurrently utilize spatial and temporal features.

Instead of relying on interest points, we can first detect faces or bodies using a generic face or body detector (Patron-Perez et al., 2012; Ryoo and Aggarwal, 2011). Given two close detections, interactions can subsequently be classified based on extracted features within the detection region (Ryoo and Aggarwal, 2011). Various attributes, including gross body movement and proximity, have been employed to classify the interaction. Patron-Perez et al. (2012) also include the relative size and orientation of each person. Khodabandeh et al. (2015) consider clusters of similar frames based on proximity and appearance of pairs of people. They find that user feedback helps to increase the purity of the clusters, in turn improving the interaction classification. The drawback

of this two-stage approach is that classification is sub-optimal when the person localization fails, for example when people partly occlude each other. This is a common situation, especially when people interact in close proximity.

This issue is mitigated when employing Deformable Parts Models (DPMs) (Felzenszwalb et al., 2010). Here, an articulated object such as a person or multiple interacting people are modeled as a set of parts and deformations between them. This allows for more flexibility in the spatial layout of the parts. As such, parts that are generally well detected, e.g. a person's head, can be coupled with parts that are traditionally more challenging to detect, such as a lower arm. Lu et al. (2015) use a DPM as a prior to localize the rough outline of a person. Optical flow is then used to propagate the outline to subsequent frames. The resulting volume is then segmented into supervoxels to refine the person's outline in each frame, and classified as action. Van Gemeren et al. (2018) use interaction-specific DPMs with poselet parts (Bourdev et al., 2010) to locate people in poses characteristic for a given interaction. Instead of encoding the orientation of (pairs of) limbs as poselets, DPMs can also include a larger number of articulations by using a mixture of parts (Yang and Ramanan, 2011). This approach has been used to describe the joint poses of two interacting people (Yang et al., 2012).

While DPMs encode a particular pose or motion spatially only, extensions have been proposed to deal with the time-varying nature of human interactions. Yao et al. (2014) focus on human–object interactions and capture the movement related to a key pose using a DPM and a linked set of motion templates that also correspond to different phases of the performance. Tian et al. (2013) have extended DPMs for action detection to model changes in pose over time. These formulations work well for the representation of coarse movements, but finer-scale movements are difficult to model because the motion is not linked to specific parts of the body.

4. Interaction detection from learned features

The hand-coded feature descriptors described in Section 3 focus on local or global spatial or spatio-temporal information. The manual selection of descriptors leaves room for improvement because the process is agnostic to the specific classification task, application domain or class of behaviors.

Based on the introduction of multiple convolutions by LeCun et al. (1998), Convolutional Neural Networks (CNNs or ConvNets) have been used for classification tasks of both image and video data. CNNs allow for the simultaneous training of a classifier, and the automated selection of informative features. Consequently, they can overcome the issue of sub-optimal feature selection. While multiple convolution kernels allow for the selection of a wide range of image or video features, the stacking of consecutive convolution operations allows for a hierarchical extraction of complex features (Simonyan and Zisserman, 2014b). Typically, the characteristics extracted in the first layers of the network correspond to low-level features such as edges and simple textures. Deeper layers of the network are targeted towards the extraction of higher-level features.

Methods based on neural networks have shown notable improvements in human action and interaction classification tasks. Deep learning benefits from extensive amounts of data without saturation in the accuracy rates equivalent to the data growth rate. This allows deep learning architectures to generalize their feature assumptions, based on the utilization of all potential information in images and videos, rather than being limited to a predefined set of features, as in the hand-crafted methods.

The purpose of this section is to present neural network architectures for human interactions that operate on single frames. We then show how temporal information can be incorporated in the convolutions and finally discuss recurrent models.

4.1. Single frame networks

CNNs have been used to classify actions and interactions in single frames (Asadi-Aghbolaghi et al., 2017; Bilen et al., 2016; Gkioxari et al., 2015). Similar to the use of handcrafted features, the focus is on characteristic joint poses. To extend this methodology to sequences of images, several approaches have been proposed.

Based on the classification of individual frames, Karpathy et al. (2014) proposed three techniques to fuse the scores of multiple frames using different convolutional configurations. In the Early Fusion strategy, the input of the network is a stack of subsequent frames. Late Fusion combines the convolutional features of the first and last frames of a sequence in the final, fully connected layers. Slow Fusion is a combination of these two approaches, that empowers a progressive fusion over frames and activation maps, with the extension of convolutional layer connections through time. All three approaches are limited in their capability to deal with subtle temporal variations between classes, and large intra-class variations. It is a challenge to deal with these variations as they have to be modeled from the typically modest number of training videos.

To partly mitigate this issue, authors have investigated the use of Transfer Learning (Bengio et al., 2011; Bengio, 2012; Caruana, 1998; Pan and Yang, 2010; Yosinski et al., 2014). This is a process in which the network is first trained on a large dataset with general examples, and subsequently re-purposed for another, more specific, classification task. In general, this means that the deeper layers are retrained for the specific domain. Consequently, fewer parameters need to be learned for the novel domain, which reduces the risk of overfitting.

4.2. Motion-based and stream networks

Two-stream CNNs combine regular images and optical flow images as input (Simonyan and Zisserman, 2014a), and are an alternative approach to model temporal information. The rationale is that still images encode the pose of an interaction, while the optical flow provides information about the motion. The network consists of two streams, branches in the network structure. The spatial-based CNN is trained on individual video frames, and the temporal stream CNN takes as input stacked optical flow fields from multiple frames. The results from the two networks are concatenated with late fusion. Different information fusion methods for each stream were explored by Park et al. (2016). Wang et al. (2016a) added a Temporal Segment Network (TSN) to the two-stream CNN architecture, applied on sporadically sampled fragments from the video, thus making a prediction on each of the snippets independently. The predicted class is then the 'point of agreement' between the video segments. This method capitalizes on information from small temporal segments rather than using the video as a single input. Following the use of selected frames (Diba et al., 2017) also proposes a representation and encoding of the sequence features in a Temporal Linear Encoding (TLE) layer, after the convolution feature extraction is performed. It is based on the aggregation of appearance features from each of the individual temporal fragments. Works have also included the use of depth data as stream inputs (Garcia et al., 2018) in which features from the depth stream are distilled in order for the depth stream to be simulated at test time as the test data does not include this supplementary modality.

Inputs in the two-stream CNN are processed independently and only fused as a last step. This approach prevents the exchange of information between the streams. As such, it is not possible to develop attention mechanisms that focus on specific parts on the input in either stream. One way of establishing these links is by using skip connections of Residual Networks (He et al., 2016; Hara et al., 2018) and additional shortcut connections between convolutional layers of the motion stream to the spatial stream. This provides benefits in optimizing the network architecture and increasing the network depth (Feichtenhofer et al., 2016). Residual learning enables the model to avoid degradation

in deep structures, which relates to the saturation of accuracy followed by a significant drop when optimizing the parameters as layers of the network are not able to effectively learn the identity map and instead "threshold" to zero mappings.

Recent advances in reinforcement learning and evolutionary algorithms have contributed to a reduction in human supervision for creating robust network architectures (Zoph and Le, 2017). This trend has further enabled the construction of architectures for specific tasks rather than general architectures (Zoph et al., 2018). With an increasing number of options for layers and connections, such techniques are welcome to avoid the slow research progress due to extensive parameter testing.

Typically, a human interaction does not occupy the entire frame. So instead of taking the entire image or image sequence as an input, the region corresponding to the actual interaction can be identified first and used as input. One technique that takes this two-step approach is Regional CNNs (R-CNN) (Girshick et al., 2014), that classify each region with a category-specific linear SVM. Notably, Peng and Schmid (2016) demonstrated a multi-regional two-stream R-CNN which uses a region-of-interest fusion layer for both appearance and motion models. Region-focused, stream-based models have also been used by Tran and Cheong (2017), who introduce cross-connections from the temporal to the spatial stream. These include convolutions that reduce the dimensionality of the temporal activation maps. The hierarchical model for features has also been used for the creation of action tubes (Gkioxari and Malik, 2015): spatio-temporal volumes centered on the performance of a particular action or interaction. Here, region proposals are found based on motion-appearance cues extracted with a two-stream CNN. The notion of using tubes for the representation of motion has also been adopted for different body parts by Mavroudi et al. (2017). Saha et al. (2016), Hou et al. (2017) have also implemented a model based on action tubes and R-CNNs as well as connections between the spatial and temporal models.

Adaptations to regional CNN models have been created by Gkioxari et al. (2015) and Mettes and Snoek (2017) to include multiple regions per example. The primary region contains the main actor or actors, while secondary regions are based on contextual cues of the scene. Similarly, Wang et al. (2016b) used a two-stream semantic region-based CNN (SR-CNNs) as an extension of Faster R-CNNs (Ren et al., 2015). The idea of using multiple independent or dependent regions for various cues, and using separate streams to encode the input, also allows to focus on discriminative regions such as a hand of one person that touches the body of another (Singh et al., 2016; Miao et al., 2017; Tu et al., 2018; Wu et al., 2016). Typically, such regions complement each other.

Instead of treating the image and motion aspects of a video in separate streams, a video sequence can be represented as a 3D volume that is composed of stacked frames. Baccouche et al. (2011) and Ji et al. (2013) use 3D convolutions to simultaneously encode the spatial and temporal features of such a volume. This approach is essentially an extension of the standard 2D convolutions to 3D. The resulting feature maps encode informative spatio-temporal patterns in the video volume. Tran et al. (2015) presented the C3D architecture and demonstrated its superiority over 2D CNNs. 3D convolutions can also be used concurrently with a two-stream network. Carreira and Zisserman (2017) have introduced a fusion of these two methodologies, two-stream inflated 3D-CNNs (I3D), that adds a temporal dimension to the kernels of both convolutional and pooling layers. The work considers the creation of two I3D models that are applied to static image and optical flow inputs, and thus allows the 3D-CNNs to benefit from the additional information of motion patterns in optical flow streams. Spatio-temporal networks can be used as a base architectures to extend the type of information processed such as queries for people regions (Girdhar et al., 2019), position and motion (Choutas et al., 2018) and feature neighborhood correspondence across time (Cao et al., 2019; Wang et al., 2018).

The larger number of parameters in 3D convolution blocks and, consequently, the demand of larger datasets for 3D-CNNs to train, have motivated the introduction of alternative convolution blocks. Notably, Qiu et al. (2017) have proposed three supplementary blocks with different configurations of a single 2D convolutional kernel for the extraction of appearance information per frame and a temporal kernel responsible for the changes of pixel values over time loosely inspired by the separable convolutions of 2D-CNNs (Chollet, 2017; Howard et al., 2017). This idea has also been used to separate spatio-temporal kernels into purely spatial and purely temporal ones by Tran et al. (2018) with the introduction of (2+1)D convolution blocks. Others have fused both solely-spatial and spatio-temporal convolutions in an effort to emphasize the spatial signal (Zhou et al., 2018). Chen et al. (2018) have also proposed the slicing of convolutional blocks in sets of fibers that are processed in parallel by the model. This significantly reduces the computation overhead, owing to the decreased size of the activation maps produced by each operation at each fiber (see Fig. 3).

4.3. Recurrent networks

While CNNs can recognize image components and learn to combine them to classify different classes, they lack the ability to recognize patterns across time. Stream-based networks and 3D convolutions can take into account motion, but do not explicitly deal with variations in the temporal performance of an action or interaction. An alternative approach is to use recurrent neural networks (RNNs) that model temporal patterns. The key idea is to use some form of recurrence in the network that allows the persistence of information through sequences of inputs. Thus the temporal variations in videos can be efficiently modeled alongside to the spatial variations.

Recurrent neural networks have been effectively used as a supplementary architecture to CNNs for extracting temporal features. In such architectures, spatial information is extracted though CNNs and is then passed to recurrent networks for learning the temporal characteristics of each interaction class (Bagautdinov et al., 2017; Deng et al., 2016). Zhao et al. (2017) proposed an approach based on the normalization of each layer of the network with batch normalization (Ioffe and Szegedy, 2015). The architecture is combined with a 3D-CNN using a two-stream fusion of the RNN and CNN. The use of multiple recurrent networks has also been extended to include tree structures (RNN-T) (Li et al., 2017), to perform a hierarchical recognition process in which each RNN is responsible for learning an action instance based on an Action Category Hierarchy (ACH). This allows for the distinction between very dissimilar classes high in the hierarchy, while subtle differences between related classes such as a handshake and a fist bump are dealt with in the lower nodes.

Recurrent Neural Networks suffer from vanishing gradients. This issue causes the updates in the network weights of the top layers to gradually diminish as the number of data-processing iterations increases. This hinders learning the temporal parameters effectively. To overcome this issue, Long Short-Term Memory (LSTM) RNNs (Hochreiter and Schmidhuber, 1997) have been introduced that include additional 'memory cell' modules that decide whether to keep the processed information. As such, they are capable of maintaining information over longer periods, which allows them to learn long-term dependencies (Chung et al., 2014). This is essential for the modeling of interaction classes as the distinctive information is often present in different phases of the interaction.

Donahue et al. (2015) and Li et al. (2018), Varol et al. (2017) have shown that the combination of convolutions and long-term recursions performs well for recognition tasks in videos. Donahue et al. (2015) was effective in both image and video description by directly connecting powerful feature extractors such as CNNs with recurrent models. Similarly, Baccouche et al. (2011) extracted features from the 3D-CNN architecture and extended the work to a two-step recognition process with a LSTM. The first step was the use of 3D convolutions for the

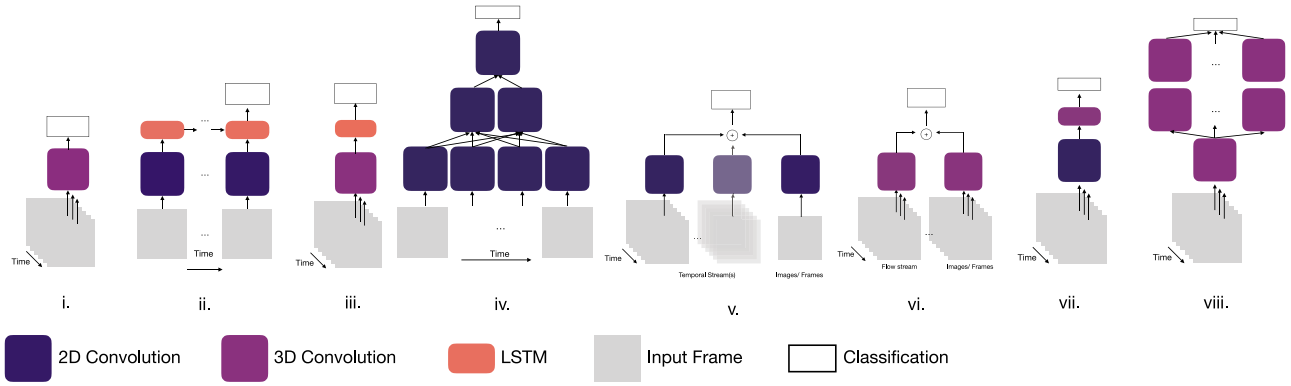

Fig. 3. Building blocks of video classification networks: (i) 3D-convolution (Ji et al., 2013), (ii) 2D-Convolutional LSTM over a sequence of frames (Donahue et al., 2015), (iii) 3D LSTM (Baccouche et al., 2011), (iv) slow-fusion (Karpathy and Fei-Fei, 2015), (v) two/multi-stream CNN (Simonyan and Zisserman, 2014a; Wang et al., 2016b; Miao et al., 2017; Tu et al., 2018), (vi) two-stream 3D-Conv network (Carreira and Zisserman, 2017), (vii) (2+1)D convolutions (Qiu et al., 2017; Tran et al., 2018) and (viii) Multi-fiber CNN (Chen et al., 2018).

extraction of spatio-temporal features. The second step is based on these learned features that are passed to the LSTM so the model can make predictions on the entire video sequence. As such, the network can benefit from both short-term and long-term temporal information.

Besides LSTMs, Highway Networks are an alternative solution to the vanishing gradient problem (Srivastava et al., 2015b). These networks allow for the direct passing of information through so-called highway modules that connect layers of the architecture similarly to LSTM's adaptive gating mechanism. Zilly et al. (2017) have extended this approach to include the spatial dimensionality in the information highways inside recurrent transitions.

Because the discriminative information of an interaction is typically found in selective parts of the input, several approaches have addressed methods for selection. In line with the multi-stream approaches (Section 4.2), Wang et al. (2017) have implemented LSTMs that consist of three branches that deal with person action, group action and scene recognition. This work is inspired by Gkioxari et al. (2017), who focused on human–object interactions instead. Multiple recurrent modules can be used to analyze human interactions. For example, Yan et al. (2017) built a model from three attention-specific LSTMs that use information from each of the two interacting actors and the overall scene of each example. Similarly, Si et al. (2019) also included spatio-temporal focused LSTMs, through a temporal hierarchy, for increasing the temporal receptive field of the network and allowing the exploration of co-occurring features in space and time. Ibrahim et al. (2016) presented a two-stage temporal model in which LSTMs are used to analyze each person in the scene while their combined outputs synthesize the relationship between them. Srivastava et al. (2015a) created an Encoder–Decoder architecture, in which the encoder LSTM maps input sequences to a delineation of specified length. The decoder LSTM then either reconstructs the inputs or creates predictions for future examples. The motivation of the work is to capture all information required to reproduce the input and therefore to select the most important features. This is achieved by minimizing the loss of the constructed sequence from the decoder LSTM and the actual input sequence. For example, in an interaction video, the decoder would focus on modeling the movement of the hands if the interaction is a handshake, or focus on the upper bodies if the interaction is a hug.

Of increasing importance for interaction recognition is the use of skeletal data, or poses. Pose data is a compact representation that is invariant to many typical image factors such as partial occlusions, low resolution and viewpoint. Consequently, the focus is mainly on modeling the temporal dynamics. Often, pose information can be regarded as a complementary input. For example, Gammulle et al. (2017) have created a spatio-temporal two-stream architecture with an addition of a LSTM with both frames and optical flow working as an attention mechanism. Attention mechanisms have also been used with pose information in recurrent structures to learn pose-related features in each time step (Du et al., 2017). This permits the analysis of the action from the collection of the per-frame human poses. Moreover, based on alternatives to LSTMs, Liu et al. (2016) have introduced gating mechanisms for creating a spatio-temporal LSTM (ST-LSTM). Given skeletal data in a tree-like structure, each ST-LSTM unit corresponds to a joint and receives spatio-temporal information from the previous and its own node. The new gating mechanism predicts the possible input based on the generated probabilities and compares it to the actual input. They implement the idea of assimilating the sequential input of videos by adjusting the effects on the context-based information stored in the network by allowing to analyze the data at each step and to decide when to update, remember of forget the contents in the memory cell with a tree-like representation of the skeleton.

Skeletal data have also been used by Zhu et al. (2016) in a fully connected LSTM model including internal gates, outputs and neurons that could be dropped by the network. Si et al. (2018) have proposed a combination of networks. The first network analyzes spatial information between frames by capturing the relationships between skeletal joins, while the second network focuses on the dynamics and the detailed temporal features that define each example. Other extensions include Lattice-LSTM (L^2STM) that enhances the capability of the memory cell to understand motion dynamics of the video sequence through individual local patterns, by leveraging both image and flow information extracted from a CNN classifier (Sun et al., 2017). Since there might be different patterns for different body parts and phases in the interaction, LSTMs have been adapted to consist of part-based sub-cells to model the long-term motion of key body parts (Du et al., 2015; Shahroudy et al., 2016). Because these models break down the interaction into meaningful blocks of motion, they can be used as the basis to learn a repertoire for action and interaction as shown by Shi et al. (2019). They introduce a directed acyclic graph (DAG) representation for the information of joints, bones and their relationship. Other approaches have targeted the dependencies among joints (Li et al., 2019) where information is also separated to actional links based on movement and structural links based on joint locations. By partitioning the interaction into sub-parts, these approaches can further reduce training cost and lead to the distinction between subtly different interaction classes.

5. Discussion

The past decades have seen impressive progress in the automated understanding of human behavior in videos. With the introduction of learned feature approaches such as CNNs, we can now analyze videos recorded in unconstrained settings. Consequently, there is a focus on more realistic video material. The result of initial works on specially recorded benchmarks datasets have largely saturated. In the meantime,

Fig. 4. Examples of ambiguous interactions. Sequence 1 shows that ambiguity can arise from an unexpected outcome: a high five that ends in holding hands. In Sequence 2, there is no contact between the two persons but their motivation for a high-five is apparent. There is comical intent in the interaction in Sequence 3. The comprehension of this scene requires deeper understanding of the interactions.

we have begun to address sustained, natural human interactions in a social context. This opens up a host of applications, from more intelligent video indexing to smart surveillance.

In Section 1.2, we discussed a number of challenges. The introduction of learned feature representations has alleviated some of the issues when dealing with variations in recording setting, person appearance and, to a lesser extent, viewpoint. The decoupling of the visual and temporal aspects of human interactions, for example using LSTMs (Alahi et al., 2016), has allowed researchers to focus more on the dynamics of interactions. Still, the promise of understanding social interactions directly from video is far from being met. Below, we discuss limitations of the state-of-the-art and highlight current trends and future directions.

Training scenarios with less data. Advances owing to CNNs come at a cost because learned feature representations require large amounts of relevant training data. While the datasets that focus on human interactions are still increasing in the number of classes and available videos, it will remain hard to harvest such datasets. Some works have exploited synthetic data generators to increase the amount and variation of the training data (Chen et al., 2017; Shotton et al., 2013). The generation of the data can also explicitly be part of the training process. Generative Adversarial Networks (GANs, Goodfellow et al. (2014)) contain a generative and a discriminative model that are jointly optimized. Recent work on the walking motion of pedestrians demonstrates the efficacy of the technique to model social behavior (Gupta et al., 2018). It remains to be investigated to what extend these results generalize to less-constrained interactions. Another line of approach is to use transfer learning (Weiss et al., 2016), to learn the parts of the network that deal with the lower-level aspects of the input from more general and more widely available training data. Despite these partial solutions, there typically is relatively few relevant data available given the complexity of the classification problem.

Increasing interaction class repertoire. Current work on the analysis of human interactions is limited by a relatively coarse division into behavior classes such as a handshake or a hug. Often, there is much more information contained in these interactions and humans have little difficulty identifying an awkward hug from a heartfelt one. *Semantically*, such interactions are very different. Yet, they can be visually very similar. With an increased focus on realistic human interactions comes a need to be able to distinguish between a larger number of classes, each of which might only subtly differ from others. These differences might originate from temporal aspects such as the coordination in time, but also from differences in poses or orientation. Completely separating the visual aspect from the temporal characteristics is likely to be sub-optimal. We consider the use of recurrent networks with more sophisticated gating functions as a promising trend.

The current practice is to consider an interaction as belonging to a single class only. But human behavior is often more open to subjectivity, and a less strict separation into classes could be beneficial for the generalization. The work on overlapping labels or behavior hierarchies (e.g., Frosst and Hinton, 2017; Yeung et al., 2018) is promising because it facilitates the focus on distinctive patterns at different levels of granularity, dependent on the type of interaction. A shift away from the one-vs-all classification can additionally facilitate the introduction of loss functions that take into account how related, visually or semantically, interactions are.

Units of interaction. Predominantly, interactions are classified directly based on the input. Some works have considered semantic mid-level features such as the action of an individual (e.g., Lan et al. (2012) and Sefidgar et al. (2015)) or the action of a body part (e.g., Chéron et al. (2015), Kong et al. (2012) and Tian et al. (2015)). Such methodologies bring some invariance in the representation, and can be learned per person. This effectively removes some of the dependencies and can facilitate the modeling of interactions as spatio-temporal patterns of these mid-level features. This approach can even be extended to deal with interactions for which no, or very little, training data is available. Specifically for human–human interactions, the coordination of pose and motion is crucial to distinguish between subtly different classes (Van Gemeren et al., 2018). Mid-level representations should take into account this coordination in both space and time, such as the distance and orientation between people, or the relative placement of a hand on the other's shoulder. Recent work on capsules by Hinton et al. (2018) and Sabour et al. (2017) appears promising in this respect. These works have shown great potential for accurately learning the pose of an object and constructing a hierarchy of parts enabling the understanding of features that is specific to a class. As such, geometric relations can be modeled in detail. An additional advantage is that capsules can be parallelized (Goyal et al., 2017), which limits the computational requirements.

Role of skeleton data. Human poses are one particular form of mid-level representation. We foresee an increased role of skeleton data, both during training and as additional input modality. Temporal patterns of interactions can be learned from skeleton data directly without having to take into account factors such as viewpoint and person appearance. Especially when units of interactions can be defined, pose and motion for an individual, as well as the coordination between people can be readily analyzed from skeleton data. Recent advances in human pose estimation from images and video (e.g., Carreira et al. (2016), Insafutdinov et al. (2017) and Yang et al. (2017)) have paved the way for effective pose-based attention mechanisms. While the computational requirements of the pose estimation task are significant, the benefit for the recognition of interactions has also been demonstrated (Du et al., 2017; Liu et al., 2016).

Detection and classification. The research on the automated analysis of human interactions has predominantly focused on recognition rather than detection. This means that interaction labels are usually not assigned to a region but to the image or video sequence as a whole. Rather, the understanding of human behavior would benefit from a link between person and interaction class. This permits us to say who interacted with whom, when. Especially in sustained or repeated social encounters, for example in public spaces, knowing the actors that interact would increase the efficacy of the analysis. A few works have addressed interaction detection (e.g., Van Gemeren et al. (2018) and Tian et al. (2013)) but usually in a two-step approach by first detection humans (e.g., Patron-Perez et al., 2012) and then considering their interactions. Especially in more crowded settings where partial occlusions are more common, such an approach is more likely to fail. An approach that focuses on the distinctive parts of the interaction is therefore favorable.

From observation to understanding. Finally, we see much potential in leveraging the recognition of interactions to the understanding of interactive human behavior. While the analysis of the observations is an essential step to understanding video contents, it often is not sufficient for our common use and demands. Often we are looking for anomalies, deviations from common practice. For example, Sequences 1 and 2 in Fig. 4 show interactions that are difficult to recognize but are more likely to be of interest to a user. Descriptive units of interactions can be instrumental in modeling anomalies. Commonly, it is the context of the behavior that is more descriptive, or gives a different meaning to our interactions. When a person is observed pushing another, it could be a playful instance between two friends or an actual act of violence. Longer-term analysis of the actors, their roles or relation to each other and knowledge of social and cultural norms can help in providing a deeper understanding of the observed social behavior. In particular, the understanding of the intentions of a person can help to analyze what a person is doing, instead of focusing on how that is achieved.

When looking at videos, we should deviate from the current agnostic perspective and treat videos not as sequences of images but as visual representations of social behavior. We foresee that datasets that target a more constrained setting, yet contain a wealth of social behavior (e.g., Alameda-Pineda et al. (2016)) are used as a step-up towards more generalized understanding of interactions from video. We identify a particular need for such datasets.

We are just scratching the surface when it comes to really understanding social behavior from video. But with the solid state-of-the-art performance in the analysis of interactions from videos, and the promising directions of research to deal with the current limitations, we expect that great strides can be made to close to the gap to the automated understanding of human interactions.

Acknowledgment

This publication is supported by the Netherlands Organization for Scientific Research (NWO) with a TOP-C2 grant for "Automatic recognition of bodily interactions" (ARBITER).

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., 2016. Social LSTM: Human trajectory prediction in crowded spaces. In: Computer Vision and Pattern Recognition, (CVPR). pp. 961–971.

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., Sebe, N., 2016. SALSA: A novel dataset for multimodal group behavior analysis. Trans. Pattern Anal. Mach. Intell. 38 (8), 1707–1720.

Anderson, D.J., Perona, P., 2014. Toward a science of computational ethology. Neuron 84 (1), 18–31.

Aran, O., Gatica-Perez, D., 2013. One of a kind: Inferring personality impressions in meetings. In: International Conference on Multimodal Interaction (ICMI). pp. 11–18.

Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H.J., Ponce-López, V., Baró, X., Guyon, I., Kasaei, S., Escalera, S., 2017. A survey on deep learning based approaches for action and gesture recognition in image sequences. In: Automatic Face & Gesture Recognition (FG). pp. 476–483.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A., 2011. Sequential deep learning for human action recognition. In: International Workshop on Human Behavior Understanding (HBU). pp. 29–39.

Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S., 2017. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Conference on Computer Vision and Pattern Recognition, (CVPR), Vol. 2. pp. 3425–3434.

Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning. In: International Conference on Machine Learning Workshops (ICML). pp. 17–36.

Bengio, Y., Bergeron, A., Boulanger-Lewandowski, N., Breuel, T., Chherawala, Y., Cisse, M., Erhan, D., Eustache, J., Glorot, X., Muller, X., 2011. Deep learners benefit more from out-of-distribution examples. In: International Conference on Artificial Intelligence and Statistics (ICAIS). pp. 164–172.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S., 2016. Dynamic image networks for action recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3034–3042.

Birdwhistell, R.L., 1952. Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture, first ed. Department of State, Foreign Service Institute, Washington, DC.

Bourdev, L., Maji, S., Brox, T., Malik, J., 2010. Detecting people using mutually consistent poselet activations. In: European Conference on Computer Vision (ECCV). pp. 168–181.

Caba Heilbron, F., Carlos Niebles, J., Ghanem, B., 2016. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1914–1923.

Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Mark Siskind, J., Wang, S., 2013. Recognize human activities from partially observed videos. In: Computer Vision and Pattern Recognition (CVPR). pp. 2658–2665.

Cao, Z., Simon, T., Wei, T., Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1302–1310.

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. arXiv preprint arXiv:1904.11492.

Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., 2016. Human pose estimation with iterative error feedback. In: Computer Vision and Pattern Recognition, (CVPR). pp. 4733–4742.

Carreira, J., Noland, E., Hillier, C., Zisserman, A., 2019. A Short Note on the Kinetics-700 Human Action Datase. arXiv preprint arXiv:1907.06987v1.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition, (CVPR). pp. 4724–4733.

Caruana, R., 1998. Learning to Learn. Springer, pp. 95–133, (Chapter) Multitask learning.

Cavazza, J., Zunino, A., San Biagio, M., Murino, V., 2016. Kernelized covariance for action recognition. In: International Conference on Pattern Recognition (ICPR). pp. 408–413.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018. Multi-fiber networks for video recognition. In: European Conference on Computer Vision, (ECCV). pp. 364–380.

Chen, Y., Shen, C., Wei, X.-S., Liu, L., Yang, J., 2017. Adversarial PoseNet: A structure-aware convolutional network for human pose estimation. In: International Conference on Computer Vision (ICCV). pp. 1221–1230.

Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P., 2015. Advances in human action recognition: A survey. arXiv preprint arXiv:1501.05964.

Chéron, G., Laptev, I., Schmid, C., 2015. P-CNN: Pose-based CNN features for action recognition. In: International Conference on Computer Vision, (ICCV). pp. 3218–3226.

Cho, N.-G., Park, S.-H., Park, J.-S., Park, U., Lee, S.-W., 2017. Compositional interaction descriptor for human interaction recognition. Neurocomputing.

Choi, W., Savarese, S., 2014. Understanding collective activities of people from videos. Trans. Pattern Anal. Mach. Intell. 36 (6), 1242–1257.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 1800–1807.

Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C., 2018. Potion: Pose motion representation for action recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 7024–7033.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V., 2011. Social interaction discovery by statistical analysis of F-formations. In: British Machine Vision Conference (BMVC), Vol. 2. p. 4.

Delaitre, V., Laptev, I., Sivic, J., 2010. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In: British Machine Vision Conference (BMVC). pp. 1–11.

Deng, Z., Vahdat, A., Hu, H., Mori, G., 2016. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 4772–4781.

Diba, A., Sharma, V., Van Gool, L., 2017. Deep temporal linear encoding networks. In: Computer Vision and Pattern Recognition, (CVPR). pp. 2329–2338.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: Computer Vision and Pattern Recognition, (CVPR). pp. 2625–2634.

Du, W., Wang, Y., Qiao, Y., 2017. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3725–3734.

Du, Y., Wang, W., Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1110–1118.

Feichtenhofer, C., Pinz, A., Wildes, R., 2016. Spatiotemporal residual networks for video action recognition. In: Advances in Neural Information Processing Systems (NIPS). pp. 3468–3476.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. Trans. Pattern Anal. Mach. Intell. 32 (9), 1627–1645.

Frosst, N., Hinton, G., 2017. Distilling a Neural Network Into a Soft Decision Tree. arXiv preprint arXiv:1711.09784.

Gammulle, H., Denman, S., Sridharan, S., Fookes, C., 2017. Two stream LSTMs: A deep fusion framework for human action recognition. In: Applications of Computer Vision (WACV). pp. 177–186.

Gao, C., Yang, L., Du, Y., Feng, Z., Liu, J., 2016. From constrained to unconstrained datasets: An evaluation of local action descriptors and fusion strategies for interaction recognition. World Wide Web 19 (2), 265–276.

Garcia, N., Morerio, P., Murino, V., 2018. Modality distillation with multiple stream networks for action recognition. In: European Conference on Computer Vision (ECCV). pp. 106–121.

Girdhar, R., Carreira, J., Doersch, C., Zisserman, A., 2019. Video action transformer network. In: Computer Vision and Pattern Recognition (CVPR).

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition, (CVPR). pp. 580–587.

Gkioxari, G., Girshick, R., Dollár, P., He, K., 2017. Detecting and Recognizing Human-Object Interactions. arXiv preprint arXiv:1704.07333.

Gkioxari, G., Girshick, R., Malik, J., 2015. Contextual action recognition with r* CNN. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1080–1088.

Gkioxari, G., Malik, J., 2015. Finding action tubes. In: Computer Vision and Pattern Recognition, (CVPR). pp. 759–768.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS). pp. 2672–2680.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, large minibatch SGD: Training Imagenet in 1 hour. arXiv preprint arXiv:1706.02677.

Gu, C., Sun, C., Ross, D., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al., 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In: Computer Vision and Pattern Recognition, (CVPR). pp. 6047–6056.

Güler, R.A., Neverova, N., Kokkinos, I., 2018. Densepose: Dense human pose estimation in the wild. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 7297–7306.

Gupta, A., Johnson, J., Li, F.-F., Savarese, S., Alahi, A., 2018. Social GAN: Socially acceptable trajectories with generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 2255–2264.

Hara, K., Kataoka, H., Satoh, Y., 2018. CaN spatiotemporal 3D CNNs retrace the history of 2d CNNs and imagenet?. In: Computer Vision and Pattern Recognition, (CVPR). pp. 18–22.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 770–778.

Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C., 2015. Activitynet: A large-scale video benchmark for human activity understanding. In: Computer Vision and Pattern Recognition, (CVPR). pp. 961–970.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image Vis. Comput. 60, 4–21.

Hinton, G., Frosst, N., Sabour, S., 2018. Matrix capsules with EM routing. In: International Conference on Learning Representations (ICLR).

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

Hou, R., Chen, C., Shah, M., 2017. Tube convolutional neural network (t-CNN) for action detection in videos. In: International Conference on Computer Vision, (ICCV). pp. 5822–5831.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G., 2016. A hierarchical deep temporal model for group activity recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1971–1980.

Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B., 2017. Arttrack: Articulated multi-person tracking in the wild. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1293–1301.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML). pp. 448–456.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. Trans. Pattern Anal. Mach. Intell. 35 (1), 221–231.

Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2015. Panoptic studio: A massively multiview system for social motion capture. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3334–3342.

Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3128–3137.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1725–1732.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., 2017. The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Khodabandeh, M., Vahdat, A., Zhou, G.-T., Hajimirsadeghi, H., Javan Roshtkhari, M., Mori, G., Se, S., 2015. Discovering human interactions in videos with limited data labeling. In: Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 9–18.

Kong, Y., Jia, Y., Fu, Y., 2012. Learning human interaction by interactive phrases. In: European Conference on Computer Vision, (ECCV). pp. 300–313.

Kong, Y., Kit, D., Fu, Y., 2014. A discriminative model with multiple temporal scales for action prediction. In: European Conference on Computer Vision, (ECCV). pp. 596–611.

Koohzadi, M., Charkari, N.M., 2017. Survey on deep learning methods in human action recognition. IET Comput. Vis. 11 (8), 623–632.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: A large video database for human motion recognition. In: International Conference on Computer Vision (ICCV). pp. 2556–2563.

Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G., 2012. Discriminative latent models for recognizing contextual group activities. Trans. Pattern Anal. Mach. Intell. 34 (8), 1549—1562.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q., 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In: Computer Vision and Pattern Recognition (CVPR).

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G., 2018. VideoLSTM Convolves, attends and flows for action recognition. Comput. Vis. Image Underst. 166, 41–50.

Li, W., Wen, L., Chang, M.-C., Nam Lim, S., Lyu, S., 2017. Adaptive RNN tree for large-scale human action recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1444–1452.

Li, S., Zhang, W., Chan, A.B., 2015. Maximum-margin structured learning with deep networks for 3D human pose estimation. In: International Conference on Computer Vision (ICCV). pp. 2848–2856.

Liu, J., Shahroudy, A., Xu, D., Wang, G., 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European Conference on Computer Vision (ECCV). pp. 816–833.

Liu, G.-H., Yang, J.-Y., Li, Z., 2015. Content-based image retrieval using computational visual attention model. Pattern Recognit. 48 (8), 2554–2566.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: International Conference on Computer Vision (ICCV), Vol. 2. pp. 1150–1157.

Lu, J., Xu, R., Corso, J.J., 2015. Human action segmentation with hierarchical supervoxel consistency. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3762–3771.

Marín-Jiménez, M.J., Yeguas, E., De La Blanca, N.P., 2013. Exploring STIP-based models for recognizing human interactions in TV videos. Pattern Recognit. Lett. 34 (15), 1819–1828.

Marszalek, M., Laptev, I., Schmid, C., 2009. Actions in context. In: Computer Vision and Pattern Recognition (CVPR). pp. 2929–2936.

Mavroudi, E., Tao, L., Vidal, R., 2017. Deep moving poselets for video based action recognition. In: Applications of Computer Vision (WACV). pp. 111–120.

Mettes, P., Snoek, C.G., 2017. Spatial-aware object embeddings for zero-shot localization and classification of actions. In: International Conference on Computer Vision, (ICCV). pp. 4443–4452.

Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., Cao, X., Liu, Z., Chai, X., Liu, Z., et al., 2017. Multimodal gesture recognition based on the ResC3D network. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3047–3055.

Mohammadi, S., Kiani, H., Perina, A., Murino, V., 2015. Violence detection in crowded scenes using substantial derivative. In: Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6.

Monfort, M., Zhou, B., Bargal, S.A., Andonian, A., Yan, T., Ramakrishnan, K., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al., 2018. Moments in Time Dataset: One million videos for event understanding. arXiv preprint arXiv:1801.03150.

Motiian, S., Siyahjani, F., Almohsen, R., Doretto, G., 2017. Online human interaction detection and recognition with multiple cameras. Trans. Circuits Syst. Video Technol. 27 (3), 649–663.

Niebles, J.C., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vis. 79 (3), 299–318.

Oneata, D., Verbeek, J., Schmid, C., 2013. Action and event recognition with fisher vectors on a compact feature set. In: Computer Vision and Pattern Recognition (CVPR). pp. 1817–1824.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Park, E., Han, X., Berg, T.L., Berg, A.C., 2016. Combining multiple sources of knowledge in deep CNNs for action recognition. In: Applications of Computer Vision (WACV). pp. 1–8.
Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A., 2012. Structured learning of human interactions in TV shows. Trans. Pattern Anal. Mach. Intell. 34 (12), 2441–2453.
Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.D., 2010. High five: Recognising human interactions in TV shows. In: British Machine Vision Conference (BMVC), Vol. 1. p. 2.
Peng, X., Schmid, C., 2016. Multi-region two-stream r-CNN for action detection. In: European Conference on Computer Vision (ECCV). pp. 744–759.
Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A., 2018. Exploiting deep residual networks for human action recognition from skeletal data. Comput. Vis. Image Underst. 170, 51–66.
Poppe, R., 2010. A survey on vision-based human action recognition. Image Vis. Comput. 28 (6), 976–990.
Poppe, R., 2017. Automatic analysis of bodily social signals. In: Burgoon, J.K., Magnenat-Thalmann, N., Pantic, M., Vinciarelli, A. (Eds.), Social Signal Processing. Cambridge University Press, pp. 155–167.
Prabhakar, K., Rehg, J.M., 2012. CaTegorizing turn-taking interactions. In: European Conference on Computer Vision, (ECCV). pp. 383–396.
Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3D residual networks. In: International Conference on Computer Vision, (ICCV). pp. 5534–5542.
Raptis, M., Sigal, L., 2013. Poselet key-framing: A model for human activity recognition. In: Computer Vision and Pattern Recognition (CVPR). pp. 2650–2657.
Reddy, K.K., Shah, M., 2013. Recognizing 50 human action categories of web videos. Mach. Vis. Appl. 24 (5), 971–981.
Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., 2013. Decoding children's social behavior. In: Computer Vision and Pattern Recognition (CVPR). pp. 3414–3421.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 91–99.
Rodriguez, M.D., Ahmed, J., Shah, M., 2008. Action MATCH a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition (CVPR). pp. 1–8.
Ronchi, M.R., Perona, P., 2015. Describing common human visual actions in images. In: British Machine Vision Conference (BMVC). 52–1.
Ryoo, M.S., 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In: International Conference on Computer Vision (ICCV). pp. 1036–1043.
Ryoo, M.S., Aggarwal, J., 2010. UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). In: Computer Vision and Pattern Recognition Workshops (CVPRW), Vol. 2. p. 4.
Ryoo, M., Aggarwal, J., 2011. Stochastic representation and recognition of high-level group activities. Int. J. Comput. Vis. 93 (2), 183–200.
Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. In: Advances in Neural Information Processing Systems (NIPS). pp. 3859–3869.
Saha, S., Singh, G., Sapienza, M., Torr, P.H., Cuzzolin, F., 2016. Deep learning for detecting multiple space-time action tubes in videos. arXiv preprint arXiv:1608.01529.
Sefidgar, Y.S., Vahdat, A., Se, S., Mori, G., 2015. Discriminative key-component models for interaction detection and recognition. Comput. Vis. Image Underst. 135, 16–30.
Sempena, S., Maulidevi, N.U., Aryan, P.R., 2011. Human action recognition using dynamic time warping. In: Electrical Engineering and Informatics (ICEEI). pp. 1–5.
Sener, F., Ikizler-Cinbis, N., 2015. Two-person interaction recognition via spatial multiple instance embedding. J. Vis. Commun. Image Represent. 32, 63–73.
Shahroudy, A., Liu, J., Ng, T.-T., Wang, G., 2016. NTU Rgb+ d: A large scale dataset for 3D human activity analysis. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1010–1019.
Shariat, S., Pavlovic, V., 2013. A new adaptive segmental matching measure for human activity recognition. In: International Conference on Computer Vision (ICCV). pp. 3583–3590.
Sheerman-Chase, T., Ong, E.-J., Bowden, R., 2011. Cultural factors in the regression of non-verbal communication perception. In: International Conference on Computer Vision Workshops (ICCV). pp. 1242–1249.
Shi, L., Zhang, Y., Cheng, J., Lu, H., 2019. Skeleton-based action recognition with directed graph neural networks. In: Computer Vision and Pattern Recognition (CVPR).
Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R., 2013. Real-time human pose recognition in parts from single depth images. Commun. ACM 56 (1), 116–124.
Shu, X., Tang, J., Qi, G.-J., Song, Y., Li, Z., Zhang, L., 2017. Concurrence-Aware Long Short-Term Sub-Memories for Person-Person Action Recognition. arXiv preprint arXiv:1706.00931.
Si, C., Chen, W., Wang, W., Wang, L., Tan, T., 2019. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: Computer Vision and Pattern Recognition (CVPR).
Si, C., Jing, Y., Wang, W., Wang, L., Tan, T., 2018. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: European Conference on Computer Vision (ECCV). pp. 106–121.
Simonyan, K., Zisserman, A., 2014a. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NIPS). pp. 568–576.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M., 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1961–1970.
Slimani, K., Benezeth, Y., Souami, F., 2014. Human interaction recognition based on the co-occurence of visual words. In: Conference on Computer Vision and Pattern Recognition Workshops, (CVPR). pp. 455–460.
Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
Srivastava, R.K., Greff, K., Schmidhuber, J., 2015b. Training very deep networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 2377–2385.
Srivastava, N., Mansimov, E., Salakhudinov, R., 2015a. Unsupervised learning of video representations using LSTMs. In: International Conference on Machine Learning (ICML). pp. 843–852.
Sun, L., Jia, K., Chen, K., Yeung, D.-Y., Shi, B.E., Savarese, S., 2017. Lattice long short-term memory for human action recognition. In: International Conference on Computer Vision (ICCV). pp. 2147–2156.
Tian, Y., Cao, L., Liu, Z., Zhang, Z., 2012. Hierarchical filtered motion for action recognition in crowded videos. IEEE Trans. Syst. Man Cybern. Part C 42 (3), 313–323.
Tian, Y., Luo, P., Wang, X., Tang, X., 2015. Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1904–1912.
Tian, Y., Sukthankar, R., Shah, M., 2013. Spatiotemporal deformable part models for action detection. In: Computer Vision and Pattern Recognition, (CVPR). pp. 2642–2649.
Tran, K.N., Bedagkar-Gala, A., Kakadiaris, I.A., Shah, S.K., 2013. Social cues in group formation and local interactions for collective activity analysis. In: International Conference on Computer Vision Theory and Applications (VISAPP). pp. 539–548.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks. In: International Conference on Computer Vision (ICCV). pp. 4489–4497.
Tran, A., Cheong, L.-F., 2017. Two-stream flow-guided convolutional attention networks for action recognition. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3110–3119.
Tran, K.N., Gala, A., Kakadiaris, I.A., Shah, S.K., 2014. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. Pattern Recognit. Lett. 44, 49–57.
Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 6450–6459.
Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R.C., Li, B., Yuan, J., 2018. Multi-stream CNN: Learning representations based on human-related regions for action recognition. Pattern Recognit. 79, 32–43.
Turchini, F., Seidenari, L., Del Bimbo, A., 2016. Understanding and localizing activities from correspondences of clustered trajectories. Comput. Vis. Image Underst..
Van Gemeren, C., Poppe, R., Veltkamp, R.C., 2016. Spatio-temporal detection of fine-grained dyadic human interactions. In: International Workshop on Human Behavior Understanding (HBU). pp. 116–133.
Van Gemeren, C., Poppe, R., Veltkamp, R.C., 2018. Hands-on: Deformable pose and motion models for spatiotemporal localization of fine-grained dyadic interactions. EURASIP J. Image Video Process. 2018 (1), 16.
Van Gemeren, C., Tan, R.T., Poppe, R., Veltkamp, R.C., 2014. Dyadic interaction detection from pose and flow. In: International Workshop on Human Behavior Understanding (HBU). pp. 101–115.
Varol, G., Laptev, I., Schmid, C., 2017. Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell..
Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social signal processing: Survey of an emerging domain. Image Vis. Comput. 27 (12), 1743–1759.
Vondrick, C., Pirsiavash, H., Torralba, A., 2016. Anticipating visual representations from unlabeled video. In: Computer Vision and Pattern Recognition, (CVPR). pp. 98–106.
Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Computer Vision and Pattern Recognition. pp. 7794–7803.
Wang, M., Ni, B., Yang, X., 2017. Recurrent modeling of interaction context for collective activity recognition. In: Computer Vision and Pattern Recognition, (CVPR). pp. 3048–3056.
Wang, H., Schmid, C., 2013. Action recognition with improved trajectories. In: Computer Vision and Pattern Recognition (CVPR). pp. 3551–3558.
Wang, Y., Song, J., Wang, L., Van Gool, L., Hilliges, O., 2016b. Two-stream SR-CNNs for action recognition in videos. In: British Machine Vision Conference (BMVC).

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016a. Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision (ECCV). pp. 20–36.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big Data 3 (1), 9.

Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X., 2016. Multi-stream multi-class fusion of deep networks for video classification. In: Multimedia Conference (ACMM). pp. 791–800.

Yan, Y., Ni, B., Yang, X., 2017. Predicting Human Interaction via Relative Attention Model. arXiv preprint arXiv:1705.09467.

Yang, Y., Baker, S., Kannan, A., Ramanan, D., 2012. Recognizing proxemics in personal photos. In: Computer Vision and Pattern Recognition (CVPR). pp. 3522–3529.

Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., 2017. Learning feature pyramids for human pose estimation. In: International Conference on Computer Vision (ICCV). pp. 1290–1299.

Yang, Y., Ramanan, D., 2011. Articulated pose estimation with flexible mixtures-of-parts. In: Computer Vision and Pattern Recognition (CVPR). pp. 1385–1392.

Yao, B.Z., Nie, B.X., Liu, Z., Zhu, S.-C., 2014. Animated pose templates for modeling and detecting human actions. IEEE Trans. Pattern Anal. Mach. Intell. 36 (3), 436–452.

Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L., 2018. Every moment counts: Dense detailed labeling of actions in complex videos. Int. J. Comput. Vis. 126 (2), 375–389.

Yi, S., Wang, X., Lu, C., Jia, J., 2014. L0 regularized stationary time estimation for crowd group analysis. In: Computer Vision and Pattern Recognition, (CVPR). pp. 2211–2218.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks?. In: Advances in Neural Information Processing Systems (NIPS). pp. 3320–3328.

Yu, G., Yuan, J., 2015. Fast action proposals for human action detection and search. In: Computer Vision and Pattern Recognition, (CVPR). pp. 1302–1311.

Yu, G., Yuan, J., Liu, Z., 2012. Propagative hough voting for human activity recognition. In: European Conference on Computer Vision, (ECCV). pp. 693–706.

Yub Jung, H., Lee, S., Seok Heo, Y., Dong Yun, I., 2015. Random tree walk toward instantaneous 3D human pose estimation. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 2467–2474.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D., 2012. Two-person interaction detection using body-pose features and multiple instance learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 28–35.

Zhang, B., De Natale, F.G., Conci, N., 2013. Recognition of social interactions based on feature selection from visual codebooks. In: Conference on Image Processing (ICIP). pp. 3557–3561.

Zhang, Y., Liu, X., Chang, M.-C., Ge, W., Chen, T., 2012. Spatio-temporal phrases for activity recognition. In: European Conference on Computer Vision (ECCV). pp. 707–721.

Zhao, R., Ali, H., Van der Smagt, P., 2017. Two-Stream RNN/CNN for Action Recognition in 3D Videos. arXiv preprint arXiv:1703.09783.

Zhao, H., Yan, Z., Torresani, L., Torralba, A., 2019. HACS: Human action clips and segments dataset for recognition and temporal localization. arXiv preprint arXiv:1712.09374.

Zhou, Y., Sun, X., Zha, Z.-J., Zeng, W., 2018. Mict: mixed 3D/2d convolutional tube for human action recognition. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 449–458.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X., 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Association for the Advancement of Artificial Intelligence (AAAI), Vol. 2. p. 8.

Ziaeefard, M., Bergevin, R., Morency, L.-P., 2015. Time-slice prediction of dyadic human activities. In: British Machine Vision Conference (BMVC). 167–161.

Zilly, J.G., Srivastava, R.K., Koutnık, J., Schmidhuber, J., 2017. Recurrent highway networks. In: International Conference on Machine Learning (ICML). pp. 4189–4198.

Zoph, B., Le, Q.V., 2017. Neural architecture search with reinforcement learning. In: International Conference of Learning Representations (ICLR).

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 8697–8710.