Understanding human-human interactions: a survey

Alexandros Stergiou

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 settings, the appearance of the people depicted and the performance of their interaction. This survey provides a summary of these challenges and datasets, followed by an in-depth discussion of relevant vision-based recognition and detection methods. We focus on recent, promising work based on convolutional neural networks (CNNs). Finally, we outline directions to overcome the limitations of the current state-of-the-art.

1. Introduction
Despite significant research progress in the automated analysis of humans and their activities [19, 50, 69, 100], the recognition of human interactions from video remains a challenging topic. Integral part of the difficulty is that understanding interactions between two or more people requires more than analyzing the actions of each person in isolation. Rather, it is the coordination, in both space and time, between the actions that reveals the true nature of their collective behavior. In this paper, we survey the research in human-human interactions, with a particular focus on recent methods based on convolutional neural networks (CNNs).

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. Examples are found in Figure 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 Figure 1 (center) could also be considered a lift interaction. The works discussed in this survey exclusively treat the interaction recognition task as deterministic. We discuss alternatives in the Discussion section.

The automated recognition of bodily interactions from video mainly benefits content-based video retrieval [79, 115], security [3] and surveillance [24, 47, 133, 161] and interactive human-computer interfaces [104, 119]. 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 for a broader use of automated measures arise when computers can understand the interactions in terms of communicative and affective intent. We discuss the evolution of the current state-of-the-art 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 image conditions

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 or fist bump greetings.

1.2.2. Variation in interaction performance

The performance in terms of body movements and coordination of the same interaction class can differ significantly, see Figure 1(left). Ronchi and Perona [107] 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 [2, 11], 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 towards the prediction of future actions have immediate impact on the improvement of scene understanding. These are shown by the work of [145], which anticipates actions and objects and is also capable of ascertaining human motions. Others build on the key idea that future actions can be predicted by classifying an action or interaction solely on its start [172]. Such an approach might work well for goal-directed interactions [13, 109], but is less successful when the variation in the performance increases (see Figure 1(right)). This is especially true when the interactions are more social and reactive in a communicative or affective way, such as a pat on the shoulder to display support.

Some works have addressed the estimation of a skeletal representation in order to circumvent having to learn interaction patterns directly from video [17, 98, 166, 167]. Recent methods rely on CNN-based approaches (e.g., Carreira et al. [14], Li et al. [75], Pishchulina et al. [99], Tompson et al. [136], Yang et al. [157, 156]) 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. 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. However, errors and inaccuracies in the pose estimation process might be propagated to the classification task.

1.2.3. Practical challenges

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 small 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 (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 Figure 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 the 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 current state-of-the-art in Section 5 and present promising avenues for further research.

2.Datasets

The availability of common, labeled datasets and the direct comparisons between methods generally leads to better understanding of the relative advantages and limitations and, consequently, progression in performance. Compared to datasets available for individual action recognition [49, 70, 106, 126], resources for human-to-human interactions are relatively scarce. Most notably, the limited amount of variation in viewpoints, application context and movement performance has thus far 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 Figure 2. A summary of the datasets appears in Table 1.

2.1. UT-Interaction

UT-Interaction [110] contains 20 sequences and six interaction classes. With almost static background, limited occlusions and a fixed viewpoint, the classification difficulty is low compared to more recent datasets. UT-Interaction is used as benchmark for many methodologies, ranging from bounding boxes techniques [89, 121] to bags-of-visual-words [118, 125]. Some works have also addressed the detection of interactions in both space and time [38].
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 [96, 95]. The dataset includes annotations of the upper body of actors, 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 a larger variation in the number of people in the scene. Because the material is taken from movies, all interactions are acted and the recording setting is highly controlled.

2.3. Hollywood2

Similar to TV Human Interaction, Hollywood2 [84] consists of clips from movies. The total number of films that were used to synthesize both the training and testing sets is close to seventy. Subtitles were used to align script data with the corresponding movie scenes, to semi-automatically find relevant clips. 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 [37]. The videos are captured in lab conditions, 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 [167]. It features eight two-person interactions: approach, depart, kick, punch, handshake, hug and pass object. The clips are segmented in time, with the interactions fully occupy the frame, so the dataset is not suitable for spatio-temporal detection of interaction. The main drawbacks are similar those of ShakeFive2.

2.6. CMU Panoptic

The CMU Panoptic dataset [61] is recorded in large geometric dome with VGA cameras distributed across the surface. This makes the recordings inherently controlled. 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.

2.7. DeepMind Kinetics

The DeepMind Kinetics dataset [64] provides the largest number of human action and interaction clips. It contains 600 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 10-second clips from YouTube videos. The video material is not professionally edited and features a large variety of background clutter, illumination settings and motion blur, which increases the difficulty of the classification task.

Table 1: Summary of datasets with types of data, number of classes and their examples, volume of interactions per sequence and inclusion of video noise

| Dataset                | Source                  | Samples   | Classes | Actors | Video duration | Scripted |
|------------------------|-------------------------|-----------|---------|--------|----------------|----------|
| UT-Interaction         | Outside recordings      | 60 interactions | 6       | 8      | 13-23s         | Yes      |
| High-Five              | TV shows                | 300       | 4       | 100+   | 1-5s           | Yes      |
| ShakeFive2             | Laboratory recordings   | 153       | 5       | 33     | 6s             | Yes      |
| Hollywood2             | Films                   | 3669      | 12      | 100+   | 11s (avrg.)    | Yes      |
| SBU Kinect             | Laboratory recordings   | 300       | 21      | 9      | 2s             | Yes      |
| CMU Panoptic           | Laboratory recordings   | Multi-view sequences | 5 vignettes | 16 | 10-15 min | Partially |
|Kinetics                | YouTube videos          | approx. 500000 | 600     | 100+   | 10s            | No       |
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 these 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 [83, 168], SIFT descriptors [25, 81] or optical flow [165]. 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.

When additional depth information is available, e.g. from RGB-D recordings, local features can take into account depth gradients [77]. For the efficient mapping of 3D points from other viewpoints, Xia and Aggarwal [154] considered the use of depth-sequences with the creation of a codebook.

To increase the robustness of local descriptors, a distribution of points is usually described as a bag-of-words (BoW) or Fisher vector [36, 92]. 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. [91] 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. [169] 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 [118, 141]. The computed histograms represent similar features in single or multiple frames. Histograms of visual words have also been utilized by Kong et al. [67] in which the words derived from the quantization of the spatio-temporal descriptors, were clustered into groups. This creates a high-level representation of dyadic interactions termed interactive phases. These phases include motion relationships between characteristic parts of the interaction motion such as two hands shaking. This idea has been extended to localize interactions by spatially clustering the phrases [140]. To allow for variation in the temporal domain, Prabhakar and Rehg [101] 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. [68] consider only body parts that characterize the interaction. Their method pools BoVW responses in a coarse grid. This allows them to identify specific motion patterns relative to a persons 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. [88] extend this approach by grouping the motion patterns as BoW vectors. Similarly, Turchini et al. [143] introduce an approach that is able to localize interactions from the trajectories of multiple local feature types. Wang et al. [146] have introduced 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) or Motion Boundary Histograms (MBH) and linked over time. The method was improved by Wang and Schmid [147] with the cancellation of camera motion by finding the homography in pairs of frames.

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 [12]. 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 [103] 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. [21] 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. [71] presented an Action Context (AC) descriptor that is based on connected action probability vectors of several people. Similarly, Choi and Savarese [22] 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 a larger region in a frame is considered, we can address the recognition of a person’s action from specific parts of the human body. Patches in an image can be described as HOG, HOF or MBH. HOGs describe the edge orientations within a grid of cells and encode spatial information such as a specific pose. HOFs are similar, but consider movement vectors from optical flow. Consequently, they encode motion direction within a spatial grid. Finally, MBH describes motion boundaries in a related way.

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. [10] to detect people engaged in specific actions, and was applied to human-human interactions.
by [102]. Sefidgar et al. [113] use the same reasoning to create a model based on discriminative key frames and consider their relative distance and timing within the interaction. Sener and Ikizler-Cinbis [116] formulate interaction detection as a multiple-instance learning problem to select 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. [39] combine HOG and HOF descriptors to encode the characteristic frame of a two-person interaction. Commonly, HOG and HOF descriptors extracted at specific locations are then classified with, either support vector machines (SVMs) or by using a BoW [72, 152]. Yu and Yuan [164] concatenate HOG and HOF and applied a Fisher Vector representation to make the detection linearly separable, thus allowing the model to concurrently utilize spatial and temporal features. While optical flow can be seen as the representation of the motion between two subsequent frames, tracklets describe the path of local key-points over longer time intervals. Mousavi et al. [90] introduce Histogram of Oriented Tracklets (HOT) that summarize tracks of local key-points.

Patches are sometimes combined with local descriptors to take advantage of potentially conjoint salient pose or motion information [60]. Yin et al. [162] employ 3D-SIFT to describe local motion events, but used a HOF to model the global motion in an image. Similarly, Lathuilière et al. [73] combined HOG descriptors and trajectory information from linked local features. Single-person and two-person interaction attributes such as “two persons are standing side-by-side” were calculated from these features.

Another approach is to first detect faces or bodies using a generic face or body detector [95, 108]. Given two close detections, interactions can subsequently be classified based on extracted features within the detection region [109]. Various attributes, including gross body movement and proximity, have been employed to classify the interaction. Patron-Perez et al. [95] also include the relative size and orientation of each person. Khodabandeh et al. [65] 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) [33]. 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. [82] 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 classify the action. van Gemeren et al. [38] use interaction-specific DPMs with poselet parts [10] 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 [159]. This approach has been used to describe the joint poses of two interacting people [158]. The parts in DPMs have a fixed relative scale. Hoai and Zisserman [52] extend the model to account for deformations in scale. Similarity with template examples of typical interactions are then used to rank the detection scores. The main constraint of this method is that it is used primarily for TV material, with an emphasis towards upper body movement.

To account for more variation in the temporal performance of interactions, authors have introduced various methods. Ji et al. [59] model the changes in HOG descriptors over time using a Hidden Markov Model (HMM). Based on the distance between two people, they consider the frame to be in the start, middle or end stage of the interaction. The HMM scores for the stages are fused for the final classification. The same rationale of splitting an interaction into phases has been adopted by Cao et al. [13], who address the task of classifying a sequence with potentially missing frames.

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. [160] 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. [135] have extended DPMs for action detection to model changes in pose over time, using spatiotemporal descriptors [66]. 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. An extension of DPMs to include part deformations not only in space and scale, but also in time is presented in van Gemeren et al. [38]. This enables the detection, in both space and time, of interactions that are characterized by a single key pose or key movement phase. Tran and Yuan [139] also address a localization task but consider linking regions over time based on HOG and HOF in a structured learning approach. A max-path algorithm is used to find the optimal volume that contains the action in space and time.

4. Interaction detection from learned features

The hand-coded feature descriptors described in Section 3 focused on local or global spatial or spatio-temporal information. Typically, low-level descriptors are used to obtain frame and video information, which are then passed either to a higher-level descriptor, responsible for mapping, or directly to a classifier, which is the final step of the process. The manual selection of features that are used for classification, could prove sub-optimal, as the process is agnostic to the specific classification task, domain or class of behaviors. Therefore, uninformative features might be selected and informative cues might be missed.
Based on the introduction of multiple convolutions by Le-Cun et al. [74], 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 suboptimal 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 as seen by [123] allows for a hierarchical extraction of complex features. Typically, the characteristics extracted in the first layers of the network correspond to low-level features such as edges and simple textures. Later layers of the network are targeted towards the extraction of more intricate features as for example, characteristic patterns and motifs.

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 defined 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 with convolutions and finally discuss the way that recurrent models deal with temporal information.

4.1. Single frame networks

CNNs have been used to classify actions and interactions in single frames [4, 9, 42]. Similar to the use of handcrafted features (Section 3), the focus is on a characteristic joint pose. To extend this methodology to sequences of image, several approaches have been proposed.

Based on the classification of individual frames, Karpathy et al. [63] 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 have proven to be insufficient as the temporal variations were not found sensitive to the different architectural connections. This was primarily due to the significant scene disparity within interaction class and the similarity of body postures across classes. It is a challenge to deal with these variations as they have to be modeled from the typically modest amount of available training videos.

To partly mitigate this issue, authors have investigated the use of Transfer Learning [8, 7, 16, 93, 163]. This is a process in which the network is first trained on a large dataset, that provides general examples, and subsequently re-purpose the features learned for another, typically 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 when using complex network architectures. An example can be seen in Figure 3 where the fully connected layers of an Inception network variant [131], trained on the ImageNet dataset [26], are re-trained on the TV Human Interaction dataset.

4.2. Motion-based and stream networks

Two-stream CNNs, is an alternative approach to model temporal information, combine regular images and optical flow images as input [122]. The rationale is that still images encode the pose of an interaction, while the optical flow provides information about the motion. Their proposed 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, which includes averaging and a linear-SVM. The different information fusion methods for each stream were explored by [94]. Wang et al. [148] 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, or clips, Diba et al. [28] has also proposed a representation and encoding of the sequence features in a Temporal Linear Encoding (TLE) layer, after the convolution feature extraction is performed, which is based on the aggregation of the appearance features from each of the individual temporal fragments.

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 and focus on specific parts on the input in either stream. One way of establishing these links is by using skip connections of Residual Networks [48, 46] 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. [32]. 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.

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) [40], that classify each region with a category-specific linear SVM. Notably, Peng and Schmid [97] demonstrated a multi-regional two-stream R-CNN which used a region-of-interest fusion layer
for both appearance and motion models. The use of region-focused, stream based models has also been used by [137], who introduce cross connections from the temporal to the spatial stream that include convolutions reducing the dimensionality of the temporal activation maps. The hierarchical model for features has also been used for the creation of action tubes [43]: spatio-temporal volumes centered on the performance of a particular action. 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. [85]. Saha et al. [112], Hou et al. [54] 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 ConvNet models have been created by Gkioxari et al. [42], Mettes and Snoek [86] 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. [150] used a two-stream semantic region based CNN (SR-CNNs) as an extension of Faster R-CNNs [105]. 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 the motion of a body part [124, 87, 142, 153]. Typically, the regions complement each other, which provide the efficient foreground extraction and localization of the per-frame motion.

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. [5] and Ji et al. [58] 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. [138] presented the C3D architecture and demonstrated it’s superiority over 2D ConvNets. 3D convolutions can also be used concurrently with a two-stream network. Carreira and Zisserman [15] have introduced a fusion of these two methodologies, two-Stream inflated 3D ConvNets (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 about motion patterns in optical flow streams.

4.3. Recurrent networks

While ConvNets 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 deal with variations in the temporal performance of an action or interaction. An alternative approach is to use recurrent neural networks (RNNs) that explicitly model temporal patterns. The key idea here is that there is 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 ConvNets 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 [6, 27]. Zhao et al. [170] proposed an approach based on the normalization of each layer of the network with batch normalization [57]. The created architecture is combined with a 3-dimensional ConvNet by using a two-stream fusion of the RNN and ConvNet, with an SVM. The use of multiple recurrent networks has also been scaled to include tree structures (RNN-T) [76], 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 [53] 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 [23]. 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. [29], Li et al. [78], Varol et al. [144] have shown that the combination of convolutions and long-term recursions performs well for recognition tasks in videos. Donahue et al. [29] was effective in both image and video descrip-
tion by directly connecting powerful feature extractors such as CNNs with recurrent models. Similarly, Baccouche et al. [5] 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 3-dimensional convolutions for the 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 both from short-scale and long-term temporal information.

Besides LSTMs, Highway Networks are an alternative solution to the vanishing gradient problem [128]. These networks allow 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. [173] has extended this approach to include the spatial dimensionality in the information highways inside recurrent transitions.

Because the discriminative information of an interaction is typically only found in selective parts of the input, several approaches have addressed the method for collecting information. In line with the multi-stream approaches (Section 4.2), Wang et al. [149] 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. [41], who focused on human-object interactions instead. Multiple recurrent modules can be used to understand human interactions, as shown by Yan et al. [155] in which the model is built from three attention-specific LSTMs that use information from each of the two interacting actors and the overall scene of each example. Ibrahim et al. [55] presented a two-stage temporal model in which LSTMs are used to analyze each person in the scene while their combined outputs synthesize the relationships that occur in the data between them. Srivastava et al. [127] created an Encoder-Decoder architecture, in which the encoder LSTM maps the 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 selecting the features that are most important to the model. This is achieved by minimizing the loss of the constructed sequence from the decoder LSTM and the actual input sequence. This way, the most descriptive features of the sequence are learned iteratively. 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 of these works is mainly on modeling the temporal dynamics. Often, pose information can be regarded as a complementary input. For example, Gammulle et al. [35] 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 [30] to learn pose-related features in each time step. This permit the robust understanding of the action from the collection of the per-frame human poses. Moreover, based on alternatives to LSTMs, Liu et al. [80] 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, forming a “trust gate”. They implement the idea of assimilating the sequential input of videos by adjusting

![Figure 4: Video classification networks: (i) 3D-convolution [58], (ii) 2D-Convolutional LSTM over a sequence of frames [29], (iii) 3D LSTM [5], (iv) slow-fusion [62], (v) two/multi-stream CNN [122, 150, 87, 142] and (vi) two-stream 3D-Conv network [15].](image-url)
the effects on the context-based information stored in the network by allowing to analyze the data at each step and make a decision on when to update, remember of forget the contents in the memory cell with a tree-like representation of the persons skeleton. Skeletal data have also been used by Zhu et al. [171] in a fully connected LSTM model including internal gates, outputs and neurons that could be dropped by the network. Other suggestions and extensions include the Lattice-LSTM (L₂STM) 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 the CNN classifier [130]. 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 [31, 117]. Because these models break down the interaction in meaningful blocks of motion, they can be used as the basis to learn a repertoire for action and interaction. This further decouples the visual input from the motion dynamics. As such, this approach can further reduce training requirements 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. While initial steps were made based on specially recorded benchmarks datasets, we can analyze 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 LSTM [1], has allowed researchers to focus more on the dynamics of interactions. Still, the promise of understanding social interactions directly from video has not been met. Below, we discuss limitations of the state-of-the-art and highlight current trends and future directions.

Computation requirements. But these advances 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 (e.g., [64]), it will be substantially hard to harvest such datasets. Some works have exploited the use of synthetic data generators to increase the amount and variation of the training data [18, 120]. Another line of approach is to use transfer learning [151], 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. This introduces technical issues. Most notably, as shown by Goodfellow et al. [44], Szegedy et al. [132], Su et al. [129] convolutions are susceptible to adversarial noise, as they can misclassify examples that look similar to humans but have slightly different pixel values.

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 with humans having little difficulty identifying an awkward hug from a heartfelt one. With an increased focus on realistic human interactions comes a need to be able to distinguish 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 suboptimal. 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 beneficial for the generalization. The work on hierarchies (e.g., [34]) is promising because it facilitates the focus on distinctive patterns at different levels of granularity, dependent on the type of interaction.

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. [71], Sefidgar et al. [113]) or the action of a body part (e.g., Chéron et al. [20], Kong et al. [67], Tian et al. [134]). 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 for distinguishing between subtly different classes [38]. 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. [51], Sabour et al. [111] 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 [45], which limits the computational requirements.

Role of skeleton data. We further 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. [14], Insafutdinov et al. [56], Yang et al. [156]) have paved the way for effective pose-based attention mechanisms. While the computational requirements of the pose estimation task are significant, the benefits for the recognition of interactions has also been demonstrated [30, 80].

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 sequences 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. 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., [38, 135]) but usually in a two-step approach by first detection humans (e.g., [95]) 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 taste. Often we are looking for anomalies, deviations from common practice. For example, Sequences 1 and 2 in Figure 5 show interactions that are difficult to recognize but are more likely to be of interest to a user. Commonly, it is the context of the behavior that is more descriptive, or 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 (see also Sequence 3 in Figure 5).

With the solid state-of-the-art performance and the promising directions of research to deal with the current limitations, the anticipated promise of the automated understanding of human interaction is coming within reach.

6. Acknowledgments

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

[1] 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, 2016. CVPR 2016, (IEEE). pp. 961–971.
[2] Anderson, D.J., Perona, P., 2014. Toward a science of computational ethology. Neuron 84, 18–31.
[3] Aran, O., Gatica-Perez, D., 2013. One of a kind: Inferring personality impressions in meetings, in: International Conference on Multimodal Interaction, 2013. ICMI 2013, (ACM). pp. 11–18.
[4] 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, 2017. FG 2017., (IEEE). pp. 476–483.
[5] 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, 2011. HBU 2011, (Springer). pp. 29–39.
Yin, Y., Yang, G., Xu, J., Man, H., 2012. Small group human activity recognition, in: International Conference on Image Processing, 2012. ICOP 2012, (IEEE). pp. 2709–2712.

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

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

Yu, G., Yuan, J., Liu, Z., 2012. Propagative hough voting for human activity recognition, in: European Conference on Computer Vision, 2012. ECCV 2012, (Springer). 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, 2015. CVPR 2015, (IEEE). 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, 2012. CVPRW 2012, (IEEE). pp. 28–35.

Zhong, B., De Natale, F.G., Conci, N., 2013. Recognition of social interactions based on feature selection from visual codebooks, in: Conference on Image Processing, 2013. ICIP 2013, (IEEE). 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, 2012. ECCV 2012, (Springer). 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.

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, 2016. AAAI 2016. p. 8.

Zaheer, M., Germain, M., Huang, M., Dyer, C., Ganguli, S., LeCun, Y., 2016. Deep sets. arXiv preprint arXiv:1605.07682.

Zilly, J.G., Srivastava, R.K., Kourtis,7k, J., Schmidhuber, J., 2017. Recurrent highway networks, in: International Conference on Machine Learning, 2017. ICML 2017., pp. 4189–4198.