International Journal of Computer Vision (2022) 130:1366–1401
https://doi.org/10.1007/s11263-022-01594-9

Human Action Recognition and Prediction: A Survey

Yu Kong[1] · Yun Fu[2]

Received: 21 February 2019 / Accepted: 13 January 2022 / Published online: 28 March 2022

Abstract
Derived from rapid advances in computer vision and machine learning, video analysis tasks have been moving from inferring the present state to predicting the future state. Vision-based action recognition and prediction from videos are such tasks, where action recognition is to infer human actions (present state) based upon complete action executions, and action prediction to predict human actions (future state) based upon incomplete action executions. These two tasks have become particularly prevalent topics recently because of their explosively emerging real-world applications, such as visual surveillance, autonomous driving vehicle, entertainment, and video retrieval, etc. Many attempts have been devoted in the last a few decades in order to build a robust and effective framework for action recognition and prediction. In this paper, we survey the complete state-of-the-art techniques in action recognition and prediction. Existing models, popular algorithms, technical difficulties, popular action databases, evaluation protocols, and promising future directions are also provided with systematic discussions.

Keywords Action recognition · Action prediction · Video data · Survey

1 Introduction

Every human action, no matter how trivial, is done for some purpose. For example, in order to complete a physical exercise, a patient is interacting with and responding to the environment using his/her hands, arms, legs, torsos, bodies, etc. An action like this denotes everything that can be observed, either with bare eyes or measured by visual sensors. Through the human vision system, we can understand the action and the purpose of the actor. We can easily know that a person is exercising, and we could guess with a certain confidence that the person's action complies with the instruction or not. However, it is way too expensive to use human labors to monitor human actions in a variety of real-world scenarios, such as smart rehabilitation and visual surveillance. Can a machine perform the same as a human?

One of the ultimate goals of artificial intelligence research is to build a machine that can accurately understand humans' actions and intentions, so that it can better serve us. Imagine that a patient is undergoing a rehabilitation exercise at home, and his/her robot assistant is capable of recognizing the patient's actions, analyzing the correctness of the exercise, and preventing the patient from further injuries. Such an intelligent machine would be greatly beneficial as it saves the trips to visit the therapist, reduces the medical cost, and makes remote exercise into reality. Other important applications including visual surveillance, entertainment, and video retrieval also need to analyze human actions in videos. In the center of these applications is the computational algorithms that can understand human actions. Similar to the human vision system, the algorithms ought to produce a label after observing the entire or part of a human action execution (Bobick & Davis, 2001; Ryoo, 2011). Building such algorithms is typically addressed in computer vision research, which studies how to make computers gain high-level understanding from digital images and videos.

The term *human action* studied in computer vision research ranges from the simple limb movement to joint complex movement of multiple limbs and the human body. This process is dynamic, and thus is usually conveyed in

Communicated by Boxin Shi.

✉ Yu Kong
yu.kong@rit.edu

Yun Fu
yunfu@ece.neu.edu

[1] B. Thomas Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA

[2] Department of ECE and College of CIS, Northeastern University, Boston, MA, USA

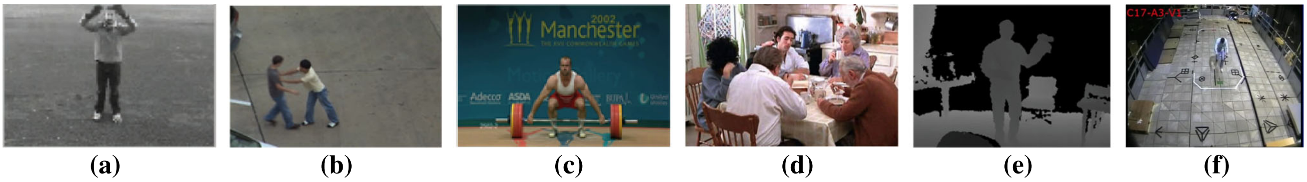
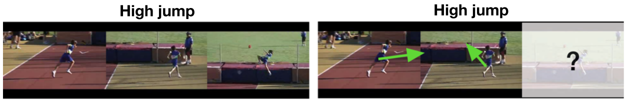


Fig. 1 Example frames of action videos used in computer vision research. **a** single person's action; **b** human interaction; **c** human-object interaction; **d** group action; **e** RGB-D action; **f** multi-view action



(a) Action recognition **(b)**Action prediction

Fig. 2 **a** Action recognition task infers an action category from a video containing complete action execution, while **b** action prediction task infers a label from temporally incomplete video. The label could be an action category (early action classification), or a motion trajectory (trajectory prediction)

a video lasting a few seconds. Though it might be difficult to give a formal definition of human action studied in the computer vision community, we provide some examples used in the community. Typical example actions are, (1) an individual action in KTH dataset (Schüldt et al., 2004) (Fig. 1(a)), which contains simple daily actions such as "clapping" and "running"; (2) a human interaction in UT-Interaction dataset (Ryoo & Aggarwal, 2009) (Fig. 1(b)), which consists of human interactions including "handshake" and "push"; (3) a human-object interaction in UCF Sports dataset (Rodriguez et al., 2008) (Fig. 1(c)), which comprises of sport actions and human-object interactions; (4) a group action in Hollywood 2 dataset (Marszałek et al., 2009) (Fig. 1(d)); (5) an action captured by a RGB-D sensor in UTKinect dataset (Xia et al., 2012) (Fig. 1(e)); and (6) a multi-view action in Multicamera dataset (Singh et al., 2010) (Fig. 1(f)) capturing human actions from multiple camera views. In all these examples, a human action attempts to achieve a certain goal, in which some of them can be achieved by simply moving arms, and the others need to be accomplished in several steps.

Technology advances in computer science and engineering have been enabling machines to understand human actions in videos. There are two basic topics in the computer vision community, vision-based human action recognition and prediction:

1. **Action recognition** recognize a human action from a video containing complete action execution.
2. **Action prediction** reason a human action from temporally incomplete video data.

Action recognition is a fundamental task in the computer vision community that recognizes human actions based on the complete action execution in a video (see Fig. 2(a)) (Bobick & Davis, 2001; Efros et al., 2003; Weinland et al., 2006; Laptev, 2005; Liu et al., 2009; Tang et al., 2012a; Tran et al., 2015). It has been studied for decades and is still a very popular topic due to broad real-world applications including video retrieval (Ciptadi et al., 2014), visual surveillance (Hu et al., 2007; Singh et al., 2010), etc. Researchers have made great efforts to create an intelligent system mimicking humans' capability that can recognize complex human actions in cluttered environments. However, to a machine, an action in a video is just an array of pixels. The machine has no idea about how to convert these pixels into an effective representation, and how to infer human actions from the representation. These two problems are considered as *action representation* and *action classification* in action recognition, and many attempts (Laptev, 2005; Raptis & Sigal, 2013; Ji et al., 2013; Carreira & Zisserman, 2017) have been proposed to address these two problems.

On the contrary, *action prediction* is a before-the-fact video understanding task and is focusing on the future state. In some real-world scenarios (*e.g.*, vehicle accidents and criminal activities), intelligent machines do not have the luxury of waiting for the entire action execution before having to react to the action contained in it. For example, being able to predict a dangerous driving situation before it occurs; opposed to recognizing it thereafter. This is referred to as the action prediction task where approaches that can recognize and infer a label from a temporally incomplete video (see Fig. 2(b)) (Ryoo, 2011; Kong et al., 2014b, 2017), different to action recognition approaches that expect to see the entire set of action dynamics extracted from a full video.

The major difference between action recognition and action prediction lies in *when to make a decision*. Human action recognition is to infer the action label *after* the entire action execution has been observed. This task is generally useful in non-urgent scenarios, such as video retrieval, entertainment, etc. Nevertheless, action prediction is to infer *before* fully observing the entire execution, which is of particular important in certain scenarios. For example, it would be very helpful if an intelligent system on a vehicle can predict a traffic accident before it happens; opposed to recognizing the dangerous accident event thereafter.

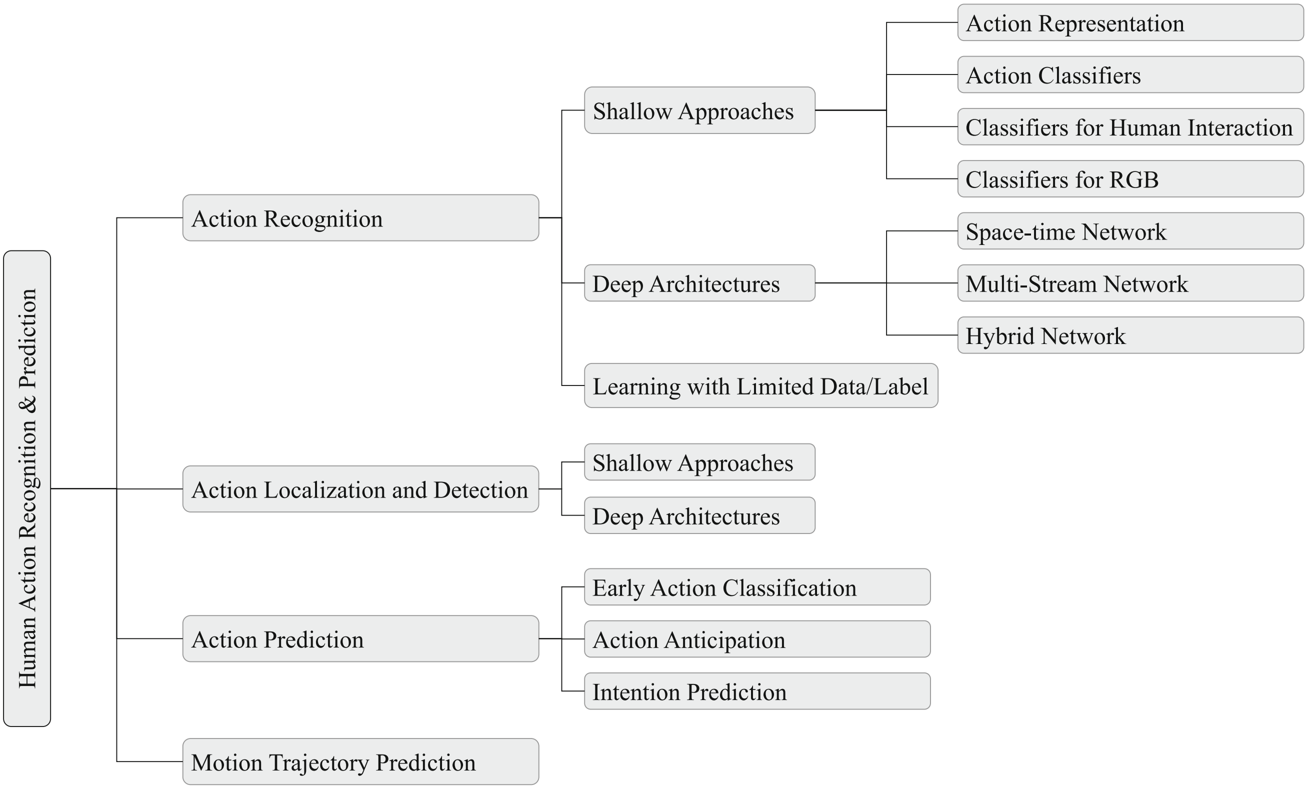


Fig. 3 Framework of the survey. The picture presents the topics discussed in the survey organized in a hierarchical tree, a list of representative works are also included for each topic

We will mainly discuss recent advance in action recognition and prediction in this survey. To ease the navigation of this paper, Fig. 3 illustrates the topics discussed in this paper and the representative works are also included. Different from recent survey papers (Herath et al., 2017; Poppe, 2010), studies in action prediction are also described in this paper. Human action recognition and prediction are closely related to other computer vision tasks such as human gesture analysis, gait recognition, and event recognition. In this survey, we focus on the vision-based recognition and prediction of actions from videos that usually involve one or more people. The input is a series of video frames and the output is an action label. We are also interested in learning human actions from RGB-D videos. Some of existing studies (Yao & Fei-Fei, 2012b, a) aim at learning actions from static images, which is not the focus of this paper. This paper will first give an overview of recent studies in action recognition and prediction, describe popular human actions datasets, and will then discuss several interesting future directions in details.

1.1 Real-World Applications

Action recognition and prediction algorithms empower many real-world applications (examples are shown in Fig. 4). State-of-the-art algorithms (Wang et al., 2016; Feichtenhofer et al., 2017; Kong et al., 2018; Ma et al., 2016) remarkably reduce the human labor in analyzing a large-scale of video data and provide understanding on the current state and future state of ongoing video data.

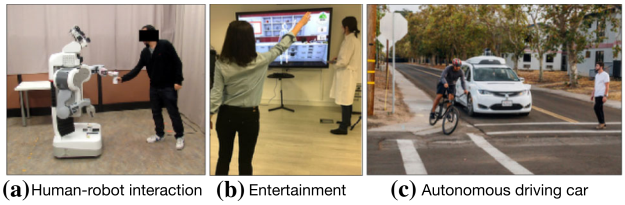

(**a**) Human-robot interaction (**b**) Entertainment (**c**) Autonomous driving car

Fig. 4 Examples of real-world applications using action recognition techniques

1.1.1 Visual Surveillance

Security issue is becoming more important in our daily life, and it is one of the most frequently discussed topics nowadays. Places under surveillance typically allow certain human actions, and other actions are not allowed (Hu et al., 2007). With the input of a network of cameras (Weinland et al., 2006; Singh et al., 2010), a visual surveillance system powered by action recognition (Ji et al., 2013; Simonyan & Zisserman, 2014; Karpathy et al., 2014) and prediction (Ryoo, 2011; Kong et al., 2014b, 2017) algorithms may increase the chances of capturing a criminal on video, and reduce the risk caused by criminal actions. For example, in Boston marathon bombing site, if we had such an intelligent visual surveillance system that can forewarn the public by looking at the criminal's suspicious action, the victims' lives could be saved. The cameras also make some people feel more secure, knowing the criminals are being watched.

1.1.2 Video Retrieval

Nowadays, due to the fast growth of technology, people can easily upload and share videos on the Internet. However, managing and retrieving videos according to video content is becoming a tremendous challenge as most search engines use the associated text data to manage video data (Ramezani & Yaghmaee, 2016). The text data, such as tags, titles, descriptions, and keywords, can be incorrect, obscure, and irrelevant, making video retrieval unsuccessful (Zhai et al., 2013). An alternative method is to analyze human actions in videos, as the majority of these videos contain such a cue. For example, in Ciptadi et al. (2014), researchers created a video retrieval framework by computing the similarity between action representations, and used the proposed framework to retrieve videos of children with autism in a classroom setting. Compared to conventional human action recognition task, the video retrieval task relies on the retrieval ranking instead of classification (Ramezani & Yaghmaee, 2016).

1.1.3 Entertainment

The gaming industry in recent years has attracted an increasingly large and diverse group of people. A new generation of games based on full body play such as dance and sports games have increased the appeal of gaming to family members of all ages. To enable accurate perception of human actions, these games use cost-effective RGB-D sensors (*e.g.*, Kinect Shotton et al., 2013) which provide an additional depth channel data (Xia & Aggarwal, 2013; Yang & Tian, 2014; Hadfield & Bowden, 2013). This depth data encode rich structural information of the entire scene, and facilitate action recognition task as it simplifies intra-class motion variations and reduces cluttered background noise (Kong & Fu, 2015, 2017; Jia et al., 2014; Liu & Shao, 2013).

1.1.4 Human-Robot Interaction

Human-robot interaction is popularly applied in home and industry environment. Imagine that a person is interacting with a robot and asking it to perform certain tasks, such as "passing a cup of water" or "performing an assembling task". Such an interaction requires communications between robots and humans, and visual communication is one of the most efficient ways (Ryoo et al., 2015; Koppula & Saxena, 2016).

1.1.5 Autonomous Driving Vehicle

Action prediction algorithms (Ryoo & Aggarwal, 2011; Kong & Fu, 2016) could be one of the potentials and maybe most important building components in an autonomous driving vehicle. Action prediction algorithms can predict a person's intention (Pei et al., 2011; Li & Fu, 2014; Koppula & Saxena, 2016) in a short period of time. In an urgent situation, a vehicle equipped with an action prediction algorithm can predict a pedestrian's future action or motion trajectory in the next few seconds, and this could be critical to avoid a collision. By analyzing human body motion characteristics at an early stage of an action using so-called interest points or convolutional neural network (Kong et al., 2017), action prediction algorithms (Kong et al., 2017; Kong & Fu, 2016) can understand the possible actions by analyzing the action evolution without the need to observe the entire action execution.

1.2 Research Challenges

Despite significant progress has been made in human action recognition and prediction, state-of-the-art algorithms still misclassify actions due to several major challenges in these tasks.

1.2.1 Intra- and Inter-Class Variations

As we all know, people behave differently for the same actions. For a given semantic meaningful action, for example, "running", a person can run fast, slow, or even jump and run. That is to say, one action category may contain multiple different styles of human movements. In addition, videos in the same action can be captured from various viewpoints. They can be taken in front of the human subject, on the side of the subject, or even on top of the subject, showing appearance variations in different views (see Fig. 5). Furthermore, different people may show different poses in executing the same action. All these factors will result in large intra-class appearance and pose variations, which confuse a lot of existing action recognition algorithms. These variations will be even

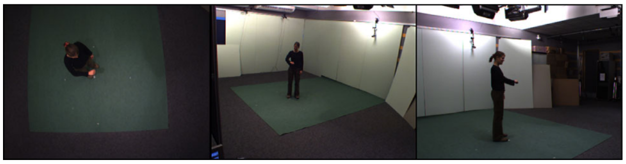


Fig. 5 Appearance variations in different camera views

larger on real-world action datasets (Karpathy et al., 2014; Caba Heilbron et al.., 2015). This triggers the investigation of more advanced action recognition algorithms that can be deployed in real-world scenarios. Furthermore, similarities exist in different action categories. For instance, "running" and "walking" involve similar human motion patterns. These similarities would also be challenging to differentiate for intelligent machines, and consequently contribute to misclassifications.

1.2.2 Cluttered Background and Camera Motion

It is interesting to see that a number of human action recognition algorithms work very well in indoor controlled environments but not in outdoor uncontrolled environments. This is mainly due to the background noise. In fact, most of the existing activity features such as histograms of oriented gradient (Laptev et al., 2008a) and interest points (Dollar et al., 2005) also encode background noise, and thus degrade the recognition performance. Camera motion is another factor that should be considered in real-world applications. Due to significant camera motion, action features cannot be accurately extracted. In order to better extract action features, camera motion should be modeled and compensated (Wang & Schmid, 2013). Other environment-related issues such as illumination conditions, viewpoint changes, dynamic background will also be the challenges that prohibit action recognition algorithms from being used in practical scenarios.

1.2.3 Insufficient Annotated Data

Even though existing action recognition approaches (Klaser et al., 2008; Liu et al., 2011; Niebles et al., 2010) have shown impressive performance on small-scale datasets in laboratory settings, it is really challenging to generalize them to real-world applications due to their inability of training on large-scale datasets. Recent deep approaches (Wang et al., 2016; Feichtenhofer et al., 2017) have shown promising results on datasets captured in uncontrolled settings, but they normally require a large amount of annotated training data. Action datasets such as HMDB51 (Kuehne et al., 2011) and UCF-101 (Khurram Soomro & Shah, 2012) contain thousands of videos, but still far from enough for training deep networks with millions of parameters. Although Youtube-8M (Abu-El-Haija et al., 2016) and Sposrts-1M datasets (Karpathy et al., 2014) provide millions of action videos, their annotations are generated by a retrieval method, and thus may not be accurate. Training on such datasets would hurt the performance of action recognition algorithms that do not have a tolerance to inaccurate labels. However, it is possible that some of the data annotations are available, which would result in a training setting with a mixture of labeled data and unlabeled data. Therefore, it is imperative to design action recognition algorithms that can learn actions from both labeled data and unlabeled data.

1.2.4 Action Vocabulary

Actions could be categorized into different levels, movements, atomic actions, composite actions, events, etc. This defines an action hierarchy, and complex actions at high levels of the hierarchy can be decomposed into a combination of actions at a lower level. How to define and analyze these different kinds of actions is very important.

1.2.5 Uneven Predictability

Not all frames are equally discriminative. As shown in Raptis and Sigal (2013), Vahdat et al. (2011), a video can be effectively represented by a small set of key frames. This indicates that lots of frames are redundant, and discriminative frames may appear anywhere in the video. However, action prediction methods (Ryoo, 2011; Kong et al., 2014b; Ma et al., 2016; Lan et al., 2014) require the beginning portions of the video to be discriminative in order to maximize predictability. To solve this problem, context information is transferred to the beginning portions of the videos (Kong et al., 2017), but the performance is still limited due to the insufficient discriminative information.

In addition, actions differ in their predictabilities (Li & Fu, 2014; Kong et al., 2017). As shown in Kong et al. (2017), some actions are instantly predictable while the other ones need more frames to be observed. However, in practical scenarios, it is necessary to predict any actions as early as possible. This requires us to create general action prediction algorithms that can make accurate and early predictions for most of or all actions.

2 Human Perception of Actions

Human actions, particularly those involving whole-body and limb (*e.g.*, arms and legs) movements, and interactions with their environment contain rich information about the performer's intention, goal, mental status, etc. Understanding the actions and intentions of other people is one of the most important social skills we have, and the human vision system provides a particularly rich source of information in support of this skill (Blake & Shiffrar, 2007). Compared to static images, human actions in videos provide even more reliable and more expressive information, and thus speak louder than images when it comes to understanding what others are doing (Darwin, 1872). There are a number of information we can tell from human actions, including the action categories (Mass et al., 1971), emotional implication (Clarke et al., 2005), identity (Cutting & Kozlowski, 1977; Troje et al., 2005), gender (Sumi, 2000; Troje, 2002), etc. The human visual system is finely optimized for the perception of human movements (Decety & Grezes, 1999).

Action understanding by humans is a complex cognitive capability performed by a complex cognitive mechanism. Such a mechanism can be decomposed into three major components, including action recognition, intention understanding, and narrative understanding (Keestra, 2015). Ricoeur (1992) suggested that actions can be approached with a set of interrelated questions including, who, what, why, how, where, and when. Three questions are prioritized, which offer different perspectives on the action: what is the action, why is the action being done, and who is the agent. Computational models for the first two questions have been extensively investigated in action recognition (Blank et al., 2005; Patron-Perez et al., 2010; Choi et al., 2009; Marszałek et al., 2009; Kuehne et al., 2011; Ji et al., 2013; Tran et al., 2015) and prediction (Ryoo, 2011; Kong et al., 2014b; Ma et al., 2016; Cao et al., 2013) research in the computer vision community. The last question "who is the agent" refers to the agent's identity, or social role, which provides a more thoroughgoing understanding of the "who" behind it, and thus is referred to as narrative understanding (Ricoeur, 1992). Few work in the computer vision community studies this question (Lan et al., 2012; Ramanathan et al., 2013).

Some of the human actions are goal-oriented, i.e., a goal is completed by performing one or a series of actions. Understanding such actions is crucial for predicting the effects or outcomes of the actions. As humans, we make inferences about the action goals of an individual by evaluating the end state that would be caused by their actions, given particular situational or environmental constraints. The inference is possibly made by a direct matching process of a mirror neuron system, which maps the observed action onto our own motor representation of that action (Rizzolatti & Craighero, 2004; Rizzolatti & Sinigaglia, 2010). According to the direct matching hypothesis, the prediction of one's action goal is heavily relying on the observer's action vocabulary or knowledge. Another cue for making action prediction is from emotional or attentional information, such as the facial expression and gaze or the other individuals. Such referential information makes the observer pay attention to the specific objects because of the particular relations that link these cues to their referents. These psychological and cognitive findings would be helpful for designing action prediction approaches.

3 Action Recognition

A typical action recognition flowchart generally contains two major components (Schüldt et al., 2004; Wang et al., 2013; Poppe, 2010), action representation and action classification. The action representation component basically converts an action video into a feature vector (Laptev, 2005; Dollar et al., 2005; Wang et al., 2015; Scovanner et al., 2007) or a series of vectors (Niebles et al., 2010; Kong et al., 2017; Morency et al., 2007), and the action classification component infers an action label from the vector (Liu et al., 2011; Sminchisescu et al., 2005; Shi et al., 2011). Recently, deep networks (Ji et al., 2013; Tran et al., 2015; Feichtenhofer et al., 2017) merge these two components into a unified end-to-end trainable framework, which further enhance the classification performance in general. In this section we will discuss recent work in action representation, action classification, and deep networks.

3.1 Shallow Approaches

3.1.1 Action Representation

The first and the foremost important problem in action recognition is *how to represent an action in a video*. Human actions appearing in videos differ in their motion speed, camera view, appearance and pose variations, etc, making action representation a really challenging problem. A successful action representation method should be efficient to compute, effective to characterize actions, and can maximize the discrepancy between actions, in order to minimize the classification error.

One of the major challenges in action recognition is large appearance and pose variations in one action category, making the recognition task difficult. The goal of action representation is to convert an action video into a feature vector, extract representative and discriminative information of human actions, and minimize the variations, thereby improving the recognition performance. Action representation approaches can be roughly categorized into holistic features and local features, which will be discussed next.

Many attempts have been made in action recognition to convert action videos into discriminative and representative features, in order to minimize with-in class variations and maximize between class variations. Here, we focus on *hand-crafted* action representation methods, which means the parameters in these methods are pre-defined by experts. This differs from deep networks, which can automatically learn parameters from data.

Holistic Representations Human action in a video generates a space-time shape in the 3D volume. This space-time shape encodes both spatial information of the human pose at various times, and dynamic information of the human body. Holistic representation methods capture the motion information of the entire human subject, providing rich and expressive motion information for action recognition. However, holis-

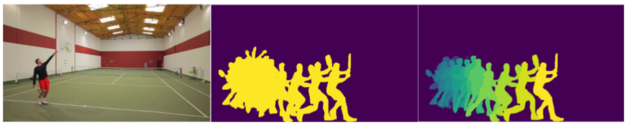


Fig. 6 Examples of an input video frame, the corresponding motion energy image and motion history image computed by Bobick and Davis (2001)

tic representations tend to be sensitive to noise. It captures the information in a certain rectangle region, and thus may introduce irrelevant information and noise from the human subject and cluttered background.

One pioneering work in Bobick and Davis (2001) presented Motion Energy Image (MEI) and Motion History Image (MHI) to encode dynamic human motion into a single image. As shown in Fig. 6, the two methods work on the silhouettes. The MEI method shows "where" the motion is occurring: the spatial distribution of motion is represented and the highlighted region suggests both the action occurring and the viewing condition. In addition to MEI, the MHI method shows both "where" and "how" the motion is occurring. Pixel intensity on a MHI is a function of the motion history at that location, where brighter values correspond to more recent motion.

Although MEI and MHI showed promising results in action recognition, they are sensitive to viewpoint changes. To address this problem, Weinland et al. (2006) generalized (Bobick & Davis, 2001) to 3D motion history volume (MHV) to remove the viewpoint dependency in the final action representation. MHV relies on the 3D voxels obtained from multiple camera views, and shows the 3D occupancy in the resulting volume. Fourier transform is then used to create features invariant to locations and rotations.

To capture space-time information in human actions, Gorelick et al. (2007), Blank et al. (2005) utilized the Poisson equation to extract various shape properties for action representation and classification. Their method takes a space-time volume as input. Then the method discovers space-time saliency of moving body parts, and locally computes the orientation using the Poisson equation. These local properties are finally converted into a global feature by weighted averaging each point inside the volume. Another method to describe shape and motion was presented in Yilmaz and Shah (2005). In this method, a spatio-temporal volume is first generated by computing correspondences between frames. Then, spatio-temporal features by analyzing differential geometric surface properties from the volume.

Instead of computing silhouette or shape for action representation, motion information can also be computed from videos. One typical motion information is computed by the so-called optical flow algorithms (Lucas & Kanade, 1981; Horn & Schunck, 1981; Sun et al., 2010), which indicate the pattern of apparent motion of objects on two consecutive frames. Under the assumption that illumination conditions do not change on the frames, optical flow computes the motion in the horizontal and vertical axis. An early work by Efros et al. (2003) split the flow field into four channels (see Fig. 7) capturing the horizontal and vertical motion in successive frames. This method was then used in Wang and Mori (2010) to describe the features of both the human body and the body parts.

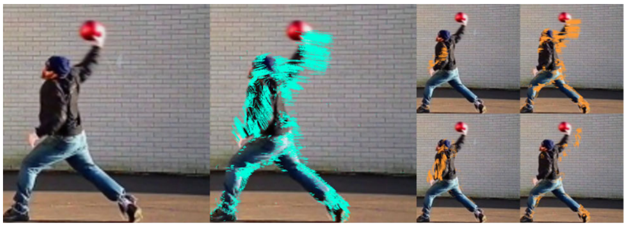


Fig. 7 Examples of the original frame, optical flow, and the flow field in four channels computed by Efros et al. (2003)

Local Representations Local representations only identify local regions having salient motion information, and thus inherently overcome the problem in holistic representations. Successful methods such as space-time interest points (Dollar et al., 2005; Laptev & Lindeberg, 2003; Klaser et al., 2008; Bregonzio et al., 2009) and motion trajectory (Wang et al., 2011, 2013) have shown their robustness to translation, appearance variation, etc. Different from holistic features, local features describe the local motion of a person in space-time regions. These regions are detected since the motion information within the regions is more informative and salient than the surrounding areas. After detection, the regions are described by extracting features in the regions.

Space-time interest points (STIPs) (Laptev & Lindeberg, 2003; Laptev, 2005)-based approaches is one of the most important local representations. Laptev's seminal work (Laptev & Lindeberg, 2003; Laptev, 2005) extended the Harris corner detector (Harris & Stephens, 1988) to space-time domain. A spatio-temporal separable Gaussian kernel is applied on a video to obtain its response function for finding large motion changes in both spatial and temporal dimensions (see Fig. 8). An alternative method was proposed in Dollar et al. (2005), which detects dense interest points. 2D Gaussian smoothing kernel is applied only along the spatial dimension, and the 1D Gabor filter is applied to the temporal dimension. Around each interest point, raw pixel values, gradient, and optical flow features are extracted and concatenated into a long vector. The principal component analysis is applied on the vector to reduce the dimensionality, and a k-means clustering algorithm is then employed to create the codebook of these feature vectors and generate one vector representation for a video (Schüldt et al., 2004). Bregonzio et al. (2009) detected spatial-temporal interest

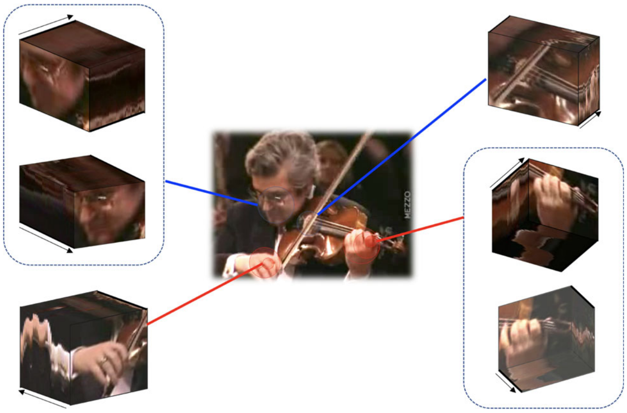


Fig. 8 Illustration of interest points detected on human body. Revised based on the original figure in Herath et al. (2017)

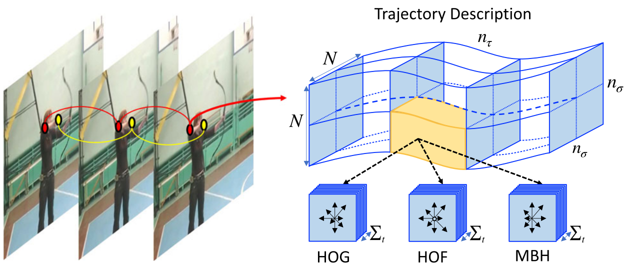


Fig. 9 Tracked point trajectories over frames, and are described by HOG, HOF and MBH features. Revised based on the original figure in Wang et al. (2013)

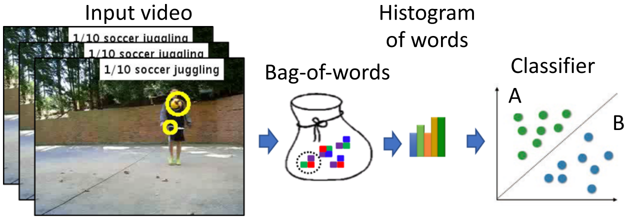


Fig. 10 A typical flowchart of the so-called bag-of-words methods. Local features detected on the input video are shown in yellow circles (Color figure online)

points using Gabor filters. Spatiotemporal interest points can also be detected by using the spatiotemporal Hessian matrix (Willems et al., 2008). Other detection algorithms detect spatiotemporal interest points by extending their counterparts of 2D detectors to spatiotemporal domains, such as 3D SIFT (Scovanner et al., 2007), HOG3D (Klaser et al., 2008), local trinary patterns (Yeffet & Wolf, 2009), etc. Several descriptors have been proposed to describe the motion and appearance information within the small region of the detected interest points such as optical flow and gradient. Optical flow feature computed in a local neighborhood is further aggregated in histograms, called histograms of optical flow (HOF) (Laptev et al., 2008a), and combined with HOG features (Dalal & Triggs, 2005; Klaser et al., 2008) to represent complex human activities (Klaser et al., 2008; Laptev et al., 2008a; Wang et al., 2009). Gradients over optical flow fields are computed to build the so-called motion boundary histograms (MBH) for describing trajectories (Wang et al., 2009).

However, spatiotemporal interest points only capture information within a short temporal duration and cannot capture long-term duration information. It would be better to track these interest points and describe their changes of motion properties. Feature trajectory is a straightforward way of capturing such long-duration information (Wang et al., 2009, 2011; Raptis & Soatto, 2010). To obtain features for trajectories, in Messing et al. (2009), interest points are first detected and tracked using Harris3D interest points with a KLT tracker (Lucas & Kanade, 1981). The method in Sun et al. (2009) finds trajectories by matching corresponding SIFT points over consecutive frames. Hierarchical context information is captured in this method to generate more accurate and robust trajectory representation. Trajectories are described by a concatenation of HOG, HOF and MBH features (Wang et al., 2011, 2013; Jain et al., 2013) (see Fig. 9), intra- and inter-trajectory descriptors (Sun et al., 2009), or HOG/HOF and averaged descriptors (Raptis & Soatto, 2010). In order to reduce the side effect of camera motion, Wang and Schmid (2013), Wang et al. (2015) find correspondences between two frames first and then use RANSAC to estimate the homography.

3.1.2 Action Classifiers

After action representations have been computed, action classifiers should be learned from training samples that determine the class boundaries for various action classes. Action classifiers can be roughly divided into the following categories:

Direct Classification This type of approaches summarize an action video into a feature vector, and then directly classify the vector into action categories using off-the-shelf classifiers such as support vector machine (Schüldt et al., 2004; Laptev et al., 2008a; Marszałek et al., 2009), k-nearest neighbor (k-NN) (Blank et al., 2005; Laptev & Perez, 2007; Tran & Sorokin, 2008), etc. In these methods, action dynamics are characterized in a holistic way using action shape (Gorelick et al., 2007; Blank et al., 2005), or using the so-called bag-of-words model, which captures local motion patterns using a histogram of visual words (Blank et al., 2005; Laptev & Perez, 2007; Schüldt et al., 2004; Laptev et al., 2008a; Marszałek et al., 2009).

In fact, bag-of-words approaches received lots of attention in the last few years. As shown in Fig. 10, these approaches first detect local salient regions using the spatiotemporal interest point detectors (Dollar et al., 2005; Schüldt et al., 2004; Laptev, 2005; Klaser et al., 2008). Features such as gradient and optical flow are extracted around each 3D interest point. The principal component analysis is adopted to reduce the dimensionality of the features. Then the so-called visual words can be computed by k-means clustering (Schüldt et al., 2004), or Fisher vector (Perronnin & Dance, 2006). Finally, an action can be represented by a histogram of visual words, and can be recognized by a classifier such as the support vector machine. The bag-of-words model has been shown to be insensitive to appearance and pose variations (Wang et al., 2009). However, it does not consider the temporal characteristics of human actions, as well as their structural information, which can be addressed by sequential approaches (Shi et al., 2011; Raptis & Sigal, 2013) and space-time approaches (Ryoo & Aggarwal, 2009), respectively.

Sequential Approaches This line of work captures temporal evolution of appearance or pose using sequential state models such as hidden Markov models (HMMs) (Duong et al., 2005; Rajko et al., 2007; Ikizler & Forsyth, 2007), conditional random fields (CRFs) (Sminchisescu et al., 2005; Wang et al., 2006; Wang & Suter, 2007; Morency et al., 2007) and structured support vector machine (SSVM) (Niebles et al., 2010; Wang et al., 2012; Tang et al., 2012a; Shi et al., 2011). These approaches treat a video as a composition of temporal segments or frames. The work in Duong et al. (2005) considers human routine trajectory in a room, and use a two-layer HMMs to model the trajectory. Recent work in Raptis and Sigal (2013) shows that representative key poses can be learned to better represent human actions. This method discards a number of non-informative poses in a temporal sequence, and builds a more compact pose sequence for classification. Nevertheless, these sequential approaches mainly use holistic features from frames, which are sensitive to background noise and generally do not perform well on challenging datasets.

Space-time Approaches Although direct approaches have shown promising results on some action datasets (Schüldt et al., 2004; Laptev et al., 2008a; Marszałek et al., 2009), they do not consider the spatiotemporal correlations between local features, and do not take the potentially valuable information about the global spatio-temporal distribution of interest points into account. This problem was addressed in Wu et al. (2011), which learns a global Gaussian mixture model (GMM) using the relative coordinates features, and uses multiple GMMs to describe the distribution of interest points over local regions at multiple scales. A global feature on top of interest points was proposed in Yuan et al. (2013) to capture the detailed geometrical distribution of interest points. The feature is computed by extended 3D discrete Radon transform. Such a feature captures the geometrical information of the interest points, and is robust to geometrical transformation and noise. The spatiotemporal distribution of interest points is described by a Directional Pyramid Co-occurrence Matrix in (DPCM) (Yuan et al., 2014). DPCM characterizes the co-occurrence statistics of local features as well as the spatio-temporal positional relationships among the concurrent features. Graph is a powerful tool for modeling structured objects, and it was used in Wu et al. (2014) to capture the spatial and temporal relationships among local features. Local features are used as the vertices of the two-graph model and the relationships among local features in the intra-frames and inter-frames are characterized by the edges. A novel family of context-dependent graph kernels (CGKs) was proposed in Wu et al. (2014) to measure the similarity between the two-graph models. Although the above methods have achieved promising results, they are limited to small datasets as the correlations between interest points in their methods which are explosive on large datasets.

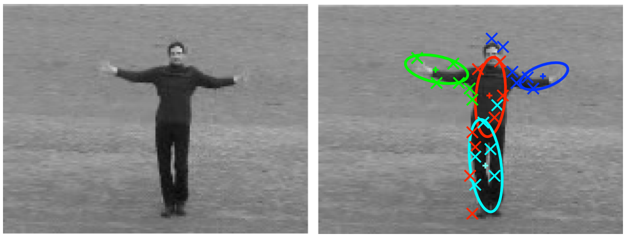


Fig. 11 Example of body parts detected by the constellation model in Niebles and Fei-Fei (2007). Revised based on the original figure in Niebles and Fei-Fei (2007)

Part-based Approaches Human bodies are structured objects, and thus it is straightforward to model human actions using motion information from body parts. Part-based approaches consider motion information from both the entire human body as well as body parts. The benefit of this line of approaches is it inherently captures the geometric relationships between body parts, which is an important cue for distinguishing human actions. A constellation model was proposed in Fanti et al. (2005), which models the position, appearance and velocity of body parts. Inspired by Fanti et al. (2005), a part-based hierarchical model was presented in Niebles and Fei-Fei (2007), in which a part is generated by the model hypothesis and local visual words are generated from a body part (see Fig. 11).

The method in Wong et al. (2007) considers local visual words as parts, and models the structure information between parts. This work was further extended in Niebles et al. (2008), where the authors assume an action is generated from a multinomial distribution, and then each visual word is generated from distribution conditioned on the action. These part-based generated models were further improved by discriminative models for better classification performance (Wang & Mori,

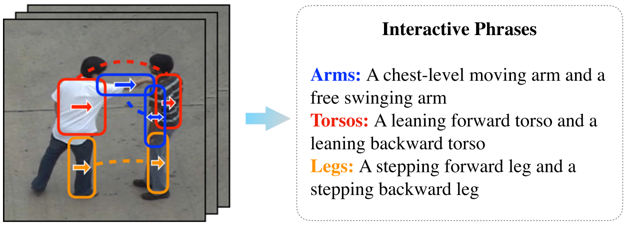

Interactive Phrases
Arms: A chest-level moving arm and a free swinging arm
Torsos: A leaning forward torso and a leaning backward torso
Legs: A stepping forward leg and a stepping backward leg

Fig. 12 Interaction recognition by learning semantic descriptions from videos. Revised based on the original figure in Kong et al. (2014a)

2008, 2010). In Wang and Mori (2008; 2010), a part is considered as a hidden variable in their models. It is corresponding to a salient region with the most positive energy.

Manifold Learning Approaches Human action videos can be described by temporally variational human silhouettes. However, the representation of these silhouettes is usually high-dimensional and prevents us from efficient action recognition. To solve this problem, manifold learning approaches were proposed in Wang and Suter (2007), Jia and Yeung (2008) to reduce the dimensionality of silhouette representation and embed them on nonlinear low-dimensional dynamic shape manifolds. The method in Wang and Suter (2007) adopts kernel PCA to perform dimensionality reduction, and discover the nonlinear structure of actions in the manifold. Then, a two-chain factorized CRF model is used to classify silhouette features in the low-dimensional space into human actions. A novel manifold embedding method was presented in Jia and Yeung (2008), which finds the optimal embedding that maximizes the principal angles between temporal subspaces associated with silhouettes of different classes. Although these methods tend to achieve very high performance in action recognition, they heavily rely on clean human silhouettes which could be difficult to obtain in real-world scenarios.

Mid-Level Feature Approaches Bag-of-words models have shown to be robust to background noise but may not be expressive enough to describe actions in the presence of large appearance and pose variations. In addition, they may not well represent actions due to the large semantic gap between low-level features and high-level actions. To address these two problems, hierarchical approaches (Wang & Mori, 2010; Choi et al., 2011; Liu et al., 2011; Kong et al., 2014a) are proposed to learn an additional layer of representations, and expect to better abstract the low-level features for classification.

Hierarchical approaches learn mid-level features from low-level features, which are then used in the recognition task. The learned mid-level features can be considered as knowledge discovered from the same database used for training or being specified by experts. Recently, semantic descriptions or attributes (see Fig.12) are popularly investigated in action recognition. These semantics are defined and further introduced into the activity classifiers in order to characterize complex human actions (Kong et al., 2012, 2014a; Liu et al., 2011). Other hierarchical approaches such as Raptis and Sigal (2013), Vahdat et al. (2011) select key poses from observed frames, which also learn better action representations during model learning. These approaches have shown superior results due to the use of human knowledge, but require extra annotations which is labor-intensive.

Feature Fusion Approaches Fusing multiple types of features from videos is a popular and effective way for action recognition. Since these features are generated from the same visual inputs, they are inter-related. However, the inter-relationship is complicated and is usually ignored in the existing fusion approaches. This problem was addressed in Luo et al. (2014), in which the maximum margin distance learning method is used to combine global temporal dynamics and local visual spatio-temporal appearance features for human action recognition. A Multi-Task Sparse Learning (MTSL) model was presented in Yuan et al. (2013) to fuse multiple features for action recognition. They assume multiple learning tasks share priors, one for each type of features, and exploit the correlations between tasks to better fuse multiple features. A multi-feature max-margin hierarchical Bayesian model (M3HBM) was proposed in Yang et al. (2015) to learn a high-level representation by combining a hierarchical generative model (HGM) and discriminative max-margin classifiers in a unified Bayesian framework. HGM represents actions by distributions over latent spatial temporal patterns (STPs) learned from multiple feature modalities. This work was further extended in Yuan et al. (2016) to combine spatial interest points with context-aware kernels for action recognition. Specifically, a video set is modeled as an optimized probabilistic hypergraph, and a robust context-aware kernel is used to measure high order relationships among videos.

3.1.3 Classifiers for Human Interactions

Human interaction is typical in daily life. Recognizing human interactions focuses on the actions performed by multiple people, such as “handshake”, “talking”, etc. Even though some of the early work such as Laptev et al. (2008a), Ryoo and Aggarwal (2009), Yu et al. (2010), Marszałek et al. (2009), Liu et al. (2009) used action videos containing human interactions, they recognize actions in the same way as single-person action recognition. Specifically, interactions are treated as a whole and are represented as a motion descriptor including all the people in a video. Then an action classifier such as a linear support vector machine is adopted to classify interactions. Despite reasonable performance has been achieved, these approaches do not explicitly consider the intrinsic methods of interactions, and fail to consider the co-occurrence information between interacting people. Furthermore, they do not extract the motion of each person from the group, and thus their methods can not infer the action label of each interacting person.

Action co-occurrence of individual person is a piece of valuable information in human interaction recognition. In Oliver et al. (2000), action co-occurrence is captured by coupling motion state of one person with the other interaction person. Human interactions such as “hug”, “push”, and “hi-five” usually involve frequent close physical contact, and thus some body parts may be occluded. To robustly find body parts, Ryoo and Aggarwal (2006) utilized body part tracker to extract each individual in videos and then applied context-free grammar to model spatial and temporal relationships between people. A human detector is adopted in Patron-Perez et al. (2012) to localize each individual. Spatial relationships between individuals are captured using the structured learning technique (Felzenszwalb et al., 2008). Spatiotemporal context of a group of people including human pose, velocity and spatiotemporal distribution of individuals is captured in Choi et al. (2011) to recognize human interactions. Their method shows promising results on collective actions without close physical contact such as “crossing the road”, “talking”, or “waiting”. They further extended their work that can simultaneously track and recognize human interactions (Choi & Savarese, 2012). A hierarchical representation of interactions is proposed in Choi and Savarese (2012) that models atomic action, interaction, and collective action. The method in Lan et al. (2012) also utilizes the idea of hierarchical representation, and studies the collective activity recognition problem using crowd context. Different from these methods, the work in Vahdat et al. (2011) represents individuals in interactions as a set of key poses, and models spatial and temporal relationships of the key poses for interaction recognition. In our earlier work (Kong et al., 2014a, 2012), a semantic description-based approach is proposed to represent complex human interactions by learned motion relationships (see Fig. 12). Instead of directly modeling action co-occurrence, we propose to learn phrases that describe the motion relationships between body parts. This will describe complex interactions in more details, and introduce human knowledge into the model. All these methods may not perform well in interactions with close physical contact due to the ambiguities in feature-to-person assignments. To address this problem, a patch-aware model was proposed in Kong and Fu (2014) to learn discriminative patches for interaction recognition, and determine the assignments at a patch level.

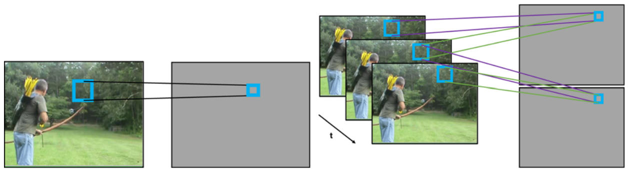
3.1.4 Classifiers for RGB-D Videos

Action recognition from RGB-D videos has been receiving a lot of attentions (Wang et al., 2012a, b; Hadfield & Bowden, 2013; Xia & Aggarwal, 2013; Liu & Shao, 2013; Oreifej & Liu, 2013) due to the advent of the cost-effective Kinect sensor (Shotton et al., 2013). RGB-D videos provide an additional depth channel compared with conventional RGB videos, allowing us to capture 3D structural information that is very useful in reducing background noise and simplifying intra-class motion variations (Ni et al., 2011; Wang et al., 2012; Oreifej & Liu, 2013; Hadfield & Bowden, 2013; Ofli et al., 2013).

Effective features have been proposed for the recognition task using depth data, such as histogram of oriented 4D normals (Oreifej & Liu, 2013; Yang & Tian, 2014) and depth spatiotemporal interest points (Xia & Aggarwal, 2013; Hadfield & Bowden, 2013). Features from depth sequences can be encoded by Luo et al. (2013), or be used to build actionlets (Wang et al., 2012) for recognition. An efficient binary range-sample feature for depth data was proposed in Lu et al. (2014). This binary depth feature is fast, and has shown to be invariant to changes in scale, viewpoint, and background. The work in Sung et al. (2012), Koppula and Saxena (2013b) built layered action graph structures to model actions and subactions in a RGB-D video. Recent work (Liu & Shao, 2013) also showed that features of RGB-D data can be learned using deep learning techniques.

The methods in Li et al. (2010, Oreifej and Liu (2013), Yang and Tian (2014), Hadfield and Bowden (2013), Wang et al. (2012a), Luo et al. (2013) only use depth data, and thus would fail if depth data were missing. Joint use of both RGB and depth data for action recognition is investigated in Hu et al. (2015), Jia et al. (2014), Lin et al. (2014), Liu and Shao (2013), Wang et al. (2012), Kong and Fu (2015). However, they only learn features shared between the two modalities and do not learn modality-specific or private features. To address this problem, shared features and privates features are jointly learned in Kong and Fu (2017), which learns extra discriminative information for classification, and demonstrate superior performance than Hu et al. (2015), Jia et al. (2014), Lin et al. (2014), Liu and Shao (2013), Wang et al. (2012), Kong and Fu (2015). The methods in Kong and Fu (2015; 2017) also show that they can achieve high recognition performance even though one modality is missing in training or testing.

Auxiliary information has also shown to be useful in RGB-D action recognition. Skeleton data provided by a Kinect sensor was used in Hu et al. (2015), Wang et al. (2012), Kong and Fu (2017), and has shown to be very effective in action recognition. The method in Hu et al. (2015) learns a shared feature space for various types of features including skeleton features and local HOG features, and project these features onto the shared space for action recognition. Different from this work, the method in Kong and Fu (2017) jointly learns RGB-D and skeleton features and action classifiers. The projection matrices in Kong and Fu (2017) are learned



(a) 2D convolution **(b)** 3D convolution

Fig. 13 Illustration of **a** 2D convolution and **b** 3D convolution

by minimizing the noise after projection and classification error using the projected features. Using auxiliary databases to improve the recognition performance was studied in Jia et al. (2014), Lin et al. (2014), in which actions are assumed to be reconstructed by entries in the auxiliary databases.

3.2 Deep Architectures

Although great success has been made by global and local features, these hand-crafted features require heavy human labor and domain expert knowledge to develop effective feature extraction methods. In addition, they normally do not generalize very well on large datasets. In recent years, feature learning using deep learning techniques has been receiving increasing attention due to their capability of learning powerful features that can be generalized very well (Ji et al., 2013; Tran et al., 2015; Donahue et al., 2015; Simonyan & Zisserman, 2014). The success of deep networks in action recognition can also be attributed to scaling up the networks to tens of millions of parameters and massive labeled datasets. Recent deep networks (Varol et al., 2017; Tran et al., 2015; Feichtenhofer et al., 2017; Kar et al., 2017) have achieved surprisingly high recognition performance on a variety of action datasets.

Action features learned by deep learning techniques has been popularly investigated (Yang & Shah, 2012; Wang et al., 2014a; Taylor et al., 2010; Sun et al., 2014; Plotz et al., 2011; Le et al., 2011; Karpathy et al., 2014; Ji et al., 2013, 2010; Hasan & Roy-Chowdhury, 2014; Bengio et al., 2013; Simonyan & Zisserman, 2014) in recent years. The two major variables in developing deep networks for action recognition are the convolution operation and temporal modeling, leading to a few lines of networks.

The convolution operation is one of the fundamental components in deep networks for action recognition, which aggregates pixel values in a small spatial (or spatiotemporal) neighborhood using a kernel matrix. **2D vs 3D Convolution** 2D convolution over images (Fig. 13(a)) is one of the basic operation in deep networks, and thus it is straightforward to use 2D convolution on video frames. The work in Karpathy et al. (2014) presented a single-frame architecture based on a 2D CNN model, and extracted a feature vector for each frame. Such a 2D convolution network (2D ConvNet) also enjoys the benefit of using the networks pre-trained on large-scale image datasets such as ImageNet. However, 2D ConvNets do not inherently model temporal information, and requires an additional aggregation or modeling of such information.

As multiple frames are presenting in videos, 3D convolution (Fig. 13(b)) is more intuitive to capture temporal dynamics in a short period of time. Using 3D convolution, 3D convolutional networks (3D ConvNets) directly create hierarchical representations of spatio-temporal data (Ji et al., 2010, 2013; Taylor et al., 2010; Tran et al., 2015). However, the issue is they have many more parameters than 2D ConvNets, making them hard to train. In addition, they are prevented from enjoying the benefits of ImageNet pre-training.

Another key variable in designing deep networks is Temporal Modeling. Generally, there are roughly three methods in temporal modeling. One straightforward way is to directly apply 3D convolution to several consecutive frames (Ji et al., 2010, 2013; Taylor et al., 2010; Tran et al., 2015; Carreira & Zisserman, 2017). As a result, the temporal dimension in the 3D convolution kernel will capture the temporal dynamics in these frames. One of the limitations of these approaches is they may not be able to reuse the 2D ConvNets pre-trained on large-scale image datasets. Another line of approaches model temporal dynamics by using multiple streams (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2016; Carreira & Zisserman, 2017; Girdhar et al., 2017; Kar et al., 2017). A stream named flow net in the networks trains on optical flow frames, which essentially capture motion information in the adjacent two frames. However, these approaches largely disregard the long-term temporal structure of videos. 2D convolution is usually used in these approaches, and thus they can easily exploit the new ultra-deep architectures and models pre-trained for still images. The third category of approaches uses temporal pooling (Kar et al., 2017; Girdhar et al., 2017) or aggregation to capture temporal information in a video. The aggregation can be performed by using a LSTM model on top of 2D ConvNets (Donahue et al., 2015; Ng et al., 2015).

3.2.1 Space-Time Networks

Space-time networks are straightforward extensions of 2D ConvNets as they capture temporal information using 3D convolutions.

The method in Ji et al. (2010) was one of the pioneering works in using convolution neural networks (CNN) for action recognition. They perform 3D convolutions over adjacent frames, and thus extract features from both spatial and temporal dimensions. Their 3D CNN network architecture starts with 5 hardwired kernels including gray, gradient-x, gradient-y, optflow-x, and optflow-y, resulting in 33 feature maps. Then the network repeats 3D convolution and

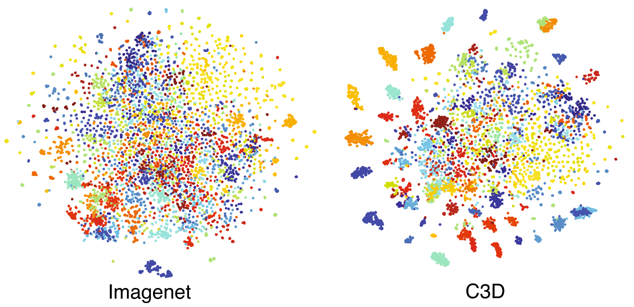


Fig. 14 Feature embedding by Imagenet and C3D. C3D features show better class separation than Imagenet, indicating its capability in learning better features for videos. Originally shown in Tran et al. (2015)

subsampling, and uses a fully-connected layer to generate a 128-dimensional feature vector for action classification. In a later extension (Ji et al., 2013), the authors regularized the network to encode long-term action information by encouraging the network to learn feature vector close to high-level motion features such as the bag-of-words representation of SIFT features.

The 3D ConvNet (Ji et al., 2010, 2013) was later extended to a modern deep architecture called C3D (Tran et al., 2015) that learns on large-scale datasets. The C3D network contains 5 convolution layers, 5 max-pooling layers, 2 fully-connected layers, and a softmax loss layer, subject to the machine memory limit and computation affordability. Their work demonstrated that C3D learns a better feature embedding for videos (see Fig. 14). Results showed that the C3D method with a linear classifier can outperform or approach the state-of-the-art methods on a variety of video analysis benchmarks including action recognition and object recognition.

Still, 3D ConvNets (Ji et al., 2010, 2013; Tran et al., 2015) for action recognition are relatively shallow with up to 8 layers. To further improve the generalization power of 3D ConvNets, Carreira and Zisserman (2017) inflated very deep networks for image classification into spatio-temporal feature extractors by repeating 2D filters along the time dimension, allowing the network to reuse 2D filters pre-trained on ImageNet. This work also shows that pre-training on the Kinetics dataset achieves better recognition accuracy on UCF-101 and HMDB51 datasets. Another solution to build a deep 3D ConvNet was proposed in Qiu et al. (2017), which uses a combination of one $1 \times 3 \times 3$ convolutional layer and one $3 \times 1 \times 1$ convolutions to take the place of a standard 3D convolution.

One limitation of 3D ConvNets is that they typically consider very short temporal intervals, such as 16 frames in Tran et al. (2015), thereby failing to capture long-term temporal information. To address this problem, Varol et al. (2017) increases the temporal extent in the 3D convolutions, and empirically shows that they can significantly improve the recognition performance.

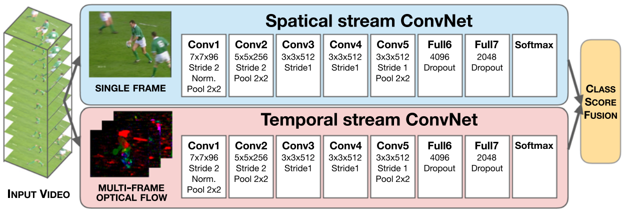


Fig. 15 Two-stream network proposed in Simonyan and Zisserman (2014) contains a spatial network and a temporal network, which are used for modeling static information in still frames and motion information in optical flow images, respectively. Revised based on the original figure in Simonyan and Zisserman (2014)

3.2.2 Multi-Stream Networks

Multi-stream networks utilize multiple convolutional networks to model both appearance and motion information in action videos. Even though the network in Karpathy et al. (2014) achieved great success, its results were significantly worse than those of the best hand-crafted shallow representations (Wang et al., 2015, 2013). To address this problem, a successful work by Simonyan and Zisserman (2014) explored a new architecture related to the two-stream hypothesis (Goodale & Milner, 1992). Their architecture contains two separate streams, a spatial ConvNet and a temporal ConvNet (see Fig. 15). The former one learns actions from still images, and the latter one performs recognition based on the optical flow field.

The two-stream network (Simonyan & Zisserman, 2014) directly fuses the outputs of the two streams generated by their respective softmax function, which may not be appropriate for gathering information over a long period of time. An improvement was proposed in Wang et al. (2015), which used the two-stream network to obtain multi-scale convolutional feature maps, and pooled the feature maps together with the detected trajectories to compute ConvNet responses centered at the trajectories. Such a scheme encodes deep features into effective descriptors constrained by sampled trajectories. Temporal feature pooling in the two-stream network was investigated in Ng et al. (2015), which is capable of making video-level predictions after the pooling layer. The work in Girdhar et al. (2017) also presented a novel pooling layer named ActionVLAD that aggregates convolutional feature descriptors in different image portions and temporal spans. They also used ActionVLAD to combine appearance and motion streams together. The network named temporal linear encoding (Diba et al., 2017) aggregates temporal features sampled from a video, and then projects onto a low-dimensional feature space. By doing so, long-range temporal structure in different frames can be captured and be encoded into a compact representation. AdaScan proposed in Kar et al. (2017) evaluated the importance of the next frame, so that only informative frames will be pooled, and non-informative frames will be disregarded in the video-level representation. Their AdaScan method uses a multilayer perceptron to compute the importance for the next frame given temporally pooled features up to the current frame. The importance score will then be used as a weight for the feature pooling operation for aggregating the next frame. Despite effective, most of the feature encoding methods lack of considering spatio-temporal information. To address this problem, the work in Duta et al. (2017) proposed a new feature encoding method for deep features. More specifically, they proposed locally max-pooling that groups features according to their similarity and then performs max-pooling. In addition, they performed max-pooling and sum-pooling over the positions of features to achieve spatio-temporal encoding.

Temporal sampling in the two-stream network was proposed in Temporal Segment Networks (TSN) (Wang et al., 2016). In TSN long-range dynamics are gathered by analyzing short video snippets formed from randomly sampled frames from segments of the full video. The idea here is that directly analyzing densely sampled video sequence makes no sense since the consecutive frames in the video contain a lot of redundancy. Moreover, some actions reveal them-self at different temporal scales, such as sprinting, which requires multiple actions over a long span of time, compared to just crouching. The original TSN network (Wang et al., 2016), was based on two-stream architecture from Simonyan and Zisserman (2014). The prediction from temporal segments was summaries by applying consensus function to frame features extracted with pre-trained Deep CNN classification network. As for consensus function was used a simple pooling operation. The advantage of this network is that it can enjoy the benefits of using big pre-trained classification networks for feature extraction. To improve the performance of temporal sampling in Zhou et al. (2018) was suggested to perform sampling at different temporal scales, and substitute pooling operation with a fully connected network, which should encode the temporal ordering of frames. The TSN can be also incorporated into another action recognition frameworks as illustrated in Qiu et al. (2019). Recently, Liu et al. (2021) attempted to use all video frames for classification by clustering the activations along the temporal dimension based on the assumption that similar frames should have similar activation values. However, this method is limited in its ability of dynamically selecting the number of clusters. Wang et al. (2021) proposed Temporal Difference Network (TDN) which aims to recognize actions from the entire video. TDN contains short-term temporal difference modules to encode local motion information and long-term temporal difference modules to capture motion across segments.

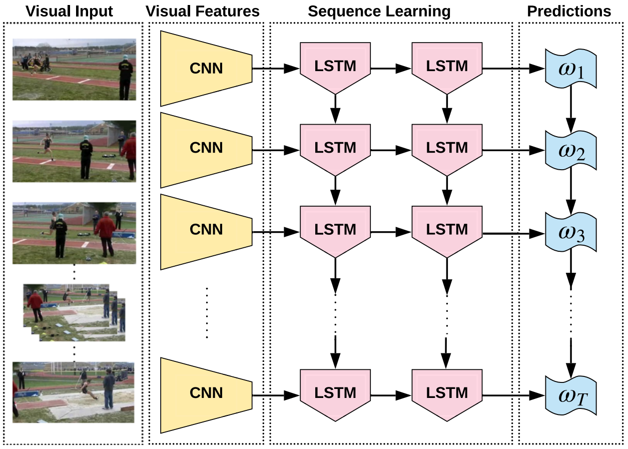


Fig. 16 Network architecture of LRCN (Donahue et al., 2015) with a hybrid of ConvNets and LSTMs. Revised based on the original figure in Donahue et al. (2015)

One of the major problems in the two-stream networks (Simonyan & Zisserman, 2014; Wang et al., 2015; Ng et al., 2015) is that they do not allow interactions between the two streams. However, such an interaction is really important for learning spatiotemporal features. To address this problem, Feichtenhofer et al. (2016) proposed a series of spatial fusion functions that make channel responses at the same pixel position be in the same correspondence. These fusion layers are placed in the middle of the two-streams allowing interactions between them. They further injected residual connections between the two streams (Feichtenhofer et al.., 2016; Feichtenhofer et al., 2017), and allow a stream to be multiplicatively scaled by the opposing stream's input (Feichtenhofer et al., 2017). Such a strategy bridges the gap between the two streams, and allows information transfer in learning spatiotemporal features.

3.2.3 Hybrid Networks

Another solution to aggregate temporal information is to add a recurrent layer on top of the CNNs, such as LSTMs, to build hybrid networks (Donahue et al., 2015; Ng et al., 2015). Such hybrid networks take the advantages of both CNNs and LSTMs, and thus have shown promising results in capturing spatial motion patterns, temporal orderings and long-range dependencies (Wang et al., 2015; Diba et al., 2017; Kar et al., 2017).

Donahue et al. (2015) explored the use of LSTM in modeling time series of frame features generated by 2D ConvNets. As shown in Fig. 16, the recurrence nature of LSTMs allows their network to generate textual descriptions of variable lengths, and recognize human actions in the videos. Ng et al. (2015) compared temporal pooling and using LSTM on top of CNNs. They discussed six types of temporal pooling methods including slow pooling and Conv pooling, and empirically showed that adding a LSTM layer generally outperforms temporal pooling by a small margin because it capture the temporal orderings of the frames. A hybrid network using CNNs and LSTMs was proposed in Wu et al. (2015). They used two-stream CNN (Simonyan & Zisserman, 2014) to extract motion features from video frames, and then fed into a bi-directional LSTM to model long-term temporal dependencies. A regularized fusion scheme was proposed in order to capture the correlations between appearance and motion features.

Hybrid networks have also been applied to skeleton-based action recognition. Skeleton data can be easily obtained by depth sensors such as Kinect or pose estimation algorithms. In these methods, hybrid deep neural networks (Shahroudy et al., 2016a; Zhu et al., 2016; Liu et al., 2016; Ke et al., 2017; Yan et al., 2018) are developed to model the structure information of various body joints as well as temporal information of body movement. Recurrent neural networks are widely used to capture the features consisting of ordered joints (Shahroudy et al., 2016a; Zhu et al., 2016; Liu et al., 2016). Temporal CNN (Ke et al., 2017) is also applied to capture the features of structured body joints. Recently, graph convolution networks have shown superior performance over RNNs and Temporal CNNs, and become the backbone for capturing the structural information of joints. Yan et al. (2018) proposed a spatio-temporal graph convolution to learn the structural and temporal information at the same time. Si et al. (2019) applied GCN-LSTM to model the temporal dependencies of skeleton and proposed an attention model to learn the importance of each joint.

3.3 Learning with Limited Data/Label

Due to the necessity of training deep neural networks, recent video are becoming extremely large. For example, Youtube-8M dataset (Abu-El-Haija et al., 2016) consists of over 8 million videos. For such large-scale datasets, it is expensive and almost impossible to annotate all the video data. Even though search engines were given action labels and were used to retrieve videos, they also make mistakes and thus the compiled video data could be noisy. One solution is to learn action models in a weakly-supervised fashion or an unsupervised fashion. Therefore, the models do not necessarily require fully-annotated video data and can learn under very limited or no supervisory signals. Few-shot learning was also recently introduced to learn in the low-sample regime.

3.3.1 Weakly-Supervised Action Learning

Weakly-supervised learning methods (Laptev et al., 2008b; Bojanowski et al., 2014; Ghadiyaram et al., 2019) are developed to deal with the scenarios where each of the videos is not fully annotated. One promising application scenario is to understand human actions in untrimmed videos, in which the temporal boundaries of various actions in the videos are not annotated. Such a learning capability enhances most of the existing action recognition methods (Tran et al., 2015; Kong et al., 2018; Simonyan & Zisserman, 2014; Wang et al., 2015; Ng et al., 2015), as require all the action videos to be trimmed which is expensive and time-consuming to achieve.

Movie with script data is a typical scenario to evaluate weakly-supervised action learning methods. Pioneering work made by Laptev et al. (2008b) presented a novel realistic action dataset from movies. Annotations were made using movie scripts. Duchenne et al. (2009) followed this work and addressed the problem of weakly-supervised learning of action models and localizing action instances in videos given the corresponding movie scripts.

Another type of work is weak-supervised action understanding given a temporally ordered list of action classes that will appear in the video. For example, Bojanowski et al. (2014) formulated the problem as a weakly supervised temporal assignment and proposed a clustering method that assigns the action labels to the temporal segments in videos. Huang et al. (2016) adapted the Connectionist Temporal classification model from speech recognition to perform weakly-supervised action labeling.

Recent works have extended weakly-supervised action representation learning to untrimmed videos with unordered action lists. Wang et al. (2017) proposed the UntrimmedNet for untrimmed video understanding by learning action models and reasoning temporal duration of action instances in an end-to-end framework. Ghadiyaram et al. (2019) took advantage of large-scale noisy labeled web videos to learn a pre-trained model for video action recognition.

3.3.2 Unsupervised and Self-Supervised Action Learning

Unsupervised or self-supervised representation learning is becoming popular in recent years as it allows deep neural networks to be pre-trained utilizing the supervisory signals within the training data, rather than given by humans. Such pre-trained models can be beneficial for downstream tasks, such as action recognition and localization. Many attempts have leveraged the temporal coherence, motion consistency and temporal continuity as supervision, which will be discussed below.

The chronological order of frames is a typical free supervision signal for videos. Action models learn to tell whether the frame sequence is ordered or not, given either shuffled or unshuffled videos (Misra et al., 2016; Fernando et al., 2017). Another related task is training the model to tell the actual order of the shuffled video frames (Lee et al., 2017a). Xu et al. (2019) extended the order prediction tasks from frames to clips. This helps to train a 3D CNN frame-

Table 1 Pros and cons of action recognition approaches

	Approaches	Pros	Cons
Shallow	Direct (Schüldt et al., 2004; Wang & Schmid, 2013)	Easy and quick to use.	Performance is limited.
	Sequential (Morency et al., 2007; Shi et al., 2011)	Models temporal evolution.	Sensitive to noise.
	Space-time (Wu et al., 2014, 2011)	Captures spatiotemporal structures.	Limited to small datasets.
	Part-based (Wang & Mori, 2010; Niebles et al., 2010)	Models body parts at a finer level.	Limited to small datasets.
	Manifold (Wang & Suter, 2007; Jia & Yeung, 2008)	Tend to achieve high performance.	Rely on human silhouettes.
	Mid-level feature (Choi et al., 2011; Liu et al., 2011)	Introduce knowledge to models.	Require extra annotations.
	Feature fusion (Yuan et al., 2013, 2016)	Tend to achieve high performance.	Slow in feature extraction.
Deep	Space-time (Ji et al., 2013; Tran et al., 2015)	Natural extension of 2D convolution.	Short temporal interval.
	Multi-stream (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2017)	Able to use pre-trained 2D ConvNets.	Int. b/w networks is difficult.
	Hybrid (Donahue et al., 2015; Ng et al., 2015)	Easy to build using existing networks.	Difficult to fine-tune.

work using chronological order supervision. Buchler et al. (2018) applied deep reinforcement learning to sample new permutations according to their expected utility to adapts to the state of the network.

The motion of objects in videos can also be used as supervision. Wang and Gupta (2015) found the corresponding pairs using visual tracking, based on Siamese-Triplet network. Purushwalkam and Gupta (2016) utilized pose as free supervision since similar pose should have similar motion. Wang et al. (2017) explored different self-supervised methods to learn the representations invariant to the variations between the object patches, which is extracted by motion cues. Gan et al. (2018) used geometry cues flow field and disparity maps to learn the video representations.

3.3.3 Few-Shot Learning

Few-shot learning aims at learning reliable models from minimalist data sets. In extreme cases, there could be no training sample for some categories which is called the zero-shot learning. Majority of few-shot works target at recognising images, while only a few address the video action recognition challenge. Zhu and Yang (2018) proposed a compound memory network (CMN) which predicts the unseen video by retrieving a similar video stored in the memory of the CMN architecture. ProtoGAN (Dwivedi et al., 2019) learns the class-prototype vectors through a feature aggregator network called *Class Prototype Transfer Network* (CPTN), then generates additional video features for the recognition classifier. Neural Graph Matching (NGM) network (Guo et al., 2018) is a graph-based approach that generates graph representations for 3D action videos and match unseen videos and seen videos by the similarity of their graph representations. Mishra et al. (2018) proposes a framework for zero-shot action recognition which models each action class as a probability distribution and the distribution parameters are a linear combination of the attributes of the action class. The weights of the attributes are learnt from the labeled samples. One challenge in few-shot action recognition is the variation of temporal lengths. Temporal Attentive Relation Network (TARN) (Bishay et al., 2019) uses attention modules to align video segments and learns a distance measure between the aligned representations for few-shot and zero-shot learning. Action Relation Network (ARN) (Zhang et al., 2020) encodes the video clips features of the query set and support set into a Power Normalized Autocorrelation Matrix (AM) from which a relation network learns to captures the relations. Similar to ARN, Ordered Temporal Alignment Module (OTAM) (Cao et al., 2020) extracts per-frame feature through an embedding network, then computes an alignment score of the distance matrix. Temporal-Relational CrossTransformers (TRX) (Perrett et al., 2021) classifies the query video by matching each sub-sequence to all sub-sequences in the support set using CrossTransformer attention modules.

3.4 Summary

Deep networks are dominant in action recognition research but shallow methods are still useful. Compared with deep networks, shallow methods are easy to train, and generally perform well on small datasets. Recent shallow methods such as improved dense trajectory with linear SVM (Wang & Schmid, 2013) have also shown promising results on large datasets, and thus they are still popularly used recently in the comparison with deep networks (Tran et al., 2015; Varol et al., 2017; Feichtenhofer et al., 2017). It would be helpful to use shallow approaches first if the datasets are small, or each video exhibits complex structures that need to be modeled. However, there are lots of pre-trained deep networks on the Internet such as C3D (Tran et al., 2015) and TSN (Wang et al., 2016) that can be easily employed. It would be also helpful to try these methods and fine-tune the models to

Table 2 Results of action detection methods on THUMOS'14 (Jiang et al., 2014). The mAP@α denotes the mean Average Precision at different threshold, α. "-" indicates the result is not reported

	mAP@0.5	mAP@0.4	mAP@0.3
End-to-End (Yeung et al., 2016)	17.1	26.4	36.0
Multi-stage (Shou et al., 2016)	19.0	28.7	36.3
TURN (Gao et al., 2017)	24.5	35.3	46.3
Temporal Context Network (Dai et al., 2017)	25.6	33.3	–
Single-stream R-C3D (Xu et al., 2017)	28.9	35.6	44.8
SSN (Zhao et al., 2017)	29.8	41.0	51.9
Single-stream R-C3D+OHEM (Xu et al., 2019)	35.8	43.1	51.1
Two-stream R-C3D (Xu et al., 2019)	36.1	43.0	51.2
BSN (Lin et al., 2018)	36.9	45.0	53.5
MGG UNet (Liu et al., 2019)	37.4	46.8	53.9
BMN (Lin et al., 2019)	38.8	47.4	56.0

particular datasets. Table 1 summarizes the pros and cons of action recognition approaches.

4 Action Localization and Detection

In order to recognize and predict an action, the machine needs to know where is the action in a video. This is achieved by action localization and detection, which find out the spatiotemporal regions containing certain human actions in videos. Both of the two tasks have attracted a large amount of research in recent years. As an analogy to object localization and detection in the image domain, action detection is additionally required to identify the action type of each action that occurs in the video compared to the action localization. Based on the feature learning paradigms, related work can be categorized into shallow and deep learning methods, for which we will make a comprehensive literature review. Table 2 summarizes some recent detection methods and compares results on thresholds of 0.3, 0.4, and 0.5. The mAP@α denotes the mean Average Precision at different IOU threshold which measures the average prevision on each action category.

4.1 Shallow Approaches

Early work (Karaman et al., 2014; Wang et al., 2014b) formulated action detection as a classification task by firstly using temporal segmentation or sliding window methods. In these work, the untrimmed video is segmented into short video clips and the multiple features are extracted for classifiers such as support vector machine (SVM) to recognize the action types. Eventually, the actions that appear in the video as well as their temporal locations are determined. Jain et al. (2014) proposed to generate a set of bounding boxes from the video which are called tubelets for action localization. However, these methods suffer from handcraft feature engineering and multi-stage model tuning, leading to quite inaccurate detection results.

4.2 Deep Architectures

Recent approaches to action localization and detection make full use of deep neural networks for learning better video feature representation. To this end, Shou et al. (2016) proposed to first generate action proposals from the long videos. Then, a localization network is introduced to fine-tune the trained action classification network to recognize the action labels. The idea of their action proposals inspired many later research (Escorci et al., 2016; Shou et al., 2017; Wang et al., 2017; Zhao et al., 2017; Xu et al., 2017; Gao et al., 2017; Chao et al., 2018). For these methods, Escorci et al. (2016) proposed a deep action proposals (DAP) method which achieves high efficiency and demonstrates to have good generalization capability. To detect human actions in frame-level granularity, Shou et al. (2017) proposed an end-to-end learning framework in which a CDC convolutional filter is designed on top of 3D ConvNet. To model the temporal structure of each action instance, Zhao et al. (2017) proposed a structured segment network (SSN) with a temporal pyramid and a dubbed temporal actionness grouping (TAG) model for action proposals generation. As the action detection is similar to the object detection, Chao et al. (2018) revisited the most widely-used object detection method Faster R-CNN and propose a temporal action localization network (TAL-Net) to address the unsolved challenges, including the large variation of action durations, temporal context modeling, and multi-stream feature fusion. Song et al. (2019b) noted that the ambiguous transition states of an action and long-term temporal context are critical for accurate action detection. Thus, they propose a transition-aware context network and it is demonstrated to be significantly effective for untrimmed video dataset. To modeling the relations among action proposals, Zeng et al. (2019) recently proposed to introduce the graph convolutional neural networks (GCN) for temporal action localization. Song et al. (2019a) introduced the action pattern tree (AP-Tree) in which the temporal information can be utilized. Inspired by the conventional idea of coarse-to-fine detection, Yang et al. (2019) proposed a spatio-temporal progressively learning method for video action detection, achieving remarkable performance on existing benchmarks. Recently, Xu et al. (2019) raised the importance of online action detection and propose a temporal recurrent network (TRN) by simultaneously performing online action detection and anticipation, significantly outperforming the state-of-the-art. Chen et al. (2021) unified the tasks of actor localization and action classification into the same backbone, which reduces model complexity and improves efficiency compared to SOTA methods. Li and Yao (2021) designed two auxiliary pretext tasks to recycle the limited labeled data and benefit both features extraction as well as prediction.

Different from previous full-supervised methods that require large-scale frame-level annotations of action instances, weakly-supervised methods need only the video- or clip-level action annotations so that they are more promising in practice. Wang et al. (2017) proposed a weakly-supervised action detection model that is directly learned on the untrimmed video data, achieving performance on-par-with those of the full-supervised action detection methods. Recently, Yu et al. (2019) introduced the temporal structure mining (TSM) approach to the weakly-supervised action detection problem. In their method, an action instance is modeled as a multi-phase process so that the phase filters can be utilized to compute the confidence score, indicating the action occurrence probability. For weakly-supervised action localization problem, it also attracts much attention in recent years. Gao et al. (2021) proposed a weakly supervised framework that consists of two modules, one module generates the pseudo ground truth of action boundaries which are used to supervise the action recognition module. Yang et al. (2021) proposed to incorporate the uncertainty for reducing the noise in the generated pseudo labels. To handle the challenge of limited temporal annotations, Yang et al. (2018) used an one-shot learning technique of matching network for temporal action localization. Narayan et al. (2019) introduced a novel loss function comprising the action classification loss, multi-label center loss, and the counting loss, setting the new state-of-the-art on weakly-supervised action localization.

In addition to using visual data, other data modalities such as skeleton and RGB-D data can also be utilized for temporal action localization and detection. To learn the features of discriminative skeleton joints, Song et al. (2018) introduced a spatio-temporal attention LSTM model for action recognition and detection. To handle the modality discrepancy in a multi-modal setting, Luo et al. (2018) proposed a graph distillation method that privileged information is learned from a large-scale multi-modal dataset in the source domain and their model can be effectively deployed to the modality-scarce target domain. For the continuous action stream scenario, Dawar and Kehtarnavaz (2018) designed a multimodal fusion system to incorporate depth camera data and wearable inertial sensor signals for action detection.

5 Action Prediction

After-the-fact action recognition has been extensively studied in the last few decades, and fruitful results have been achieved. State-of-the-art methods (Donahue et al., 2015; Girdhar et al., 2017; Wang et al., 2016) are capable of accurately giving action labels after observing the entire action executions. However, in many real-world scenarios (*e.g.*, vehicle accident and criminal activity), intelligent systems do not have the luxury of waiting for the entire video before having to react to the action contained in it. For example, being able to predict a dangerous driving situation before it occurs; opposed to recognizing it thereafter. In addition, it would be great if an autonomous driving vehicle could predict the motion trajectory of a pedestrian on the street and avoid the crash, rather than identify the trajectory after the crash into the pedestrian. Unfortunately, most of the existing action recognition approaches are unsuitable for such early classification tasks as they expect to see the entire set of action dynamics from a full video, and then make decisions.

Different from action recognition approaches, action or motion prediction[1] approaches reason about the future and infer labels before action executions end. These labels could be the discrete action categories, or continuous positions on a motion trajectory. The capability of making a prompt reaction makes action/motion prediction approaches more appealing in time-sensitive tasks. However, action/motion prediction is really challenging because accurate decisions have to be made on partial action videos.

5.1 Action Prediction

Action prediction tasks can be roughly categorized into two types, *short-term prediction* and *long-term prediction*. The former one, short-term prediction focuses on short-duration action videos, which generally last for several seconds, such as action videos in UCF-101 and Sports-1M datasets. The goal of this task is to infer action labels based

[1] In this paper, action prediction refers to the task of predicting action category, and motion prediction refers to the task of predicting motion trajectory. Video prediction is not discussed in this paper as it focuses on motion in videos rather than motion of human.

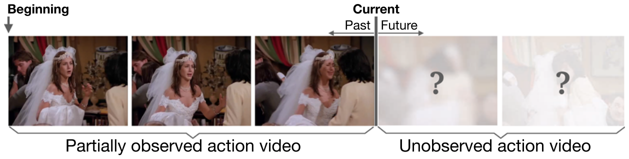


Fig. 17 Early action classification methods predicts action label given a partially observed video. Revised based on the original figure in Kong et al. (2014b)

upon temporally incomplete action videos. Formally, given an incomplete action video $\mathbf{x}_{1:t}$ containing t frames, i.e., $\mathbf{x}_{1:t} = \{f_1, f_2, \cdots, f_t\}$, the goal is to infer the action label y: $\mathbf{x}_{1:t} \rightarrow y$. Here, the incomplete action video $\mathbf{x}_{1:t}$ contains the beginning portion of a complete action execution $\mathbf{x}_{1:T}$, which only contains one single action. The latter one, long-term prediction or intention prediction, infers the future actions based on current observed human actions. It is intended for modeling action transition, and thus focuses on long-duration videos that last for several minutes. In other words, this task predicts the action that is going to happen in the future. More formally, given an action video $\mathbf{x}_a$, where $\mathbf{x}_a$ could be a complete or an incomplete action execution, the goal is to infer the next action $\mathbf{x}_b$. Here, $\mathbf{x}_a$ and $\mathbf{x}_b$ are two independent, semantically meaningful, and temporally correlated actions.

5.1.1 Early Action Classification

This task aims at recognizing a human action at an early stage, i.e., based on a temporally incomplete video (see Fig. 17). The goal is to achieve high recognition accuracy when only the beginning portion of a video is observed. The observed video contains an unfinished action, and thus making the prediction task challenging. Although this task may be solved by action recognition methods (Raptis & Sigal, 2013; Vahdat et al., 2011; Yao & Fei-Fei, 2012b, a), they were developed for recognizing complete action executions, and were not optimized for partial action observations, making action recognition approaches unsuitable for predicting actions at an early stage. Table 3 provides some results of early action classification on four datasets.

Most of the short-term action prediction approaches follow the problem setup described in Kong et al. (2014b) shown in Fig. 18. To mimic sequential data arrival, a complete video $\mathbf{x}$ with T frames is segmented into $K = 10$ segments. Consequently, each segment contains $\frac{T}{K}$ frames. Video lengths T may vary for different videos, thereby causing different lengths in their segments. For a video of length T, its k-th segment ($k \in \{1, \cdots, K\}$) contains frames starting from the $[(k-1) \cdot \frac{T}{K} + 1]$-th frame to the $(\frac{kT}{K})$-th frame. A temporally *partial video* or *partial observation* $\mathbf{x}^{(k)}$ is defined as a temporal subsequence that consists of the beginning k segments of the video. The *progress level* g of the partial video $\mathbf{x}^{(k)}$ is defined by the number of the segments contained in the partial video $\mathbf{x}^{(k)}$: $g = k$. The *observation ratio* r of a partial video $\mathbf{x}^{(k)}$ is $\frac{k}{K}$: $r = \frac{k}{K}$.

Action prediction approaches aim at recognizing unfinished action videos. Ryoo (2011) proposed the integral bag-of-words (IBoW) and dynamic bag-of-words (DBoW) approaches for action prediction. The action model of each

Table 3 Results of early action classification methods on various datasets. X@Y denotes the prediction results at Y dataset when observation ratio is set to X. "-" indicates the result is not reported

Methods	Year	0.1@BIT	0.5@BIT	0.1@UTI-1	0.5@UTI-1	0.1@UCF-101	0.5@UCF-101	0.1@Sports-1M	0.5@Sports-1M
Integral BoW (Ryoo, 2011)	2011	22.66%	48.44%	18.00%	48.00%	36.29%	74.39%	43.47%	55.99%
MSSC (Cao et al., 2013)	2013	21.09%	48.44%	28.00%	70.00%	34.05%	61.79%	46.70%	57.16%
Poselet (Raptis & Sigal, 2013)	2013	–	–	–	73.33%	–			
HM (Lan et al., 2014)	2014	–	–	38.33%	83.10%	–		–	
MTSSVM (Kong et al., 2014b)	2014	28.12%	60.00%	36.67%	78.33%	40.05%	82.39%	49.92%	66.90%
MMAPM (Kong & Fu, 2016)	2016	32.81%	67.97%	46.67%	78.33%	–	–	–	
DeepSCN (Kong et al., 2017)	2017	37.50%	78.13%	–	–	45.02%	85.75%	55.02%	70.23%
GLTSD (Lai et al., 2018)	2018	26.60%	79.40%	–	–	–		–	–
Mem-LSTM (Kong et al., 2018)	2018	–	–	–	–	51.02%	88.37%	57.60%	71.63%

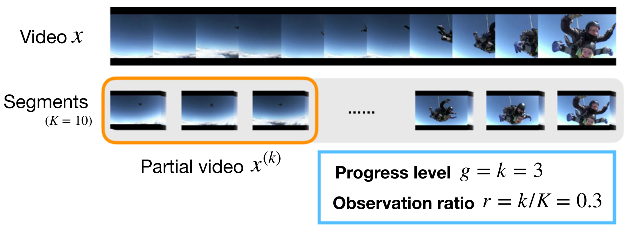


Fig. 18 Example of a temporally partial video, and graphical illustration of progress level and observation ratio. Revised based on the original figure in Kong et al. (2017)

Instantly Predictable	Early Predictable	Late Predictable
Billiards	Fencing	JavelinThrow
IceDancing	FrisbeeCatch	HighJump
RockClimbingIndoor	SoccerPenalty	FrontCrawl
PlayingPiano	VolleyballSpiking	HeadMassage
PommelHorse	HulaHoop	Haircut
Rowing	FieldHockeyPenalty	PlayingViolin
Skijet	BasketballDunk	HandstandWalking
JugglingBalls	CliffDiving	PoleVault
SoccerJuggling	Bowling	CricketBowling
TaiChi	TennisSwing	ThrowDiscus

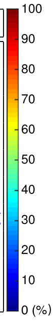


Fig. 19 Top 10 instantly, early, and late predictable actions in UCF101 dataset. Action names are colored and sorted according to the percentage of their testing samples falling in the category of instant predictable, early predictable, or late predictable. Originally shown in Kong et al. (2017)

progress level is computed by averaging features of a particular progress level in the same category. However, the learned model may not be representative if the action videos of the same class have large appearance variations, and it is sensitive to outliers. To overcome these two problems, Cao et al. (2013) built action models by learning feature bases using sparse coding and used the reconstruction error in the likelihood computation. Li et al. (2012) explored long-duration action prediction problem. However, their work detects segments by motion velocity peaks, which may not be applicable to complex outdoor datasets. Compared with Cao et al. (2013), Li et al. (2012), Ryoo (2011), Kong et al. (2014b) incorporates an important prior knowledge that informative action information is increasing when new observations are available. In addition, the method in Kong et al. (2014b) models label consistency of segments, which is not presented in their methods. From a perspective of interfering social interaction, Lan et al. (2014) developed "hierarchical movements" for action prediction, which is able to capture the typical structure of human movements before an action is executed. An early event detector (Hoai & la Torre, 2012) was proposed to localize the starting and ending frames of an incomplete event. Their method first introduces a monotonically increasing scoring function in the model constraint, which has been popularly used in a variety of action prediction methods (Kong et al., 2014b; Kong & Fu, 2016; Ma et al., 2016). Different from the aforementioned methods, Ryoo et al. (2015) studied the action prediction problem in a first-person scenario, which allows a robot to predict a person's action during human-computer interactions.

Deep learning methods have also shown in action prediction. The work in Ma et al. (2016) proposed a new monotonically decreasing loss function in learning LSTMs for action prediction. Inspired by that, the work in Kong et al. (2017) adopted an autoencoder to model sequential context information for action prediction. This method learns such information from fully-observed videos, and transfer it to partially observed videos. We enforced that the amount of the transferred information is temporally ordered for the purpose of modeling the temporal orderings of inhomogeneous action segments. We demonstrated that actions differ in their predictability, and show the top 10 instantly, early, and late predictable actions in Fig. 19. We also studied the action prediction problem following the popular two-stream framework (Simonyan & Zisserman, 2014). In Kong et al. (2018), we proposed to use memory to store hard-to-predict training samples in order to improve the prediction performance at the early stage. The memory module used in Kong et al. (2018) measures the predictability of each training sample, and will store those challenging ones. Such a memory retains a large pool of samples, and allows us to create complex classification boundaries, which are particularly useful for discriminating partial videos at the beginn ing stage.

5.1.2 Action Anticipation

Action anticipation aims to anticipate future actions from a history of actions (Gao et al., 2017). This task is fundamental to many real world applications. For example, surveillance cameras can raise an alarm before a road accident happens, robots can make better plans and decisions by anticipating human actions (Koppula & Saxena, 2013a). Action anticipation is a challenging task because the models not only need to detect the actions, but also infer future actions from the seen actions. RED (Gao et al., 2017) uses an encoder-decoder LSTM structure to predict the future video representations from the extracted representations of the historical video frames. Similarly, two LSTMs were used in Furnari and Farinella (2020) to summarize the past and infer the future for egocentric videos. The work in Vondrick et al. (2016) trains a CNN to regress the future representations from the past ones in an unsupervised way. Three similarity metrics between the past and future video representations were presented in Fernando and Herath (2021), namely Jaccard vector similarity, Jaccard cross-correlation, and Jaccard Frobenius inner product over covariances for early action anticipation. Future actions are predicted in Mehrasa et al. (2019) by learning a distribution of future actions using Variational

Auto-Encoder. The work in Ke et al. (2019) predicts actions at different future timestamps in one-shot by incorporating a temporal parameter and skip connections. Hyperbolic space is used in Surís et al. (2021) to predict future actions because it can represent actions through a compact hierarchy. In Ke et al. (2021), the authors proposed a model that consists of a conditional VAE for modeling the uncertainty of the action starting time and a MLP to predict whether the action will happen. Recently, Rohit and Kristen (2021) presents a new model called Anticipative Video Transformer and a self-supervised future prediction loss for action anticipation.

5.1.3 Intention Prediction

In practice, there are certain types of actions that contain several primitive action patterns and exhibit complex temporal arrangements, such as “make a dish”. Typically, the length of these complex actions is longer than that of short-term actions. Prediction of these long-term actions is receiving a surge of interest as it allows us to understand “what is going to happen”, including the final goal of complex human action and the person’s plausible intended action in the near future.

However, long-term action prediction is extremely challenging due to the large uncertainty in human future actions. Cognitive science shows that context information is critical to action understanding, as they typically occur with certain object interactions under particular scenes. Therefore, it would be helpful to consider the interacting objects together with the human actions, in order to achieve accurate long-term action prediction. Such knowledge can provide valuable clues for two questions “what is happening now?” and “what is going to happen next?”. It also limits the search space for potential actions using the interacting object. For example, if an action “a person grabbing a cup” is observed, most likely the person is going to “drink a beverage”, rather than going to “answering a phone”. Therefore, a prediction method considering such context is expected to provide opportunities to benefit from contextual constraints between actions and objects.

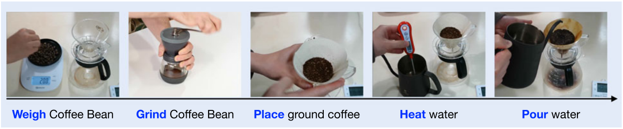


Fig. 20 A complex action can be decomposed into a series of action primitives. Revised based on the original figure in Li and Fu (2014)

Pei et al. (2011) addressed the problem of goal inference and intent prediction using an And-Or-Graph method, in which the Stochastic Context Sensitive Grammar is embodied. They modeled agent-object interactions, and generated all possible parse graphs of a single event. Combining all the possibilities generates the interpretation of the input video and achieves the global maximum posterior probability. They also show that ambiguities in the recognition of atomic actions can be reduced largely using hierarchical event contexts. Li et al. (2012) proposed a long-term action prediction method using Probabilistic Suffix Tree (PST), which captures variable Markov dependencies between action primitives in complex action. For example, as shown in Fig. 20, a wedding ceremony can be decomposed into primitives of “hold-hands”, “kneel”, “kiss”, and “put-ring-on”. In their extension (Li & Fu, 2014), object context is added to the prediction model, which enables the prediction of human-object interactions occurring in actions such as “making a dish”. Their work first introduced a concept “predictability”, and used the Predictive Accumulative Function (PAF) to show that some actions can be early predictable while others cannot be early predicted. Prediction of human action and object affordance was investigated in Koppula and Saxena (2016). They proposed an anticipatory temporal conditional random field (ATCRF) to model three types of context information, including the hierarchical structure of action primitives, the rich spatial-temporal correlations between objects and their affordances, and motion anticipation of objects and humans. In order to find the most likely motion, ATCRFs are considered as particles, which are propagated over time to represent the distribution of possible actions in the future. The work in Girase et al. (2021) introduces a new dataset called LOKI (LOng term and Key Intentions) for autonomous driving. The authors also proposed a long-term goal proposal network and a scene graph refinement and trajectory decoder module for jointly predicting the future trajectory and intention of pedestrians. In Bhattacharyya et al. (2021), the authors provided a new dataset for pedestrian trajectory prediction in dense urban scenarios. A Joint-β-cVAE is further designed to effectively model the interaction between pedestrians and vehicles, the model is trained by optimizing the ELBO The authors in Rasouli et al. (2021) proposed a multi-task learning framework which predicts both trajectories and actions of pedestrians conditioned on multi-modal data. They proposed a bi-fold feature fusion to effectively fuse multiple modalities, also a semantic map as an additional input to the model for categorical interaction modeling during training.

5.2 Summary

The availability of big data and recent advance in computer vision and machine learning enable the reasoning about the future. The key in this research is how to discover temporal correlations in large-scale data and how to model such correlations. Results shown in Table 19 demonstrate the predictability of actions that can be used as a prior and inspiring more powerful action prediction methods. There are still some unexplored opportunities in this research, such as interpretability of temporal extent, how to model long-

Fig. 21 Motion trajectory prediction is essential for practical applications such as visual surveillance and self-driving cars

term temporal correlations, and how to utilize multi-modal data to enrich the prediction model, which will be discussed in Sect. 9.

6 Motion Trajectory Prediction

Besides predicting human actions, the other key aspect in human-centered prediction is motion trajectory prediction, which aims at predicting a pedestrian's moving path. Motion trajectory prediction, an inherent capability of us, reasons the possible destination and motion trajectory of the target person. We can predict with high confidence that a person is going to walk on sidewalks than streets, and will avoid any obstacles during walking. Therefore, it is interesting to study how to make machines do the same job. Table 4 shows the ADE and FDE results on ETH/UCF dataset. ADE and FDE are standard metrics on motion trajectory prediction. Some works do not report the FDE result.

Vision-based motion trajectory prediction is essential for practical applications such as visual surveillance and self-driving cars (see Fig. 21), in which reasons about the future motion patterns of a pedestrian is critical. A large body of work learns motion patterns by clustering trajectories (Zhou et al., 2011; Morrisand & Trivedi, 2011; Kim et al., 2011; Hu et al., 2007). However, forecasting future motion trajectory of a person is really challenging as the prediction cannot be predicted in isolation. In a crowded environment, humans adapt their motion according to the behaviors of neighboring people. They may stop, or alter their paths to accommodate other people or the environment in the vicinity. Jointly modeling such complex dependencies is really difficult in dynamic environments. In addition, the predicted trajectories should not only be *physically acceptable*, but also *socially acceptable* (Gupta et al., 2018). Pedestrians always respect personal space while walking, and thus yield the right-of-way. Human-human and human-object interactions are typically subtle and complex in crowded environments, making the problem even more challenging. Furthermore, there are multiple future predictions in a crowded environment, which are all socially acceptable. Thus uncertainty estimation for the multimodal predictions is desired.

Forecasting trajectory and destination by understanding the physical scene was investigated in Kitani et al. (2012), which was one of the pioneering work in trajectory prediction in the computer vision community. The proposed method models the effect of the physical environment on the choice of human actions. The authors integrate state-of-the-art semantic scene understanding with the ideas from inverse optimal control (IOC) or inverse reinforcement learning (Abbeel & Ng, 2004; Ziebart et al., 2008). In this work, human motion is modeled as a sequence of decision-making process, and a prediction is made by maximizing the reward. Lee and Kitani (2016) extends (Kitani et al., 2012) to a dynamic environment. The state reward function is extended to a linear combination of static and dynamic state functions to update the forecasting distribution in a dynamic environment. However, IOC is limited to controlled settings as the goal state of the pedestrian's destination requires a priori. To relax this assumption, the concept of *goal set* was introduced in Mainprice et al. (2016), Dragan et al. (2011), which defines a target task space. The work in Alahi and Fei-Fei (2014) introduced a large-scale dataset of 42 million trajectories and studied the problem of trajectory prediction by modeling the social interactions of pedestrians. They captured the spatial positions of the neighboring trajectories of a person by a so-called *social affinity map*. The trajectory prediction task is formulated as a maximum a-posterior estimation problem, and the origin and destination prior knowledge is introduced to the model. The method in Ballan et al. (2016) takes a step further and generalizes trajectory prediction by considering human-scene interactions. Instead of just using semantic labels of the scene (*e.g.*, grass, street, etc), functional properties of a scene map (Turek et al., 2010) are learned in Ballan et al. (2016), which allows the prediction model to understand how agents of the same class move from one patch to another. This provides us with rich navigation patterns to the final destination. Scene semantics was also used to predict the dynamics of multiple objects (Fouhey & Zitnick, 2014; Huang & Kitani, 2008; Kooij et al., 2014; Kretzschmar et al., 2014). Kooij et al. (2014) focused on predicting pedestrians' path intention of crossing the street from the viewpoint of an approaching vehicle. Their method is built upon the dynamic Bayesian network (DBN), which considers the pedestrian's decision to stop by three cues, including the existence of an approaching vehicle, the pedestrian's awareness, and the spatial layout of the scene. Walker *et al.*in Walker et al. (2014) predicted the behavior of agents (*e.g.*, a car) in a visual scene. Ziebart et al. (2009) presented a planning-based approach for trajectory prediction.

Thanks to the recent advance in deep networks, motion trajectory prediction problem can be solved using RNN/LSTM networks (Alahi et al., 2016; Lee et al., 2017b; Su et al., 2017; Gupta et al., 2018), which have the capability of generating long sequences. More specifically, a single LSTM model was

Table 4 Results of motion trajectory prediction on ETH/UCF datasets. ADE is the minimum average displacement error, and FDE denotes the final displacement error. "–" indicates the result is not reported

	ADE	FDE
Social GAN (Gupta et al., 2018)	0.58	–
Sophie (Sadeghian et al., 2019)	0.54	–
CGNS (Li et al., 2019)	0.49	–
Social BiGAT (Kosaraju et al., 2019)	0.48	–
Next (Liang et al., 2019)	0.46	–
Social-STCNN (Mohamed et al., 2020)	0.44	–
MANTRA (Marchetti et al., 2020)	0.32	0.65
Transformer TF (Giuliari et al., 2021)	0.31	–
PECNet (Mangalam et al., 2020)	0.29	0.48
Social-NCE (Liu et al., 2020)	0.19	0.40
SGNet (Wang et al., 2021)	0.18	0.35
AgentFormer (Yuan et al., 2021)	0.18	0.29
Y-Net (Mangalam et al., 2020)	0.18	0.27

used to account for one single person's trajectory, and a social pooling layer in LSTMs was proposed to model dependencies between LSTMs, and preserve the spatial information (Alahi et al., 2016). Compared to previous work (Kitani et al., 2012; Lee & Kitani, 2016; Alahi & Fei-Fei, 2014; Ballan et al., 2016; Kooij et al., 2014), the method in Alahi et al. (2016) is end-to-end trainable, and generalizes well in complex scenes. An encoder-decoder framework was proposed in Lee et al. (2017b) for path prediction in more natural scenarios where agents interact with each other and dynamically adapt their future behaviors. Past trajectories are encoded in a RNN and then future trajectory hypotheses are generated using another decoder implemented by a separate RNN. This method also extends inverse optimal control (IOC) (Lee & Kitani, 2016; Kitani et al., 2012) to a deep model, which has shown promising results in robot control (Finn et al., 2016) and driving (Wulfmeier et al., 2016) tasks. The proposed Deep IOC is used to rank all the possible hypotheses. The scene context is captured using a CNN model, which is part of the input to the RNN encoder. A Social-GAN network in Gupta et al. (2018) was proposed to address the limitation of L2 loss in Lee et al. (2017b). Using an adversarial loss, Gupta et al. (2018) can potentially learn the distribution of multiple socially acceptable trajectories, rather than learning the average trajectories in the training data. The work in Dendorfer et al. (2021) proposes a Multi-Generator Model (MGM) to address the problem of out-of-distribution samples generated using a single generator. A categorical distribution over different trajectory types is first predicted by a Path Module Network, from which the generator is chosen to sample the future trajectories. Thus, the model can select scene-specific generators and deactivate unsuitable ones. A divide and conquer method was proposed in Narayanan et al. (2021) which prevents mode collapse problems in trajectory prediction under the winner-takes-all objective. The work in Zhao and Wildes (2021) proposes a model for goal-conditioned trajectory prediction which exploits nearest examples for goal position query and considers multi-modality and physical constraints.

7 Datasets

This section discusses some of the popular action video datasets, including actions captured in a controlled and uncontrolled environment. A detailed list is shown in Table 5. These datasets differ in the number of human subjects, background noise, appearance and pose variations, camera motion, etc., and have been widely used for the comparison of various algorithms.

7.1 Controlled Action Video Datasets

We first describe individual action datasets captured in controlled settings, and then list datasets with two or more people involved in actions. We also discuss some of the RGB-D action datasets captured using a cost-effective Kinect sensor.

7.1.1 Individual Action Datasets

Weizmann Dataset (Blank et al., 2005) is a popular video dataset for human action recognition. The dataset contains 10 action classes such as "walking", "jogging", "waving" performed by 9 different subjects, to provide a total of 90 video sequences. The videos are taken with a static camera under a simple background.

Table 5 A list of popular action video datasets used in action recognition research

Datasets	Year	#Videos	#Views	#Actions	#Subjects	#Modality	Env.
KTH (Schüldt et al., 2004)	2004	600	1	6	25	RGB	Controlled
Weizmann (Blank et al., 2005)	2005	90	1	10	9	RGB	Controlled
INRIA XMAS (Weinland et al., 2006)	2006	390	5	13	10(3 times)	RGB	Controlled
IXMAS (Yuan et al., 2009)	2006	1,148	5	11	–	RGB	Controlled
UCF Sports (Rodriguez et al., 2008)	2008	150	–	10	–	RGB	Uncontrolled
Hollywood (Laptev et al., 2008a)	2008	–	–	8	–	RGB	Uncontrolled
Hollywood2 (Marszałek et al., 2009)	2009	3,669	–	12	10	RGB	Uncontrolled
UCF 11 (Jingen Liu & Shah, 2009)	2009	1,100+	–	11	–	RGB	Uncontrolled
CA (Choi et al., 2009)	2009	44	–	5	–	RGB	Uncontrolled
MSR-I (Yuan et al., 2009)	2009	63	–	3	10	RGB	Controlled
MSR-II (Yuan et al., 2010)	2010	54	–	3	–	RGB	Crowded
MHAV (Singh & Ragheb, 2010)	2010	238	8	17	14	RGB	Controlled
UT-I (Ryoo & Aggarwal, 2010)	2010	60	2	6	10	RGB	Uncontrolled
TV-I (Patron-Perez et al., 2010)	2010	300	–	4	–	RGB	Uncontrolled
MSR-A (Li et al., 2010)	2010	567	–	20	1	RGB-D	Controlled
Olympic (Niebles et al., 2010)	2010	783	–	16	–	RGB	Uncontrolled
HMDB51 (Kuehne et al., 2011)	2011	6849	–	51	–	RGB	Uncontrolled
CAD-60 (Sung et al., 2011)	2011	60	–	12	4	RGB-D	Controlled
BIT-I (Kong et al., 2012)	2012	400	–	8	50	RGB	Controlled
LIRIS (Wolf et al., 2014)	2012	828	1	10	–	RGB	Controlled
MSRDA (Wang et al., 2012b)	2012	320	–	16	10	RGB-D	Controlled
UCF50 (Reddy & Shah, 2012)	2012	50	–	50	–	RGB	Uncontrolled
UCF101 (Khurram Soomro & Shah, 2012)	2012	13,320	–	101	–	RGB	Uncontrolled
MSR-G (Kurakin et al., 2012)	2012	336	–	12	1	RGB-D	Controlled
UTKinect-A (Xia et al., 2012)	2012	200	–	10	–	RGB-D	Controlled
ASLAN (Kliper-Gross et al., 2012)	2012	3,698	–	432	–	RGB	Uncontrolled
MSRAP (Oreifej & Liu, 2013)	2013	360	–	6 pairs	10	RGB-D	Controlled
CAD-120 (Koppula et al., 2013)	2013	120	–	12	4	RGB-D	Controlled
THUMOS’14 (Jiang et al., 2014)	2014	413	1	20	–	RGB	Uncontrolled
Sports-1M (Karpathy et al., 2014)	2014	1,133,158	-	487	–	RGB	Uncontrolled
3D Online (Yu et al., 2014)	2014	567	–	20	–	RGB-D	Uncontrolled
FCVID (Jiang et al., 2018)	2015	91,233	–	239	–	RGB	Uncontrolled
ActivityNet (Caba Heilbron et al.., 2015)	2015	28,000	–	203	–	RGB	Uncontrolled
YouTube-8M (Abu-El-Haija et al., 2016)	2016	8,000,000	–	4,716	–	RGB	Uncontrolled
Charades (Shahroudy et al., 2016b)	2016	9,848	2	157	–	RGB	Controlled
NTU-RGB+D (Shahroudy et al., 2016a)	2016	56,680	–	120	106	RGB+D+IR+Skeleton	Controlled
PKU-MMD (Phase 1) (Chunhui et al., 2017)	2017	1076	3	51	66	RGB+D+IR+Skeleton	Uncontrolled
PKU-MMD (Phase 2) (Chunhui et al., 2017)	2017	2000	3	49	13	RGB+D+IR+Skeleton	Uncontrolled
NEU-UB	2017	600	–	6	20	RGB-D	Controlled
Kinetics (Kay et al., 2017)	2017	500,000	–	600	–	RGB	Uncontrolled
AVA (Gu et al., 2017)	2017	57,600	–	80	–	RGB	Uncontrolled
20BN-Something-Something (Goyal et al., 2017)	2017	108,499	-	174	-	RGB	Uncontrolled
SLAC (Zhao et al., 2017)	2017	520,000	–	200	–	RGB	Uncontrolled
Moments in Time (Monfort et al., 2019)	2017	1,000,000	-	339	–	RGB	Uncontrolled
EPIC-Kitchens (Damen et al., 2018)	2018	90,000+	–	397	32	RGB	Uncontrolled
COIN (Tang et al., 2019)	2019	11,827	1	180	–	RGB	Uncontrolled

Table 5 continued

Datasets	Year	#Videos	#Views	#Actions	#Subjects	#Modality	Env.
HACS Segments (Zhao et al., 2019)	2019	50,000+	1	200	–	RGB	Uncontrolled
HAA00 (Chung et al., 2021)	2021	10,000	-	500	–	RGB	Uncontrolled
MultiSports (Li et al., 2021)	2021	3200	–	4	–	RGB	Uncontrolled

KTH Dataset (Schüldt et al., 2004) consists of 6 types of human actions (boxing, hand clapping, hand waving, jogging, running and walking) repeated several times by 25 different subjects in 4 scenarios (outdoors, outdoors with scale variation, outdoors with different clothes and indoors). There are 600 action videos in the dataset.

INRIA XMAS Multiview Dataset (Weinland et al., 2006) was complied for multi-view action recognition. It contains videos captured from 5 views including a top-view camera. This dataset consists of 13 actions, each of which is repeated 3 times by 10 actors.

7.1.2 Group Action Datasets

UT-Interaction Dataset (Ryoo & Aggarwal, 2010) is comprised of 2 sets of 10 videos with different background and camera settings. The videos contain 6 classes of human-human interactions: handshake, hug, kick, point, punch, and push.

BIT-Interaction Dataset (Kong et al., 2012) consists of 8 classes of human interactions (bow, boxing, handshake, high-five, hug, kick, pat, and push), with 50 videos per class. Videos are captured in realistic scenes with cluttered backgrounds, partially occluded body parts, moving objects, and variations in subject appearance, scale, illumination condition, and viewpoint.

TV-Interaction Dataset (Patron-Perez et al., 2010) contains 300 videos clips with human interactions. These videos are categorized into 4 interaction categories: handshake, high five, hug, and kiss, and annotated with the upper body of people, discrete head orientation and interaction.

MultiSports Dataset (Li et al., 2021) is a multi-person dataset that contains 3200 video clips of 4 sport classes. The dataset contains 37701 action instances with 902, 000 bounding boxes, which helps for more fine-grained spatio-temporal action detection and localization.

7.2 Unconstrained Datasets

Although the aforementioned datasets lay a solid foundation for action recognition research, they were captured in controlled settings, and may not be able to train approaches that can be used in real-world scenarios. To address this problem, researchers collected action videos from the Internet, and compiled large-scale action datasets, which will be discussed in the following.

UCF101 Dataset (Khurram Soomro & Shah, 2012) has been widely used in action recognition research. It comprises of realistic videos collected from Youtube. It contains 101 action categories, with 13320 videos in total. UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. The dataset can be roughly divided into 5 categories: (1) Human-Object Interaction (2) Body-Motion Only (3) Human-Human Interaction (4) Playing Musical Instruments (5) Sports. It should be noted that many clips are collected from the same video. Consequently, different clips may have the same person or the same scenario, or the same lighting, etc. This seems different from practical scenarios, and thus its difficulty is limited.

HMDB51 Dataset (Kuehne et al., 2011) contains a total of about 6849 video clips distributed in a large set of 51 action categories. Each category contains a minimum of 101 video clips. In addition to the label of the action category, each clip is annotated with an action label as well as a meta-label describing the property of the clip, such as visible body parts, camera motion, camera viewpoint, number of people involved in the action, and video quality. The actions can be grouped into five categories, including general facial actions (*e.g.*smile, chew, talk), Facial actions with object manipulation (*e.g.*smoke, eat, drink), General body movements (*e.g.*cartwheel, clap hands, climb), Body movements with object interaction (*e.g.*brush hair, catch, draw sword), Body movements for human interaction (*e.g.*fencing, hug, kick someone). The dataset also has two distinct categories namely “no motion” and “camera motion”. The dataset is extremely challenging mainly due to the presence of a significant camera/background motion. To remove camera motion, standard image stitching techniques to can be used to align frames of a clip.

Kinetics (Carreira & Zisserman, 2017) dataset comprises of 700 human action classes and approximately 650, 000 video clips, including human-object interactions and human-human interactions. The videos were compiled from YouTube by matching its title and the prepared Kinetics actions list. After that, the videos were segmented by tracking actions on Google Image Search, and then labeled by

Amazon's Mechanical Turk(AMT). In the end, this dataset is cleaned and de-noised using machine learning techniques. Different from previous datasets, one clip in this dataset may contain several different actions in sequence, but it is only classified into one action category. This means these clips don't have complete action labels. As described in their work, the top-5 measure is supposed to be used because of the incomplete labels. Within the same action category, clips are captured from different videos, including TV and film videos. Consequently, there is a large appearance variation, for example, people in clips may have different age, height, clothes, *etc.*, and there are various types of camera motion/shake, background clutter. Besides, each clip lasts around 10s and has a variable resolution.

Sports-1M Dataset (Karpathy et al., 2014) contains 1, 133, 158 video URLs, which have been annotated automatically with 487 labels. It is one of the largest video datasets. Very diverse sports videos are included in this dataset, such as Shaolin Kung Fu, Wing Chun, etc. The dataset is extremely challenging due to very large appearance and pose variations, significant camera motion, noisy background motion, etc.

THUMOS'14 Dataset (Jiang et al., 2014) contains more than 20 hours of sport videos. Though the training sets are trimmed videos labeled with 20 action classes, the validation and testing sets include 200 and 213 untrimmed videos, respectively. This dataset has been the most widely used dataset for action detection and localization.

ActivityNet Dataset (Heilbron et al., 2015) has two versions for action detection and localization. The first is Activity v1.2, which covers 100 activity classes and contains 4,819 training videos and 2,383 videos for validation. The other version is Activity v1.3, which consists of 10,024 videos for training and 4,926 videos for validation with 200 activity classes.

PKU-MMD Dataset (Chunhui et al., 2017) is a large-scale multi-modal datasets focusing on long continuous sequences action detection and multi-modality action analysis. The first phase contains 51 action categories, performed by 66 distinct subjects in 3 camera views. Each video lasts about $3 \sim 4$ minutes and contains approximately 20 action instances. The second phase contains 2,000 short video sequences in 49 action categories, performed by 13 subjects in 3 camera views. Each video lasts about $1 \sim 2$ minutes and contains approximately 7 action instances.

AVA Dataset (Gu et al., 2018) provides audio-visual annotations for about 15 minute long movie clips. For the AVA Action subsets, it contains 430 videos split into 235 for training, 64 for validation, and 131 for test. Each video has 15 minutes annotated in 1-second intervals.

COIN Dataset (Tang et al., 2019) is a recently released large-scale dataset to address instruction video analysis problems. It contains 11,827 daily activity videos of 180 different classes. Different from other action datasets, human actions in COIN dataset are hierarchically structured with practical semantics.

HACS Dataset (Zhao et al., 2019) is also a recently released large-scale dataset for action localization and recognition. For the HACS Segments subset, it contains 139K action segments densely annotated in 50K untrimmed videos spanning 200 action categories.

20BN-SOMETHING-SOMETHING dataset (Goyal et al., 2017) is a dataset shows human interaction with everyday objects. In the dataset, human performs pre-defined action with a daily object. It contains 108, 499 video clips across 174 classes. The dataset enables the learning of visual representations for the physical properties of the objects and the world.

Moments-in-Time Dataset (Monfort et al., 2019) is a large-scale video dataset for action understanding. It contains over 1, 000, 000 3-second labeled video clips distributed in 339 categories. The visual elements of the videos include people, animals, objects or natural phenomena. The dataset is dedicated to building models that are capable of abstracting and reasoning complex human actions.

EPIC-Kitchens dataset (Damen et al., 2018) is one of the largest first-person vision dataset. It consists of 55 hours videos and 125 verb classes and 300 noun classes recorded by head-mounted camera. These videos are shot at different cities and different styles kitchens and divided to 39, 600 action segments with object bounding boxes. Besides, these videos contain human doing different kitchen tasks at the same time. To better annotate these actions, voice notes for the actions are collected in the dataset.

HAA500 dataset (Chung et al., 2021) is a human-centric atomic action dataset. It consists of 500 atomic classes, where 212 are sport/athletics, 51 are playing musical instruments, 82 are games and hobbies, and 155 are daily actions.

7.3 RGB-D Action Video Datasets

All the datasets described above were captured by RGB video cameras. Recently, there is an increasing interest in using cost-effective Kinect sensors to capture human actions due to the extra depth data channel. Compared to RGB data channels, the extra depth data channel elegantly provides scene structure, which can be used to simplify intra-class motion variations and reduce cluttered background noise (Kong & Fu, 2017). Popular RGB-D action datasets are listed in the following.

MSR Daily Activity Dataset (Wang et al., 2012b): there are 16 categories of actions: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. All these actions are performed by 10 subjects. There are 320 RGB samples and 320 depth samples available.

3D Online Action Dataset (Yu et al., 2014) was compiled for three evaluation tasks: same-environment action recognition, cross-environment action recognition and continuous action recognition. The dataset contains human action or human-object interaction videos captured from RGB-D sensors. It contains 7 action categories, such as drinking, eating, and reading cellphone.

CAD-120 Dataset (Koppula et al., 2013) comprises of 120 RGB-D action videos of long daily activities. It is also captured using the Kinect sensor. Action videos are performed by 4 subjects. The dataset consists of 12 action types, such as rinsing mouth, talking on the phone, cooking, writing on whiteboard, etc. Tracked skeletons, RGB images, and depth images are provided in the dataset.

UTKinect-Action Dataset (Xia et al., 2012) was captured by a Kinect device. There are 10 high-level action categories contained in the dataset such as walk, sit down, etc. The dataset comprises 200 action vidos and three channels were recorded: RGB, depth and skeleton joint locations.

NTU-RGB+D (Shahroudy et al., 2016b; Liu et al., 2020) dataset contains 60 action classes and 56, 880 video samples. Recently, it has been extended to 120 action classes and another 114, 480 video samples in Liu et al. (2020). All the samples were collected from 106 distinct subjects by Kinect sensors. RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos are provided for each sample. There is higher variation of environmental conditions compared with previous datasets, including 96 different backgrounds with illumination variations.

8 Evaluation Protocols for Action Recognition and Prediction

Due to different application purposes, action recognition and prediction techniques are evaluated in different ways.

Shallow action recognition methods such as Schüldt et al. (2004), Niebles and Fei-Fei (2007), Wu et al. (2011) were usually evaluated on small-scale datasets, for example, Weizmann dataset (Blank et al., 2005), KTH dataset (Schüldt et al., 2004), UCF Sports dataset (Rodriguez et al., 2008). The leave-one-out training scheme is popularly used on these datasets, and the confusion matrix is usually adopted to show the recognition accuracy of each action category. For sequential approaches such as Wang and Mori (2008; 2010), per-frame recognition accuracy is often used. In Marszałek et al. (2009), Tang et al. (2012a), average precision that approximates the area under the precision-recall curve is also adopted for each individual action class. Deep networks (Carreira & Zisserman, 2017; Tran et al., 2015; Varol et al., 2017) are generally evaluated on large-scale datasets such as UCF-101 (Khurram Soomro & Shah, 2012) and HMDB51 (Kuehne et al., 2011) and thus can only report overall recognition performance on each dataset. Please refer to Herath et al. (2017) for a list of performance of recent action recognition methods on various datasets.

Most of action prediction methods (Ryoo, 2011; Cao et al., 2013; Kong et al., 2014b, 2017) were evaluated on existing action datasets. Different from the evaluation method used in action recognition, recognition accuracy at each observation ratio (ranging from 10% to 100%) is reported for action prediction methods. As described in Kong et al. (2017), the goal of these methods is to achieve high recognition accuracy at the beginning stage of action videos, in order to accurately recognize actions as early as possible. Table 3 summarizes the performance of action prediction methods on various datasets.

There are several popular metrics for evaluating motion trajectory prediction methods, including *Average Displacement Error* (ADE), *Final Displacement Error* (FDE), and *Average Non-linear Displacement Error* (ANDE). ADE is the mean square error computed over all estimated points of a trajectory and the ground-truth points. FDE is defined as the distance between the predicted final destination and the ground-true final destination. ANDE is the MSE at the non-linear turning regions of a trajectory arising from human-human interactions.

Vairous metrics exist to evaluate action detection and localization methods. Recall that Recall (R) measures the number of true positives over the total number of true positives and false negatives. Average Recall (AR) is the average of recalls over multiple Intersection over Union (IoU) values. Area under the AR vs. AN curve (AUC) measures how well the detection method is able to distinguish between positive and negative proposals. Another metric called mean Average Precision (mAP) @ α where α denotes different IoU threshold which measures the Average Prevision (AP) on each action category.

9 Future Directions

In this section, we discuss some future directions in action recognition and prediction research that might be interesting to explore.

Dataset Significant efforts have been made to collect different types of action video datasets in recent years in order to advance the research of action recognition and prediction. Nevertheless, existing action recognition and prediction models trained on these datasets are still difficult to be generalized to real-world scenarios, possibly because the incapability of these datasets in covering all the aspects that may happen in practical scenarios. First of all, majority of the video datasets were collected under good lighting and weather conditions. However, this assumption may not hold in practice. A visual surveillance system may need to run 24 hours a day whatever the weather is. Unfortunately, existing methods are still difficult to be generalized to poor lighting conditions or extreme weather. Second of all, some datasets are restricted to certain scenarios, for example, UCF101 contains sports videos and EPIC-Kitchens dataset captured in kitchens. Although one well-trained model may perform well in one scenario, it may perform poorly in a new scenario. This could be attributed to the new environment, camera motion, appearance changes, *etc.*that have not been seen in the previous scenario. Last but not least, existing deep neural networks based methods require a significant amount of data for training. However, video data could be limited in some research areas, such as biomedical research or human rehabilitation research. Is it possible to create and render virtual training video data using game engines such as UnReal (Unreal engine, UnrealCV) based on existing small-scale data? This could serve as an alternative solution to directly generalizing deep neural networks to small-scale data. All of these challenges bring new problems to action recognition research and prompt us to collect new datasets to advance the research.

Benefitting from image models. Deep architectures are dominating the action recognition research lately like the trend of other developments in the computer vision community. However, training deep networks on videos is difficult, and thus benefiting from deep models pre-trained on images or other sources would be a better solution to explore. In addition, image models have done a good job of capturing spatial relationships of objects, which could also be exploited in action understanding. It is interesting to explore how to transfer knowledge from image models to video models using the idea of inflation (Carreira & Zisserman, 2017) or domain adaptation (Tang et al., 2012b).

Interpretability on temporal extent. Interpretability of image models has been discussed but it has not been extensively discussed in video models. As shown in Satkin and Hebert (2010), Raptis and Sigal (2013), not all frames are equally important for action recognition; only few of them are critical. Therefore, there are a few things that require a deep understanding of the temporal interpretability of video models. First of all, actions, especially long-duration actions can be considered as a sequence of primitives. It would be interesting to have interpretability of these primitives, such as how are these primitives organized in the temporal domain in actions, how do they contribute to the classification task, can we only use few of them without sacrificing recognition performance in order to achieve fast training? In addition, actions differ in their temporal characteristics. Some actions can be understood at their early stage and some actions require more frames to be observed. It would be interesting to ask why these actions can be early predicted, and what are the salient signals that are captured by the machine. Such an understanding would be useful in developing more efficient action prediction models.

Learning from multi-modal data. Humans are observing multi-modal data everyday, including visual, audio, text, etc. These multi-modal data help the understanding of each type of data. For example, reading a book helps us to reconstruct the corresponding part of the visual scene. However, little work is paying attention to action recognition/prediction using multi-modal data. It is beneficial to use multi-modal data to help visual understanding of complex actions because the multi-modal data such as text data contain rich semantic knowledge given by humans. In addition to action labels, which can be considered as verbs, textual data may include other entities such as nouns (objects), prepositions (spatial structure of the scene), adjectives and adverbs, etc. Although learning from nouns and prepositions have been explored in action recognition and human-object interaction, few studies have been devoted to learning from adjectives and adverbs. Such learning tasks provide more descriptive information about human actions such as motion strength, thereby making fine-grained action understanding into reality.

Learning long-term temporal correlations. Multi-modal data also enable the learning of long-term temporal correlations between visual entities from the data, which might be difficult to directly learn from visual data. Long-term temporal correlations characterize the sequential order of actions occurring in a long sequence, which is similar to what our brain stores. When we want to recall something, one pattern evokes the next pattern, suggesting the associations spanning in long-term videos. Interactions between visual entities are also critical to understanding long-term correlations. Typically, certain actions occur with certain object interactions under particular scene settings. Therefore, it needs to involve not only actions, but also an interpretation of objects, scenes and their temporal arrangements with actions, since this knowledge can provide a valuable clue for "what's happening now" and "what's going to happen next". This learning task also allows us to predict actions in a long-duration sequence.

Physical aspect of actions. Action recognition and prediction are tasks fairly targeting at high-level aspects of videos, and not focusing on finding action primitives that encode basic physical properties. Recently, there has been an increasing interest in learning the physical aspects of the world, which studies fine-grained actions. One example is the something-something dataset introduced in Goyal et al. (2017) that studies human-object interactions. Interestingly, this dataset provides labels or textual description templates such as "Dropping [something] into [something]", to describe the interaction between humans and objects, and an object and an object. This allows us to learn models that can understand physical aspects of the world including human actions, object-object interactions, spatial relationships, etc.

Even though we can infer a large amount of information from action videos, there are still some physical aspects that are challenging to be inferred. We are wondering that can we make a step further, saying understanding more physical aspects, such as the motion style, force, acceleration, etc, from videos? Physics-101 (Wu et al., 2015) studied this problem in objects, but can we extend it to actions? A new action dataset containing such fine-grained information is needed. To achieve this goal, our ongoing work is providing a dataset containing human actions with EMG signals, which we hope to benefit fine-grained action recognition.

Learning actions without annotations. For increasingly large action datasets such as Something-Something (Goyal et al., 2017) and Sports-1M (Karpathy et al., 2014), manual labeling becomes prohibitive. Automatic labeling using search engines (Karpathy et al., 2014; Abu-El-Haija et al., 2016), video subtitles and movie scripts (Marszałek et al., 2009; Laptev et al., 2008a) is possible in some domains, but still requires manual verification. Crowdsourcing (Goyal et al., 2017) would be a better option but still suffers from labeling diversity problem, and may generate incorrect action labels. In addition, videos in almost all the action datasets are temporally segmented, with only one action in each of the videos. However, this assumption does not hold as videos may be streaming and it is difficult to know the exact starting and ending frames of an action execution in streaming videos. This prompts us to develop more robust and efficient action recognition/prediction approaches that can automatically learn from unlabeled videos or untrimmed videos.

Actions in open-world. Human action recognition in real-world is essentially an open set problem, which requires the model to simultaneously recognize the known action classes and reject the unknown actions (Geng et al., 2020; Bao et al., 2021). However, existing open set recognition (OSR) research works mainly focus on image modality (Scheirer et al., 2012, 2014; Zhang & Patel, 2016; Bendale & Boult, 2016; Oza & Patel, 2019; Perera et al., 2020; Chen et al., 2020), except for a few works on videos (Shu et al., 2018; Roitberg et al., 2020) and other modalities (Yang et al., 2019). These works typically do not work well on video data due to the following challenges. First, the temporal nature of videos leads to high diversity of human actions, which is challenging for an OSR model to be aware of *what it does not know* when given human actions with unknown temporal dynamics. Besides, the static bias (i.e., appearance of the video background and foreground actor) in video data could be easily over-fitted by deep learning models. The model finally could hardly identify unknown actions in an unbiased open vision world. These challenges motivate recent work (Bao et al., 2021) to build an uncertainty-aware and unbiased model for open set action recognition (OSAR). Since open-world actions can be regarded as out-of-distribution (OOD) data, developing more advanced OOD detection methods to tackle the distributional shift of human actions under OSAR setting is promising in the future.

10 Conclusion

The availability of big data and powerful models diverts the research focus on human actions from understanding the present to reasoning the future. We have presented a complete survey of state-of-the-art techniques for action recognition and prediction from videos. These techniques became particularly interesting in recent decades due to their promising and practical applications in several emerging fields focusing on human movements. We investigate several aspects of the existing attempts including hand-crafted feature design, models and algorithms, deep architectures, datasets, and system performance evaluation protocols. Future research directions are also discussed in this survey.

References

Abbeel, P., & Ng, A. (2004). Apprenticeship learning via inverse reinforcement learning. In: ICML.

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675.

Alahi, A., & Fei-Fei, V.R.L. (2014). Socially-aware large-scale crowd forecasting. In: CVPR.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In: CVPR.

Ballan, L., Castaldo, F., Alahi, A., Palmieri, F., & Savarese, S. (2016). Knowledge transfer for scene-specific motion prediction. In: ECCV.

Bao, W., Yu, Q., & Kong, Y. (2021). Evidential deep learning for open set action recognition. In: ICCV.

Bendale, A., & Boult, T.E. (2016). Towards open set deep networks. In: CVPR.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Bhattacharyya, A., Reino, D.O., Fritz, M., & Schiele, B. (2021). Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In: CVPR.

Bishay, M., Zoumpourlis, G., & Patras, I. (2019). Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In: BMVC.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*, 47–73.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In: Proc. ICCV.

Bobick, A., & Davis, J. (2001). The recognition of human movement using temporal templates. *IEEE Trans Pattern Analysis and Machine Intelligence, 23*(3), 257–267.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In: European Conference on Computer Vision, pp. 628–643. Springer.

Bregonzio, M., Gong, S., & Xiang, T. (2009). Recognizing action as clouds of space-time interest points. In: CVPR.

Buchler, U., Brattoli, B., & Ommer, B. (2018). Improving spatiotemporal self-supervision by deep reinforcement learning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–786.

Cao, K., Ji, J., Cao, Z., Chang, C.Y., & Niebles, J.C. (2020). Few-shot video classification via temporal alignment. In: CVPR.

Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J., & Wang, S. (2013). Recognizing human activities from partially observed videos. In: CVPR.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR.

Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., & Sukthankar, R. (2018). Rethinking the Faster R-CNN architecture for temporal action localization. In: CVPR.

Chen, G., Qiao, L., Shi, Y., Peng, P., Li, J., Huang, T., Pu, S., & Tian, Y. (2020). Learning open set network with discriminative reciprocal points. In: ECCV.

Chen, S., Sun, P., Xie, E., Ge, C., Wu, J., Ma, L., Shen, J., & Luo, P. (2021). Watch only once: An end-to-end video action detection framework. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 8178–8187.

Choi, W., & Savarese, S. (2012). A unified framework for multi-target tracking and collective activity recognition. In: ECCV, pp. 215–230. Springer.

Choi, W., Shahid, K., & Savarese, S. (2009). What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: computer vision workshops (ICCV Workshops), 2009 IEEE 12th international conference on, pp. 1282 –1289.

Choi, W., Shahid, K., & Savarese, S. (2011). Learning context for collective activity recognition. In: CVPR.

Chung, J., hsin Wuu, C., ru Yang, H., Tai, Y.W., & Tang, C.K. (2021). Haa500: Human-centric atomic action dataset with curated videos. In: ICCV.

Chunhui, L., Yueyu, H., Yanghao, L., Sijie, S., & Jiaying, L. (2017). Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475.

Ciptadi, A., Goodwin, M. S., & Rehg, J. M. (2014). Movement pattern histogram for action recognition and retrieval. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision - ECCV 2014* (pp. 695–710). Springer International Publishing.

Clarke, T., Bradshaw, M., Field, D., Hampson, S., & Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception, 24*, 1171–80.

Cutting, J., & Kozlowski, L. (1977). Recognition of friends by their work: Gait perception without familarity cues. *Bulletin of the Psychonomic Society, 9*, 353–56.

Dai, X., Singh, B., Zhang, G., Davis, L., & Chen, Y. (2017). Temporal context network for activity localization in videos. 2017 IEEE International conference on computer vision (ICCV) pp. 5727–5736.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In: CVPR.

Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

Dawar, N., & Kehtarnavaz, N. (2018). Action detection and recognition in continuous action streams by deep learning-based sensing fusion. *IEEE Sensors Journal, 18*(23), 9660–9668.

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Neural Mechanisms of Perception and Action, 3*(5), 172–178.

Dendorfer, P., Elflein, S., & Leal-Taixé, L. (2021). Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In: ICCV.

Diba, A., Sharma, V., & Gool, L.V. (2017). Deep temporal linear encoding networks. In: CVPR.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In: ICCV VS-PETS.

Donahue, J., Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In: CVPR.

Dragan, A., Ratliff, N., & Srinivasa, S. (2011). Manipulation planning with goal sets using constrained trajectory optimization. In: ICRA.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., & Ponce, J. (2009). Automatic annotation of human actions in video. In: 2009 IEEE 12th International conference on computer vision, pp. 1491–1498. IEEE.

Duong, T.V., Bui, H.H., Phung, D.Q., & Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In: CVPR.

Duta, I.C., Ionescu, B., Aizawa, K., & Sebe, N. (2017). spatio-temporal vector of locally max pooled features for action recognition in videos. In: CVPR.

Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., & Jain, A. (2019). Protogan: Towards few shot learning for action recognition. In: ICCVW.

Efros, A., Berg, A., Mori, G., & Malik, J. (2003). Recognizing action at a distance. *ICCV, 2*, 726–733.

Escorcia, V., Caba Heilbron, F., Niebles, J.C., & Ghanem, B. (2016). DAPs: Deep action proposals for action understanding. In: ECCV.

Fabian Caba Heilbron Victor Escorcia, B.G., & Niebles, J.C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 961–970.

Fanti, C., Zelnik-Manor, L., & Perona, P. (2005). Hybrid models for human motion recognition. In: CVPR.

Feichtenhofer, C., Pinz, A., & Wildes, R.P. (2016). Spatiotemporal residual networks for video action recognition. In: NIPS.

Feichtenhofer, C., Pinz, A., & Wildes, R.P. (2017). Spatiotemporal multiplier networks for video action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7445–7454. IEEE.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In: CVPR.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In: CVPR.

Fernando, B., Bilen, H., Gavves, E., & Gould, S. (2017). Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3636–3645.

Fernando, B., & Herath, S. (2021). Anticipating human actions by correlating past with the future with jaccard similarity measures. In: CVPR.

Finn, C., Levine, S., & Abbeel, P. (2016). Guided cost learning: deep inverse optimal control via policy optimization. In: arXiv preprint arXiv:1603.00448.

Fouhey, D.F., & Zitnick, C.L. (2014). Predicting object dynamics in scenes. In: CVPR.

Furnari, A., & Farinella, G.M. (2020). Rolling-unrolling lstms for action anticipation from first-person video. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).

Gan, C., Gong, B., Liu, K., Su, H., & Guibas, L.J. (2018). Geometry guided convolutional neural networks for self-supervised video representation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5589–5597.

Gao, J., Yang, Z., Chen, K., Sun, C., & Nevatia, R. (2017). TURN TAP: Temporal unit regression network for temporal action proposals. In: ICCV.

Geng, C., Huang, S.j., & Chen, S. (2020). Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence.

Ghadiyaram, D., Tran, D., & Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 12046–12055.

Girase, H., Gang, H., Malla, S., Li, J., Kanehara, A., Mangalam, K., & Choi, C. (2021). Loki: Long term and key intentions for trajectory prediction. In: ICCV.

Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B. (2017). Actionvlad: Learning spatio-temporal aggregation for action classification. In: CVPR.

Giuliari, F., Hasan, I., Cristani, M., & Galasso, F. (2021). Transformer networks for trajectory forecasting. In: 2020 25th international conference on pattern recognition (ICPR), pp. 10335–10342. IEEE.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence, 29*(12), 2247–2253.

Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The" something something" video database for learning and evaluating visual common sense. In: Proc. ICCV.

Gu, C., Sun, C., Ross, D.A., 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: CVPR.

Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D.A., Toderici, G., Li, Y., Ricco, S., Sukthankar, R., Schmid, C., et al. (2017). Ava: A video dataset of spatio-temporally localized atomic visual actions. arXiv preprint arXiv:1705.08421.

Guo, M., Chou, E., Huang, D.A., Song, S., Yeung, S., & Fei-Fei, L. (2018). Neural graph matching networks for fewshot 3d action recognition. In: ECCV.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In: CVPR.

Hadfield, S., & Bowden, R. (2013). Hollywood 3d: Recognizing actions in 3d natural scenes. In: CVPR. Portland, Oregon.

Harris, C., & Stephens., M. (1988). A combined corner and edge detector. In: Alvey vision conference.

Hasan, M., & Roy-Chowdhury, A.K. (2014). Continuous learning of human activity models using deep nets. In: ECCV.

Heilbron, F.C., Escorcia, V., Ghanem, B., & Niebles, J.C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In: CVPR.

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and Vision Computing.

Hoai, M., & la Torre, F.D. (2012). Max-margin early event detectors. In: CVPR.

Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence, 17*, 185–203.

Hu, J.F., Zheng, W.S., Lai, J., & Zhang, J. (2015). Jointly learning heterogeneous features for rgb-d activity recognition. In: CVPR.

Hu, W., Xie, D., Fu, Z., Zeng, W., & Maybank, S. (2007). Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on, 16*(4), 1168–1181.

Huang, D.A., Fei-Fei, L., & Niebles, J.C. (2016). Connectionist temporal modeling for weakly supervised action labeling. In: European conference on computer Vision, pp. 137–153. Springer.

Huang, D.A., & Kitani, K.M. (2008). Action-reaction: Forecasting the dynamics of human interaction. In: ECCV.

Ikizler, N., & Forsyth, D. (2007). Searching video for complex activities with finite state models. In: CVPR.

Jain, M., van Gemert, J., Jegou, H., Bouthemy, P., & Snoek, C.G. (2014). Action localization with tubelets from motion. In: CVPR.

Jain, M., Jégou, H., & Bouthemy, P. (2013). Better exploiting motion for better action recognition. In: CVPR.

Ji, S., Xu, W., Yang, M., & Yu, K. (2010). 3d convolutional neural networks for human action recognition. In: ICML.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). *3d convolutional neural networks for human action recognition*. Pattern Analysis and Machine Intelligence: IEEE Trans.

Jia, C., Kong, Y., Ding, Z., & Fu, Y. (2014). Latent tensor transfer learning for rgb-d action recognition. In: ACM Multimedia.

Jia, K., & Yeung, D.Y. (2008). Human action recognition using local spatio-temporal discriminant embedding. In: CVPR.

Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., & Sukthankar, R. (2014). THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/.

Jiang, Y. G., Wu, Z., Wang, J., Xue, X., & Chang, S. F. (2018). Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(2), 352–364. https://doi.org/10.1109/TPAMI.2017.2670560

Jingen Liu, J.L., & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In: CVPR.

Gao, Jiyang., Yang, Zhenheng., & N, R. (2017). Red: Reinforced encoder-decoder networks for action anticipation. In: BMVC.

Kar, A., Rai, N., Sikka, K., & Sharma, G. (2017). Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In: CVPR.

Karaman, S., Seidenari, L., & Bimbo, A.D. (2014). Fast saliency based pooling of fisher encoded dense trajectories. In: ECCV THUMOS Workshop.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In: CVPR.

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

Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2017). A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3288–3297.

Ke, Q., Fritz, M., & Schiele, B. (2019). Time-conditioned action anticipation in one shot. In: CVPR.

Ke, Q., Fritz, M., & Schiele, B. (2021). Future moment assessment for action query. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision.

Keestra, M. (2015). Understanding human action. integraiting meanings, mechanisms, causes, and contexts. TRANSDISCIPLINARITY IN PHILOSOPHY AND SCIENCE: APPROACHES, PROBLEMS, PROSPECTS pp. 201–235.

Khurram Soomro, A.R.Z., & Shah, M. (2012). Ucf101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01.

Kim, K., Lee, D., & Essa, I. (2011). Gaussian process regression flow for analysis of motion trajectories. In: ICCV.

Kitani, K.M., Ziebart, B.D., Bagnell, J.A., & Hebert, M. (2012). Activity forecasting. In: ECCV.

Klaser, A., Marszalek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In: BMVC.

Kliper-Gross, O., Hassner, T., & Wolf, L. (2012). The action similarity labeling challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3).

Kong, Y., & Fu, Y. (2014). Modeling supporting regions for close human interaction recognition. In: ECCV workshop.

Kong, Y., & Fu, Y. (2015). Bilinear heterogeneous information machine for rgb-d action recognition. In: CVPR.

Kong, Y., & Fu, Y. (2016). Max-margin action prediction machine. *TPAMI, 38*(9), 1844–1858.

Kong, Y., & Fu, Y. (2017). Max-margin heterogeneous information machine for rgb-d action recognition. *International Journal of Computer Vision (IJCV), 123*(3), 350–371.

Kong, Y., Gao, S., Sun, B., & Fu, Y. (2018). Action prediction from videos via memorizing hard-to-predict samples. In: AAAI.

Kong, Y., Jia, Y., & Fu, Y. (2012). Learning human interaction by interactive phrases. In: Proceedings of European conference on computer vision.

Kong, Y., Jia, Y., & Fu, Y. (2014). Interactive phrases: Semantic descriptions for human interaction recognition. In: PAMI.

Kong, Y., Kit, D., & Fu, Y. (2014). A discriminative model with multiple temporal scales for action prediction. In: ECCV.

Kong, Y., Tao, Z., & Fu, Y. (2017). Deep sequential context networks for action prediction. In: CVPR.

Kong, Y., Tao, Z., & Fu, Y. (2018). Adversarial action prediction networks. IEEE TPAMI.

Kooij, J.F.P., Schneider, N., Flohr, F., & Gavrila, D.M. (2014). Context-based pedestrian path prediction. In: European Conference on Computer Vision, pp. 618–633. Springer.

Koppula, H.S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. International Journal of Robotics Research.

Koppula, H.S., & Saxena, A. (2013). Anticipating human activities for reactive robotic response. In: IROS.

Koppula, H.S., & Saxena, A. (2013). Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In: ICML.

Koppula, H. S., & Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(1), 14–29.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, S.H., & Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. arXiv preprint arXiv:1907.03395.

Kretzschmar, H., Kuderer, M., & Burgard, W. (2014). Learning to predict trajecteories of cooperatively navigation agents. In: International conference on robotics and automation.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: A large video database for human motion recognition. In: ICCV.

Kurakin, A., Zhang, Z., & Liu, Z. (2012). A real-time system for dynamic hand gesture recognition with a depth sensor. In: EUSIPCO.

Lai, S., Zhang, W. S., Hu, J. F., & Zhang, J. (2018). Global-local temporal saliency action prediction. *IEEE Transactions on Image Processing, 27*(5), 2272–2285.

Lan, T., Chen, T.C., & Savarese, S. (2014). A hierarchical representation for future action prediction. In: European conference on computer vision, pp. 689–704. Springer.

Lan, T., Sigal, L., & Mori, G. (2012). Social roles in hierarchical models for human activity. In: CVPR.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *TPAMI, 34*(8), 1549–1562.

Laptev, I. (2005). On space-time interest points. *IJCV, 64*(2), 107–123.

Laptev, I., & Lindeberg, T. (2003). Space-time interest points. In: ICCV, pp. 432–439.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In: CVPR.

Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies.

Laptev, I., & Perez, P. (2007). Retrieving actions in movies. In: ICCV.

Le, Q.V., Zou, W.Y., Yeung, S.Y., & Ng, A.Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR.

Lee, H.Y., Huang, J.B., Singh, M., & Yang, M.H. (2017). Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE international conference on computer vision, pp. 667–676.

Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., & Chandraker, M. (2017). Desire: Distant future prediction in dynamic scenes with interacting agents. In: CVPR.

Lee, N., & Kitani, K.M. (2016). Predicting wide receiver trajectories in american football. In: WACV2016.

Li, J., Ma, H., & Tomizuka, M. (2019). Conditional generative neural system for probabilistic trajectory prediction. In: 2019 IEEE/RSJ International conference on intelligent robots and systems (IROS), pp. 6150–6156. IEEE.

Li, K., & Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(8), 1644–1657.

Li, K., Hu, J., & Fu, Y. (2012). Modeling complex temporal composition of actionlets for activity prediction. In: ECCV.

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3d points. In: CVPR workshop.

Li, Y., Chen, L., He, R., Wang, Z., Wu, G., & Wang, L. (2021). Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In: ICCV.

Li, Z., & Yao, L. (2021). Three birds with one stone: Multi-task temporal action detection via recycling temporal annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4751–4760.

Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., & Fei-Fei, L. (2019). Peeking into the future: Predicting future person activities and locations in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5725–5734.

Lin, T., Liu, X., Li, X., Ding, E., & Wen, S. (2019). Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3889–3898.

Lin, T., Zhao, X., Su, H., Wang, C., & Yang, M. (2018). Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Lin, Y.Y., Hua, J.H., Tang, N.C., Chen, M.H., & Liao, H.Y.M. (2014). Depth and skeleton associated action recognition without online accessible rgb-d cameras. In: CVPR.

Liu, J., Kuipers, B., & Savarese, S. (2011). Recognizing human actions by attributes. In: CVPR.

Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos “in the wild”. In: Proceedings of IEEE conference on computer vision and pattern recognition.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2020). Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(10), 2684–2701.

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, pp. 816–833. Springer.

Liu, L., & Shao, L. (2013). Learning discriminative representations from rgb-d video data. In: IJCAI.

Liu, X., Pintea, S.L., Nejadasl, F.K., Booij, O., & van Gemert, J.C. (2021). No frame left behind: Full video action recognition. In: CVPR.

Liu, Y., Ma, L., Zhang, Y., Liu, W., & Chang, S.F. (2019). Multi-granularity generator for temporal action proposal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3604–3613.

Liu, Y., Yan, Q., & Alahi, A. (2020). Social nce: Contrastive learning of socially-aware motion representations. arXiv preprint arXiv:2012.11717.

Lu, C., Jia, J., & Tang, C.K. (2014). Range-sample depth feature for action recognition. In: CVPR.

Lucas, B.D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In: Proceedings of imaging understanding workshop.

Luo, G., Yang, S., Tian, G., Yuan, C., Hu, W., & Maybank, S. J. (2014). Learning human actions by combining global dynamics and local appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(12), 2466–2482.

Luo, J., Wang, W., & Qi, H. (2013). Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: ICCV.

Luo, Z., Hsieh, J.T., Jiang, L., Carlos Niebles, J., & Fei-Fei, L. (2018). Graph distillation for action detection with privileged modalities. In: ECCV.

Ma, S., Sigal, L., & Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In: CVPR.

Mainprice, J., Hayne, R., & Berenson, D. (2016). Goal set inverse optimal control and iterative re-planning for predicting human reaching motions in shared workspace. In: arXiv preprint arXiv:1606.02111.

Mangalam, K., An, Y., Girase, H., & Malik, J. (2020). From goals, waypoints & paths to long term human trajectory forecasting. arXiv preprint arXiv:2012.01526.

Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., & Gaidon, A. (2020). It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: European conference on computer vision, pp. 759–776. Springer.

Marchetti, F., Becattini, F., Seidenari, L., & Bimbo, A.D. (2020). Mantra: Memory augmented networks for multiple trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7143–7152.

Marszałek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In: IEEE conference on computer vision & pattern recognition.

Mass, J., Johansson, G., Jason, G., & Runeson, S. (1971). *Motion perception I and II [film]*. Houghton Mifflin.

Mehrasa, N., Jyothi, A.A., Durand, T., He, J., Sigal, L., & Mori, G. (2019). A variational auto-encoder model for stochastic point processes. In: CVPR.

Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In: ICCV.

Gao, Mingfei., Zhou, Yingbo., X, R., S, R., X, C. (2021). Woad: Weakly supervised online action detection in untrimmed videos. In: CVPR.

Mishra, A., Verma, V., Reddy, M.K.K., Subramaniam, A., Rai, P., & Mittal, A. (2018). A generative approach to zero-shot and few-shot action recognition.

Misra, I., Zitnick, C.L., & Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In: European conference on computer vision, pp. 527–544. Springer.

Mohamed, A., Qian, K., Elhoseiny, M., & Claudel, C. (2020). Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14424–14432.

Monfort, M., Zhou, B., Bargal, S. A., Yan, T., Andonian, A., Ramakrishnan, K., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 502–508.

Morency, L.P., Quattoni, A., & Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR.

Morrisand, B., & Trivedi, M. (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33*(11), 2287–2301.

Narayan, S., Cholakkal, H., Khan, F.S., & Shao, L. (2019). 3C-Net: Category count and center loss for weakly-supervised action localization. In: ICCV.

Narayanan, S., Moslemi, R., Pittaluga, F., Liu, B., & Chandraker, M. (2021). Divide-and-conquer for lane-aware diverse trajectory prediction. In: CVPR.

Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In: CVPR.

Ni, B., Wang, G., & Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: ICCV Workshop on CDC3CV.

Niebles, J.C., Chen, C.W., & Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV.

Niebles, J.C., & Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In: CVPR.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision, 79*(3), 299–318.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2013). Berkeley mhad: A comprehensive multimodal human action database. In: Proceedings of the IEEE workshop on applications on computer vision.

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *PAMI, 22*(8), 831–843.

Oreifej, O., & Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: CVPR.

Oza, P., & Patel, V.M. (2019). C2AE: Class conditioned auto-encoder for open-set recognition. In: CVPR.

Patron-Perez, A., Marszalek, M., Reid, I., & Zissermann, A. (2012). Structured learning of human interaction in tv shows. *PAMI, 34*(12), 2441–2453.

Patron-Perez, A., Marszalek, M., Zisserman, A., & Reid, I. (2010). High five: Recognising human interactions in tv shows. In: Proceedings of British conference on machine vision.

Pei, M., Jia, Y., & Zhu, S.C. (2011). Parsing video events with goal inference and intent prediction. In: ICCV, pp. 487–494. IEEE.

Perera, P., Morariu, V.I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., & Patel, V.M. (2020). Generative-discriminative feature representations for open-set recognition. In: CVPR.

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., & Damen, D. (2021). Temporal-relational crosstransformers for few-shot action recognition. In: CVPR.

Perronnin, F., & Dance, C. (2006). Fisher kernels on visual vocabularies for image categorization. In: CVPR.

Plotz, T., Hammerla, N.Y., & Olivier, P. (2011). Feature learning for activity recognition in ubiquitous computing. In: IJCAI.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*, 976–990.

Purushwalkam, S., & Gupta, A. (2016). Pose from action: Unsupervised learning of pose features based on motion. arXiv preprint arXiv:1609.05420.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual network. In: ICCV.

Qiu, Z., Yao, T., Ngo, C.W., Tian, X., & Mei, T. (2019). Learning spatio-temporal representation with local and global diffusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 12056–12065.

Rajko, S., Qian, G., Ingalls, T., & James, J. (2007). Real-time gesture recognition with minimal training requirements and on-line learning. In: CVPR.

Ramanathan, V., Yao, B., & Fei-Fei, L. (2013). Social role discovery in human events. In: CVPR.

Ramezani, M., & Yaghmaee, F. (2016). A review on human action analysis in videos for retrieval applications. *Artificial Intelligence Review, 46*(4), 485–514.

Raptis, M., & Sigal, L. (2013). Poselet key-framing: A model for human activity recognition. In: CVPR.

Raptis, M., & Soatto, S. (2010). Tracklet descriptors for action modeling and video analysis. In: ECCV.

Rasouli, A., Rohani, M., & Luo, J. (2021). Bifold and semantic reasoning for pedestrian behavior prediction. In: CVPR.

Reddy, K.K., & Shah, M. (2012). Recognizing 50 human action categories of web videos. Machine Vision and Applications Journal.

Ricoeur, P. (1992). Oneself as another (K. Blamey, Trans.). Chicago: University of Chicago Press.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nat. Rev. Neurosci., 11*, 264–274.

Rodriguez, M.D., Ahmed, J., & Shah, M. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In: CVPR.

Rohit, G., & Kristen, G. (2021). Anticipative video transformer. In: ICCV.

Roitberg, A., Ma, C., Haurilet, M., & Stiefelhagen, R. (2020). Open set driver activity recognition. In: IVS.

Ryoo, M., & Aggarwal, J. (2006). Recognition of composite human activities through context-free grammar based representation. *CVPR, 2*, 1709–1718.

Ryoo, M., & Aggarwal, J. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV, pp. 1593–1600.

Ryoo, M., & Aggarwal, J. (2011). Stochastic representation and recognition of high-level group activities. *IJCV, 93*, 183–200.

Ryoo, M., Fuchs, T.J., Xia, L., Aggarwal, J.K., & Matthies, L. (2015). Robot-centric activity prediction from first-person videos: What will they do to me? In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, pp. 295–302. ACM.

Ryoo, M.S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV.

Ryoo, M.S., & Aggarwal, J.K. (2010). UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

S Singh, S.V., & Ragheb, H. (2010). Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: 2nd Workshop on Activity monitoring by multi-camera surveillance systems (AMMCSS), pp. 48–55.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., & Savarese, S. (2019). Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1349–1358.

Satkin, S., & Hebert, M. (2010). Modeling the temporal extent of actions. In: ECCV.

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(11), 2317–2324.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2012). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(7), 1757–1772.

Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local svm approach. In: IEEE ICPR.

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In: Proc. ACM Multimedia.

Shahroudy, A., Liu, J., Ng, T.T., & Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: IEEE conference on computer vision and pattern recognition.

Shahroudy, A., Liu, J., Ng, T.T., & Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: CVPR.

Shi, Q., Cheng, L., Wang, L., & Smola, A. (2011). Human action segmentation and recognition using discriminative semi-markov models. *IJCV, 93*, 22–32.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., & Blake, A. (2013). Efficient human pose estimation from single depth images. PAMI.

Shou, Z., Chan, J., Zareian, A., Miyazawa, K., & Chang, S.F. (2017). CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: CVPR.

Shou, Z., Wang, D., & Chang, S.F. (2016). Temporal action localization in untrimmed videos via multi-stage CNNs. In: CVPR.

Shu, Y., Shi, Y., Wang, Y., Zou, Y., Yuan, Q., & Tian, Y. (2018). ODN: Opening the deep network for open-set action recognition. In: ICME.

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1227–1236.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In: NIPS.

Singh, S., Velastin, S.A., & Ragheb, H. (2010). Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE international conference on, pp. 48–55. IEEE.

Sminchisescu, C., Kanaujia, A., Li, Z., & Metaxas, D. (2005). Conditional models for contextual human motion recognition. In: International conference on computer vision.

Song, H., Wu, X., Zhu, B., Wu, Y., Chen, M., & Jia, Y. (2019). Temporal action localization in untrimmed videos using action pattern trees. *IEEE Transactions on Multimedia (TMM), 21*(3), 717–730.

Song, L., Zhang, S., Yu, G., & Sun, H. (2019). TACNet: Transition-aware context network for spatio-temporal action detection. In: CVPR.

Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2018). Spatio-temporal attention-based LSTM networks for 3d action recognition and detection. *IEEE Transactions on Image Processing (TIP), 27*(7), 3459–3471.

Su, H., Zhu, J., Dong, Y., & Zhang, B. (2017). Forecast the plausible paths in crowd scenes. In: IJCAI.

Sumi, S. (2000). Perception of point-light walker produced by eight lights attached to the back of the walker. *Swiss Journal of Psychology, 59*, 126–32.

Sun, D., Roth, S., & Black, M.J. (2010). Secrets of optical flow estimation and their principles. In: CVPR.

Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., & Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In: CVPR.

Sun, L., Jia, K., Chan, T.H., Fang, Y., Wang, G., & Yan, S. (2014). Dl-sfa: Deeply-learned slow feature analysis for action recognition. In: CVPR.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). Human activity detection from rgbd images. In: AAAI workshop on pattern, activity and intent recognition.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from rgbd images. In: ICRA.

Surís, D., Liu, R., & Vondrick, C. (2021). Learning the predictability of the future. In: CVPR.

Tang, K., Fei-Fei, L., & Koller, D. (2012). Learning latent temporal structure for complex event detection. In: CVPR.

Tang, K., Ramanathan, V., Fei-Fei, L., & Koller, D. (2012). Shifting weights: Adapting object detectors from image to video. In: Advances in Neural Information Processing Systems.

Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., & Zhou, J. (2019). COIN: A large-scale dataset for comprehensive instructional video analysis. In: CVPR.

Taylor, G.W., Fergus, R., LeCun, Y., & Bregler, C. (2010). Convolutional learning of spatio-temporal features. In: ECCV.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In: ICCV.

Tran, D., & Sorokin, A. (2008). Human activity recognition with metric learning. In: ECCV.

Troje, N. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2*, 371–87.

Troje, N., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception Psychophys, 67*, 667–75.

Turek, M., Hoogs, A., & Collins, R. (2010). Unsupervised learning of functional categories in video scenes. In: ECCV.

Unreal engine. https://www.unrealengine.com/.

UnrealCV. https://unrealcv.org.

Vahdat, A., Gao, B., Ranjbar, M., & Mori, G. (2011). A discriminative key pose sequence model for recognizing human interactions. In: ICCV Workshops, pp. 1729 –1736.

Varol, G., Laptev, I., & Schmid, C. (2017). Long-term temporal convolutions for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Anticipating visual representations from unlabeled video. In: CVPR.

Walker, J., Gupta, A., & Hebert, M. (2014). Patch to the future: Unsupervised visual prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3302–3309.

Wang, C., Wang, Y., Xu, M., & Crandall, D.J. (2021). Stepwise goal-driven networks for trajectory prediction. arXiv preprint arXiv:2103.14107.

Wang, H., Kläser, A., Schmid, C., & Liu, C.L. (2013). Dense trajectories and motion boundary descriptors for action recognition. IJCV **103**(60–79).

Wang, H., Kläser, A., Schmid, C., & Liu, C.L. (2011). Action Recognition by Dense Trajectories. In: IEEE conference on computer vision & pattern recognition, pp. 3169–3176. Colorado Springs, United States. http://hal.inria.fr/inria-00583818/en.

Wang, H., Oneata, D., Verbeek, J., & Schmid, C. (2015). A robust and efficient video representation for action recognition. IJCV.

Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision. Sydney, Australia. http://hal.inria.fr/hal-00873267.

Wang, H., Ullah, M.M., Kläser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In: BMVC.

Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012). Robust 3d action recognition with random occupancy patterns. In: ECCV.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In: CVPR.

Wang, K., Wang, X., Lin, L., Wang, M., & Zuo, W. (2014). 3d human activity recognition with reconfigurable convolutional neural networks. In: ACM Multimedia.

Wang, L., Qiao, Y., & Tang, X. (2014). Action recognition and detection by combining motion and appearance features. In: ECCV THUMOS Workshop.

Wang, L., Qiao, Y., & Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR.

Wang, L., & Suter, D. (2007). Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: CVPR.

Wang, L., Tong, Z., Ji, B., & Wu, G. (2021). Tdn: Temporal difference networks for efficient action recognition. In: CVPR, pp. 1895–1904.

Wang, L., Xiong, Y., Lin, D., & Van Gool, L. (2017). UntrimmedNets for weakly supervised action recognition and detection. In: CVPR.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L.V. (2016). Temoral segment networks: Toward good practices for deep action recognition. In: ECCV.

Wang, S.B., Quattoni, A., Morency, L.P., Demirdjian, D., & Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In: CVPR.

Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE international conference on computer vision, pp. 2794–2802.

Wang, X., He, K., & Gupta, A. (2017). Transitive invariance for self-supervised visual representation learning. In: Proceedings of the IEEE international conference on computer vision, pp. 1329–1338.

Wang, Y., & Mori, G. (2008). Learning a discriminative hidden part model for human action recognition. In: NIPS.

Wang, Y., & Mori, G. (2010). Hidden part models for human action recognition: Probabilistic vs. max-margin. PAMI.

Wang, Z., Wang, J., Xiao, J., Lin, K.H., & Huang, T.S. (2012). Substructural and boundary modeling for continuous action recognition. In: CVPR.

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding, 104*(2–3), 249–257.

Willems, G., Tuytelaars, T., & Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest poing detector. In: ECCV.

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

Wong, S.F., Kim, T.K., & Cipolla, R. (2007). Learning motion categories using both semantic and structural information. In: CVPR.

Wu, B., Yuan, C., & Hu, W. (2014). Human action recognition based on context-dependent graph kernels. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2609–2616.

Wu, J., Yildirim, I., Lim, J.J., Freeman, W.T., & Tenenbaum, J.B. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In: Advances in Neural Information Processing Systems, pp. 127–135.

Wu, X., Xu, D., Duan, L., & Luo, J. (2011). Action recognition using context and appearance distribution features. In: CVPR.

Wu, Z., Wang, X., Jiang, Y.G., Ye, H., & Xue, X. (2015). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: ACM Multimedia.

Wulfmeier, M., Wang, D., & Posner, I. (2016). Watch this: Scalable cost function learning for path planning in urban environment. In: arXiv preprint arXiv:1607:02329.

Xia, L., & Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: CVPR.

Xia, L., Chen, C., & Aggarwal, J. (2012). View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE computer society conference on, pp. 20–27. IEEE.

Xia, L., Chen, C.C., & Aggarwal, J.K. (2012). View invariant human action recognition using histograms of 3d joints. In: CVPRW.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., & Zhuang, Y. (2019). Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 10334–10343.

Xu, H., Das, A., & Saenko, K. (2017). R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision, pp. 5783–5792.

Xu, H., Das, A., & Saenko, K. (2019). Two-stream region convolutional 3d network for temporal activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(10), 2319–2332.

Xu, M., Gao, M., Chen, Y.T., Davis, L.S., & Crandall, D.J. (2019). Temporal recurrent networks for online action detection. In: ICCV.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI conference on artificial intelligence.

Yang, H., He, X., & Porikli, F. (2018). One-shot action localization by learning sequence matching network. In: CVPR.

Yang, S., Yuan, C., Wu, B., Hu, W., & Wang, F. (2015). Multi-feature max-margin hierarchical bayesian model for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1610–1618.

Yang, W., Zhang, T., Yu, X., Qi, T., Zhang, Y., & Wu, F. (2021). Uncertainty guided collaborative training for weakly supervised temporal action detection. In: CVPR.

Yang, X., & Tian, Y. (2014). Super normal vector for activity recognition using depth sequences. In: CVPR.

Yang, X., Yang, X., Liu, M.Y., Xiao, F., Davis, L.S., & Kautz, J. (2019). STEP: Spatio-temporal progressive learning for video action detection. In: CVPR.

Yang, Y., Hou, C., Lang, Y., Guan, D., Huang, D., & Xu, J. (2019). Open-set human activity recognition based on micro-doppler signatures. *Pattern Recognition, 85*, 60–69.

Yang, Y., & Shah, M. (2012). Complex events detection using data-driven concepts. In: ECCV.

Yao, B., & Fei-Fei, L. (2012). Action recognition with exemplar based 2.5d graph matching. In: ECCV.

Yao, B., & Fei-Fei, L. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *TPAMI, 34*(9), 1691–1703.

Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. In: CVPR.

Yeung, S., Russakovsky, O., Mori, G., & Fei-Fei, L. (2016). End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2678–2687.

Yilmaz, A., & Shah, M. (2005). Actions sketch: A novel action representation. In: CVPR.

Yu, G., Liu, Z., & Yuan, J. (2014). Discriminative orderlet mining for real-time recognition of human-object interaction. In: ACCV.

Yu, T., Ren, Z., Li, Y., Yan, E., Xu, N., & Yuan, J. (2019). Temporal structure mining for weakly supervised action detection. In: ICCV.

Yu, T.H., Kim, T.K., & Cipolla, R. (2010). Real-time action recognition by spatiotemporal semantic and structural forests. In: BMVC.

Yuan, C., Hu, W., Tian, G., Yang, S., & Wang, H. (2013). Multi-task sparse learning with beta process prior for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 423–429.

Yuan, C., Li, X., Hu, W., Ling, H., & Maybank, S.J. (2013). 3d r transform on spatio-temporal interest points for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 724–730.

Yuan, C., Li, X., Hu, W., Ling, H., & Maybank, S. J. (2014). Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition. *IEEE Transactions on Image Processing, 23*(2), 658–672.

Yuan, C., Wu, B., Li, X., Hu, W., Maybank, S. J., & Wang, F. (2016). Fusing r features and local features with context-aware kernels for action recognition. *International Journal of Computer Vision, 118*(2), 151–171.

Yuan, J., Liu, Z., & Wu, Y. (2009). Discriminative subvolume search for efficient action detection. In: IEEE conference on computer vision and pattern recognition.

Yuan, J., Liu, Z., & Wu, Y. (2010). Discriminative video pattern search for efficient action detection. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Yuan, Y., Weng, X., Ou, Y., & Kitani, K. (2021). Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. arXiv preprint arXiv:2103.14023.

Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., & Gan, C. (2019). Graph convolutional networks for temporal action localization. In: ICCV.

Zhai, X., Peng, Y., & Xiao, J. (2013). Cross-media retrieval by intra-media and inter-media correlation mining. *Multimedia Systems, 19*(5), 395–406.

Zhang, H., & Patel, V. M. (2016). Sparse representation-based open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(8), 1690–1696.

Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H.S., & Koniusz, P. (2020). Few-shot action recognition with permutation-invariant attention. In: ECCV.

Zhao, H., Torralba, A., Torresani, L., & Yan, Z. (2019). HACS: Human action clips and segments dataset for recognition and temporal localization. In: ICCV.

Zhao, H., & Wildes, R.P. (2021). Where are you heading? dynamic trajectory prediction with expert goal examples. In: ICCV.

Zhao, H., Yan, Z., Wang, H., Torresani, L., & Torralba, A. (2017). Slac: A sparsely labeled dataset for action classification and localization. arXiv preprint arXiv:1712.09374.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., & Lin, D. (2017). Temporal action detection with structured segment networks. In: ICCV.

Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. In: Proceedings of the European conference on computer vision (ECCV), pp. 803–818.

Zhou, B., Wang, X., & Tang, X. (2011). Random field topic model for semantic region analysis in crowded scenes from tracklets. In: CVPR.

Zhu, L., & Yang, Y. (2018). Compound memory networks for few-shot video classification. In: ECCV.

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: Thirtieth AAAI conference on artificial intelligence.

Ziebart, B., Maas, A., Bagnell, J., & Dey, A. (2008). Maximum entropy inverse reinforcement learning. In: AAAI.

Ziebart, B., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J., Hebert, M., Dey, A., & Srinivasa, S. (2009). Planning-based prediction for pedestrians. In: IROS.