Human Action Recognition and Prediction: A Survey

Yu Kong, *Member, IEEE,* and Yun Fu, *Senior Member, IEEE*

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 the 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.

Index Terms—Action recognition, Action prediction, Human interaction, RGB-D videos, Heterogeneous data, Feature learning, Deep networks, hand-crafted Features, Action dataset

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 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 is complied 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 human vision system, the algorithms ought to produce a label after observing the entire or part of a human action execution [1], [2]. 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.

- *Yu Kong is with Department of ECE, Northeastern University, Boston, MA. E-mail: yukong@ece.neu.edu.*
- *Yun Fu is with Department of ECE and College of CIS, Northeastern University, Boston, MA. E-mail: yunfu@ece.neu.edu.*

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 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 [3] (Fig. 1(a)), which contains simple daily actions such as "clapping" and "running"; 2) a human interaction in UT-Interaction dataset [4] (Fig. 1(b)), which consists of human interactions including "handshake" and "push"; 3) a human-object interaction in UCF Sports dataset [5] (Fig. 1(c)), which comprises of sport actions and human-object interactions; 4) a group action in Hollywood 2 dataset [6] (Fig. 1(d)); 5) an action captured by a RGB-D sensor in UTKinect dataset [7] (Fig. 1(e)); and 6) a multi-view action in Multicamera dataset [8] (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.

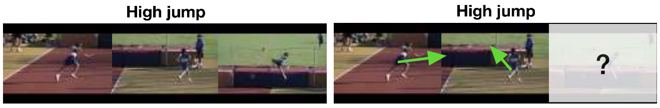
Technology advances in computer science and engineering have been enabling machines to understand human actions in videos. There are two basic topics in the computer vision community, vision-based human action recognition and prediction:

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 Figure 2(a)) [1], [9], [10], [11], [12], [13], [14]. In other words, action recognition is an after-the-fact video analysis task that focuses on the present state. It

(a) (b) (c) (d) (e) (f)

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).

has been studied for decades and is still a very popular topic due to broad real-world applications including video retrieval [15], visual surveillance [8], [16], 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 [11], [17], [18], [19] 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 Figure 2(b)) [2], [20], [21], 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.

We will mainly discuss recent advance in action recognition and prediction in this survey. Different from recent survey papers [22], [23], 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 [24], [25] 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 Figure 3). State-of-the-art algorithms [26], [27], [28], [29] remarkably reduce the human labor in analyzing a large-scale of video data and provide understanding on the current state and future state of an ongoing video data.

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 [16]. With the input of a network of cameras [8], [10], a visual surveillance system powered by action recognition [18], [30], [31] and prediction [2], [20], [21] 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 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 [32]. The text data, such as tags, titles, descriptions and keywords, can be incorrect, obscure, and irrelevant, making video retrieval unsuccessful [33]. An alternative method is to analyze human actions in videos, as the majority of these videos contain such a cue. For example, in [15], 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 [32].

1.1.3 Entertainment

The gaming industry in recent years has attracted an increasing 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.

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

Fig. 3. Examples of real-world applications using action recognition techniques.

To enable accurate perception of human actions, these games use cost-effective RGB-D sensors (e.g. Kinect [34]) which provide an additional depth channel data [35], [36], [37]. 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 [38], [39], [40], [41].

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 [42], [43].

1.1.5 Autonomous Driving Vehicle

Action prediction algorithms [44], [45] could be one of the potential and may be most important building components in an autonomous driving vehicle. Action prediction algorithms can predict a person's intention [43], [46], [47] 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 [21], action prediction algorithms [21], [45] 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 Figure 4). 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 larger on real-world action datasets [31], [48]. 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.

Fig. 4. Appearance variations in different camera views.

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 existing activity features such as histograms of oriented gradient [49] and interest points [50] 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 [51]. 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 [52], [53], [54] 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 [26], [27] have shown promising results on datasets captured in uncontrolled settings, but they normally require large amount of annotated training data. Action datasets such as HMDB51 [55] and UCF-101 [56] contain thousands of videos, but still far from enough for training deep networks with millions of parameters. Although Youtube-8M [57] and Sposrts-1M datasets [31] provide millions of action videos, their annotations are generated by 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 Uneven Predictability

Not all frames are equally discriminative. As shown in [17], [58], 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 [2], [20], [29], [59] 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 [21], but the performance is still limited due to the insufficient discriminative information.

In addition, actions differ in their predictabilities [21], [47]. As shown in [21], 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 most important social skills we have, and human vision system provides a particularly rich source of information in support of this skill [60]. 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 [61]. There are a number of information we can tell from human actions, including the action categories [62], emotional implication [63], identity [64], [65], gender [66], [67], etc. Human visual system is finely optimized for the perception of human movements [68].

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 [69]. Ricoeur [70] 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 [6], [14], [18], [55], [71], [72], [73] and prediction [2], [20], [29], [74] 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 [70]. Few work in the computer vision community studies this question [75], [76].

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 outcome 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 [77], [78]. 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.

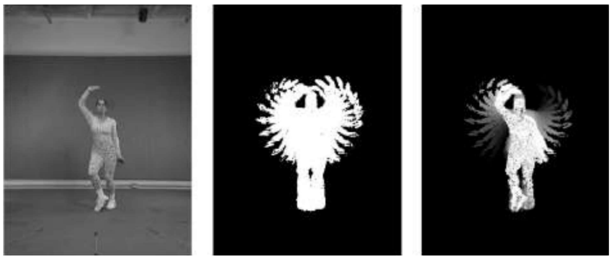


Fig. 5. Examples from [1] of an input video frame, the corresponding motion energy image and motion history image.

3 Action Recognition

A typical action recognition flowchart generally contains two major components [3], [23], [79], action representation and action classification. The action representation component basically converts an action video into a feature vector [11], [80], [81], [82] or a series of vectors [21], [54], [83], and the action classification component infers an action label from the vector [53], [84], [85]. Recently, deep networks [14], [18], [27] 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.

3.1.1.1 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 human body. Holistic representation methods capture the motion information of the entire human subject, providing rich and expressive motion information for action recognition. However, holistic 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.

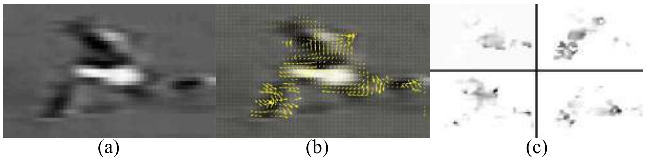


Fig. 6. Examples of (a) the original frame, (b) optical flow, and (c) flow field in four channels. Originally shown in [9].

One pioneering work in [1] presented Motion Energy Image (MEI) and Motion History Image (MHI) to encode dynamic human motion into a single image. As shown in Figure 5, the two methods work on the silhouettes. The MEI method shows "where" the motion is occurring: the spatial distribution of motion is represented and bright 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, [10] generalized [1] 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, [71], [86] 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 [87]. 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 [88], [89], [90], 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 horizontal and vertical axis. An early work by Efros *et al.* [9] split the flow field into four channels (see Figure 6) capturing the horizontal and vertical motion in successive frames. This method was then used in [91] to describe features of both human body and the body parts.

3.1.1.2 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 [50], [52], [92], [93] and motion trajectory [79], [94] have shown their robustness to translation, appearance variation, etc. Different from holistic features, local features describe local motion of a person in space-time regions. These regions are detected since the motion information within the regions are more informative and salient than surrounding areas. After detection, the regions are described by extracting features in the regions.

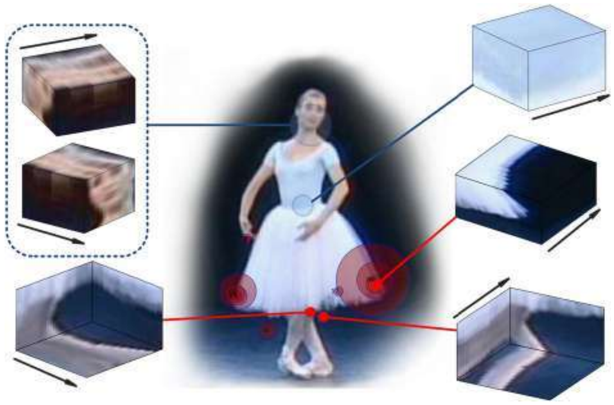


Fig. 7. Illustration of interest points detected on human body. Originally shown in [22].

Space-time interest points (STIPs) [11], [92]-based approaches is one of the most important local representations. Laptev's seminal work [11], [92] extended the Harris corner detector [95] 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 Figure 7). An alternative method was proposed in [80], which detects dense interest points. 2D Gaussian smoothing kernel is applied only along the spatial dimension, and the 1D Gabor filter is applied on 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 [3]. Bregonzio *et al.* [93] detected spatial-temporal interest points using Gabor filters. Spatiotemporal interest points can also be detected by using the spatiotemporal Hessian matrix [96]. Other detection algorithms detect spatiotemporal interest points by extending their counterparts of 2D detectors to spatiotemporal domains, such as 3D SIFT [82], HOG3D [52], local trinary patterns [97], 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) [49], and combined with HOG features [52], [98] to represent complex human activities [49], [52], [99]. Gradients over optical flow fields are computed to build the so-called motion boundary histograms (MBH) for describing trajectories [99].

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 [94], [100], [101]. To obtain features for trajectories, in [102], interest points are first detected and tracked using Harris3D interest points with a KLT tracker [88]. The method in [103] finds trajectories by matching corresponding SIFT points over consecutive frames. A 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 [79], [94], [104] (see Figure 8),

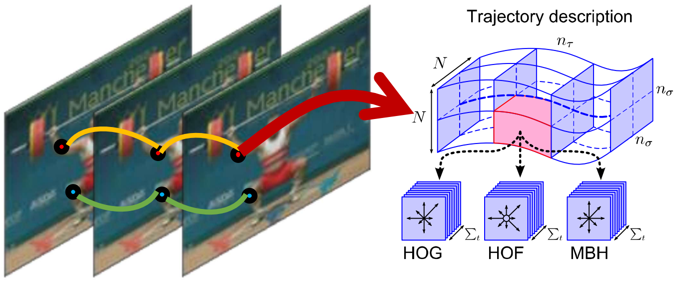


Fig. 8. Tracked point trajectories over frames, and are described by HOG, HOF and MBH features. Part of the figure was originally shown in [51].

intra- and inter-trajectory descriptors [103], or HOG/HOF and averaged descriptors [101]. In order to reduce the side effect of camera motion, [51], [81] 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:

3.1.2.1 Direct Classification: This type of approaches summarize an action video into a feature vector, and then directly recognize actions using off-the-shelf classifiers such as support vector machine [3], [6], [49], k-nearest neighbor (k-NN) [71], [105], [106], etc. In these methods, action dynamics is characterized in a holistic way using action shape [71], [86], or using the so-called bag-of-words model, which encodes the distribution of local motion patterns using a histogram of visual words [3], [6], [49], [71], [105].

In fact, bag-of-words approaches received lots of attentions in the last few years. As shown in Figure 9, these approaches first detect local salient regions using the spatiotemporal interest point detectors [3], [11], [50], [52]. Features such as gradient and optical flow are extracted around each 3D interest point. 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 [3], or Fisher vector [107]. Finally, an action can be represented by a histogram of visual words, and can be classified by a classifier such as support vector machine. The bag-of-words approaches have been shown to be insensitive to appearance and pose variations [100]. However, they do not consider the temporal characteristics in human actions, as well as the structural information of human actions, which can be addressed by sequential approaches [17], [85] and space-time approaches [4], respectively.

3.1.2.2 Sequential Approaches: This line of work captures temporal evolution of appearance or pose using sequential state models such as hidden Markov models (HMMs) [108], [109], [110], conditional random fields (CRFs) [83], [84], [111], [112] and structured support vector machine (SSVM) [13], [54], [85], [113]. These approaches treat a video as a composition of temporal segments or frames. The work in [108] considers human routine trajectory in a room, and use a two-layer HMMs to model the trajectory. Recent work in [17] 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

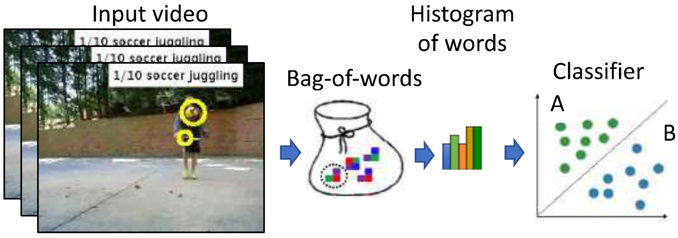


Fig. 9. A typical flowchart of the so-called bag-of-words methods. Local features detected on the input video are shown in yellow circles.

features from frames, which are sensitive to background noise and generally do not perform well on challenging datasets.

3.1.2.3 Space-time Approaches: Although direct approaches have shown promising results on some action datasets [3], [6], [49], 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 [114], 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 is proposed in [115] to capture the detailed geometrical distribution of interest points. The feature is calculated by using the transform which is defined as an extended 3D discrete Radon transform. Such feature captures the geometrical information of the interest points and keeps invariant to geometry transformation and robust to noise. The spatiotemporal distribution of interest points is described by a Directional Pyramid Co-occurrence Matrix in (DPCM) [116]. 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 is used in [117] 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) is proposed in [117] to measure the similarity between the two-graph models. Although the above methods have achieved promising results, they are limited to small datasets as they need to model the correlations between interest points which are explosive on large datasets.

3.1.2.4 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 [118], which models the position, appearance and velocity of body parts. Inspired by [118], a part-based hierarchical model was presented in [119], in which a part is generated by the model hypothesis and local visual words are generated from a body part (see Figure 10).

The method in [120] considers local visual words as parts, and models the structure information between parts. This work was further extended in [121], 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

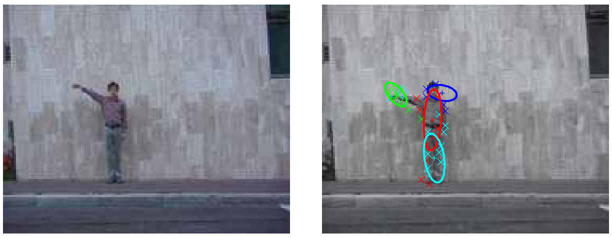


Fig. 10. Example of body parts detected by the constellation model in [119]. Originally shown in [119].

discriminative models for better classification performance [91], [122]. In [91], [122], a part is considered as a hidden variable in their models. It is corresponding to a salient region with the most positive energy.

3.1.2.5 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 [112], [123] to reduce the dimensionality of silhouette representation and embed them on nonlinear low-dimensional dynamic shape manifolds. The method in [112] adopts kernel PCA to perform dimensionality reduction, and discover the non-linear 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 [123], 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.

3.1.2.6 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 [53], [91], [124], [125] 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 Figure11) are popularly investigated in action recognition. These semantics are defined and further introduced into the activity classifiers in order to characterize complex human actions [53], [125], [126]. Other hierarchical approaches such as [17], [58] 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.

3.1.2.7 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 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 [127], in which

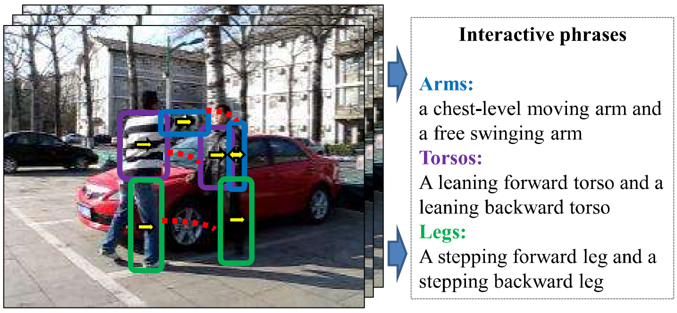


Fig. 11. Interaction recognition by learning semantic descriptions from videos. The figure was originally shown in [125].

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 [128] 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 In [129] to learn a high-level representation by combining a hierarchical generative model (HGM) and discriminative maxmargin classifiers in a unified Bayesian framework. HGM was proposed to represent actions by distributions over latent spatial temporal patterns (STPs) which are learned from multiple feature modalities and shared among different classes. This work was further extended in [130] to combine spatio 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 [4], [6], [12], [49], [131] 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 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 valuable information in human interaction recognition. In [132], 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 [133] 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 [134] to localize each individual. Spatial relationships between individuals is captured using the structured learning technique [135]. Spatiotemporal context of a group of people including human pose, velocity and spatiotemporal distribution of individuals is captured in [124] 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 [136]. A hierarchical representation of interactions is proposed in [136] that models atomic action, interaction, and collective action. The method in [137] also utilizes the idea of hierarchical representation, and studies the collective activity recognition problem using crowd context. Different from these methods, the work in [58] 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 [125], [126], a semantic description-based approach is proposed to represent complex human interactions by learned motion relationships (see Figure 11). 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 the ambiguities in feature-to-person assignments. To address this problem, a patch-aware model was proposed in [138] to learn discriminative patches for interaction recognition, and determine the assignments at a patch level.

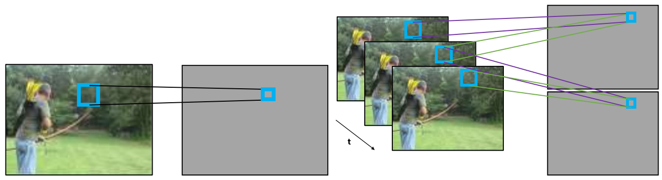
3.1.4 Classifiers for RGB-D Videos

Action recognition from RGB-D videos has been receiving a lot of attentions [35], [37], [41], [139], [140], [141] due to the advent of the cost-effective Kinect sensor [34]. 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 [37], [113], [141], [142], [143].

Effective features have been proposed for the recognition task using depth data, such as histogram of oriented 4D normals [36], [141] and depth spatiotemporal interest points [35], [37]. Features from depth sequences can be encoded by [144], or be used to build actionlets [113] for recognition. An efficient binary range-sample feature for depth data was proposed in [145]. This new type of depth feature has shown to be invariant to possible changes in scale, viewpoint, and background, and it is fast due to the binary property. The work in [146], [147] built layered action graph structures to model actions and subactions in a RGB-D video. Recent work [41] also showed that features of RGB-D data can be learned using deep learning techniques.

The methods in [36], [37], [139], [141], [144], [148] 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 [38], [40], [41], [113], [149], [150]. 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 [39], which learns extra discriminative information for classification, and demonstrate superior performance than [38], [40], [41], [113], [149], [150]. The methods in [38], [39] 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 [39], [113], [149], and has shown to be very effective in action recognition. The method in [149] projects various types of features including skeleton features and local HOG features into a shared feature space, which is learned by minimizing the reconstruction loss. Different from this work, the method in [39] jointly learns RGB-D and skeleton features and action classifiers. The projection matrices in [39] are learned by minimizing the noise after projection and classification error using the projected features. Using auxiliary databases to improve the recognition performance was studied in [40], [150]. Their methods assume actions can be reconstructed by entries in the auxiliary databases.



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

Fig. 12. Illustration of (a) 2D convolution and (b) 3D convolution.

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 ability of designing powerful features that can be generalized very well [14], [18], [30], [151]. 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 [14], [27], [152], [153] have achieved surprisingly high recognition performance on a variety of action datasets.

Action features learned by deep learning techniques has been popularly investigated [18], [30], [31], [154], [155], [156], [157], [158], [159], [160], [161], [162] in recent years. The two major variables in developing deep networks for action recognition are convolution operation and temporal modeling, leading to a few lines of networks.

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 (Figure 12(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 [31] presented a single-frame architecture based on 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 pretrained 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 (Figure 12(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 [14], [18], [156], [160]. 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.

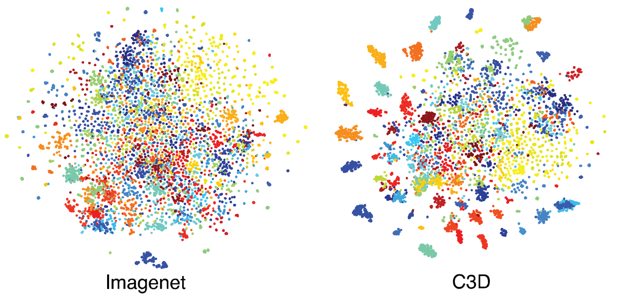


Fig. 13. Feature embedding by Imagenet and C3D. C3D features show better class separation than Imagenet, indicating its capability in learning better features for videos. Originall shown in [14].

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 [14], [18], [19], [156], [160]. As a result, the temporal dimension in the 3D convolution kernel will capture the temporal dynamics in these frames. One of the limitation of this type of 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 [19], [30], [153], [163], [164]. 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 pretrained for still images. The third category of approaches use temporal pooling [153], [164] or aggregation to capture temporal information in a video. The aggregation can be performed by using a LSTM model on top of 2D ConvNets [151], [165].

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 [160] 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 subsampling, and uses a fully-connected layer to generate a 128-dimensional feature vector for action classification. In a later extension [18], the authors regularized the network to encode long-term action information by encourage the network to learn feature vector close to high-level motion features such as bag-of-words representation of SIFT features.

The 3D ConvNet [18], [160] was later extended to a modern deep architecture called C3D [14] 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 Figure 13). Results showed that 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.

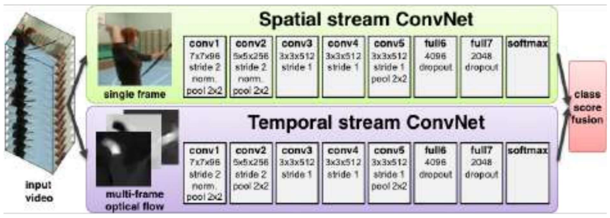


Fig. 14. Two-stream network proposed in [30] contains a spatial network and a temporal network, which are used for modeling static information in still frames and motion information in optical flow images, respectively. This figure was originally shown in [30].

Still, 3D ConvNets [14], [18], [160] for action recognition are relatively shallow with up to 8 layers. To further improve the generalization power of 3D ConvNets, [19] inflated very deep networks for image classification into spatio-temporal feature extractors by repeating 2D filters along time dimension, allowing the network to reuse 2D filters pretrained 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 [166], 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 [14], thereby failing to capture long-term temporal information. To address this problem, [152] increases the temporal extent in the 3D convolutions, and empirically shows that they can significantly improve the recognition performance.

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 [31] achieved great success, its results were significantly worse than those of the best hand-crafted shallow representations [79], [81]. To address this problem, a successful work by [30] explored a new architecture related to the two-stream hypothesis [167]. Their architecture contains two separate streams, a spatial ConvNet and a temporal ConvNet (see Figure 14). The former one learns actions from still images, and the later one performs recognition based on optical flow field.

The two-stream network [30] 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 [168], 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 [165], which is capable of making video-level predictions after the pooling layer. The work in [164] 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 [169] 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

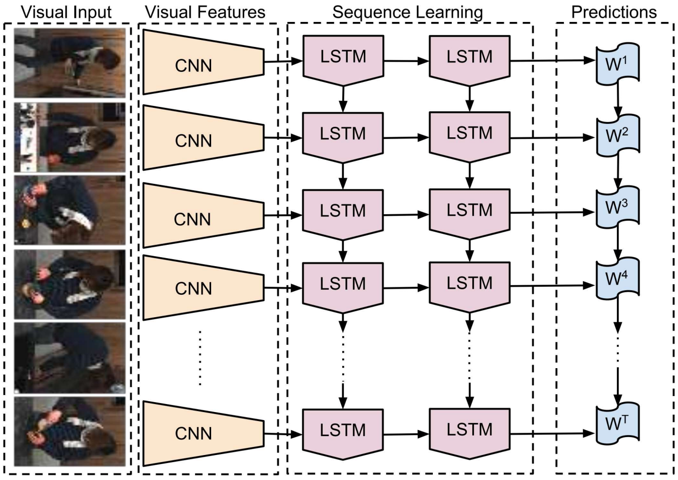


Fig. 15. Network architecture of LRCN [151] with a hybrid of ConvNets and LSTMs. Originally shown in [151].

and be encoded into a compact representation. AdaScan proposed in [153] 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 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 encodign methods lacks of considering spatio-temporal information. To address this problem, the work in [170] 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.

One of the major problems in the two-stream networks [30], [165], [168] 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.* [163] proposed a seriel 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 [27], [171], and allow a stream to be multiplicatively scaled by the opposing stream's input [27]. 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 [151], [165]. 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 [153], [168], [169].

Donahue *et al.* [151] explored the use of LSTM in modeling time series of frame features generated by 2D ConvNets. As shown in Figure 15, 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.* [165] 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 [172]. They used two-stream CNN [30] 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.

TABLE 1
Pros and cons of action recognition approaches.

	Approaches	Pros	Cons
Shallow	Direct [3], [51]	Easy and quick to use.	Performance is limited.
	Sequential [83], [85]	Models temporal evolution.	Sensitive to noise.
	Space-time [114], [117]	Captures spatiotemporal structures.	Limited to small datasets.
	Part-based [54], [91]	Models body parts at a finer level.	Limited to small datasets.
	Manifold [112], [123]	Tend to achieve high performance.	Rely on human silhouettes.
	Mid-level feature [53], [124]	Introduce knowledge to models.	Require extra annotations.
	Feature fusion [128], [130]	Tend to achieve high performance.	Slow in feature extraction.
Deep	Space-time [14], [18]	Natural extension of 2D convolution.	Short temporal interval.
	Multi-stream [27], [30]	Able to use pre-trained 2D ConvNets.	Int. b/w networks is difficult
	Hybrid [151], [165]	Easy to build using existing networks.	Difficult to fine-tune.

3.3 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 [51] have also shown promising results on large datasets, and thus they are still popularly used recently in the comparison with deep networks [14], [27], [152]. 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 [14] and TSN [26] that can be easily employed. It would be also helpful to try these methods and fine-tune the models to particular datasets. Table 1 summarizes the pros and cons of action recognition approaches.

4 Action Prediction and Motion 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 [26], [151], [164] 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.

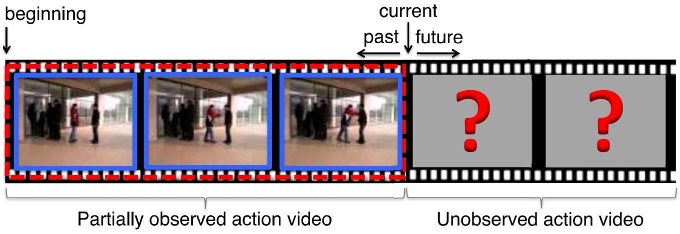


Fig. 16. Early action classification methods predicts action label given a partially observed video. Originally shown in [20].

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.

4.1 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 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.

4.1.1 *Early Action Classification*

This task aims at recognizing a human action at an early stage, i.e., based on a temporally incomplete video (see Figure 16). 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 [17], [24], [25], [58], 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.

Most of the short-term action prediction approaches follow the problem setup described in [20] shown in Figure 17. 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 $[(k\ddagger\ 1)\cdot\frac{T}{K}+1]$-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 $\frac{k}{K}$: $r = \frac{k}{K}$.

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.

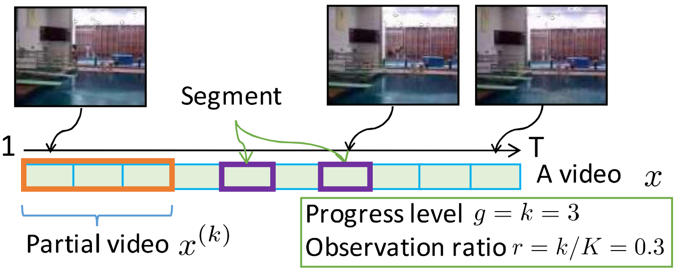


Fig. 17. Example of a temporally partial video, and graphical illustration of progress level and observation ratio. Originally shown in [21].

Action prediction approaches aim at recognizing unfinished action videos. Ryoo [2] proposed the integral bag-of-words (IBoW) and dynamic bag-of-words (DBoW) approaches for action prediction. The action model of each 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.* [74] built action models by learning feature bases using sparse coding and used the reconstruction error in the likelihood computation. Li *et al.* [173] explored long-duration action prediction problem. However, their work detects segments by motion velocity peaks, which may not be applicable on complex outdoor datasets. Compared with [2], [74], [173], [20] incorporates an important prior knowledge that informative action information is increasing when new observations are available. In addition, the method in [20] models label consistency of segments, which is not presented in their methods. From a perspective of interfering social interaction, Lan *et al.* [59] 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 [174] 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 [20], [29], [45]. Different from the aforementioned methods, [42] 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 [29] proposed a new monotonically decreasing loss function in learning LSTMs for action prediction. Inspired by that, we adopted an autoencoder to model sequential context information for action prediction [21]. Our 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

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. 18. 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 [21].

demonstrated that actions differ in their predictability, and show top 10 instantly, early, and late predictable actions in Figure 18. We also studied the action prediction problem following the popular two-stream framework [30]. In [28], we proposed to use memory to store hard-to-predict training samples in order to improve the prediction performance at early stage. The memory module used in [28] 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 beginning stage.

4.1.2 *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 as it allows us to understand "what is going to happen", including the final goal of a complex human action and the person's plausible intended action in the near future.

However, long-term action prediction is extremely challenging due to 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.

Pei *et al.* [46] 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.* [173] proposed a long-term action prediction method using Probabilistic Suffix Tree (PST), which captures variable Markov dependencies between action primitives in a complex action. For example, as shown in Figure 19, a wedding ceremony can be decomposed into primitives of "hold-hands", "kneel", "kiss", and "put-ring-on". In their extension [47], 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 on human action and object affordance was investigated in [43]. They proposed an anticipatory temporal conditional random field (ATCRF) to model three types of context information, including 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.

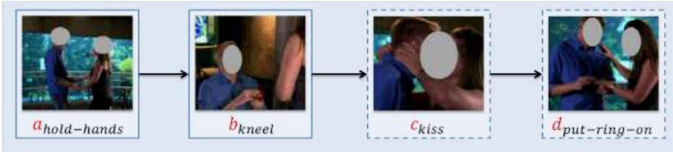


Fig. 19. A complex action can be decomposed into a series of action primitives. Originally shown in [47].

4.2 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.

Vision-based motion trajectory prediction is essential for practical applications such as visual surveillance and self-driving cars (see Figure 20), in which reasons about the future motion patterns of a pedestrian is critical. A large body of work learns motion patterns by clustering trajectories [16], [175], [176], [177]. 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* [178]. Pedestrians always respect personal space while walking, and thus yield 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 physical scene was investigated in [181], which was one of the pioneering work in trajectory prediction in the computer vision

Fig. 20. Motion trajectory prediction is essential for practical applications such as visual surveillance and self-driving cars. Originally shown in [179], [180].

community. The proposed method models the effect of the physical environment on the choice of human actions. This is accomplished by the use of state-of-the-art semantic scene understanding combined with ideas from inverse optimal control (IOC) or inverse reinforcement learning [182], [183]. 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 [184] extends [181] 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 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 [185], [186], which defines a target task space. The work in [187] introduced a large-scale dataset of 42 million trajectories and studied the problem of trajectory prediction by modeling 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 [188] 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 [189] are learned in [188], 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 dynamics of multiple objects [180], [190], [191], [192]. Kooji *et al.* [180] 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 [193] predicted the behavior of agents (e.g. a car) in a visual scene. Ziebart *et al.* [194] 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 [178], [179], [195], [196], which have the capability of generating long sequences. More specifically, a single LSTM model was 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 [195]. Compared to previous work [180], [181], [184], [187], [188], the method in [195] is end-to-end trainable, and generalizes well in complex scenes. An encoder-decoder framework was proposed in [196] 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) [181], [184] to a deep model, which has shown promising results in robot control [197] and driving [198] tasks. The proposed Deep IOC is used to rank all the possible hypotheses. Scene context is captured using a CNN model, which is part of the input to the RNN encoder. A Social-GAN network in [178] was proposed to address the limitation of L2 loss in [196]. Using an adversarial loss, [178] can potentially learn the distribution of multiple socially acceptable trajectories, rather than learning the average trajectories in the training data.

4.3 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 fig:actionpredictability demonstrate the predictabilties 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-term temporal correlations, and how to utilize multi-modal data to enrich the prediction model, which will be discussed in Section 7.

5 Datasets

This section discusses some of the popular action video datasets, including actions captured in controlled and uncontrolled environment. A detailed list is shown in Table 2. 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.

5.1 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.

5.1.1 Individual Action Datasets

Weizmann dataset [71] 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.

KTH dataset [3] 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 599 action videos in the dataset.

INRIA XMAS multiview dataset [199] 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.

TABLE 2
A list of popular action video datasets used in action recognition research.

Datasets	Year	#Videos	#Views	#Actions	#Subjects	#Modality	Env.
KTH [3]	2004	599	1	6	25	RGB	Controlled
Weizmann [71]	2005	90	1	10	9	RGB	Controlled
INRIA XMAS [199]	2006	390	5	13	10(3 times)	RGB	Controlled
IXMAS [200]	2006	1,148	5	11	-	RGB	Controlled
UCF Sports [5]	2008	150	-	10	-	RGB	Uncontrolled
Hollywood [49]	2008	-	-	8	-	RGB	Uncontrolled
Hollywood2 [6]	2009	3,669	-	12	10	RGB	Uncontrolled
UCF 11 [201]	2009	1,100+	-	11	-	RGB	Uncontrolled
CA [73]	2009	44	-	5	-	RGB	Uncontrolled
MSR-I [200]	2009	63	-	3	10	RGB	Controlled
MSR-II [202]	2010	54	-	3	-	RGB	Crowded
MHAV [203]	2010	238	8	17	14	RGB	Controlled
UT-I [204]	2010	60	2	6	10	RGB	Uncontrolled
TV-I [72]	2010	300	-	4	-	RGB	Uncontrolled
MSR-A [148]	2010	567	-	20	1	RGB-D	Controlled
Olympic [54]	2010	783	-	16	-	RGB	Uncontrolled
HMDB51 [55]	2011	7,000	-	51	-	RGB	Uncontrolled
CAD-60 [205]	2011	60	-	12	4	RGB-D	Controlled
BIT-I [126]	2012	400	-	8	50	RGB	Controlled
LIRIS [206]	2012	828	1	10	-	RGB	Controlled
MSRDA [140]	2012	320	-	16	10	RGB-D	Controlled
UCF50 [207]	2012	50	-	50	-	RGB	Uncontrolled
UCF101 [56]	2012	13,320	-	101	-	RGB	Uncontrolled
MSR-G [208]	2012	336	-	12	1	RGB-D	Controlled
UTKinect-A [7]	2012	10	-	10	-	RGB-D	Controlled
ASLAN [209]	2012	3,698	-	432	-	RGB	Uncontrolled
MSRAP [141]	2013	360	-	6 pairs	10	RGB-D	Controlled
CAD-120 [210]	2013	120	-	10	4	RGB-D	Controlled
Sports-1M [31]	2014	1,133,158	-	487	-	RGB	Uncontrolled
3D Online [211]	2014	567	-	20	-	RGB-D	Uncontrolled
FCVID [212]	2015	91,233	-	239	-	RGB	Uncontrolled
ActivityNet [213]	2015	28,000	-	203	-	RGB	Uncontrolled
YouTube-8M [57]	2016	8,000,000	-	4,716	-	RGB	Uncontrolled
Charades [214]	2016	9,848	2	157	-	RGB	Controlled
NEU-UB	2017	600	-	6	20	RGB-D	Controlled
Kinetics [215]	2017	500,000	-	600	-	RGB	Uncontrolled
AVA [216]	2017	57,600	-	80	-	RGB	Uncontrolled
20BN-Something-Something [217]	2017	108,499	-	174	-	RGB	Uncontrolled
SLAC [218]	2017	520,000	-	200	-	RGB	Uncontrolled
Moments in Time [219]	2017	1,000,000	-	339	-	RGB	Uncontrolled

5.1.2 Group Action Datasets

UT-Interaction dataset [204] is comprised of 2 sets with different environments. Each set consists of 6 types of human interactions: handshake, hug, kick, point, punch and push. Each type of interactions contains 10 videos, to provide 60 videos in total. Videos are captured at different scales and illumination conditions. Moreover, some irrelevant pedestrians are present in the videos.

BIT-Interaction dataset [126] 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 [72] 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.

5.2 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 Internet, and compiled large-scale action datasets, which will be discussed in the following.

UCF101 dataset [56] 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.

HMDB51 dataset [55] contains a total of about 7000 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.

Sports-1M dataset [31] 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.

20BN-SOMETHING-SOMETHING dataset [217] is a dataset shows human interaction with everyday objects. In the dataset, human performs pre-defined action with daily object. It contains $108,499$ video clips across 174 classes. The dataset enables the learning of visual representations for physical properties of the objects and the world.

Moments-in-Time dataset [219] 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.

5.3 RGB-D Action Video Datasets

All the datasets described above were captured by RGB video cameras. Recently, there is an increasing interests 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 [39]. Popular RGB-D action datasets are listed in the following.

MSR Daily Activity dataset [140]: 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 [211] 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 [210] 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 10 action types, such as rinsing mouth, talking on the phone, cooking, and writing on whiteboard. Tracked skeletons, RGB images, and depth images are provided in the dataset.

UTKinect-Action dataset [7] was captured by a Kinect device. There are 10 high-level action categories contained in the dataset, such as making cereal, taking medicine, stacking objects, and unstacking objects. Each high-level action can be comprised of 10 sub-activities such as reaching, moving, eating, and opening. 12 object affordable labels are also annotated in the dataset, including pourable, drinkable, and openable.

6 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 [3], [114], [119] were usually evaluated on small-scale datasets, for example, Weizmann dataset [71], KTH dataset [3], UCF Sports dataset [5]. Leave-one-out training scheme is popularly used on these datasets, and confusion matrix is usually adopted to show the recognition accuracy of each action category. For sequential approaches such as [91], [122], per-frame recognition accuracy is often used. In [6], [13], average precision that approximates the area under the precision-recall curve is also adopted for each individual action class. Deep networks [14], [19], [152] are generally evaluated on large-scale datasets such as UCF-101 [56] and HMDB51 [55] and thus can only report overall recognition performance on each dataset. Please refer to [22] for a list of performance of recent action recognition methods on various datasets.

Most of action prediction methods [2], [20], [21], [74] 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 [21], 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.

7 FUTURE DIRECTIONS

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

Benefitting from image models. Deep architectures are dominating the action recognition research lately like the trend of other developments in 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 on 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 [19] or domain adaptation [221].

Interpretability on temporal extent. Interpretability of image models has been discussed but it has not been extensively discussed in video models. As shown in [17], [222], 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 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 an 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

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
Dynamic BoW [2]	2011	22.66%	46.88%	25.00%	51.65%	36.29%	53.16%	43.47%	45.46%
Integral BoW [2]	2011	22.66%	48.44%	18.00%	48.00%	36.29%	74.39%	43.47%	55.99%
MSSC [74]	2013	21.09%	48.44%	28.00%	70.00%	34.05%	61.79%	46.70%	57.16%
Poselet [17]	2013	-	-	-	73.33%	-			
HM [59]	2014	-	-	38.33%	83.10%	-		-	
MTSSVM [20]	2014	28.12%	60.00%	36.67%	78.33%	40.05%	82.39%	49.92%	66.90%
MMAPM [45]	2016	32.81%	67.97%	46.67%	78.33%	-	-	-	
DeepSCN [21]	2017	37.50%	78.13%	-	-	45.02%	85.75%	55.02%	70.23%
GLTSD [220]	2018	26.60%	79.40%	-	-	-		-	-
mem-LSTM [28]	2018	-	-	-	-	51.02%	88.37%	57.60%	71.63%

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 goanna 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 interests in learning physical aspects of the world, which studies fine-grained actions. One example is the something-something dataset introduced in [217] 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 human 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 [223] 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 labels. For increasingly large action datasets such as Something-Something [217] and Sports-1M [31], manual labeling becomes prohibitive. Automatic labeling using search engines [31], [57], video subtitles and movie scripts [6], [49] is possible in some domains, but still requires manual verification. Crowdsourcing [217] would be a better option but still suffers from labeling diversity problem, and may generate incorrect action labels. This prompts us to develop more robust and efficient action recognition/prediction approaches that can automatically learn from unlabeled videos [224].

8 CONCLUSION

The availability of big data and powerful models diverts the research focus about 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

[1] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[2] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV*, 2011.

[3] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *IEEE ICPR*, 2004.

[4] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, 2009, pp. 1593–1600.

[5] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: A spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[6] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[7] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[8] S. Singh, S. A. Velastin, and H. Ragheb, "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*. IEEE, 2010, pp. 48–55.

[9] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *ICCV*, vol. 2, 2003, pp. 726 –733.

[10] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[11] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2, pp. 107–123, 2005.

[12] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[13] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *CVPR*, 2012.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[15] A. Ciptadi, M. S. Goodwin, and J. M. Rehg, "Movement pattern histogram for action recognition and retrieval," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 695–710.

[16] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *Image Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 1168–1181, 2007.

[17] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *CVPR*, 2013.

[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2013.

[19] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[20] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *ECCV*, 2014.

[21] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *CVPR*, 2017.

[22] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing*, 2017.

[23] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.

[24] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *TPAMI*, vol. 34, no. 9, pp. 1691–1703, 2012.

[25] ——, "Action recognition with exemplar based 2.5d graph matching," in *ECCV*, 2012.

[26] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temoral segment networks: Toward good practices for deep action recognition," in *ECCV*, 2016.

[27] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 7445–7454.

[28] Y. Kong, S. Gao, B. Sun, and Y. Fu, "Action prediction from videos via memorizing hard-to-predict samples," in *AAAI*, 2018.

[29] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *CVPR*, 2016.

[30] K. Simonyan and A. Zisserman, "two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[31] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[32] M. Ramezani and F. Yaghmaee, "A review on human action analysis in videos for retrieval applications," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 485–514, 2016.

[33] X. Zhai, Y. Peng, and J. Xiao, "Cross-media retrieval by intra-media and inter-media correlation mining," *Multimedia Systems*, vol. 19, no. 5, pp. 395–406, 2013.

[34] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *PAMI*, 2013.

[35] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *CVPR*, 2013.

[36] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014.

[37] S. Hadfield and R. Bowden, "Hollywood 3d: Recognizing actions in 3d natural scenes," in *CVPR*, Portland, Oregon, 2013.

[38] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for rgb-d action recognition," in *CVPR*, 2015.

[39] ——, "Max-margin heterogeneous information machine for rgb-d action recognition," *International Journal of Computer Vision (IJCV)*, vol. 123, no. 3, pp. 350–371, 2017.

[40] C. Jia, Y. Kong, Z. Ding, and Y. Fu, "Latent tensor transfer learning for rgb-d action recognition," in *ACM Multimedia*, 2014.

[41] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data," in *IJCAI*, 2013.

[42] M. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies, "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*. ACM, 2015, pp. 295–302.

[43] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[44] M. Ryoo and J. Aggarwal, "Stochastic representation and recognition of high-level group activities," *IJCV*, vol. 93, pp. 183–200, 2011.

[45] Y. Kong and Y. Fu, "Max-margin action prediction machine," *TPAMI*, vol. 38, no. 9, pp. 1844 – 1858, 2016.

[46] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *ICCV*. IEEE, 2011, pp. 487–494.

[47] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1644–1657, Aug 2014.

[48] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[49] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[50] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005.

[51] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013. [Online]. Available: http://hal.inria.fr/hal-00873267

[52] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008.

[53] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *CVPR*, 2011.

[54] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, 2010.

[55] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011.

[56] A. R. Z. Khurram Soomro and M. Shah, "Ucf101: A dataset of 101 human action classes from videos in the wild," 2012, cRCV-TR-12-01.

[57] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[58] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori, "A discriminative key pose sequence model for recognizing human interactions," in *ICCV Workshops*, 2011, pp. 1729 –1736.

[59] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *European Conference on Computer Vision*. Springer, 2014, pp. 689–704.

[60] R. Blake and M. Shiffrar, "Perception of human motion," *Annu. Rev. Psychol.*, vol. 58, pp. 47–73, 2007.

[61] C. Darwin, *The Expression of the Emotions in Man and Animals*. London: John Murray, 1872.

[62] J. Mass, G. Johansson, G. Jason, and S. Runeson, *Motion perception I and II [film]*. Boston: Houghton Mifflin, 1971.

[63] T. Clarke, M. Bradshaw, D. Field, S. Hampson, and D. Rose, "The perception of emotion from body movement in point-light displays of interpersonal dialogue," *Perception*, vol. 24, pp. 1171–80, 2005.

[64] J. Cutting and L. Kozlowski, "Recognition of friends by their work: gait perception without familarity cues," *Bull. Psychon. Soc.*, vol. 9, pp. 353–56, 1977.

[65] N. Troje, C. Westhoff, and M. Lavrov, "Person identification from biological motion: effects of structural and kinematic cues," *Percept. Psychophys*, vol. 67, pp. 667–75, 2005.

[66] S. Sumi, "Perception of point-light walker produced by eight lights attached to the back of the walker," *Swiss J. Psychol.*, vol. 59, pp. 126–32, 2000.

[67] N. Troje, "Decomposing biological motion: a framework for analysis and synthesis of human gait patterns," *J. Vis.*, vol. 2, pp. 371–87, 2002.

[68] J. Decety and J. Grezes, "Neural mechanisms subserving the perception of human actions," *Neural mechanisms of perception and action*, vol. 3, no. 5, pp. 172–178, 1999.

[69] M. Keestra, "Understanding human action. integraiting meanings, mechanisms, causes, and contexts," *TRANSDISCIPLINARITY*

IN PHILOSOPHY AND SCIENCE: APPROACHES, PROBLEMS, PROSPECTS, pp. 201–235, 2015.

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

[71] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. ICCV*, 2005.

[72] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in tv shows," in *Proc. British Conference on Machine Vision*, 2010.

[73] W. Choi, K. Shahid, and S. Savarese, "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*, 2009, pp. 1282 –1289.

[74] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Siskind, and S. Wang, "Recognizing human activities from partially observed videos," in *CVPR*, 2013.

[75] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity," in *CVPR*, 2012.

[76] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *CVPR*, 2013.

[77] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, pp. 169–192, 2004.

[78] G. Rizzolatti and C. Sinigaglia, "The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations," *Nat. Rev. Neurosci.*, vol. 11, pp. 264–274, 2010.

[79] H. Wang, A. Kla aser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 60-79, 2013.

[80] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *ICCV VS-PETS*, 2005.

[81] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *IJCV*, 2015.

[82] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. ACM Multimedia*, 2007.

[83] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *CVPR*, 2007.

[84] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *International Conference on Computer Vision*, 2005.

[85] Q. Shi, L. Cheng, L. Wang, and A. Smola, "Human action segmentation and recognition using discriminative semi-markov models," *IJCV*, vol. 93, pp. 22–32, 2011.

[86] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[87] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *CVPR*, 2005.

[88] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of Imaging Understanding Workshop*, 1981.

[89] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[90] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *CVPR*, 2010.

[91] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic vs. max-margin," *PAMI*, 2010.

[92] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, 2003, pp. 432–439.

[93] M. Bregonzio, S. Gong, and T. Xiang, "Recognizing action as clouds of space-time interest points," in *CVPR*, 2009.

[94] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176. [Online]. Available: http://hal.inria.fr/inria-00583818/en

[95] C. Harris and M. Stephens., "A combined corner and edge detector," in *Alvey Vision Conference*, 1988.

[96] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest poing detector," in *ECCV*, 2008.

[97] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *CVPR*, 2009.

[98] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[99] H. Wang, M. M. Ullah, A. Kl aser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2008.

[100] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[101] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *ECCV*, 2010.

[102] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV*, 2009.

[103] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *CVPR*, 2009.

[104] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *CVPR*, 2013.

[105] I. Laptev and P. Perez, "Retrieving actions in movies," in *ICCV*, 2007.

[106] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *ECCV*, 2008.

[107] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2006.

[108] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-markov model," in *CVPR*, 2005.

[109] S. Rajko, G. Qian, T. Ingalls, and J. James, "Real-time gesture recognition with minimal training requirements and on-line learning," in *CVPR*, 2007.

[110] N. Ikizler and D. Forsyth, "Searching video for complex activities with finite state models," in *CVPR*, 2007.

[111] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *CVPR*, 2006.

[112] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *CVPR*, 2007.

[113] Z. Wang, J. Wang, J. Xiao, K.-H. Lin, and T. S. Huang, "Substructural and boundary modeling for continuous action recognition," in *CVPR*, 2012.

[114] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *CVPR*, 2011.

[115] C. Yuan, X. Li, W. Hu, H. Ling, and S. J. Maybank, "3d r transform on spatio-temporal interest points for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 724–730.

[116] ——, "Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 658–672, 2014.

[117] B. Wu, C. Yuan, and W. Hu, "Human action recognition based on context-dependent graph kernels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2609–2616.

[118] C. Fanti, L. Zelnik-Manor, and P. Perona, "Hybrid models for human motion recognition," in *CVPR*, 2005.

[119] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *CVPR*, 2007.

[120] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *CVPR*, 2007.

[121] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[122] Y. Wang and G. Mori, "Learning a discriminative hidden part model for human action recognition," in *NIPS*, 2008.

[123] K. Jia and D.-Y. Yeung, "Human action recognition using local spatio-temporal discriminant embedding," in *CVPR*, 2008.

[124] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011.

[125] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," in *PAMI*, 2014.

[126] ——, "Learning human interaction by interactive phrases," in *Proc. European Conf. on Computer Vision*, 2012.

[127] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. J. Maybank, "Learning human actions by combining global dynamics and local appearance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2466–2482, 2014.

[128] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 423–429.

[129] S. Yang, C. Yuan, B. Wu, W. Hu, and F. Wang, "Multi-feature max-margin hierarchical bayesian model for action recognition," in *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1610–1618.

[130] C. Yuan, B. Wu, X. Li, W. Hu, S. J. Maybank, and F. Wang, "Fusing r features and local features with context-aware kernels for action recognition," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 151–171, 2016.

[131] T.-H. Yu, T.-K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forests," in *BMVC*, 2010.

[132] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *PAMI*, vol. 22, no. 8, pp. 831–843, 2000.

[133] M. Ryoo and J. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *CVPR*, vol. 2, 2006, pp. 1709–1718.

[134] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zissermann, "Structured learning of human interaction in tv shows," *PAMI*, vol. 34, no. 12, pp. 2441–2453, 2012.

[135] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[136] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*. Springer, 2012, pp. 215–230.

[137] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *TPAMI*, vol. 34, no. 8, pp. 1549–1562, 2012.

[138] Y. Kong and Y. Fu, "Modeling supporting regions for close human interaction recognition," in *ECCV workshop*, 2014.

[139] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *ECCV*, 2012.

[140] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[141] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013.

[142] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *ICCV Workshop on CDC3CV*, 2011.

[143] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Proceedings of the IEEE Workshop on Applications on Computer Vision*, 2013.

[144] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *ICCV*, 2013.

[145] C. Lu, J. Jia, and C.-K. Tang, "Range-sample depth feature for action recognition," in *CVPR*, 2014.

[146] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *ICRA*, 2012.

[147] H. S. Koppula and A. Saxena, "Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation," in *ICML*, 2013.

[148] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPR workshop*, 2010.

[149] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *CVPR*, 2015.

[150] Y.-Y. Lin, J.-H. Hua, N. C. Tang, M.-H. Chen, and H.-Y. M. Liao, "Depth and skeleton associated action recognition without online accessible rgb-d cameras," in *CVPR*, 2014.

[151] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[152] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[153] A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *CVPR*, 2017.

[154] Y. Yang and M. Shah, "Complex events detection using data-driven concepts," in *ECCV*, 2012.

[155] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *ACM Multimedia*, 2014.

[156] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *ECCV*, 2010.

[157] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, "Dl-sfa: Deeply-learned slow feature analysis for action recognition," in *CVPR*, 2014.

[158] T. Pl otz, N. Y. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *IJCAI*, 2011.

[159] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *CVPR*, 2011.

[160] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," in *ICML*, 2010.

[161] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *ECCV*, 2014.

[162] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

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

[164] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Action-vlad: Learning spatio-temporal aggregation for action classification," in *CVPR*, 2017.

[165] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015.

[166] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual network," in *ICCV*, 2017.

[167] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[168] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015.

[169] A. Diba, V. Sharma, and L. V. Gool, "Deep temporal linear encoding networks," in *CVPR*, 2017.

[170] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe, "spatio-temporal vector of locally max pooled features for action recognition in videos," in *CVPR*, 2017.

[171] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition," in *NIPS*, 2016.

[172] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM Multimedia*, 2015.

[173] K. Li, J. Hu, and Y. Fu, "Modeling complex temporal composition of actionlets for activity prediction," in *ECCV*, 2012.

[174] M. Hoai and F. D. la Torre, "Max-margin early event detectors," in *CVPR*, 2012.

[175] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *CVPR*, 2011.

[176] B. Morrisand and M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2287–2301, 2011.

[177] K. Kim, D. Lee, and I. Essa, "Gaussian process regression flow for analysis of motion trajectories," in *ICCV*, 2011.

[178] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[179] H. Su, J. Zhu, Y. Dong, and B. Zhang, "Forecast the plausible paths in crowd scenes," in *IJCAI*, 2017.

[180] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *European Conference on Computer Vision*. Springer, 2014, pp. 618–633.

[181] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, 2012.

[182] P. Abbeel and A. Ng, "Apprenticeship learning via inverse reinforcement learning," in *ICML*, 2004.

[183] B. Ziebart, A. Maas, J. Bagnell, and A. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI*, 2008.

[184] N. Lee and K. M. Kitani, "Predicting wide receiver trajectories in american football," in *WACV2016*.

[185] J. Mainprice, R. Hayne, and D. Berenson, "Goal set inverse optimal control and iterative re-planning for predicting human reaching motions in shared workspace," in *arXiv preprint arXiv:1606.02111*, 2016.

[186] A. Dragan, N. Ratliff, and S. Srinivasa, "Manipulation planning with goal sets using constrained trajectory optimization," in *ICRA*, 2011.

[187] A. Alahi and V. R. L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *CVPR*, 2014.

[188] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *ECCV*, 2016.

[189] M. Turek, A. Hoogs, and R. Collins, "Unsupervised learning of functional categories in video scenes," in *ECCV*, 2010.

[190] D. F. Fouhey and C. L. Zitnick, "Predicting object dynamics in scenes," in *CVPR*, 2014.

[191] D. A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *ECCV*, 2008.

[192] H. Kretzschmar, M. Kuderer, and W. Burgard, "Learning to predict trajecteories of cooperatively navigation agents," in *International Conference on Robotics and Automation*, 2014.

[193] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3302–3309.

[194] B. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. Bagnell, M. Hebert, A. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *IROS*, 2009.

[195] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.

[196] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *CVPR*, 2017.

[197] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: deep inverse optimal control via policy optimization," in *arXiv preprint arXiv:1603.00448*, 2016.

[198] M. Wulfmeier, D. Wang, and I. Posner, "Watch this: scalable cost function learning for path planning in urban environment," in *arXiv preprint arXiv:1607:02329*, 2016.

[199] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, 2006.

[200] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume search for efficient action detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[201] J. L. Jingen Liu and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *CVPR*, 2009.

[202] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[203] S. V. S Singh and H. Ragheb, "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)*, 2010, pp. 48–55.

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

[205] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images," in *AAAI workshop on Pattern, Activity and Intent Recognition*, 2011.

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

[207] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications Journal*, 2012.

[208] A. Kurakin, Z. Zhang, and Z. Liu, "A real-time system for dynamic hand gesture recognition with a depth sensor," in *EUSIPCO*, 2012.

[209] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, 2012.

[210] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *International Journal of Robotics Research*, 2013.

[211] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *ACCV*, 2014.

[212] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 352–364, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2670560

[213] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[214] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

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

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

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

[218] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba, "Slac: A sparsely labeled dataset for action classification and localization," *arXiv preprint arXiv:1712.09374*, 2017.

[219] M. Monfort, B. Zhou, S. A. Bargal, T. Yan, A. Andonian, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding."

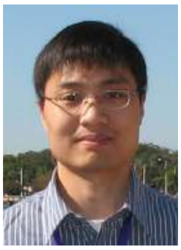
[220] S. Lai, W.-S. Zhang, J.-F. Hu, and J. Zhang, "Global-local temporal saliency action prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2272–2285, 2018.

[221] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Advances in Neural Information Processing Systems*, 2012.

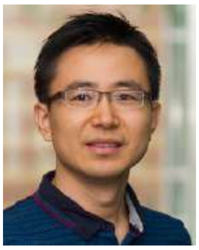
[222] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *ECCV*, 2010.

[223] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," in *Advances in Neural Information Processing Systems*, 2015, pp. 127–135.

[224] J. Hou, X. Wu, J. Chen, J. Luo, and Y. Jia, "Unsupervised deep learning of mid-level video representation for action recognition," in *AAAI*, 2018.



Yu Kong received B.Eng. degree in automation from Anhui University in 2006, and PhD degree in computer science from Beijing Institute of Technology, China, in 2012. He was a visiting student at the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Science from 2007 to 2009. He is now a postdoctoral research associate in the Electrical and Computer Engineering, Northeastern University, Boston, MA. Dr. Kong's research interests include computer vision, social media analytics, and machine learning. He is a member of the IEEE.



Yun Fu (S'07-M'08-SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. Prior to joining the Northeastern faculty, he was a tenure-track Assistant Professor of the Department of Computer Science and Engineering, State University of New York, Buffalo, during 2010-2012. His research interests are Machine Learning, Computer Vision, Social Media Analytics, and Big Data Mining. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He is the recipient of 5 best paper awards (SIAM SDM 2014, IEEE FG 2013, IEEE ICDM-LSVA 2011, IAPR ICFHR 2010, IEEE ICIP 2007), 3 young investigator awards (2014 ONR Young Investigator Award, 2014 ARO Young Investigator Award, 2014 INNS Young Investigator Award), 2 service awards (2012 IEEE TCSVT Best Associate Editor, 2011 IEEE ICME Best Reviewer), etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS), and IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He is a Lifetime Member of ACM, AAAI, SPIE, and Institute of Mathematical Statistics, member of INNS and Beckman Graduate Fellow during 2007-2008.