(To appear. ACM Computing Surveys.)

Human Activity Analysis: A Review

J. K. Aggarwal[1] and M. S. Ryoo[1,2]
[1]The University of Texas at Austin
[2]Electronics and Telecommunications Research Institute

Human activity recognition is an important area of computer vision research. Its applications include surveillance systems, patient monitoring systems, and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces. Most of these applications require an automated recognition of high-level activities, composed of multiple simple (or atomic) actions of persons. This paper provides a detailed overview of various state-of-the-art research papers on human activity recognition. We discuss both the methodologies developed for simple human actions and those for high-level activities. An approach-based taxonomy is chosen, comparing the advantages and limitations of each approach.

Recognition methodologies for an analysis of simple actions of a single person are first presented in the paper. Space-time volume approaches and sequential approaches that represent and recognize activities directly from input images are discussed. Next, hierarchical recognition methodologies for high-level activities are presented and compared. Statistical approaches, syntactic approaches, and description-based approaches for hierarchical recognition are discussed in the paper. In addition, we further discuss the papers on the recognition of human-object interactions and group activities. Public datasets designed for the evaluation of the recognition methodologies are illustrated in our paper as well, comparing the methodologies' performances. This review will provide the impetus for future research in more productive areas.

Categories and Subject Descriptors: I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*motion*; I.4.8 [**Image Processing**]: Scene Analysis; I.5.4 [**Pattern Recognition**]: Applications—*computer vision*

General Terms: Algorithms

Additional Key Words and Phrases: computer vision; human activity recognition; event detection; activity analysis; video recognition

1. INTRODUCTION

Human activity recognition is an important area of computer vision research today. The goal of human activity recognition is to automatically analyze ongoing activities from an unknown video (i.e. a sequence of image frames). In a simple case where a video is segmented to contain only one execution of a human activity, the objective

This work was supported partly by Texas Higher Education Coordinating Board under award no. 003658-0140-2007.
Authors' addresses: J. K. Aggarwal, Computer and Vision Research Center, Department of Electrical and Computer Engineering, the University of Texas at Austin, Austin, TX 78705, U.S.A.; M. S. Ryoo, Robot Research Department, Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea; Correspondence e-mail: mryoo@etri.re.kr

of the system is to correctly classify the video into its activity category. In more general cases, the continuous recognition of human activities must be performed, detecting starting and ending times of all occurring activities from an input video.

The ability to recognize complex human activities from videos enables the construction of several important applications. Automated surveillance systems in public places like airports and subway stations require detection of abnormal and suspicious activities as opposed to normal activities. For instance, an airport surveillance system must be able to automatically recognize suspicious activities like 'a person leaving a bag' or 'a person placing his/her bag in a trash bin'. Recognition of human activities also enables the real-time monitoring of patients, children, and elderly persons. The construction of gesture-based human computer interfaces and vision-based intelligent environments becomes possible as well with an activity recognition system.

There are various types of human activities. Depending on their complexity, we conceptually categorize human activities into four different levels: gestures, actions, interactions, and group activities. Gestures are elementary movements of a person's body part, and are the atomic components describing the meaningful motion of a person. 'Stretching an arm' and 'raising a leg' are good examples of gestures. Actions are single person activities that may be composed of multiple gestures organized temporally, such as 'walking', 'waving', and 'punching'. Interactions are human activities that involve two or more persons and/or objects. For example, 'two persons fighting' is an interaction between two humans and 'a person stealing a suitcase from another' is a human-object interaction involving two humans and one object. Finally, group activities are the activities performed by conceptual groups composed of multiple persons and/or objects. 'A group of persons marching', 'a group having a meeting', and 'two groups fighting' are typical examples of them.

The objective of this paper is to provide a complete overview of state-of-the-art human activity recognition methodologies. We discuss various types of approaches designed for the recognition of different levels of activities. The previous review written by Aggarwal and Cai [1999] has covered several essential low-level components for the understanding of human motion, such as tracking and body posture analysis. However, the motion analysis methodologies themselves were insufficient to describe and annotate ongoing human activities with complex structures, and most of approaches in 1990s focused on the recognition of gestures and simple actions. In this new review, we concentrates on high-level activity recognition methodologies designed for the analysis of human actions, interactions, and group activities, discussing recent research trends in activity recognition.

Figure 1 illustrates an overview of the tree-structured taxonomy that our review follows. We have chosen an approach-based taxonomy. All activity recognition methodologies are first classified into two categories: single-layered approaches and hierarchical approaches. Single-layered approaches are approaches that represent and recognize human activities directly based on sequences of images. Due to their nature, single-layered approaches are suitable for the recognition of gestures and actions with sequential characteristics. On the other hand, hierarchical approaches represent high-level human activities by describing them in terms of other simpler activities, which they generally call *sub-events*. Recognition systems composed of

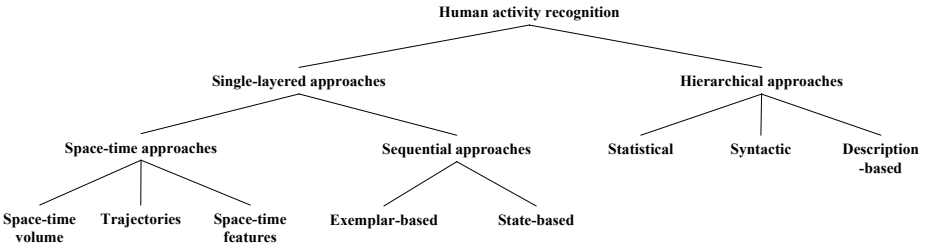


Fig. 1. The hierarchical approach-based taxonomy of this review.

multiple layers are constructed, making them suitable for the analysis of complex activities.

Single-layered approaches are again classified into two types depending on how they model human activities: space-time approaches and sequential approaches. Space-time approaches view an input video as a 3-dimensional (XYT) volume while sequential approaches interpret it as a sequence of observations. Space-time approaches are further divided into three categories based on what features they use from the 3-D space-time volumes: volumes themselves, trajectories, or local interest point descriptors. Sequential approaches are classified depending on whether they use exemplar-based recognition methodologies or model-based recognition methodologies. Figure 2 shows a detailed taxonomy used for single-layered approaches covered in the review, together with a number of publications corresponding to each category.

Hierarchical approaches are classified based on the recognition methodologies they use: statistical approaches, syntactic approaches, and description-based approaches. Statistical approaches construct statistical state-based models concatenated hierarchically (e.g. layered hidden Markov models) to represent and recognize high-level human activities. Similarly, syntactic approaches use a grammar syntax such as stochastic context-free grammar (SCFG) to model sequential activities. Essentially, they are modeling a high-level activity as a string of atomic-level activities. Description-based approaches represent human activities by describing sub-events of the activities and their temporal, spatial, and logical structures. Figure 3 presents lists of representative publications corresponding to categories.

In addition, in Figures 2 and 3, we have indicated previous works that recognize human-object interactions and group activities by using different colors and by attaching 'O' (object) and 'G' (group) tags to the right-hand side. The recognition of human-object interactions requires the analysis of interplays between object recognition and activity analysis. This paper provides a survey on the methodologies focusing on the analysis of such interplays for the improved recognition of human activities. Similarly, the recognition of groups and the analysis of their structures is necessary for group activity detection, and we cover them as well in this review.

This review paper is organized as follows: Section 2 covers single-layered approaches. In Section 3, we review hierarchical recognition approaches for the analysis of high-level activities. Subsection 4.1 discusses recognition methodologies for interactions between humans and objects, while especially concentrating on how

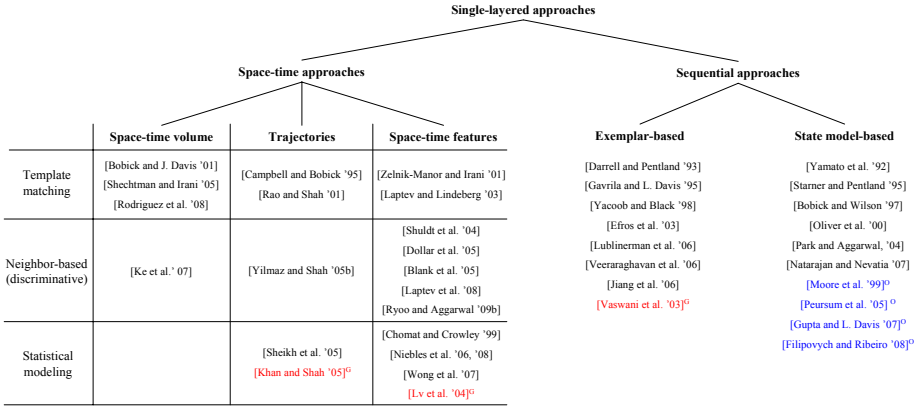


Fig. 2. Detailed taxonomy for single-layered approaches and the lists of selected publications corresponding to each category.

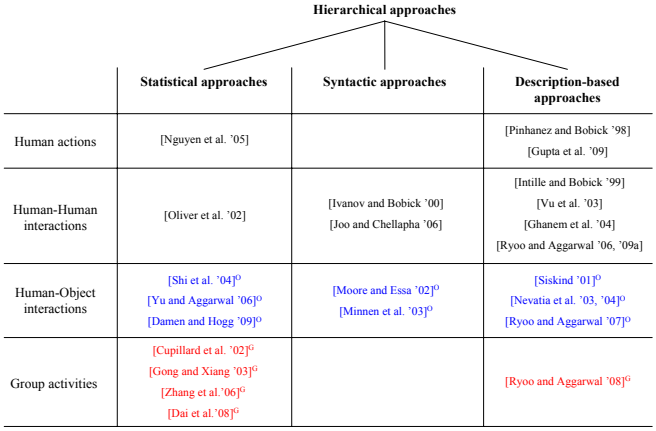


Fig. 3. Detailed taxonomy for hierarchical approaches and the lists of publications corresponding to each category.

previous works handled interplays between object recognition and motion analysis. Subsection 4.2 presents works on group activity recognition. In Subsection 5.1, we review public datasets available and compare systems tested on them. In addition, Subsection 5.2 covers real-time systems for human activity recognition. Section 6 concludes the paper.

1.1 Comparison with previous review papers

There have been other related surveys on human activity recognition. Several previous reviews on human motion analysis [Cedras and Shah 1995; Gavrila 1999; Aggarwal and Cai 1999] discussed human action recognition approaches as a part of their review. Kruger et al. [2007] reviewed human action recognition approaches while classifying them based on the complexity of features involved in the action

recognition process. Their review especially focused on the planning aspect of human action recognitions, considering their potential application to robotics. Turaga et al. [2008]'s survey covered human activity recognition approaches, similar to ours. In their paper, approaches are first categorized based on the complexity of the activities that they want to recognize, and then classified in terms of the recognition methodologies they use.

However, most of the previous reviews have focused on the introduction and summarization of activity recognition methodologies, and are lacking in the aspect of comparing different types of human activity recognition approaches. In this review, we present inter-class and intra-class comparisons between approaches, while providing an overview of human activity recognition approaches categorized based on the approach-based taxonomy presented above. Comparisons among abilities of recognition methodologies are essential for one to take advantage of them. Our goal is to enable a reader (even who is from a different field) to understand the context of human activity recognition's developments, and comprehend advantages and disadvantages of different approach categories.

We use a more elaborate taxonomy and compare and contrast each approach category in detail. For example, differences between single-layered approaches and hierarchical approaches are discussed in the highest-level of our review, while space-time approaches are compared with sequential approaches in an intermediate level. We present a comparison among abilities of previous systems within each class as well, pointing out what they are able to recognize and what they are not. Furthermore, our review covers recognition methodologies for complex human activities including human-object interactions and group activities, which previous reviews have not focused on. Finally, we discuss the public datasets used by the systems, and compare the recognition methodologies' performances on the datasets.

2. SINGLE-LAYERED APPROACHES

Single-layered approaches recognize human activities directly from video data. These approaches consider an activity as a particular class of image sequences, and recognize the activity from an unknown image sequence (i.e. an input) by categorizing it into its class. Various representation methodologies and matching algorithms have been developed to enable the recognition system to make an accurate decision whether an image sequence belongs to a certain activity class or not. For the recognition from continuous videos, most single-layered approaches have adopted a sliding windows technique that classifies all possible sub-sequences. Single-layered approaches are most effective when a particular sequential pattern describing an activity can be captured from training sequences. Due to their nature, the main objective of the single-layered approaches has been to analyze relatively simple (and short) sequential movements of humans, such as walking, jumping, and waving.

In this review, we categorize single-layered approaches into two classes: space-time approaches and sequential approaches. Space-time approaches model a human activity as a particular 3-D volume in a space-time dimension or a set of features extracted from the volume. The video volumes are constructed by concatenating image frames along a time axis, and are compared to measure their similarities. On the other hand, sequential approaches treat a human activity as a sequence

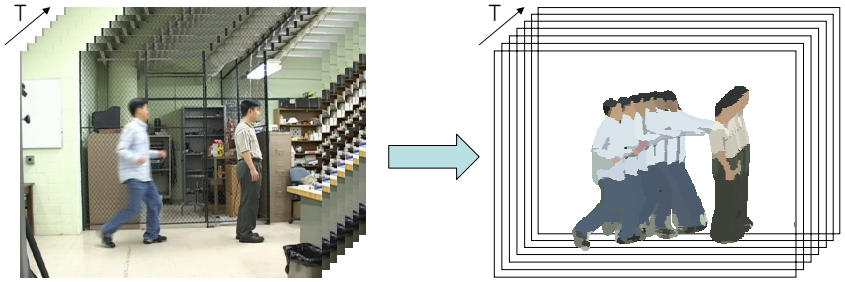


Fig. 4. Example XYT volumes constructed by concatenating (a) entire images and (b) foreground blob images obtained from a 'punching' sequence.

of particular observations. More specifically, they represent a human activity as a sequence of feature vectors extracted from images, and recognize activities by searching for such sequence. We discuss space-time approaches in Subsection 2.1, and compare sequential approaches in Subsection 2.2.

2.1 Space-time approaches

An image is 2-dimensional data formulated by projecting a 3-D real-world scene, and it contains spatial configurations (e.g. shapes and appearances) of humans and objects. A video is a sequence of those 2-D images placed in chronological order. Therefore, a video input containing an execution of an activity can be represented as a particular 3-D XYT space-time volume constructed by concatenating 2-D (XY) images along time (T).

Space-time approaches are approaches that recognize human activities by analyzing space-time volumes of activity videos. A typical space-time approach for human activity recognition is as follows. Based on the training videos, the system constructs a model 3-D XYT space-time volume representing each activity. When an unlabeled video is provided, the system constructs a 3-D space-time volume corresponding to the new video. The new 3-D volume is compared with each activity model (i.e. template volume) to measure the similarity in shape and appearance between the two volumes. The system finally deduces that the new video corresponds to the activity which has the highest similarity. This example can be viewed as a typical space-time methodology using the '3-D space-time volume' representation and the 'template matching' algorithm for the recognition. Figure 4 shows example 3-D XYT volumes corresponding to a human action of 'punching'.

In addition to the pure 3-D *volume* representation, there are several variations of the space-time representation. First, the system may represent an activity as *trajectories* (instead of a volume) in a space-time dimension or other dimensions. If the system is able to track feature points such as estimated joint positions of a human, the movements of the person performing an activity can be represented more explicitly as a set of trajectories. Secondly, instead of representing an activity with a volume or a trajectory, the system may represent an action as a set of *features* extracted from the volume or the trajectory. 3-D volumes can be viewed as rigid objects, and extracting common patterns from them enables their representations.

Researchers have also focused on developing various recognition algorithms using space-time representations to correctly match volumes, trajectories, or their features. We already have seen a typical example of an approach using a *template matching*, which constructs a representative model (i.e. a volume) per action using training data. Activity recognition is done by matching the model with the volume constructed from inputs. *Neighbor-based matching* algorithms (i.e. discriminative methods) have also been applied widely. In the case of neighbor-based matching, the system maintains a set of sample volumes (or trajectories) to describe an activity. The recognition is performed by matching the input with all (or a portion) of them. Finally, *statistical modeling* algorithms have been developed, which match videos by explicitly modeling a probability distribution of an activity.

Accordingly, we have classified space-time approaches into several categories. A representation-based taxonomy and a recognition-based taxonomy have been jointly applied for the classification. That is, each of the activity recognition publications with space-time approaches are assigned to a slot corresponding to a specific (representation, recognition) pair. The left part of Figure 2 shows a detailed hierarchy tree of space-time approaches.

2.1.1 *Action recognition with space-time volumes.* The core of the recognition using space-time volumes is in the similarity measurement between two volumes. The system must be able to compute how similar humans' movements described in two volumes are. In order to calculate the correct similarities, various types of space-time volume representations and recognition methodologies have been developed. Instead of concatenating entire images along time, some approaches only stack foreground regions of a person (i.e. silhouettes) to track shape changes explicitly [Bobick and Davis 2001]. An approach to compare volumes in terms of their patches has been proposed as well [Shechtman and Irani 2005]. Ke et al. [2007] used over-segmented volumes, automatically calculating a set of 3-D XYT volume segments that corresponds to a moving human. Rodriguez et al. [2008] generated filters capturing characteristics of volumes, in order to match volumes more reliably and efficiently. In this subsection, we cover each of these approaches while focusing on our taxonomy of 'what types of space-time volume they use' and 'how they match volumes to recognize activities'.

Bobick and Davis [2001] constructed a real-time action recognition system using template matching. Instead of maintaining the 3-dimensional space-time volume of each action, they have represented each action with a template composed of two 2-dimensional images: a 2-dimensional binary *motion-energy image* (MEI) and a scalar-valued *motion-history image* (MHI). The two images are constructed from a sequence of foreground images, which essentially are weighted 2-D (XY) projections of the original 3-D XYT space-time volume. By applying a traditional template matching technique to a pair of (MEI, MHI), their system was able to recognize simple actions like sitting, arm waving, and crouching. Further, their real-time system has been applied to the interactive play environment of children called 'Kids-Room'. Figure 5 shows example MHIs.

Shechtman and Irani [2005] have estimated motion flows from a 3-D space-time volume to recognize human actions. They have computed a 3-D space-time video-template correlation, measuring the similarity between an observed video volume

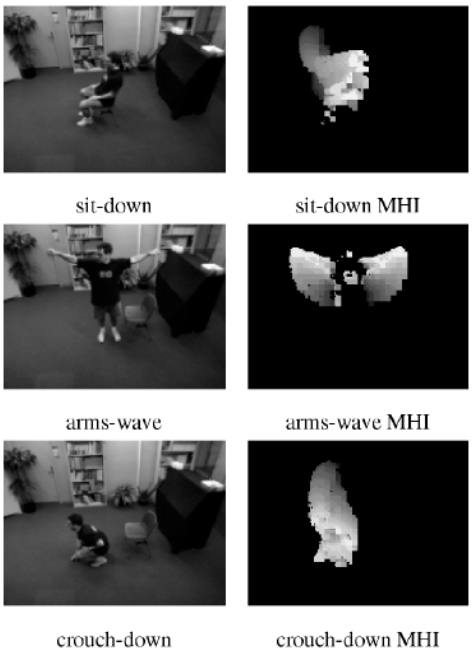


Fig. 5. Examples of space-time action representation: *motion-history images* from [Bobick and Davis 2001] (©2001 IEEE). This representation can be viewed as an weighted projection of a 3-D XYT volume into 2-D XY dimension.

and maintained template volumes. Their similarity measurement can be viewed as a hierarchical space-time volume correlation. At every location of the volume (i.e. (x, y, t)), they extracted a small space-time patch around the location. Each volume patch captures the flow of a particular local motion, and the correlation between a patch in a template and a patch in video at the same location gives a local match score to the system. By aggregating these scores, the overall correlation between the template volume and a video volume is computed. When an unknown video is given, their system searches for all possible 3-D volume segments centered at every (x, y, t) that best matches with the template (i.e. sliding windows). Their system was able to recognize various types of human actions, including ballet movements, pool dives, and waving.

Ke et al. [2007] used segmented spatio-temporal volumes to model human activities. Their system applies a hierarchical meanshift to cluster similarly colored voxels, and obtains several segmented volumes. The motivation is to find the actor volume segments automatically, and measure their similarity to the action model. Recognition is done by searching for a subset of over-segmented spatio-temporal volumes that best matches the shape of the action model. Support vector machines (SVM) have been applied to recognize human actions while considering both shapes and flows of the volumes. As a result, their system recognized simple actions such as hand waving and boxing from the KTH action database [Schuldt et al. 2004] as well as tennis plays in TV broadcast videos with more complex backgrounds.

Rodriguez et al. [2008] have analyzed 3-D space-time volumes by synthesizing filters: They adopted the maximum average correlation height (MACH) filters that have been used for an analysis of images (e.g. object recognition), to solve the action recognition problem. That is, they have generalized the traditional 2-D MACH filter for 3-D XYT volumes. For each action class, one synthesized filter that fits the observed volume is generated, and the action classification is performed by applying the synthesized action MACH filter and analyzing its response on the new observation. They have further extended the MACH filters to analyze vector-valued data using the Clifford Fourier transform. They not only have tested their system on the existing KTH dataset and the Weizmann dataset [Blank et al. 2005], but also on their own dataset constructed by gathering clips from movie scenes. Actions such as 'kissing' and 'hitting' have been recognized.

Table I compares the abilities of the space-time volume-based action recognition approaches. The major disadvantage of space-time volume approaches is the difficulty in recognizing actions when multiple persons are present in the scene. Most of the approaches apply the traditional sliding window algorithm to solve this problem. However, this requires a large amount of computations for the accurate localization of actions. Furthermore, they have difficulty recognizing actions which cannot be spatially segmented.

2.1.2 *Action recognition with space-time trajectories.* Trajectory-based approaches are recognition approaches that interpret an activity as a set of space-time trajectories. In trajectory-based approaches, a person is generally represented as a set of 2-dimensional (XY) or 3-dimensional (XYZ) points corresponding to his/her joint positions. Human body part estimation methodologies, especially the stick figure modeling, have widely been used to extract the joint positions of a person at each image frame. As a human performs an action, his/her joint position changes are recorded as space-time trajectories, constructing 3-D XYT or 4-D XYZT representations of the action. Figure 6 shows example trajectories. The early work done by Johansson [1975] suggested that the tracking of joint positions itself is sufficient for humans to distinguish actions, and this paradigm has been studied for the recognition of activities in depth [Webb and Aggarwal 1982; Niyogi and Adelson 1994].

Several approaches used the trajectories themselves (i.e. sets of 3-D points) to represent and recognize actions directly [Sheikh et al. 2005; Yilmaz and Shah 2005b]. Sheikh et al. [2005] represented an action as a set of 13 joint trajectories in a 4-D XYZT space. They have used an affine projection to obtain normalized XYT trajectories of an action, in order to measure the view-invariant similarity between two sets of trajectories. Yilmaz and Shah [2005b] presented a methodology to compare action videos obtained from moving cameras, also using a set of 4-D XYZT joint trajectories.

Campbell and Bobick [1995] recognized human actions by representing them as curves in low-dimensional *phase spaces*. In order to track joint positions, they took advantage of 3-D body-part models of a person. Based on the 3-D XYZ models estimated for each frame, they have defined body phase space as a space where each axis represents an independent parameter of the body (e.g. ankle-angle or knee-angle) or its first derivative. In their phase space, a person's static state at

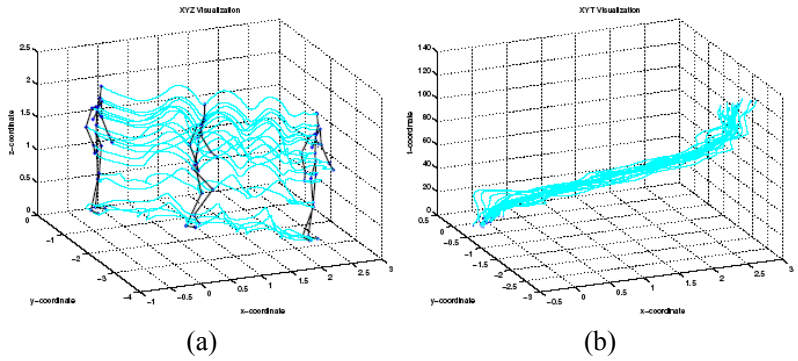


Fig. 6. An example trajectories of human joint positions when performing a human action 'walking' [Sheikh et al. 2005] (©2005 IEEE). Figure (a) shows trajectories in XYZ space, and (b) shows those in XYT space.

each frame corresponds to a point and an action corresponds to a set of points (i.e. curve). They have projected the curve in the phase space into multiple 2-D subspaces, and maintained the projected curves to represent the action. Each curve is modeled to have a cubic polynomial form, indicating that they assume the actions to be relatively simple in the projected subspace. Among all possible curves of 2-D subspaces, their system automatically selects the top k stable and reliable ones to be used for the recognition process.

Once an action representation, a set of projected curves, has been constructed, Campbell and Bobick recognized the action by converting an unseen video also into a set of points in the phase space. Without explicitly analyzing the dynamics of the points from the unseen video, their system simply verifies whether the points are on the maintained curves (i.e. trajectories in the subspaces) when projected. Various types of basic ballet movements have been recognized successfully with markers attached to a subject to track joint positions.

Instead of maintaining trajectories to represent human actions, Rao and Shah [2001]'s methodology extracts meaningful curvature patterns from the trajectories. They have tracked the position of a hand in 2-D image space using the skin pixel detection, obtaining a 3-D XYT space-time curve. Their system extracts the positions of peaks of trajectory curves, representing an action as a set of peaks and intervals between them. They have verified that these peak features are view-invariant. Automated learning of the human actions is possible for their system, incrementally constructing several action prototypes as representations of human actions. These prototypes can be considered action templates, and the overall recognition process can be regarded as a template matching process. As a result, by analyzing peaks of trajectories, their system was able to recognize human actions in an office environment such as 'opening a cabinet' and 'picking up an object'.

Again, Table I compares the trajectory-based approaches. The major advantage of the trajectory-based approaches is their ability to analyze detailed levels of human movements. Furthermore, most of these methods are view invariant. However, in order to do so, they generally require a strong low-level component which

is able to correctly estimate the 3-D XYZ joint locations of persons appearing in a scene. The problem of the 3-D body-part detection and tracking is still an unsolved problem, and researchers are actively working in this area.

2.1.3 *Action recognition using space-time local features.* The approaches discussed in this subsection are approaches using local features extracted from 3-dimensional space-time volumes to represent and recognize activities. The motivation behind these approaches is in the fact that a 3-D space-time volume essentially is a rigid 3-D object. This implies that if a system is able to extract appropriate features describing characteristics of each action's 3-D volumes, the action can be recognized by solving an object matching problem.

In this subsection, we discuss each of the approaches using 3-D space-time features, while especially focusing on three aspects: what 3-D local features the approaches extract, how they represent an activity in terms of the extracted features, and what methodology they use to classify activities. In general, we are able to describe the activity recognition approaches using local features by presenting the above three components. Similar to the object recognition process, the system first extracts specific local features that have been designed to capture the local motion information of a person from a 3-D space-time volume. These features are then combined to represent the activities while considering their spatio-temporal relationships or ignoring their relations. Finally, recognition algorithms are applied to classify the activities.

We use the terminology 'local features', 'local descriptors', and 'interest points' interchangeably, similar to the case of object recognition problems. Several approaches extract these local features at every frame and concatenate them temporally to describe the overall motion of human activities [Chomat and Crowley 1999; Zelnik-Manor and Irani 2001; Blank et al. 2005]. The other approaches extract sparse spatio-temporal local interest points from 3-D volumes [Laptev and Lindeberg 2003; Dollar et al. 2005; Niebles et al. 2006; Yilmaz and Shah 2005a; Ryoo and Aggarwal 2009b]. Example 3-D local interest points are illustrated in Figure 7. These features have been particularly popular because of their reliability under noise, camera jitter, illumination changes, and background movements.

Chomat and Crowley [1999] proposed an idea of using local appearance descriptors to characterize an action, thereby enabling the action classification. Motion energy receptive fields together with Gabor filters are used to capture motion information from a sequence of images. More specifically, local spatio-temporal appearance features describing motion orientations are detected per frame. Multi-dimensional histograms are constructed based on the detected local features, and the posterior probability of an action occurring given the detected features is calculated by applying the Bayes rule to the histograms. Their system first calculates the local probability of an activity occurring at each pixel location, and integrates them for the final recognition of the actions. Even though only simple gestures such as 'come', 'go', 'left', and 'right' are recognized due to the simplicity of their motion descriptors, they have shown that local appearance detectors may be utilized for the recognition of human activities.

Zelnik-Manor and Irani [2001] proposed an approach utilizing local spatio-temporal features at multiple temporal scales. Multiple temporally scaled video volumes are

analyzed to handle execution speed variations of an action. For each point in a 3-D XYT volume, their system estimates a normalized local intensity gradient. Similar to [Chomat and Crowley 1999], they have computed a histogram of these space-time gradient features per video, and presented a histogram-based distance measurement ignoring the positions of the extracted features. An unsupervised clustering algorithm has been applied to these histograms to learn actions, and human activities including outdoor sports video sequences like basketball and tennis plays have been automatically recognized.

Similarly, Blank et al. [2005] also calculated local features at each frame. Instead of utilizing optical flows for the calculation of local features, they calculated appearance-based local features at each pixel by constructing a space-time volume whose pixel values are solutions to the Poisson equation. The solution to the Poisson equation has proved to be able to extract a wide variety of useful local shape properties, and their system has extracted local features capturing space-time saliency and space-time orientation using the equation. Each sequence of an action is represented as a set of global features, which are the weighted moments of the local features. They have applied a simple nearest neighbor classification with a Euclidean distance to recognize the actions. Simple actions such as 'walking', 'jumping', and 'bending' in their Weizmann dataset as well as basic ballet movements have been recognized successfully.

On the other hands, there are approaches extracting sparse local features from video volumes to represent activities. Laptev and Lindeberg [2003] recognized human actions by extracting sparse spatio-temporal interest points from videos. They have extended the previous local feature detectors [Harris and Stephens 1988] commonly used for object recognition, in order to detect interest points in a space-time volume. This scale-invariant interest point detector searches for spatio-temporal corners in a 3-dimensional space (XYT), which captures various types of non-constant motion patterns. Motion patterns such as a direction change of an object, splitting and merging of an image structure, and/or collision and bouncing of objects, are detected as a result (Figure 7 (a) and (b)). In their work, these features have been used to distinguish a walking person from complex backgrounds. Furthermore, Schuldt et al. [2004] classified multiple actions by applying SVMs to Laptev and Lindeberg [2003]'s features, illustrating their applicability for the activity recognition. A new database called 'KTH actions dataset' containing action videos (e.g. 'jogging' and 'hand waving') was introduced, and has been widely adopted. We discuss more about this dataset in Subsection 5.1.1.

This paradigm of recognizing actions by extracting sparse local interest points from a 3-dimensional space-time volume has been adopted by several researchers. They have focused on the fact that sparse local features characterizing local motion are sufficient to represent actions, as [Laptev and Lindeberg 2003] have suggested. These approaches are particularly motivated by the success of the object recognition methodologies using sparse local appearance features, such as SIFT descriptors [Lowe 1999]. Instead of extracting features at every frame, these approaches extract features only when there exists a salient appearance or shape change in 3-D space-time volume. Most of these features have been verified to be invariant to scale, rotation, and translations, similar to object recognition descriptors.

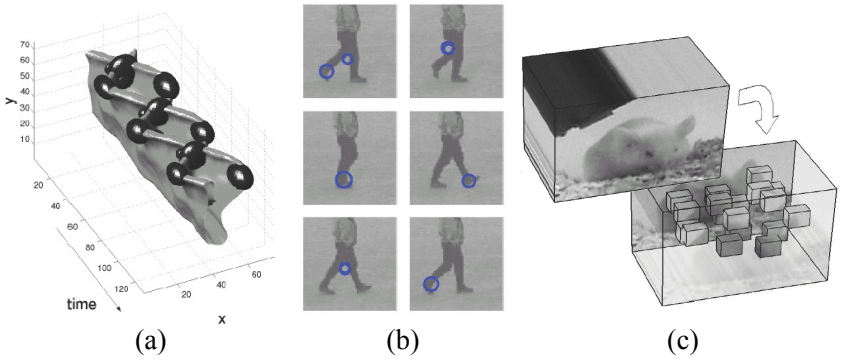


Fig. 7. Example 3-D space-time local features extracted from a video of a human action 'walking' [Laptev and Lindeberg 2003] (©2003 IEEE), and those from a mouse movement video [Dollar et al. 2005] (©2005 IEEE). Figure (a) shows a concatenated XYT surfaces of legs of a person and detected interest points using [Laptev and Lindeberg 2003]. Figure (b) shows the same interest points placed on a sequence of original images. Figure (c) shows cuboid features extracted using [Dollar et al. 2005].

Dollar et al. [2005] proposed a new spatio-temporal feature detector for the recognition of human (and animal) actions. Their detector is especially designed to extract space-time points with local periodic motions, obtaining a sparse distribution of interest points from a video. Once detected, their system associates a small 3-D volume called *cuboid* to each interest point (Figure 7 (c)). Each cuboid captures pixel appearance values of the interest point's neighborhoods. They have tested various transformations to be applied to cuboids to extract final local features, and have chosen the flattened vector of brightness gradients that shows the best performance. A library of cuboid prototypes is constructed per each dataset by clustering cuboid appearances with k-means. As a result, each action is modeled as a histogram of cuboid types detected in 3-D space-time volume while ignoring their locations (i.e. bag-of-words paradigm). They have recognized facial expressions, mouse behaviors, and human activities (i.e. the KTH dataset) using their method.

Niebles et al. [2006][Niebles et al. 2008] presented an unsupervised learning and classification method for human actions using the above-mentioned feature extractor [Dollar et al. 2005]. Their recognition method is a generative approach, modeling an action class as a collection of spatio-temporal feature appearances. A probabilistic Latent Semantic Analysis (pLSA) commonly used in the field of text mining has been applied to recognize actions statistically. Each feature in the scene is categorized into an action class by calculating its posterior probability of being generated by the action. As a result, they were able to recognize simple actions from public datasets [Schuldt et al. 2004; Blank et al. 2005] as well as figure skating actions.

In this context, various spatio-temporal feature extractors have been developed recently. Yilmaz and Shah [2005a] proposed an action recognition approach to extract sparse features called *action sketches* from a 3-D contour concatenation, which have been confirmed to be view-invariant. Scovanner et al. [2007] designed the 3-D version of the SIFT descriptor, similar to the cuboid features [Dollar et al. 2005]. Liu et al. [2009] presented a methodology to prune cuboid features to choose

important and meaningful features. Bregonzio et al. [2009] proposed an improved detector for extracting cuboid features, and presented a feature selection method similar to [Liu et al. 2009]. Rapantzikos et al. [2009] extended the cuboid features to utilized color and motion information as well, in contrast to previous features only using intensities (e.g. [Laptev and Lindeberg 2003; Dollar et al. 2005]).

In most approaches using sparse local features, spatial and temporal relationships among detected interest points are ignored. The approaches that we have discussed above have shown that simple actions can successfully be recognized even without any spatial and temporal information among features. This is similar to the success of object recognition techniques ignoring local features' spatial relationships, typically called as *bag-of-words*. The bag-of-words approaches were particularly successful for simple periodic actions.

Recently, action recognition approaches considering spatial configurations among the local features are getting an increasing amount of interests. Unlike the approaches following the bag-of-words paradigm, these approaches attempt to model spatio-temporal distribution of the extracted features for better recognition of actions. Wong et al. [2007] extended the basic pLSA, constructing a pLSA with an implicit shape model (pLSA-ISM). In contrast to the pLSA used by [Niebles et al. 2006], their pLSA-ISM captures the relative spatio-temporal location information of the features from the activity center, successfully recognizing and localizing activities in the KTH dataset.

Savarese et al. [2008] proposed a methodology to capture spatio-temporal proximity information among features. For each action video, they have measured feature co-occurrence patterns in a local 3-D region, constructing histograms called *ST-correlograms*. Liu and Shah [2008] also considered correlations among features. Similarly, Laptev et al. [2008] constructed spatio-temporal histograms by dividing an entire space-time volume into several grids. The method roughly measures how local descriptors are distributed in the 3-D XYT space, by analyzing which feature falls into which grid. Both methods have been tested on the KTH dataset as well, obtaining successful recognition results. Notably, similar to [Rodriguez et al. 2008], [Laptev et al. 2008] has been tested on realistic videos obtained from various movie scenes.

Ryoo and Aggarwal [2009b] introduced the *spatio-temporal relationship match* (STR match), which explicitly considers spatial and temporal relationships among detected features to recognize activities. Their method measures structural similarity between two videos by computing pair-wise spatio-temporal relations among local features (e.g. *before* and *during*), enabling the detection and localization of complex-structured activities. Their system not only classified simple actions (i.e. those from the KTH datasets), but also recognized interaction-level activities (e.g. hand shaking and pushing) from continuous videos.

The space-time approaches extracting local descriptors have several advantages. By its nature, background subtraction or other low-level components are generally not required, and the local features are scale, rotation, and translation invariant in most cases. They were particularly suitable for recognizing simple periodic actions such as 'walking' and 'waving', since periodic actions will generate feature patterns repeatedly.

Approach Type	Authors	Required low-levels	Structural consideration	Scale invariant	Localization	View invariant	Multiple activities
Space-time volume	Bobick and J. Davis '01	Background	Volume-based	Templates needed	√		
	Shechtman and Irani '05	None	Volume-based	Scaling required	√		
	Ke et al. '07	None	Volume-based	Templates needed	√		
	Rodriguez et al. '08	None	Volume-based	√	√		
Space-time trajectories	Campbell and Bobick '95	Body-part estimation		√	√	√	
	Rao and Shah '01	Skin detection	Ordering only	√	√	√	
	Sheikh et al. '05	Body-part estimation	Ordering only	√	√	√	
Space-time features	Chomat and Crowley '99	None		√	√		
	Zalnik-Manor and Irani '01	None		√			
	Laptev and Lindeberg '03	None		√	√		
	Shuldt et al. '04	None		√			
	Dollar et al. '05	None		√			
	Yilmaz and Shah '05a	Background	Ordering only	√	√	√	
	Blank et al. '05	Background		√	√	Δ	
	Niebles et al. '06	None		√	√		√
	Wong et al. '07	None	√	√	√		
	Savarese et al. '08	None	Proximity-based	√	√		√
	Liu and Shah '08	None	Co-occur only	√			
	Laptev et al. '08	None	Grid-based	√			
	Ryoo and Aggarwal '09b	None	√	√	√		√

Table I. A table comparing the abilities of the important space-time approaches. The column 'required low-levels' specifies the low-level components necessary for the approach to be applicable. 'Structural consideration' shows temporal patterns the approach is able to capture. 'Scale invariant' and 'view invariant' columns describe whether the approaches are invariant to scale and view changes in videos, and 'localization' indicates the ability to correctly locate where the activity is occurring spatially and temporally. 'Multiple activities' indicates that the system is designed to consider multiple activities in the same scene.

2.1.4 *Comparison.* Table I compares the abilities of the space-time approaches reviewed in this paper. Space-time approaches are suitable for recognition of periodic actions and gestures, and many have been tested on public datasets (e.g. the KTH dataset [Schuldt et al. 2004] and the Weizmann dataset [Blank et al. 2005]). Basic approaches using space-time volumes provide a straight-forward solution, but often have difficulties handling speed and motion variations inherently. Recognition approaches using space-time trajectories are able to perform detailed-level analysis and are view-invariant in most cases. However, 3-D modeling of body parts

from videos, which still is an unsolved problem, is required for a trajectory-based approach to be applied.

The spatio-temporal local feature-based approaches are getting an increasing amount of attention because of their reliability under noise and illumination changes. Furthermore, some approaches [Niebles et al. 2006; Ryoo and Aggarwal 2009b] are able to recognize multiple activities without background subtraction or body-part modeling. The major limitation of the space-time feature-based approaches is that they are not suitable for modeling more complex activities. The relations among features are important for a non-periodic activity that takes a certain amount of time, which most of the previous approaches ignored. Several researchers have worked on approaches to overcome such limitations [Wong et al. 2007; Savarese et al. 2008; Laptev et al. 2008; Ryoo and Aggarwal 2009b]. Viewpoint invariance is another issue that space-time local feature-based approaches must handle.

2.2 Sequential approaches

Sequential approaches are the single-layered approaches that recognize human activities by analyzing sequences of features. They consider an input video as a sequence of observations (i.e. feature vectors), and deduce that an activity has occurred in the video if they are able to observe a particular sequence characterizing the activity. Sequential approaches first convert a sequence of images into a sequence of feature vectors by extracting features (e.g. degrees of joint angles) describing the status of a person per image frame. Once feature vectors have been extracted, sequential approaches analyze the sequence to measure how likely the feature vectors are produced by the person performing the activity. If the likelihood between the sequence and the activity class (or the posterior probability of the sequence belonging to the activity class) is high enough, the system decides that the activity has occurred.

We classify the sequential approaches into two categories using a methodology-based taxonomy: exemplar-based recognition approaches and state model-based recognition approaches. Exemplar-based sequential approaches describe classes of human actions using training samples directly. They maintain either a representative sequence per class or a set of training sequences per activity, and match them with a new sequence to recognize its activity. On the other hand, state model-based sequential approaches are approaches that represent a human action by constructing a model which is trained to generate sequences of feature vectors corresponding to the activity. By calculating the likelihood (or posterior probability) that a given sequence is generated by each activity model, the state model-based approaches are able to recognize the activities.

2.2.1 *Exemplar-based approaches.* Exemplar-based approaches represent human activities by maintaining a template sequence or a set of sample sequences of action executions. When a new input video is given, the exemplar-based approaches compare the sequence of feature vectors extracted from the video with the template sequence (or sample sequences). If their similarity is high enough, the system is able to deduce that the given input contains an execution of the activity. Humans may perform an identical activity in different styles and/or different rates, and the similarity must be measured considering such variations. The dynamic time warp-

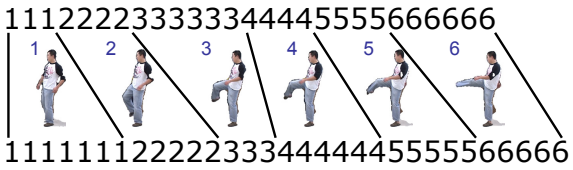


Fig. 8. An example matching between two 'stretching a leg' sequences with different non-linear execution rates. Each number represents a particular status (i.e. pose) of the person.

ing (DTW) algorithm, originally developed for speech processing, has been widely adopted for matching two sequences with variations [Darrell and Pentland 1993; Gavrila and Davis 1995; Veeraraghavan et al. 2006]. The DTW algorithm finds an optimal nonlinear match between two sequences with a polynomial amount of computations. Figure 8 shows a conceptual matching between two sequences (i.e. strings) with different execution rates.

Darrell and Pentland [1993] proposed a DTW-based gesture recognition methodology using *view* models to represent the dynamics of articulated objects. Their system maintains multiple models (i.e. template images) of an object in different conditions, which they called views. Each view-model abstracts a particular status (e.g. rotation and scale) of an articulated object such as a hand. Given a video, the correlation scores between image frames and each view are modeled as a function of time. Means and variations of these scores of training videos are used as a gesture template. The templates are matched with a new observation using the DTW algorithm, so that speed variations of action executions are handled. Their system successfully recognized 'hello' and 'good-bye' gestures, and was able to distinguish them from other gestures such as a 'come closer' gesture.

Gavrila and Davis [1995] also developed the DTW algorithm to recognize human actions, utilizing a 3-dimensional (XYZ) model-based body-part tracking methodology. The motivation is to estimate a 3-D skeleton model at each image frame and to analyze his/her movement by tracking them. Multiple cameras have been used to obtain 3-D body-part models of a human, which is composed of a collection of segments and their joint angles (i.e. the stick figure). This stick figure model with 17 degree-of-freedom (DOF) is tracked throughout the frames, recording the values of joint angles. These angle values are treated as features characterizing human movement at each frame. The sequences of angle values are analyzed using the DTW algorithm to compare them with a reference sequence pre-trained per action, similar to [Darrell and Pentland 1993]. Gestures including 'waving hello', 'waving-to-come', and 'twisting' have been recognized with their system.

Yacoob and Black [1998] have treated an input as a set of signals (instead of discrete sequences) describing sequential changes of feature values. Instead of directly matching the sequences (e.g. DTW), they have decomposed signals using singular value decompositions (SVD). That is, they used principle component analysis (PCA)-based modeling to represent an activity as a linear combination of a set of *activity basis* that essentially is a set of eigen vectors. When a new input is provided to the system, their system calculates the coefficients of the activity basis while considering transformation parameters such as scale and speed variations.

The similarity between the input and an action template is measured by comparing the coefficients of the two. Their approach showed successful recognition results for walking-related actions and lip movements, utilizing different types of features.

Efros et al. [2003] presented a methodology for recognizing actions at a distance, where each human is around 30 pixels tall. In order to recognize actions in such environments where the detailed motion of humans is unclear, they used motion descriptors based on optical flows obtained per frame. Their system first computes the space-time volume of each person being tracked, and then calculates 2-D (XY) optical flows at each frame by tracking humans using a temporal difference image similar to [Yacoob and Black 1998]. They used blurry motion channels as a motion descriptor, converting optical flows into a spatio-temporal motion descriptor per frame. That is, they are interpreting a video of a human action as a sequence of motion descriptors obtained from optical flows of a human. The basic nearest neighbor classification method has been applied to a sequence of motion descriptors for the recognition of actions. First, frame-to-frame similarities between all possible pairs of frames from two sequences (i.e. a frame-to-frame similarity matrix) are calculated. The recognition is done by detecting diagonal patterns in the frame-to-frame similarity matrix. Their system was able to classify ballet movements, tennis plays, and soccer plays even from moving cameras.

Lublinerman et al. [2006] presented a methodology that recognizes human activities by modeling them as linear time invariant (LTI) systems. Their system converts a sequence of images into a sequence of silhouettes, extracting two types of contour representations: silhouette width and Fourier descriptors. An activity is represented as a LTI system capturing the dynamics of changes in silhouette features. SVMs have been applied to classify a new input which has been converted to the parameters of a LTI model. Four types of simple actions, 'slow walk', 'fast walk', 'walk on an incline' and 'walk with a ball' have been correctly recognized as a consequence.

Veeraraghavan et al. [2006] described an activity as a function of time describing parameter changes similar to [Yacoob and Black 1998]. The main contribution of Veeraraghavan et al.'s system is in the explicit modeling of inter- and intra-personal speed variations of activity executions and the consideration of them for matching activity sequences. Focusing on the fact that humans may be able to change the speed of an execution of a part of the activity while it may not be possible for other parts, they learn non-linear characteristics of activity speed variations. More specifically, their system learns the nature of time warping transformation per activity. They are modeling an action execution with two functions: (i) a function of feature changes over time and (ii) a function space of possible time warping. They have developed an extension of a DTW matching algorithm to take the time warping function into account when matching two sequences. Human actions including 'picking up an object', 'throwing', 'pushing', and 'waving' have been recognized with high recognition accuracy.

2.2.2 *State model-based approaches.* State model-based approaches are the sequential approaches which represent a human activity as a model composed of a set of states. The model is statistically trained so that it corresponds to sequences of feature vectors belonging to its activity class. More specifically, the statistical

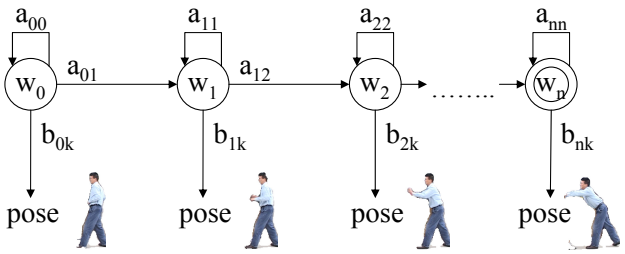


Fig. 9. An example hidden Markov model for the action 'stretching an arm'. The model is one of the most simple case among HMMs, which is designed to be strictly sequential. Each actor image in the figure represents a pose with the highest observation probability b_{jk} for its state w_j.

model is designed to generate a sequence with a certain probability. Generally, one statistical model is constructed for each activity. For each model, the probability of the model generating an observed sequence of feature vectors is calculated to measure the likelihood between the action model and the input image sequence. Either the maximum likelihood estimation (MLE) or the maximum a posteriori probability (MAP) classifier is constructed as a result, in order to recognize activities.

Hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs) have been widely used for state model-based approaches. In both cases, an activity is represented in terms of a set of hidden states. A human is assumed to be in one state at each time frame, and each state generates an observation (i.e. a feature vector). In the next frame, the system transitions to another state considering the transition probability between states. Once transition and observation probabilities are trained for the models, activities are commonly recognized by solving the 'evaluation problem'. The evaluation problem is a problem of calculating the probability of a given sequence (i.e. new input) generated by a particular state-model. If the calculated probability is high enough, the state model-based approaches are able to decide that the activity corresponding to the model occurred in the given input. Figure 9 shows an example of a sequential HMM.

Yamato et al. [1992]'s work is the first work applying standard HMMs to recognize activities. They adopted HMMs which originally have been widely used for speech recognition. At each frame, their system converts a binary foreground image into an array of meshes. The number of pixels in each mesh is considered a feature, thereby extracting a feature vector per frame. These feature vectors are treated as a sequence of observations generated by the activity model. Each activity is represented by constructing one HMM that probabilistically corresponds to particular sequences of feature vectors (i.e. meshes). More specifically, parameters of HMMs (transition probabilities and observation probabilities) are trained with a labeled dataset with the standard learning algorithm for HMMs. Once each of the HMMs is trained, they are used for the recognition of activities by measuring the likelihoods between a new input and the HMMs by solving the 'evaluation problem'. As a result, various types of tennis plays, such as 'backhand stroke', 'forehand stroke', 'smash', and 'serve', have been recognized with Yamato et al.'s system. They have shown that the HMMs are able to model feature changes during human activities reliably, encouraging other researchers to pursue further investigations.

Starner and Pentland [1995] also used standard HMMs, in order to recognize American Sign Language (ASL). Their method tracks the location of hands, and extracts features describing shapes and positions of the hands. Each word of ASL is modeled as one HMM generating a sequence of features describing hand shapes and positions, similar to the case of [Yamato et al. 1992]. Their method uses the Viterbi algorithm for each HMM, to estimate the probability the HMM generated the observations. The Viterbi algorithm provides an efficient approximation of the likelihood distance, enabling an unknown observation sequence to be classified into the most suitable word.

Bobick and Wilson [1997] also recognized gestures using state models. They represented a gesture as a 2-D XY trajectory describing the location changes of a hand. Each curve is decomposed into sequential vectors, which can be interpreted as a sequence of states computed from a training example. Furthermore, each state is made to be fuzzy, in order to consider speed and motion variance in executions of the same gesture. This is similar to a fuzzy version of a sequential Markov model (MM). Transition costs between states, which correspond to the transition probabilities in the case of HMMs, are also defined in their system. For the recognition of gestures with their model, a dynamic programming algorithm is designed. Their system measures an optimal matching cost between the given observation (i.e. motion trajectory) and each prototype using the dynamic programming algorithm. Applying their framework, they have successfully recognized two different types of gestures: 'wave' and 'point'.

In addition, approaches using variants of HMMs also have been developed for human activity recognition [Oliver et al. 2000; Park and Aggarwal 2004; Natarajan and Nevatia 2007]. Similar to previous frameworks for action recognition using HMMs [Yamato et al. 1992; Starner and Pentland 1995; Bobick and Wilson 1997], they construct one model (HMM) for each activity they want to recognize, and use visual features from the scene as observations directly generated by the model. The methods with extended HMMs are designed to handle more complex activities (usually combinations of multiple simple actions) by extending the structure of the basic HMM.

Oliver et al. [2000] constructed a variant of the basic HMM, the coupled HMM (CHMM), to model human-human interactions. The major limitation of the basic HMM is its inability to represent activities composed of motions of two or more agents. A HMM is a sequential model and only one state is activated at a time, preventing it from modeling the activities of multiple agents. Oliver et al. introduced the concept of the CHMM to model complex interactions between two persons. Basically, a CHMM is constructed by coupling multiple HMMs, where each HMM models the motion of one agent. They have coupled two HMMs to model human-human interactions. More specifically, they coupled the hidden states of two different HMMs by specifying their dependencies. As a result, their system was able to recognize complex interactions between two persons, such as concatenation of 'two persons approaching, meeting, and continuing together'.

Park and Aggarwal [2004] used a DBN to recognize gestures of two interacting persons. They have recognized gestures such as 'stretching an arm' and 'turning a head left', by constructing a tree-structured DBN to take advantage of the de-

pendent nature among body parts' motion. A DBN is an extension of a HMM, composed of multiple conditionally independent hidden nodes that generate observations at each time frame directly or indirectly. In the Park and Aggarwal's work, a gesture is modeled as state transitions of hidden nodes (i.e. body-part poses) in one time point to the next time point. Each pose is designed to generate a set of features associated with the corresponding body part. Features including locations of skin regions, maximum curvature points, and the ratio and orientation of each body-part have been used to recognize gestures.

Natarajan and Nevatia [2007] developed an efficient recognition algorithm using coupled hidden semi-Markov models (CHSMMs), which extend previous CHMMs by explicitly modeling the duration of an activity staying in each state. In the case of basic HMMs and CHMMs, the probability of a person staying in an identical state decays exponentially as time increases. In contrast, each state in a CHSMM has its own duration that best models the activity the CHSMM is representing. As a result, they were able to construct a statistical model that captures the characteristics of activities that the system wants to recognize better compared to HMMs and CHMMs. Similar to [Oliver et al. 2000], they tested their system for the recognition of human-human interactions. Because of the CHSMMs' ability to model the duration of the activity, the recognition accuracy using CHSMMs was better than other simpler statistical models. Lv and Nevatia [2007] also designed a CHMM-like structure called the *Action Net* to construct a view-invariant recognition system using synthetic 3-D human poses.

2.2.3 *Comparison.* In general, sequential approaches consider sequential relationships among features in contrast to most of the space-time approaches, thereby enabling detection of more complex activities (i.e. non-periodic activities such as sign languages). Particularly, the recognition of the interactions of two persons, whose sequential structure is important, has been attempted in [Oliver et al. 2000; Natarajan and Nevatia 2007].

Compared to the state model-based sequential approaches, exemplar-based approaches provide more flexibility for the recognition system, in the sense that multiple sample sequences (which may be completely different) can be maintained by the system. Further, the dynamic time warping algorithm generally used for the exemplar-based approaches provides a non-linear matching methodology considering execution rate variations. In addition, exemplar-based approaches are able to cope with less training data than the state model-based approaches.

On the other hand, state-based approaches are able to make a probabilistic analysis on the activity. A state-based approach calculates a posterior probability of an activity occurring, enabling it to be easily incorporated with other decisions. One of the limitations of the state-based approaches is that they tend to require a large amount of training videos, as the activity they want to recognize gets more complex. Table II is provided for the comparison of the systems.

3. HIERARCHICAL APPROACHES

The main idea of hierarchical approaches is to enable the recognition of high-level activities based on the recognition results of other simpler activities. The motivation is to let the simpler sub-activities (also called sub-events) which can be

Type	Approaches	Required low-levels	Execution variations	Probabilistic	Target activities
Exemplar-based	Darrell and Pentland '93	None	√		Gesture-level
	Gavrila and L. Davis '95	Body-part estimation	√		Gesture-level
	Yacoob and Black '98	Body-part estimation	√		Gesture-level
	Efros et al. '03	Tracking	Linear only		Action-level
	Lublinerman et al. '06	Background subtraction	Linear only		Action-level
	Veeraraghavan et al. '06	Background subtraction	√		Action-level
State model-based	Yamato et al. '92	Background subtraction	Model-based	√	Action-level
	Starner and Pentland '95	Tracking	Model-based	√	Gesture-level
	Bobick and Wilson '97	Tracking	Model-based		Gesture-level
	Oliver et al. '00	Background subtraction	Model-based	√	Interaction-level
	Park and Aggarwal '04	Background subtraction	Model-based	√	Gesture-level
	Natarajan and Nevatia '07	Action recognition	Model-based	√	Interaction-level
	Lv and Nevatia '07	3-D pose model	Model-based	√	Action-level

Table II. Comparison among sequential approaches. The column 'required low-levels' specifies the low-level components necessary for the approach to be applicable. 'Execution variations' shows whether the system is able to handle variations in the execution of human activities (e.g. speed variations). 'Probabilistic' indicates that the system makes a probabilistic inference, and 'target activity' shows the type of human activities the system aims to recognize. Notably, [Lv and Nevatia 2007]'s system is view-invariant.

modeled relatively easily to be recognized first, and then to use them for the recognition of higher-level activities. For example, a high-level interaction of 'fighting' may be recognized by detecting a sequence of several 'punching' and 'kicking' interactions. Therefore, in hierarchical approaches, a high-level human activity (e.g. fighting) that the system aims to recognize is represented in terms of its sub-events (e.g. punching), which themselves may be decomposable until the atomicity is obtained. That is, sub-events serve as observations generated by a higher-level activity. The paradigm of hierarchical representation not only makes the recognition process computationally tractable and conceptually understandable, but also reduces redundancy in the recognition process by re-using recognized sub-events multiple times.

In general, common activity patterns of motion that appear frequently during high-level human activities are modeled as atomic-level (or primitive-level) actions, and high-level activities are represented and recognized by concatenating them hierarchically. In most hierarchical approaches, these atomic actions are recognized by adopting single-layered recognition methodologies which we presented in the previous section. For example, the gestures 'stretching hand' and 'withdrawing hand'

occur often in human activities, implying that they can become good atomic actions to represent human activities such as 'shaking hands' or 'punching'. Single-layered approaches such as sequential approaches using HMMs can safely be adopted for recognition of those gestures.

The major advantage of hierarchical approaches over non-hierarchical approaches (i.e. single-layered approaches) is their ability to recognize high-level activities with more complex structures. Hierarchical approaches are especially suitable for a semantic-level analysis of interactions between humans and/or objects as well as complex group activities. This advantage is a result of two abilities of hierarchical approaches: the ability to cope with less training data, and the ability to incorporate prior knowledge into the representation.

First, the amount of training data required to recognize activities with hierarchical models is significantly less than that with single-layered models. Even though it may also possible for non-hierarchical approaches to model complex human activities in some cases, they generally require a large amount of training data. For example, single-layered HMMs need to learn a large number of transition and observation probabilities, since the number of hidden states increases as the activities get more complex. By encapsulating structurally redundant sub-events shared by multiple high-level activities, hierarchical approaches model the activities with a lesser amount of training and recognize them more efficiently.

In addition, the hierarchical modeling of high-level activities makes recognition systems to incorporate human knowledge (i.e. prior knowledge on the activity) much easier. Human knowledge can be included in the system by listing semantically meaningful sub-activities composing a high-level activity and/or by specifying their relationships. As mentioned above, when modeling high-level activities, non-hierarchical techniques tend to have complex structures and observation features which are not easily interpretable, preventing a user from imposing prior knowledge. On the other hand, hierarchical approaches model a high-level activity as an organization of semantically interpretable sub-events, making the incorporation of prior knowledge much easier.

Using our approach-based taxonomy, we categorize hierarchical approaches into three groups: statistical approaches, syntactic approaches, and description-based approaches. Figure 3 illustrates our taxonomy tree as well as the lists of selected previous works corresponding to the categories.

3.1 Statistical approaches

Statistical approaches use statistical state-based models to recognize activities. In the case of hierarchical statistical approaches, multiple layers of state-based models (usually two layers) such as HMMs and DBNs are used to recognize activities with sequential structures. At the bottom layer, atomic actions are recognized from sequences of feature vectors, just as in single-layered sequential approaches. As a result, a sequence of feature vectors are converted to a sequence of atomic actions. The second-level models treat this sequence of atomic actions as observations generated by the second-level models. For each model, a probability of the model generating a sequence of observations (i.e. atomic-level actions) is calculated to measure the likelihood between the activity and the input image sequence. Either the maximum likelihood estimation (MLE) or the maximum a posteriori probabil-

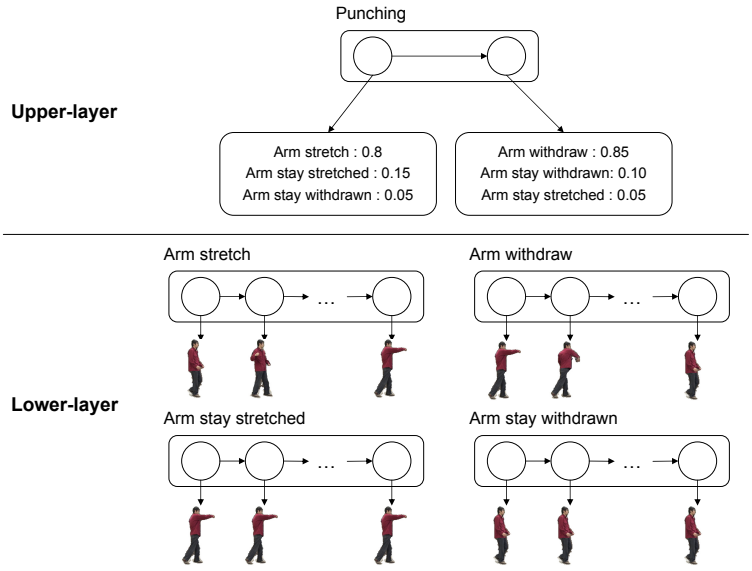


Fig. 10. An example hierarchical hidden Markov model (HHMM) for recognizing an activity 'punching'. The model is composed of two layers. In the lower layer, HMMs are used to recognize various atomic-level activities, such as 'stretching' and 'withdrawing'. The upper layer HMM treats recognition results of the lower layer HMMs are an input, recognizing 'punching' is 'stretching' and 'withdrawing' occurred in a sequence.

ity (MAP) classifier is constructed as a result. Figure 10 shows an example model of a statistical hierarchical approach, which is designed to recognize 'punching'.

Oliver et al. [2002] presented layered hidden Markov models (LHMMs), one of the most fundamental forms of the hierarchical statistical approaches (e.g. Figure 10). In this approach, the bottom layer HMMs recognize atomic actions of a single person by matching the models with the sequence of feature vectors extracted from videos. The upper layer HMMs treats recognized atomic actions as observations generated by the upper layer HMMs. That is, they essentially are representing a high-level activity as a sequence of atomic actions by making each state in the upper layer HMM to probabilistically correspond to one atomic action. By its nature, all sub-events of an activity are required to be strictly sequential in each LHMM. Human-human interactions in a conference room environment including 'a person giving a presentation' and 'face-to-face conversation' have been recognized based on the detection of atomic-level actions (e.g. 'nobody', 'one active person', and 'multiple persons present'). Each layer of the HMM is designed to be trained separately with fully labeled data, enabling a flexible retraining.

The paradigm of multi-layered HMMs has been explored by various researchers. Nguyen et al. [2005] also constructed hierarchical HMMs of two layers to recognize complex sequential activities. Similar to [Oliver et al. 2002], they have constructed two-levels of HMMs to recognize human activities such as 'a person having a meal' and 'a person having a snack'. Zhang et al. [2006] constructed multi-layered HMMs to recognize group activities occurring in a meeting room. Their framework is

also composed of two-layered HMMs. Their system recognized atomic actions of 'speaking', 'writing', and 'idling' using the lower-layer HMMs. With the upper-layer HMMs, group activities such as 'monologue', 'discussion', and 'presentation' have been represented and recognized with the atomic actions. Yu and Aggarwal [2006] used a block-based HMM for the recognition of a person climbing a fence. This block-based HMM can also be interpreted as a 2-layered HMM.

In addition, hierarchical approaches using DBNs have been studied for the recognition of complex activities. DBNs may contain multiple levels of hidden states, suggesting that they can be formulated represent hierarchical human activities. Gong and Xiang [2003] have extended traditional HMMs to construct dynamic probabilistic networks (DPNs) to represent activities of multiple participants. Their method was able to recognize group activities of trucks loading and unloading cargo. Dai et al. [2008] constructed DBNs to recognize group activities in a conference room environment similar to [Zhang et al. 2006]. High-level activities such as 'break', 'presentation', and 'discussion' were recognized based on the atomic actions 'talking', 'asking', and so on. Damen and Hogg [2009] constructed Bayesian networks using a Markov chain Monte Carlo (MCMC) for hierarchical analysis of bicycle-related activities (e.g. 'drop-and-pick'). They used Bayesian networks to model relations between atomic-level actions, and these Bayesian networks were iteratively updated using the MCMC to search for the structure that best explains ongoing observations.

Shi et al. [2004] proposed a hierarchical approach using a propagation network (*P-net*). The structure of a P-net is similar to that of a HMM: an activity is represented in terms of multiple state nodes, their transition probabilities, and the observation probabilities. Their work also decomposes actions into several atomic actions, and constructs a network describing the temporal order needed among them. The main difference between a P-net and a HMM is that the P-net allows activation of multiple state nodes simultaneously. This implies that a P-net is able to model a high-level activity composed of concurrent as well as sequential sub-events. If the sub-events are activated in a particular temporal order specified through the graph, the system is able to deduce that the activity occurred. They have represented an activity of a person performing a chemical experiment using a P-net, and have successfully recognized it.

Statistical approaches are especially suitable when recognizing sequential activities. With enough training data, statistical models are able to reliably recognize corresponding activities even in the case of noisy inputs. The major limitation of statistical approaches are their inherent inability to recognize activities with complex temporal structures, such as an activity composed of concurrent sub-events. For example, HMMs and DBNs have difficulty modeling the relationship of an activity A occurred 'during', 'started with', or 'finished with' an activity B. The edges of HMMs or DBNs specify the sequential order between two nodes, suggesting that they are suitable for modeling sequential relationships, not concurrent relationships.

3.2 Syntactic approaches

Syntactic approaches model human activities as a string of symbols, where each symbol corresponds to an atomic-level action. Similar to the case of hierarchical statistical approaches, syntactic approaches also require atomic-level actions to be

```
Fighting  ->  Punching              : 0.3
          |   Punching Fighting      : 0.7

Punching  ->  stretch withdraw      : 0.8
          |   stretch stay_withdrawn : 0.1
          |   stay_stretched withdraw : 0.1
```

Fig. 11. The figure shows a simplified example of production rules of a SCFG used for representing and recognizing 'fighting' interaction. The 'fighting' is defined as any number of consecutive 'punching' action which itself can be decomposed into 'stretching' and 'withdrawal' similar to Figure 10.

recognized first, using any of the previous techniques. Human activities are represented as a set of production rules generating a string of atomic actions, and they are recognized by adopting parsing techniques from the field of programming language. Context-free grammars (CFGs) and stochastic context-free grammars (SCFGs) have been used by previous researchers to recognize high-level activities. Production rules of CFGs naturally lead to a hierarchical representation and recognition of the activities. Figure 11 shows an example SCFG.

Ivanov and Bobick [2000] proposed a hierarchical approach for the recognition of high-level activities using SCFGs. They divided the framework into two layers: the lower layer using HMMs for the recognition of simple (i.e. atomic) actions, and the higher layer using stochastic parsing techniques for the recognition of high-level activities. They have encoded a large number of stochastic productions rules which are able to explain all activity possibilities. The higher layer parses a string of atomic actions generated by the lower layer, recognizing activities probabilistically. The Earley-Stolcke parsing algorithm is extended to handle uncertain observations. Moore and Essa [2002] also used SCFGs for the recognition of activities, focusing on multi-task activities. By extending [Ivanov and Bobick 2000], they have introduced more reliable error detection and recovery techniques for the recognition. They were able to recognize human activities happening in a Blackjack card game, such as 'a dealer dealt a card to a player', with a high accuracy.

Minnen et al. [2003] adopted SCFGs for the activity recognition as well. Their system focuses on the segmentation problem of multiple objects. They have shown that the semantic-level processing of activities using CFGs may help the segmentation and the tracking of objects. The concept of the *hallucinations* is introduced to compensate for the failures of atomic-level recognition explicitly. Taking advantage of the CFG parsing techniques while considering hallucinations, they have recognized the activity of a person working on the 'Tower of Hanoi' problem. They were able to correctly recognize the activities without any appearance information om the objects, depending solely on the motion information of the activities.

Joo and Chellappa [2006] designed an attribute grammar for recognition, which is an extension of the SCFG. Their grammar attaches semantic tags and conditions to the production rules of the SCFG, enabling the recognition of more descriptive activities. That is, their grammar is able to describe feature constraints as well as temporal constraints of atomic actions. Only when the observations satisfy the syntax of the SCFG (i.e. only when the string can be generated by following the production rules) and the feature constraints are satisfied, their system decides that the activity has occurred. As a result, they have recognized events in a parking lot by tracking cars and humans. Atomic actions including 'parking', 'picking up', and

'walk though' are first detected based on location changes of cars and humans. By representing the typical activity in a parking lot, normal and abnormal activities are distinguished.

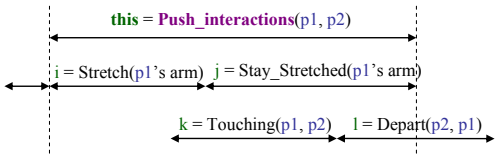
One of the limitations of syntactic approaches is in the recognition of concurrent activities. Syntactic approaches are able to probabilistically recognize hierarchical activities composed of sequential sub-events, but are inherently limited on activities composed of concurrent sub-events. Since syntactic approaches are modeling a high-level activity as a string of atomic-level activities composing them, the temporal ordering of atomic-level activities has to be strictly sequential. In addition, synthetic approaches assume that all observations are parsed by applying their production rules. For these systems, a user must provide a set of production rules for all possible events, even for a large domain. Therefore, they tend to have difficulty when an unknown observation (e.g. a pedestrian) interferes with the system. In order to overcome such limitation, there was an attempt by Kitani et al. [2007] to develop an algorithm to automatically learn grammar rules from observations.

3.3 Description-based approaches

A description-based approach is a recognition approach that explicitly maintains human activities' spatio-temporal structures. They represent a high-level human activity in terms of simpler activities composing the activity (i.e. sub-events), describing their temporal, spatial, and logical relationships. That is, description-based approaches model a human activity as an occurrence of its sub-event (which might be composed of their own sub-events) that satisfies certain relations. Therefore, the recognition of the activity is performed by searching the sub-events satisfying the relations specified in its representation. All description-based approaches are inherently hierarchical (since they use sub-events to represent human activities), and they are able to handle activities with concurrent structures.

In description-based approaches, a *time interval* is usually associated with an occurring sub-event to specify necessary temporal relationships among sub-events. Allen's temporal predicates [Allen 1983; Allen and Ferguson 1994] have been widely adopted for these approaches to specify relationships between time intervals [Pinhanez and Bobick 1998; Siskind 2001; Nevatia et al. 2003; Vu et al. 2003; Ryoo and Aggarwal 2006a; Gupta et al. 2009]. Seven basic predicates that Allen has defined are: *before*, *meets*, *overlaps*, *during*, *starts*, *finishes*, and *equals*. Note that the predicates *before* and *meets* describe sequential relationships while the other predicates are used to specify concurrent relationships. Figure 12 (a) illustrates a conceptual temporal structure of a human-human interaction 'pushing' represented in terms of time intervals.

In a description-based approach, a CFG is often used as a formal syntax for the representation of human activities [Nevatia et al. 2004; Ryoo and Aggarwal 2006a]. Notice that the description-based approaches' usage of CFGs is completely different from that of syntactic approaches: Syntactic approaches directly use CFGs for the recognition, implying that the CFGs themselves describe the semantics of the activities. On the other hand, a description-based approach adopts a CFG as a syntax to represent the activities formally. The activities' semantics are usually encoded in a structure similar to that of a programming language (e.g. Figure 12 (b)), and the CFG only plays a role to ensure that the activity representation fits its



(a)

```
Push_interaction(p1, p2) = {
  list( def('i', Stretch(p1's arm)),
        list( def('j', Stay_stretched(p2's arm)),
              list( def('k', Touching(p1, p2)),
                    def('l', Depart(p2, p1)) ))),
  and( meets('i', 'j'),
       and( during('k', 'this'),
            and( meets('k', 'l'),
                 and( starts('i', 'this'),
                      finishes('j', 'this') ))))
};
```

(b)

Fig. 12. (a) Time intervals of an interaction 'push' and its sub-events, and (b) its programming language-like representation following Ryoo and Aggarwal [2006]'s syntax (©2009 Springer). The figure (a) is a conceptual illustration describing the activity's temporal structure, whose sub-events are organized sequentially as well as concurrently. Following the CFG, we convert this into a formal representation as shown in the figure (b).

grammar. In general, the recognition is performed by developing an approximation algorithm to solve the constraint satisfaction problem (which is NP-hard).

Pinhanez and Bobick [1998] directly adopted the concept of Allen's interval algebra constraint network (*IA-network*) [Allen 1983] to describe the temporal structure of activities. In an IA-network, sub-events are specified as nodes and their temporal relationships are described with typed edges between them. Pinhanez and Bobick have developed a methodology to convert an IA-network into a {past, now, future} network (*PNF-network*). The PNF-network that they have proposed is able to describe the identical temporal information contained in the IA-network, while making it computationally tractable.

They have developed a polynomial time algorithm to process the PNF-network. Their system recognizes the top-level activity by checking which sub-events have already occurred and which have not. They have shown that their representation is expressive enough to recognize cooking activities occurring in a kitchen environment, such as 'picking up a bowl'. Atomic-level actions were manually labeled from the video in their experiments, and their system was able to recognize the activities even when one of the atomic actions was not provided. One of the drawbacks of their system is that a sub-network corresponding to a sub-event has to be specified redundantly if it is used multiple times. Another limitation is that they require all sub-event relations to be expressed in a network form.

Intille and Bobick [1999] designed a description-based recognition approach to analyze plays in American football. Even though their system was limited to use conjunctions of relatively simple temporal predicates (*before* and *around*), they have shown that complex human activities can be represented by listing the temporal constraints in a format similar to those of programming languages, instead of a network form. They have represented human activities with three levels of hierarchy: atomic-level, individual-level, and team-level activities.

A Bayesian belief network is constructed for the recognition of the activity, based on its temporal structure representation. The root node of the belief network corresponds to the high-level activity that the system aims to recognize. The other nodes correspond to the occurrence of the sub-events or describe the temporal

relationships between the sub-events. The nodes become 'true' if the sub-events occur and the relationships are satisfied. Only when all nodes are probabilistically satisfied and propagated to the root node, the activity is said to be detected. Similar to [Pinhanez and Bobick 1998], they have used manually labeled data.

Siskind [2001] also proposed a hierarchical description-based approach for human activity recognition. Notably, it was able to represent and recognize high-level activities with more than three levels. Siskind's methodology uses force dynamics for the recognition of simple actions, and uses the description-based approach called *event logic* to recognize high-level activities. It particularly focused on the recognition of an activity with a *liquid* characteristic, whose occurrences are true for all sub-intervals of a particular time interval. The approach computes the recognized activity's time interval by calculating 'union' and 'intersection' of sub-events' time intervals, assuming liquidity. This suggests that the recognized activity itself can be used as a sub-event of another activity, but is permitted to be used only once.

Nevatia et al. [2003] designed a representation language called 'VERL' to describe human activities. They classified human activities into three categories similar to [Intille and Bobick 1999], enabling the representation of human activities having three levels of hierarchy: primitive events, single-thread composite events, and multi-thread composite events. Allen's temporal predicates, spatial predicates, and logical predicates were used to represent human activities by specifying their necessary conditions. Bayesian networks are used for primitive event recognition, and HMMs are used for the recognition of single-thread composite events (i.e. they are strictly sequential). A heuristic algorithm is designed for the constraint satisfaction problem, recognizing interactions between multiple persons. Their system was probabilistic, but was not able to overcome the failures of low-level components. Vu et al. [2003]'s approach was similar to [Nevatia et al. 2003], while extending the representation to describe activities with any levels of hierarchy. However, unlike Nevatia et al.'s system, only conjunctive predicates are allowed when concatenating multiple temporal relationships (i.e. only *and* allowed, not *or*). Hakeem et al. [2004] have designed a representation language, 'CASEE', which also represent an activity as a conjunction of necessary temporal and causal relations.

Several researchers utilized the *Petri nets* to represent and recognize human activities [Zaidi 1999; Nam et al. 1999; Ghanem et al. 2004]. Petri nets specify the temporal ordering of an activity's sub-events in terms of a graph representation. The recognition is done by sequentially handing tokens in the graph, where each node corresponds to a state before (or after) the completion of particular sub-events. Zaidi [1999] showed that the Petri nets are able to fully represent temporal relationships described by Allen's temporal predicates. Nam et al. [1999] applied the Petri nets for the recognition of hand gestures from videos. Ghanem et al. [2004] took advantage of the Petri nets to represent and recognize interactions between humans and vehicles similar to [Ivanov and Bobick 2000]. Because of the Petri net's characteristic that the tokens cannot describe multiple possibilities and are non-reversible (i.e. the recognition process is strictly sequential), these deterministic systems have limitations in terms of processing complex scenes.

In order to overcome the limitations of the previous approaches, Ryoo and Aggarwal [2006a] proposed a description-based approach using a CFG as a syntax of

their representation language. Their formal grammar enables the representation of human-human interactions with any levels of hierarchy, which are described as logical concatenations (*and*, *or*, and *not*) of complex temporal and spatial relationships among their sub-events. As a result, they have represented high-level human interactions composed of concurrent sub-events (e.g. 'hand shaking' and 'pushing') in terms of time interval variables and predicates (e.g. Allen's temporal predicates). They have developed a hierarchical semantic matching between the observations and the representations for the activity recognition. In the lowest level, Bayesian networks and HMMs are used for the recognition of atomic actions from a sequence of raw image frames. The recognition of represented high-level activities is done by performing a hierarchical matching from the bottom to the top.

In addition, their approach was extended to recognize recursive activities with a continuous nature, such as 'fighting' and 'greeting' [Ryoo and Aggarwal 2006b]. Even though the representation of recursive activities with sequential sub-events has been possible with syntactic approaches, the recognition of recursive activities with complex concurrent sub-events has been limitedly studied. They have introduced the special time interval 'this', which always corresponds to the activity being represented, and proposed an iterative algorithm to recognize activities described using 'this'. With the proposed approach, the recursive activity 'fighting' was represented as a single negative interaction (e.g. 'punching' and 'pushing') followed by a shorter 'fighting', and has successfully been recognized.

Furthermore, Ryoo and Aggarwal [2009a] proposed a probabilistic extension of their recognition framework which is able to compensate for the failures of its low-level components. One of the limitations of description-based approaches is that they are mostly deterministic, and are fragile when their low-level components are noisy. Ryoo and Aggarwal have overcome such limitations. They have used a logistic regression to model the probability distribution of an activity, and used it to detect the activity even when some of its sub-events have been mis-classified. In order to compensate for the complete failure of the atomic-level components (i.e. no atomic action detected at all), they took advantage of the concept of the *hallucination* time intervals, similar to the ones used in [Minnen et al. 2003].

There also has been attempts to adopt symbolic artificial intelligence techniques to recognize human activities. Tran and Davis [2008] adopted Markov logic networks (MLNs) to probabilistically infer events in a parking lot. This 2-layered approach successfully handled uncertainties in human activities. However, their MLNs relied on the assumption that an identical sub-event occurs only once during interactions, limiting itself from being applied to dynamically interacting actors.

Gupta et al. [2009] recently presented a description-based approach for a probabilistic analysis as well. Unlike other description-based approaches designed to recognize complex activities, their approach aims to recognize atomic-level actions more reliably by modeling causality among the actions. A tree structured AND-OR graph similar to [Hongeng et al. 2004] has been used to represent a storyline of a sports game (e.g. baseball), labeling each action (e.g. hitting) that fits the storyline. Their system iteratively searches for the best explaining AND-OR graph structures and the best video-action associations by taking advantage of captions and video trajectories. That is, a representation fitting algorithm has been developed.

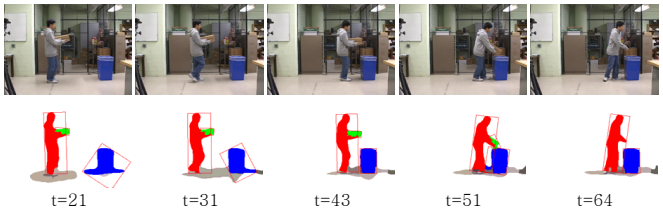
Type	Approaches	Levels of hierarchy	Complex temporal relations	Complex logical concatenations	Recognition of recursive activities	Handle imperfect low-levels
Statistical	Oliver et al. '02	limited (2-levels)				√
	Shi et al. '04	limited (2-levels)	one relation: 'before'			√
	Damen and Hogg '09	limited (2-levels)				√
Syntactic	Ivanov and Bobick '00	unlimited			√	√
	Joo and Chellappa '06	unlimited		conjunctions only	√	√
Description-based	Pinhanez and Bobick '98	limited (redundant nodes required)	network form only	network form only		compensates 1 error
	Intille and Bobick '99	unlimited	two relations: 'before' and 'around'			√
	Siskind '01	unlimited	a sub-event participates only once	√		
	Nevatia et al. '03	limited (3-levels)	√	√		
	Vu et al. '03	unlimited	√	conjunctions only		
	Ghanem et al. '04	unlimited	time intervals of an activity do not overlap	√		
	Ryoo and Aggarwal '09a	unlimited	√	√	√	√
	Gupta et al. '09	limited (2-levels)	√	network form only		√

Table III. A table comparing the abilities of the hierarchical approaches. The column 'levels of hierarchy' describes the possible levels of the activity hierarchy. 'Complex temporal relations' suggests that the approach is able to represent and recognize activities with a complex temporal structure. Similarly, 'complex logical concatenations' shows whether the system is able to represent activities with complex logical concatenations.

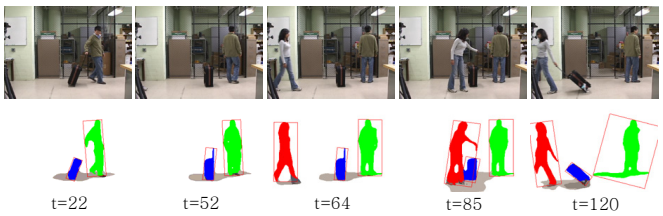
3.4 Comparison

Hierarchical approaches are suitable for recognizing high-level activities which can be decomposed into simpler sub-events. Because of their nature, they can more easily incorporate human knowledge into the systems and require less training data as pointed out by many researchers [Oliver et al. 2002; Nevatia et al. 2003; Ryoo and Aggarwal 2006a]. Statistical and syntactic approaches provide a probabilistic framework for reliable recognition with noisy inputs. However, they have difficulties representing and recognizing activities with concurrently organized sub-events.

Description-based approaches are able to represent and recognize human activities with complex temporal structures. Not only sequentially occurring, but also concurrent organized sub-events are handled with description-based approaches. The major drawback of description-based approaches are their inability to compensate for the failures of low-level components (e.g. gesture detection failure). That is, most of the description-based approaches have a deterministic high-level



(a) A human-object interaction of a person placing a box in a trash bin.



(a) A human-object interaction of a person stealing another's suitcase.

Fig. 13. Example human-object interactions that Ryoo and Aggarwal [2007] have recognized (©2007 IEEE).

component. Pinhanez and Bobick [1998] showed that the high-level system has the potential to compensate for a single low-level detection failure, and a couple of recent works have proposed probabilistic frameworks for description-based approaches [Ryoo and Aggarwal 2009a; Gupta et al. 2009]. Table III compares the abilities of important hierarchical approaches.

4. HUMAN-OBJECT INTERACTIONS AND GROUP ACTIVITIES

In this section, we present and summarize previous papers on the recognition of human-object interactions and those on the recognition of group activities. These approaches fall into different categories if the approach-based taxonomy of the previous sections is applied as shown in Figures 2 and 3. However, even though they use different methodologies for the recognition, they exhibit interesting common properties and characteristics since they share the same objective. In the first subsection, we discuss approaches for analyzing interplays between humans and objects. Next, we compare various recognition approaches for group activities.

4.1 Recognition of interactions between humans and objects

In order to recognize interactions between humans and objects, an integration of multiple components is required. The identification of objects and motion involved in an activity as well as analysis of their interplays is essential for the reliable recognition of human activities involving humans and objects. While we provide an overview of general human-object recognition approaches, we particularly focus on the approaches that analyzed interplays between object recognition, motion estimation, and activity-level analysis toward robust recognition of human-object interactions. Figure 13 shows an example of human-object interactions.

The most typical human-object interaction recognition approaches are the ap-

proaches ignoring interplays between object recognition and motion estimation. In those works, objects are generally recognized first, and activities involving them are recognized by analyzing the objects' motion. They have made the object recognition and motion estimation independent or made it so that the motion estimation is strictly dependent on the object recognition. Most of the previous recognition approaches fall into this category, including the approaches that we have discussed in previous sections [Siskind 2001; Vu et al. 2003; Nevatia et al. 2003; Shi et al. 2004; Yu and Aggarwal 2006; Damen and Hogg 2009].

On the other hand, several researchers have studied relationships and dependencies between objects, motion, and human activities to improve object recognitions as well as activity recognitions [Moore et al. 1999; Gupta and Davis 2007; Ryoo and Aggarwal 2007]. In principle, these components are highly dependent on each other: objects have their own roles, suggesting that the way humans interact with an object depends on the identity of the object. For example, an object 'water bottle' is expected to be involved in a particular type of interaction: 'drinking'. Therefore, the motion related to the water bottle must be different from that of 'spray bottle', even though their appearances are similar. Several researchers have designed a probabilistic model describing mutual information between objects and humans' inherent motion with the objects. The results suggest that the recognition of objects can benefit activity recognition while activity recognition helps the classification of objects, and we discuss these approaches one by one.

Moore et al. [1999] constructed the system that compensates for the failures of object classification with the recognition results of simple actions. Most of the time, their system performs the object recognition first, and then estimates human activities with objects depending on the object recognition results as most of the other researchers have done. However, when an object component fails to make a concrete decision, their systems uses action information of objects to compensate for the object recognition. In order to recognize actions, positions of hands and their tracking results are used. HMMs are applied to characterize actions based on the tracking results. Finally, an object-based evidence is integrated with an action-based evidence using a Bayesian network to decide the final class of the object, making the system recover from the failure of the object recognition. They have tested their system with various objects in office, kitchen, and automobile environments (such as books, phones, bowls, cups, and steering wheels). They focused on the recognition of simple activities of a single person.

Peursum et al. [2005] proposed a Bayesian framework for better labeling of objects based on activity context. Similar to [Moore et al. 1999], they have focused on the fact that humans interact with objects in many different ways, depending on the function of the objects. They have pointed out that appearance (i.e. shape) cues of objects are unreliable due to scale and view point variations, and presented an object recognition solely based on the activity information. The system calculates an *interaction signature* per object, which essentially is a concatenation of activity recognition results involving the object. HMMs are adopted for the action recognition: a 3-D pose skeleton of a person as well as relative locations of objects are analyzed to recognize activities, where each object candidate is computed by region segmentation based on colors. They have recognized objects such as 'floor',

'chair', and 'keyboard', by recognizing printing-related activities.

Gupta and Davis [2007] proposed a probabilistic model integrating an objects' appearance, human motion with objects, and reactions of objects. Similar to [Moore et al. 1999], a Bayesian network is constructed to combine cues. Two types of motion in which humans interact with objects, 'reach motion' and 'manipulation motion', are estimated using trajectories as well as HMMs. Reactions of objects, i.e. the effect of human activity in relation to their interaction with objects such as 'a light going on after pressing the switch', are considered as well for the classification. The Bayesian network integrates all of these information, and makes a final decision to recognize objects and human activities. Human-object interactions involving cups, spray bottles, phones, and flash lights have been recognized in their experiments.

Ryoo and Aggarwal [2007] designed and implemented a recognition system for high-level human-object interactions such as 'a person stealing another's suitcase'. Similar to the above-mentioned approaches [Moore et al. 1999; Gupta and Davis 2007], their object recognition and motion estimation components were constructed to help each other. Furthermore, their system is designed to compensate for object recognition failures or motion estimation failures using high-level activity recognition results probabilistically. That is, their object recognition and motion estimation components not only help each other, but also get feedback from the high-level activity recognition results for improved recognition. For example, by observing a person pulling an object in an airport environment, their system was able to deduce that it is the activity of 'a person carrying a suitcase' and provide feedback that the object in the scene is a 'suitcase'. With experiments, they have shown that the feedback generated by the high-level activity recognitions may benefit object recognition, motion estimation, and low-level tracking of objects.

4.2 Recognition of group activities

Group activities are the activities whose actors are one or more conceptual groups. 'A group of soldiers marching' and 'a group of persons carrying a large object' are examples of simple group activities. In order to recognize group activities, the analysis of activities of individuals as well as their overall relations becomes essential. In this subsection, we discuss the recognition approaches on group activities, while focusing on the types of activities that they have recognized. There are various types of group activities and most of the works specialize in recognizing a particular type among them. Figure 14 illustrates example snapshots of various group activities.

First of all, researchers have focused on the recognition of group activities where each group member has its own role different from the others [Cupillard et al. 2002; Gong and Xiang 2003; Lv et al. 2004; Zhang et al. 2006; Dai et al. 2008]. The goal of these approaches is to recognize an activity of a single group with a limited number of members who exhibit non-uniform behaviors. A group activity of 'presentation' with a fixed number of participants in a meeting room is an example of this type: the presenter will be 'talking' while the other members will be 'taking notes', 'asking questions', and/or 'listening'. For this type of group activity, the system must recognize activities of each individual member and then analyze their structures. By nature, most of these approaches are hierarchical approaches since

Fig. 14. This figure shows example group activities from Ryoo and Aggarwal [2008]'s dataset. From left to right, figures are the snapshots of group activities of 'group carrying', 'group stealing' in an office, 'group stealing' in a shop, 'group fighting', and 'group arresting'. 'Group stealing' indicates a situation which a thief takes an object while the other thieves are distracting its owners.

there exist at least two-levels of activities: activities of the group and activities of individual persons. Statistical hierarchical approaches have been especially popular, which use state models that we discussed in Subsection 3.1. Essentially, this type of group activity is equivalent to multi-agent interactions recognized by [Intille and Bobick 1999; Ivanov and Bobick 2000; Vu et al. 2003; Nevatia et al. 2003; Joo and Chellappa 2006; Ryoo and Aggarwal 2007].

Cupillard et al. [2002] have recognized a group activity using a finite state machine, which is equivalent to a fully observable Markov model. They have used multiple cameras, and were able to recognize an activity 'a group is fighting' which essentially is intra-group fighting of a group composed of two members. Similarly, as presented in Subsection 3.1, Gong and Xiang [2003] used variations of dynamic Bayesian networks to recognize group activities. With their system, they have recognized 'a group of trucks loading (or unloading) baggage on an airplane' which is a group activity of a fixed number of trucks and an airplane. Zhang et al. [2006] recognized a group activity occurring in a meeting room using DBNs, similar to [Gong and Xiang 2003]. Sequentially organized group activities including 'monologues', 'discussion', 'presentation', and 'note-taking' have been successfully recognized. Similarly, Dai et al. [2008] have recognized 'break', 'presentation', and 'discussion' using DBNs with hierarchical structures.

The second type of group activity is the activities which are characterized by the overall motion of entire group members. A group of people 'parading' or 'marching' is a typical example of this type. In contrast to the first type of group activity where the individual activities of specific members are important, the analysis of overall motion and formation changes of entire group members are important for the second type of group activity. By their nature, single-layered approaches are appropriate for their recognition since the entire motion of group members must be considered simultaneously [Vaswani et al. 2003; Khan and Shah 2005].

Vaswani et al. [2003] have recognized group activities of people interacting with an airplane. Their approach corresponds to the category of single-layered exemplar-based sequential approaches that we presented in Subsubsection 2.1.3. They have represented a group activity as a shape change over time frames. At each frame, they have extracted k point objects, and constructed a polygon by treating the extracted points as corners. The points are tracked, and the dynamics of shape changes following the statistical shape theory are maintained. Their system was able to distinguish normal and abnormal activities by comparing the activity shape extracted from an input with a maintained model in a tangent space. Similarly,

Khan and Shah [2005] have recognized a group of people 'parading' by analyzing the overall motion of group members. Their approach is a single-layered space-time approach using trajectory features, discussed in Subsubsection 2.1.2. They have extracted the trajectory of each group member, and analyzed their activities by fitting a 3-D polygon to check the rigidity formation of the group.

Finally, Ryoo and Aggarwal [2008] have developed a general representation and recognition methodology that is able to handle various types of group activities. Their approach is a description-based approach (Subsection 3.3), and various classes of group activities including group actions (e.g. marching), group-group interactions (e.g. group stealing), group-persons interactions (e.g. march by signal), and intra-group interactions (e.g. intra-group fighting) have been represented and recognized with their system. They took advantage of the universal ($\forall$) and existential ($\exists$) quantifiers to describe sub-events (usually activities of individuals) that need to be performed by any one member of a group or by all members of the group. By attaching the universal and existential quantifiers to the participating group members, their system was able to represent most of group activities that previous researchers have recognized. The first class of activities that we discussed above is represented by attaching an existential quantifier to each actor of the group activity. The second class of activities is represented by applying the universal quantifier and by posing spatial constraints to the group. In addition, high-level group activities with complex structures that the previous methods had difficulty representing and recognizing, such as 'a thief stealing an object while other thieves are distracting the owners' or 'policemen arresting a group of criminals', have successfully been represented and recognized with their system.

5. DATASETS AND REAL-TIME APPLICATIONS

In this section, we discuss public datasets available for the performance evaluation of the approaches, and review real-time human activity recognition systems.

5.1 Datasets

Public datasets provide common criterion to measure and compare accuracies of proposed approaches. Therefore, a construction of a dataset containing videos of human activities plays a vital role in the advancement of human activity recognition research. In this subsection, we describe the existing human activity datasets which are currently available, and discuss the characteristics of the datasets. We also compare the performance of the systems tested on an identical dataset.

Existing datasets that have been made publicly available can be categorized into three groups as follows. The first type of datasets includes the KTH dataset [Schuldt et al. 2004] and the Weizmann dataset [Blank et al. 2005], which are designed to test general purpose action recognition systems academically. They contain videos of different participants performing simple actions such as 'walking' and 'waving', which are taken by the authors in a controlled environment. The second type is a class of more application-oriented datasets obtained from realistic environments (e.g. airport). The PETS datasets containing activities like 'baggage stealing' and 'fighting' are typical examples of this type, targeted for surveillance applications. In addition, datasets collected from real video medias such as TV broadcasts and movies have been constructed and presented recently.

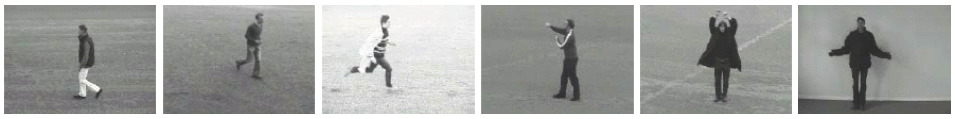


Fig. 15. Example snapshots from the KTH dataset [Schuldt et al. 2004] (©2004 IEEE).

5.1.1 *Action recognition datasets.* A large number of researchers have tested their system on the KTH dataset [Schuldt et al. 2004; Dollar et al. 2005; Jiang et al. 2006; Niebles et al. 2006; Yeo et al. 2006; Ke et al. 2007; Kim et al. 2007; Jhuang et al. 2007; Savarese et al. 2008; Laptev et al. 2008; Liu and Shah 2008; Bregonzio et al. 2009; Rapantzikos et al. 2009; Ryoo and Aggarwal 2009b] and the Weizmann dataset [Blank et al. 2005; Niebles et al. 2006; Scovanner et al. 2007; Rodriguez et al. 2008; Bregonzio et al. 2009]. The KTH dataset is a large scale dataset which contains 2391 videos of six actions performed by 25 subjects. 'Walking', 'jogging', 'running', 'boxing', 'hand waving', and 'hand clapping' are the six actions the dataset contains (Figure 14). Videos are taken at slightly different scales with various backgrounds in indoor and outdoor environments (yet mostly uniform backgrounds). Each video contains repeated executions of a single action in a resolution of 160*120, 25fps. Similarly, the Weizmann dataset consists of 10 action categories with 9 people, resulting in 90 videos. In the Weizmann dataset, a static and simple background is used throughout the videos. Simple human actions of 'running', 'walking', 'jumping-jack', 'jumping forward on two legs', 'jumping in place on two legs', 'galloping sideways', 'waving one hand', 'waving two hand', and 'bending' are performed by the actors. The resolutions of the videos are 180*144, 25fps. Both dataset are composed of relatively simple action-level activities, and only one participant appears in the scene.

What we must note is that these datasets are designed to verify the 'classification' ability of the systems on simple actions. Each video of the datasets contains executions of only one simple action, performed by a single actor. That is, entire motion-related features extracted from each video corresponds to a single action, and the goal is to identify the label of the video while knowing that the video belongs to one of a limited number of known action classes. Further, all actions in both datasets except for the 'bend' action of the Weizmann dataset are periodic actions (e.g. walking), making the videos suitable for action-level classification systems.

Because of its nature, methodologies utilizing spatio-temporal local features (Section 2.1.3) have been popularly tested. As discussed in previous sections, these approaches do not require background subtraction and are robust to scale changes. Further, they are particularly suitable for recognition of periodic actions, since spatio-temporal features will be extracted repeatedly from the periodic actions. Figure 16 compares the classification accuracies of the systems. The X axis corresponds to the time of the publication, while the Y axis shows the classification performance of the systems. Most of the systems tested on the Weizmann dataset have obtained successful results, mainly because of the simplicity of the dataset. Particularly, [Blank et al. 2005; Niebles et al. 2006; Rodriguez et al. 2008; Bregonzio et al. 2009] have obtained more than 0.95 classification accuracy.

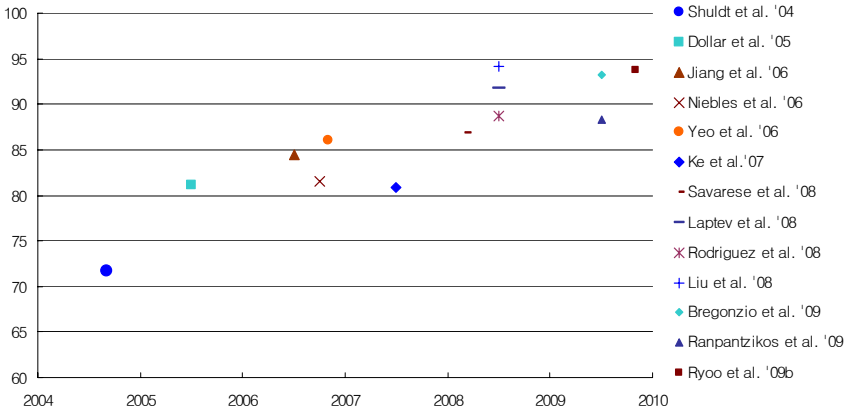


Fig. 16. The classification accuracies of various systems tested on the KTH dataset. The X axis corresponds to the time of publications. Only the results of the systems with the common experimental settings, the original 16 training-9 testing setting [Schuldt et al. 2004; Laptev et al. 2008] or the leave-one actor-out cross validation setting (i.e. 25-fold cross validation), are shown. Results of other systems using non-trivial settings, such as Wong et al. [2007]'s system tested with 100-fold cross validations, Kim et al. [2007]'s system using manually labeled bounding boxes, and Jhuang et al. [2007]'s system tested with the subsets, are not presented. The results of [Niebles et al. 2006; Savarese et al. 2008] are from the system trained with unlabeled data (i.e. unsupervised learning). [Dollar et al. 2005; Niebles et al. 2006; Savarese et al. 2008; Ryoo and Aggarwal 2009b] used the same *cuboid* features, and most of the other systems developed their own features to recognize the actions.

5.1.2 *Surveillance datasets.* On the other hands, the PETS datasets (i.e. the datasets provided at the PETS workshops on 2004, 2006, 2007) and other similar datasets including the i-Lids dataset are composed of realistic videos in uncontrolled environments, such as crowded subway stations and airports. Their camera view points are similar to those of typical CCTVs, and even multiple camera view points are provided in some of the datasets. Cameras are fixed, implying that the backgrounds are static and the scales of persons are mostly constant. Multiple persons and objects appear in the scene simultaneously, and occlusion among them occurs frequently. The goal of these surveillance videos is to test the ability of recognition systems to analyze realistic and specific (e.g. 'baggage abandonment' and 'baggage theft') activities, which are of practical interests. These datasets are closely related to real-time applications, which we will discuss in the following subsection.

The PETS 2004 dataset (also known as the CAVIAR dataset) contains 6 categories of activities where each category is composed of one or more actions: 'walking', 'browsing', 'resting-slumping-fainting', 'leaving bags behind', 'people meeting, walking together, and splitting up', and 'fighting'. Each class has 3 to 6 videos, with a total of 28 videos. Videos are of the 384*288 spatial resolution, 25fps. Background images are provided, and the videos were taken in a shop environment. It is a single view-point dataset (i.e. only one camera was installed).

In the PETS 2006 dataset, 7 long video sequences were provided from every one of the 4 different view points. The PETS 2006 dataset focused on the baggage abandonment problem: each sequence contains an event in which a bag is abandoned in

a train station. Either one person or two persons participated in the activity, and several other pedestrians were presented in the scene. All 4 cameras have a high spatial resolution of 768*576 with 25 fps. The PETS 2007 have a similar setup to the PETS 2006. The videos were taken in an airport hall with 4 cameras, providing 8 executions of human activities. They focused on human-baggage interactions: 2 sequences of general 'loitering', 4 sequences of 'baggage theft', and 2 sequences of 'baggage abandonment' similar to those of the PETS 2006 are provided. The resolution of the images are identical to the PETS 2006. Actors, objects, and pedestrians were severely occluded in the videos.

Similar to the PETS datasets, the recently introduced i-Lids dataset focuses on the baggage abandonment problem. Videos were taken from a single view point in a London subway station, in a crowded environment. The videos not only contain persons and objects, but also a moving subway train in which people get out and get in. Humans and objects are severely occluded by themselves, and pedestrians were easily occluded by pillars in the station. Three videos were provided for training and validation purposes, and a lengthy video containing 6 baggage abandonment activities were given for the testing. Videos have the resolution of 720*576 with 25 fps. They had a real-time abandoned baggage detection competition in AVSS 2007 conference with the dataset [Venetianer et al. 2007; Bhargava et al. 2007].

Several works testing their system on these surveillance datasets have been published [Lv et al. 2004; Kitani et al. 2005; Ribeiro et al. 2007]. In contrast to the datasets mentioned in 5.1.1, these datasets are motivated by the practical needs for the construction of surveillance systems for public safety. They provide more realistic videos in practical environments. However, they lack generality in a certain aspect, since they are highly oriented toward surveillance applications. That is, they are focused on particular types of activities.

5.1.3 *Movie datasets.* Movie datasets are challenging datasets obtained from real movie videos (or from TV broadcasts). Unlike the datasets of 5.1.1, they are not taken in a controlled environment. They are different from the datasets of 5.1.2 as well, since camera view points are moving frequently, and background information is seldom provided. Most of the movie datasets [Ke et al. 2007; Laptev and Perez 2007; Laptev et al. 2008; Rodriguez et al. 2008] focused on relatively simple actions such as 'kissing' and 'hitting'. Even though the actions are simple, each video of an action exhibits person dependent, view-point dependent, and situation dependent variations. Thus, the major challenge is in handling those variations rather than recognizing complex structured activities, and space-time feature-based approaches have been applied to solve the problem.

5.2 Real-time applications

In this subsection, we review several computer vision systems designed to recognize activities in real-time. Even though the approaches that we have discussed in the previous sections have shown results on various types of human activities, most of the proposed algorithms are far from being real-time. In order for an activity recognition methodology to be applicable for real-world applications including surveillance systems, human-computer interfaces, intelligent robots, and autonomous vehicles, this computational gap must be overcome.

Recently, various real-time human activity recognition systems have been proposed, and we review some of them here. The general idea of most of the approaches is to increase the efficiency of the algorithms by simplifying them. They sacrifice the detailed analysis of activities and focus on simple but effective features. Lv et al. [2004] used a traditional Bayesian posterior probability calculation for the recognition of actions. In order to detect activities reliably without spending too much computational cost, their approach searches for an optimal set of features from a large number of features. They have proposed a dynamic programming algorithm to find a set of features that maximizes the detection accuracy on the training data. PETS 2004 dataset has been used for the testing.

Yeo et al. [2006] focused on a frame-to-frame similarity measurement based on optical flow calculations. The key to their system is the fact that the modern video compression technology takes advantage of the optical flows to encode the videos. That is, optical flows are naturally embedded in those videos, and are easily extractable. The similarity between frames are measured based on the optical flow distribution, and they are aggregated to measure the similarities between two videos. Their approach can be viewed as a sequential exemplar-based approach similar to [Efros et al. 2003], and the KTH dataset has been applied to test their system.

Li et al. [2008]'s approach is a space-time trajectory analysis approach. They used the principle component analysis (PCA) to compress the trajectories from a high dimensional space to a low dimension. Several learning algorithms were applied on these low-dimensional trajectories, and the Gaussian mixture model was adopted for the classification. The reduction in the dimensionality provided them the ability to process videos in real-time. Their system was also tested with the KTH dataset.

Notably, Rofouei et al. [2008] have utilized graphical processing units (GPUs) of computer systems to enable the real-time recognition of human activities. Instead of making the algorithms simpler, they focused on the fact that modern hardwares are able to support computationally expensive processing. The state-of-the-art graphic cards are composed of GPUs with many cores, suggesting that they are able to compute repetitive computations in parallel. This implies that they are suitable for the parallel processing of many computer vision algorithms analyzing images and videos (e.g. a GPU-based SIFT feature extraction). Rofouei et al. [2008] have designed a GPU-version algorithm of [Dollar et al. 2005], which is 50 times faster than the CPU implementation of the algorithm without sacrificing the performance. They have illustrated the potential that the use of GPUs (or multi-core CPUs) will greatly improve the speed of computer vision systems, enabling the real-time implementation of existing activity recognition algorithms.

6. CONCLUSION

Computer recognition of human activities is an important area of research in computer vision with applications in many diverse fields. The application to surveillance is natural in today's environment where the tracking and monitoring people is becoming an integral part of everyday activities. Other applications include human-computer interaction, biometrics based on gait or face, and hand and face gesture recognition. We have provided an overview of the current approaches to

human activity recognition. The approaches are diverse and they are yielding a spectrum of results. The senior author of the paper has been involved in the study of motion since the early 1970s [Aggarwal and Duda 1975] and human activity recognition since the early 1980s [Webb and Aggarwal 1982]. The impetus for the study of human motion and human activities was provided by Johansson [1975]'s pioneering work in the early 1970's. Human activity research came to the forefront in the early 1990's.

In this review, we have summarized the methodologies that have previously been explored for the recognition of human activities, and discussed advantages and disadvantages of those approaches. An approach-based taxonomy is designed and applied to categorize previous works. We have discussed non-hierarchical approaches developed for the recognition of gestures and actions as well as hierarchical approaches for the analysis of high-level interactions between multiple humans and objects. Non-hierarchical approaches are again divided into space-time approaches and sequential approaches, and we have discussed the similarities and differences of the two approaches thoroughly. Previous publications following statistical, syntactic, and description-based approaches have been compared for hierarchical approaches.

In 1999, human activity recognition was in its infancy as Aggarwal and Cai [1999] pointed out. A significant amount of progress on human activity recognition has been made in the past 10 years, but it is still far from being an off the shelf technology. We are at a stage where experimental systems are deployed at airports and other public places. It is likely that more and more, such systems will be deployed. There is a strong interaction between the surveillance-authorities and computer vision researchers. For example, Professor Mubarak Shah of the University of Central Florida and the Orlando Police Department are joining forces to develop a system to monitor downtown Orlando: http://server.cs.ucf.edu/~vision/projects/Knight/Knight.html.

Further, today's environment for human activity recognition is significantly different from the scenario at the end of the last decade. The cameras were mostly fixed cameras and without pan-tilt-zoom adjustments. Today's cameras may be mounted on several types of moving platforms ranging from a moving car or a truck to an unmanned aerial vehicle (UAV). A global positioning system may be attached to the camera system to pin-point its location. The recognition of activity from a moving platform poses many more challenges. Noise, tracking, and segmentation issues arising out of stabilization of video add to the difficulty of the problem of the recognition of activities. Tracking is a difficult problem though animals and human do it almost effortlessly. If the tracking algorithm does not extract the object of the focus of attention, recognition of the activity being performed becomes enormously more difficult. Designing an activity recognition system which is able to compensate for such low-level failures in those environments (i.e. moving platforms) is an extremely challenging task.

The future direction of research is obviously encouraged and dictated by applications. The pressing applications are the surveillance and monitoring of public facilities like train stations, underground subways or airports, monitoring patients in a hospital environment or other health care facilities, monitoring activities in the

context of UAV surveillance, and other similar applications. All of these applications are trying to understand the activities of an individual or the activities of a crowd as a whole and as subgroups. These problems will occupy us for a number of years and several generations of graduate students.

As pointed out above, segmenting and tracking multiple persons in videos is harder than it appears. This difficulty is partly due to poor lighting, crowded environments, noisy images, and camera movements. For example, lighting in subways is almost universally poor. Further, it is difficult to segment individuals or their body parts when occlusion is present in the scene. Alternative approaches to segmenting body parts based on analyzing 3-D XYT volumes by extracting gross features are being developed. In particular, 3-D local patch features described in terms of histogram of gradient (HOG) and/or histogram of optical flow (HOOF), such as cuboids [Dollar et al. 2005] and 3-D SIFT [Scovanner et al. 2007], are gaining popularity. These approaches are motivated by the success of object recognition using 2-D local descriptors (e.g. SIFT [Lowe 1999]).

However, they involve long feature vectors obtained from a large 3-D XYT volume created by concatenating image frames, and are likely to have an impact on real time analysis. The 3-D search space is much larger than its 2-D versions. Further, the existing local space-time features generally require a non-textured background for reliable recognition, such as the ones in the KTH and Weizmann datasets [Schuldt et al. 2004; Blank et al. 2005]. Also, a limited amount of work has been published on the 3-D feature-based approaches for analysis of complex human activities. What one needs is an approach which exploits the easy computation of SIFT, HOG, and HOOF operators and avoids the difficulties of segmentation of body parts and/or combines the two approaches in a meaningful way.

One promising direction for enabling real-time implementation is the study of hardware supports. Rofouei et al. [2008] have implemented a GPU-based version of the cuboid feature extractor, utilizing graphical processing units (GPUs) with tens of cores running thousands of threads. The GPU-version turned out to be 50 times faster than the CPU counterpart of it, while obtaining the same results. Modern CPUs and GPUs are composed of multiple cores, and the number of cores is likely to continually increase for the next few years, suggesting computer vision researchers to explore the utilization of them.

There are a number of other innovative approaches being explored. One such approach is exploiting the fact that images, high dimensional signals, are possibly residing in low dimensional manifolds. Several researchers are pursuing issues relating to characterizing the manifolds and exploring the relationships of the manifolds of different activities of the same person or the same activity of different persons [Veeraraghavan et al. 2006]. The temporal segmentation of activities and gestures is still a difficult issue. The inability to simultaneously register rigid and non-rigid parts of a face (in general parts of the human body) contributes to this difficulty. In certain activities, parts of the body move fairly rigidly whereas other parts undergo non-rigid motion, for example, the movement of the head/face. Shape deformations may be modeled as a linear combination of unknown shape bases [la Torre Frade et al. 2007], providing another approach to the recognition of facial expressions.

Hierarchical recognition approaches are being studied intensively especially for

the recognition of complex multi-person activities. Particularly, description-based approaches are gaining an increasing amount of popularity because of their ability to represent and recognize human interactions with complex spatio-temporal structures. Activities with structured scenarios (e.g. most of surveillance scenarios) require hierarchical approaches, and they are showing the potential to make a reliable decision probabilistically. In the near future, hierarchical approaches together with strong action-level detectors such as the ones mentioned above will be explored for reliable recognition of complex activities. As we have covered in previous sections, hierarchical approaches have their advantages in recognition of high-level activities performed by multiple persons, and they must be explored further in the future to support demands from surveillance systems and other applications.

The above areas of research, the space-time feature-based approaches, manifold learning, rigid/non-rigid motion analysis, and hierarchical approaches briefly mentioned are but a small glimpse into the large number of approaches being pursued today. Hopefully, a review in another ten years will document significant progress in human activity recognition to the extent that off the shelf systems would be readily available.

REFERENCES

Aggarwal, J. K. and Cai, Q. 1999. Human motion analysis: A review. *Computer Vision and Image Understanding (CVIU) 73,* 3, 428–440.

Aggarwal, J. K. and Duda, R. O. 1975. Computer analysis of moving polygonal images. *IEEE Transactions on Computers 24,* 10, 966–976.

Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM 26,* 11, 832–843.

Allen, J. F. and Ferguson, G. 1994. Actions and events in interval temporal logic. *Journal of Logic and Computation 4,* 5, 531–579.

Bhargava, M., Chen, C.-C., Ryoo, M. S., and Aggarwal, J. K. 2007. Detection of abandoned objects in crowded environments. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS).*

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. 2005. Actions as space-time shapes. In *IEEE International Conference on Computer Vision (ICCV).* 1395–1402.

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

Bobick, A. F. and Wilson, A. D. 1997. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19,* 12, 1325–1337.

Bregonzio, M., Gong, S., and Xiang, T. 2009. Recognising action as clouds of space-time interest points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Campbell, L. W. and Bobick, A. F. 1995. Recognition of human body motion using phase space constraints. In *IEEE International Conference on Computer Vision (ICCV).* 624–630.

Cedras, C. and Shah, M. 1995. A motion-based recognition: A survey. *Image and Vision Computing 13,* 2, 129–155.

Chomat, O. and Crowley, J. 1999. Probabilistic recognition of activity using local appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Vol. 2.

Cupillard, F., Bremond, F., and Thonnat, M. 2002. Group behavior recognition with multiple cameras. In *IEEE Workshop on Applications of Computer Vision (WACV).* 177–183.

Dai, P., H. Di, H., Dong, L., Tao, L., and Xu, G. 2008. Group interaction analysis in dynamic context. *IEEE Transactions on Systems, Man, and Cybernetics, Part B 38,* 1 (Feb), 275–282.

Damen, D. and Hogg, D. 2009. Recognizing linked events: Searching the space of feasible explanations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Darrell, T. and Pentland, A. 1993. Space-time gestures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 335–340.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*. 65–72.

Efros, A., Berg, A., Mori, G., and Malik, J. 2003. Recognizing action at a distance. In *IEEE International Conference on Computer Vision (ICCV)*. Vol. 2. 726–733.

Gavrila, D. and Davis, L. 1995. Towards 3-D model-based tracking and recognition of human movement. In *International Workshop on Face and Gesture Recognition*. 272–277.

Gavrila, D. M. 1999. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU) 73,* 1, 82–98.

Ghanem, N., DeMenthon, D., Doermann, D., and Davis, L. 2004. Representation and recognition of events in surveillance video using Petri nets. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Gong, S. and Xiang, T. 2003. Recognition of group activities using dynamic probabilistic networks. In *IEEE International Conference on Computer Vision (ICCV)*. 742.

Gupta, A. and Davis, L. S. 2007. Objects in action: An approach for combining action understanding and object perception. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gupta, A., Srinivasan, P., Shi, J., and Davis, L. S. 2009. Understanding videos, constructing plots Learning a visually grounded storyline model from annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hakeem, A., Sheikh, Y., and Shah, M. 2004. CASEE: A hierarchical event representation for the analysis of videos. In *Proceedings of the 20th national conference on Artificial intelligence (AAAI)*. 263–268.

Harris, C. and Stephens, M. 1988. A combined corner and edge detector. In *Alvey Vision Conference*. 147–152.

Hongeng, S., Nevatia, R., and Bremond, F. 2004. Video-based event recognition: Activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding (CVIU) 96,* 2, 129–162.

Intille, S. S. and Bobick, A. F. 1999. A framework for recognizing multi-agent action from visual evidence. In *AAAI/IAAI*. 518–525.

Ivanov, Y. A. and Bobick, A. F. 2000. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22,* 8, 852–872.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. 2007. A biologically inspired system for action recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Jiang, H., Drew, M., , and Li, Z. 2006. Successive convex matching for action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johansson, G. 1975. Visual motion perception. *Scientific American 232,* 6, 76–88.

Joo, S.-W. and Chellappa, R. 2006. Attribute grammar-based event recognition and anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. 107.

Ke, Y., Sukthankar, R., and Hebert, M. 2007. Spatio-temporal shape and flow correlation for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Khan, S. M. and Shah, M. 2005. Detecting group activities using rigidity of formation. In *ACM International Conference on Multimedia (ACM MM)*. 403–406.

Kim, T.-K., Wong, S.-F., and Cipolla, R. 2007. Tensor canonical correlation analysis for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kitani, K. M., Sato, Y., and Sugimoto, A. 2005. Deleted interpolation using a hierarchical Bayesian grammar network for recognizing human activity. In *Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*.

Kitani, K. M., Sato, Y., and Sugimoto, A. 2007. Recovering the basic structure of human activities from a video-based symbol string. In *IEEE Workshop on Motion and Video Computing (WMVC)*.

Kruger, V., Kragic, D., Ude, A., and Geib, C. 2007. The meaning of action: a review on action recognition and mapping. *Advanced Robotics 21,* 13, 1473–1501(29).

la Torre Frade, F. D., Campoy, J., Cohn, J., and Kanade, T. 2007. Simultaneous registration and clustering for temporal segmentation. In *International Conference on Computer Vision Theory and Applications*. 110–115.

Laptev, I. and Lindeberg, T. 2003. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)*. 432.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Laptev, I. and Perez, P. 2007. Retrieving actions in movies. In *IEEE International Conference on Computer Vision (ICCV)*.

Li, Z., Fu, Y., Huang, T., and Yan, S. 2008. Real-time human action recognition by luminance field trajectory analysis. In *ACM International Conference on Multimedia (ACM MM)*. 671–676.

Liu, J., Luo, J., and Shah, M. 2009. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, J. and Shah, M. 2008. Learning human actions via information maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision ICCV*. 1150–1157.

Lublinerman, R., Ozay, N., Zarpalas, D., and Camps, O. 2006. Activity recognition from silhouettes using linear systems and model (in)validation techniques. In *International Conference on Pattern Recognition (ICPR)*. 347–350.

Lv, F., Kang, J., Nevatia, R., Cohen, I., and Medioni, G. 2004. Automatic tracking and labeling of human activities in a video sequence. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Lv, F. and Nevatia, R. 2007. Single view human action recognition using key pose matching and Viterbi path searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minnen, D., Essa, I. A., and Starner, T. 2003. Expectation grammars: Leveraging high-level expectations for activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 626–632.

Moore, D. J. and Essa, I. A. 2002. Recognizing multitasked activities from video using stochastic context-free grammar. In *AAAI/IAAI*. 770–776.

Moore, D. J., Essa, I. A., and Hayes, M. H. 1999. Exploiting human actions and object context for recognition tasks. In *IEEE International Conference on Computer Vision (ICCV)*. Vol. 1. 80–86.

Nam, Y., Wohn, K., and Lee-Kwang, H. 1999. Modeling and recognition of hand gesture using colored Petri nets. *IEEE Transactions on Systems, Man and Cybernetics 29,* 5, 514–521.

Natarajan, P. and Nevatia, R. 2007. Coupled hidden semi Markov models for activity recognition. In *IEEE Workshop on Motion and Video Computing (WMVC)*.

Nevatia, R., Hobbs, J., and Bolles, B. 2004. An ontology for video event representation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. Vol. 7.

Nevatia, R., Zhao, T., and Hongeng, S. 2003. Hierarchical language-based representation of events in video streams. In *IEEE Workshop on Event Mining*.

Nguyen, N. T., Phung, D. Q., Venkatesh, S., and Bui, H. H. 2005. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 955–960.

Niebles, J. C., Wang, H., and Fei-Fei, L. 2006. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference (BMVC)*.

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

Niyogi, S. and Adelson, E. 1994. Analyzing and recognizing walking figures in XYT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 469–474.

Oliver, N., Horvitz, E., and Garg, A. 2002. Layered representations for human activity recognition. In *IEEE International Conference on Multimodal Interfaces (ICMI)*. 3–8.

Oliver, N. M., Rosario, B., and Pentland, A. P. 2000. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22,* 8, 831–843.

Park, S. and Aggarwal, J. K. 2004. A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems 10,* 2, 164–179.

Peursum, P., West, G., and Venkatesh, S. 2005. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *IEEE International Conference on Computer Vision (ICCV)*.

Pinhanez, C. S. and Bobick, A. F. 1998. Human action detection using PNF propagation of temporal constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 898.

Rao, C. and Shah, M. 2001. View-invariance in action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 316–322.

Rapantzikos, K., Avrithis, Y., and Kollias, S. 2009. Dense saliency-based spatiotemporal feature points for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ribeiro, P. C., Moreno, P., and Santos-Victor, J. 2007. Detecting luggage related behaviors using a new temporal boost algorithm. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Rodriguez, M. D., Ahmed, J., and Shah, M. 2008. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rofouei, M., Moazeni, M., and Sarrafzadeh, M. 2008. Fast GPU-based space-time correlation for activity recognition in video sequences. In *IEEE/ACM/IFIP Workshop on Embedded Systems for Real-Time Multimedia (ESTImedia)*. 33–38.

Ryoo, M. S. and Aggarwal, J. K. 2006a. Recognition of composite human activities through context-free grammar based representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1709–1718.

Ryoo, M. S. and Aggarwal, J. K. 2006b. Semantic understanding of continued and recursive human activities. In *International Conference on Pattern Recognition (ICPR)*. 379–382.

Ryoo, M. S. and Aggarwal, J. K. 2007. Hierarchical recognition of human activities interacting with objects. In *2nd International Workshop on Semantic Learning Applications in Multimedia (SLAM), in Proceedings of CVPR*.

Ryoo, M. S. and Aggarwal, J. K. 2008. Recognition of high-level group activities based on activities of individual members. In *IEEE Workshop on Motion and Video Computing (WMVC)*.

Ryoo, M. S. and Aggarwal, J. K. 2009a. Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision (IJCV) 32,* 1, 1–24.

Ryoo, M. S. and Aggarwal, J. K. 2009b. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *IEEE International Conference on Computer Vision (ICCV)*.

Savarese, S., DelPozo, A., Niebles, J., and Fei-Fei, L. 2008. Spatial-temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing (WMVC)*.

Schuldt, C., Laptev, I., and Caputo, B. 2004. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition (ICPR)*. Vol. 3. 32–36.

Scovanner, P., Ali, S., and Shah, M. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *ACM International Conference on Multimedia (ACM MM)*. 357–360.

Shechtman, E. and Irani, M. 2005. Space-time behavior based correlation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 405–412.

Sheikh, Y., Sheikh, M., and Shah, M. 2005. Exploring the space of a human action. In *IEEE International Conference on Computer Vision (ICCV)*. Vol. 1. 144–149.

Shi, Y., Huang, Y., Minnen, D., Bobick, A. F., and Essa, I. A. 2004. Propagation networks for recognition of partially ordered sequential action. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 862–869.

Siskind, J. M. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research (JAIR) 15*, 31–90.

Starner, T. and Pentland, A. 1995. Real-time American Sign Language recognition from video using hidden Markov models. *International Symposium on Computer Vision*, 265.

Tran, S. D. and Davis, L. S. 2008. Event modeling and recognition using markov logic networks. In *Proceedings of European Conference on Computer Vision (ECCV)*. 610–623.

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. 2008. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology 18,* 11 (Nov), 1473–1488.

Vaswani, N., Roy Chowdhury, A., and Chellappa, R. 2003. Activity recognition using the dynamics of the configuration of interacting objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2.

Veeraraghavan, A., Chellappa, R., and Roy-Chowdhury, A. 2006. The function space of an activity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 959–968.

Venetianer, P., Zhang, Z., Yin, W., and Lipton, A. 2007. Stationary target detection using the ObjectVideo surveillance system. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 242–247.

Vu, V.-T., Brémond, F., and Thonnat, M. 2003. Automatic video interpretation: A novel algorithm for temporal scenario recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 1295–1302.

Webb, J. A. and Aggarwal, J. K. 1982. Structure from motion of rigid and jointed objects. *Artificial Intelligence 19*, 107–130.

Wong, S.-F., Kim, T.-K., and Cipolla, R. 2007. Learning motion categories using both semantic and structural information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yacoob, Y. and Black, M. 1998. Parameterized modeling and recognition of activities. In *IEEE International Conference on Computer Vision (ICCV)*. 120–127.

Yamato, J., Ohya, J., and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 379–385.

Yeo, C., Ahammad, P., Ramachandran, K., and Shankar Sastry, S. 2006. Compressed domain real-time action recognition. In *IEEE Workshop on Multimedia Signal Processing*. 33–36.

Yilmaz, A. and Shah, M. 2005a. Actions sketch: a novel action representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 984–989.

Yilmaz, A. and Shah, M. 2005b. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *IEEE International Conference on Computer Vision (ICCV)*.

Yu, E. and Aggarwal, J. K. 2006. Detection of fence climbing from monocular video. In *International Conference on Pattern Recognition (ICPR)*. 375–378.

Zaidi, A. K. 1999. On temporal logic programming using Petri nets. *IEEE Transactions on Systems, Man and Cybernetics 29,* 3, 245–254.

Zelnik-Manor, L. and Irani, M. 2001. Event-based analysis of video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. 2006. Modeling individual and group actions in meetings with layered hmms. *IEEE Transactions on Multimedia 8,* 3, 509–520.