Motion Pattern Analysis for Modeling and Recognition of Complex Human Activities

Nandita M. Nayak, Ricky J.Sethi, Bi Song, Amit K. Roy-Chowdhury

1

Abstract Activity recognition is a field of computer vision which has shown great progress in the past decade. Starting from simple single person activities, research in activity recognition is moving towards more complex scenes involving multiple objects and natural environments. The main challenges in the task include being able to localize and recognize events in a video and deal with the large amount of variation in viewpoint, speed of movement and scale. This chapter gives the reader an overview of the work that has taken place in activity recognition, especially in the domain of complex activities involving multiple interacting objects. We begin with a description of the challenges in activity recognition and give a broad overview of the different approaches. We go into the details of some of the feature descriptors and classification strategies commonly recognized as being the state of the art in this field. We then move to more complex recognition systems, discussing the challenges in complex activity recognition and some of the work which has taken place in this respect. Finally, we provide some examples of recent work in complex activity recognition. The ability to recognize complex behaviors involving multiple

Nandita M Nayak
University of California, Riverside, 900 University Ave. Riverside, CA 92521, e-mail: nandita.nayak@email.ucr.edu

Ricky J. Sethi
University of California, Los Angeles, 4532 Boelter Hall, Los Angeles, CA 90095-1596, e-mail: rickys@sethi.org

Bi Song
University of California, Riverside, 900 University Ave. Riverside, CA 92521,e-mail: bsong@ee.ucr.edu

Amit K. Roy-Chowdhury,
University of California, Riverside, 900 University Ave. Riverside, CA 92521, e-mail: amitrc@ee.ucr.edu

[1] This work has been partially supported by the DARPA VIRAT program and NSF award IIS-0905671.

interacting objects is a very challenging problem and future work needs to study its various aspects features, recognition strategies, models, robustness issues, context, to name a few.

1 Introduction

Activity recognition is the task of interpretation of the activities of objects in video over a period of time. The goal of an system is to extract information on the movements of objects and/or their surroundings from the video data so as to conclude on the events and context in the video in an automated manner. In a simple scenario where the video is segmented to contain only one execution of a human activity, the objective of the system is to correctly classify the activity into its category, whereas in a more complex scenario of a long video sequence containing multiple activities, it may also involve the detection of the starting and ending points of all occurring activities in the video[1].

Activity recognition has been a core area of study in computer vision due to its usefulness in diverse fields such as surveillance, sports analysis, patient monitoring and human computer interaction systems. These diverse applications in turn lead to several kinds of activity recognition systems. For example, in a surveillance system, the interest could be in being able to identify an abnormal activity - such as abandoning of a baggage, unusual grouping of people, unusual movements in a public place, etc. A patient monitoring system might require the system to be familiar with the movements of a patient. A sports analysis system would aim at the detection of certain known events, such as a goal detection or kick detection in a soccer game or the statistical learning of the semantics of the play. A traffic monitoring system would require a detection of events such as congestion, accidents or violation of rules. A gesture based human computer interface such as in video games would require posture and gesture recognition.

Although there is no formal classification of activities into different categories, for the sake of understanding, we will divide activities into simple and complex activities based on the complexity of the recognition task. An activity which involves a single person and lasts only a few seconds can be termed as a . Such video sequences consist of a single action to be categorized. Some examples of simple activities are running, walking, waving, etc. and do not contain much extraneous noise or variations. Although it is uncommon to find such data in the real world, these video sequences are useful in the learning and testing of new models for activity recognition. Popular examples of such activities are found in the Weizmann [6] and KTH [53] datasets.

The task of understanding the behaviors of multiple interacting objects (for eg. people grouping, entering and exiting facilities, group sports) by visual analysis will be termed as complex activity recognition in this chapter. Recognition of complex behaviors makes the analysis of high-level events possible. For example, complex activity recognition can help to build automated recognition systems for suspicious

multi-person behaviors such as group formations and dispersal. Some examples of complex activity datasets are the UT-Interaction dataset [50] and the UCR videoweb dataset [12].

In this chapter, we will look at some techniques of activity recognition. We will start with an overview of the description and classification techniques in activity recognition and take a brief glimpse at abnormal activity recognition. Next, we will show some examples of features used in activity recognition followed by some common recognition strategies. We will then discuss what complex activities are and look at some of the challenges in complex activity recognition. Finally, we will discuss some examples of recent approaches used in the modeling of complex activities.

1.1 Overview of Activity Recognition Methods

The basic steps involved in an activity recognition system are the extraction of features from the video frames and inference of activities from features. A popular approach to activity recognition has been the use of local interest points. Each interest point has a local descriptor to describe the characteristics of the point. Motion analysis is thus brought about by the analysis of feature vectors. Some researchers used spatial interest points to describe a scene [16] [19] [31]. Such approaches are termed as local approaches [8]. Over time, researchers described other robust spatio-temporal feature vectors. SIFT (scale invariant feature transform) [34] and STIP (space time interest points) [28] are commonly used local descriptors in videos. A more recent approach is to combine multiple features in a multiple instance learning (MIL) framework to improve accuracy [9].

Another approach to action recognition is global analysis of the video [8]. This involves a study of the overall motion characteristics of the video. Many of these methods use optical flow to represent motion in a video frame. One example of this method is in [3]. These approaches often involve modeling of the flow statistics over time. Optical flow histograms have commonly been used to compute and model flow statistics like in [15] which demonstrates the use of optical flow histograms in the analysis of soccer matches. In some other cases, human activities have been represented by 3-D space-time shapes where classification is performed by comparing geometric properties of these shapes against training data [6][60].

Methods which have been used for modeling activities can be classified as non-parametric, volumetric and parametric time-series approaches [57]. Non-parametric approaches typically extract a set of features from each frame of the video. Non parametric approaches could involve generation of 2D templates from image features [44][7], 3D object models using shape descriptors or object contours [6] or manifold learning by dimensionality reduction methods such as PCA, locally linear embedding (LLE) [45] and Laplacian eigenmaps [4]. Parametric methods involve learning the parameters of a model for the action using training data. These could involve Hidden Markov Models (HMM) [33] and linear [11] and non-linear dynam-

ical systems [41]. Volumetric methods of action recognition perform sub volume matching or use filter banks for spatio-temporal filtering [61]. Some researchers have used multiple 2D videos to arrive at a 3D model which is then used for view invariant action recognition [40].

1.2 Abnormal Activity Recognition

One of the important applications of activity recognition is the detection of suspicious or anomalous activities. This is the task of being able to detect anything "out of ordinary" in an automated manner. Abnormal activity recognition is about separating events which contain large deviations from the expected. The main challenge in the task is to define normalcy or anomaly. Since it is not always easy to define anomaly, a practical approach to abnormal activity detection is to detect normal events and treat the rest as anomaly [22].

When training data is available, a common approach is to model the normal activities using the training data. Graphical models have popularly been used in such cases. For example, Hidden Markov models (HMMs) have been used in a weakly supervised learning framework in [59] for anomaly detection in industrial processes. Here, the whole frame is taken as a feature vector and reduced to a lower dimension before modeling using HMMs. The authors in [58] use HMMs to model the configuration of multiple point objects in a scene. Abnormal activities are identified as a change in this model. Anomalies are detected by modeling the co-occurrence of pixels during an action using Markov Random Fields (MRFs) in [5].

When training data is not available, attempts have been made to detect anomalous activities by an unsupervised clustering of the data in a given video [35]. Dense clusters are classified as normal whereas rare ones could be anomalous. Anomalous activities have been detected using spatio-temporal context or the surrounding motion in [22]. Crowd anomalies have been detected by crowd motion modeling using Helmholtz decomposition of flow in [38].

2 Feature Descriptors

The first step in an activity recognition system is to extract a set of which constitute the description of motion in the video. These features are the input to the recognition system. Researchers have used several feature descriptors in activity recognition. These can be broadly categorized into two kinds - which represent small spatio-temporal regions of the video and which are used to represent motion in an entire segment of the video [8]. In this section, we will briefly look into some of the popular image descriptors.

2.1 Local Features

We will begin with some examples of local features. The idea behind these features is that they represent points or regions of high interest in the video. It is believed that there is similarity between the local features extracted for similar actions. Thereafter, activity matching is achieved by a comparison of the feature set of the given videos. Examples of local features for different activities are seen in Figure 1

2.1.1 Spatio-temporal Interest Points

Spatio-temporal interest points [28] are points of high gradient in the space-time volume of a video. These are inspired from the SIFT points [34] which are popularly used by the object recognition community. It was found that these points are fairly invariant to scale, rotation and change in illumination. Given a video sequence, the are extracted by first computing a scale-space representation $\mathbf{L}$ by convolution with a spatio-temporal Gaussian kernel $g(x,y,t;\sigma,\tau) = 1/(2\pi\sigma^2\sqrt{2\pi\tau})\exp(-(x^2+y^2)/2\sigma^2 - t^2/2\tau^2)$ with spatial and temporal scale parameters σ and τ. At any point $p(x,y,t)$ in the video, a second moment matrix μ is defined as

$$\mu(\mathbf{p}) = \int_{q\in\mathbb{R}^3} (\nabla\mathbf{L}(\mathbf{q}))(\nabla\mathbf{L}(\mathbf{q}))^T g(p-q;\sigma_i,\tau_i)dq \tag{1}$$

where $\nabla\mathbf{L} = (L_x, L_y, L_z)^T$ denotes the spatio-temporal gradient vector and $(\sigma_i = \gamma\sigma, \tau_i = \gamma\tau)$ are spatial and temporal integration scales with $\gamma = \sqrt{2}$. In other words, μ denotes the gradient of point p in its neighborhood. The interest points are detected as the which are given by significant variations of image value both over space and time. These would correspond to the points where the eigenvalues of μ attain a maxima. The interest points are found by computing the maxima of the interest point operator

$$\begin{aligned}\mathbf{H} &= det(\mu) - k(trace(\mu))^2 \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3\end{aligned} \tag{2}$$

over (x,y,t) subject to $\mathbf{H} \geq 0$ with $k \approx 0.005$. Here, λ_1, λ_2 and λ_3 are the eigenvalues of μ.Next, a is defined for each spatio-temporal interest point. The authors of [28] have shown the use of several kinds of descriptors, some of which are give below:

1. Output of a combination of space-time derivative filters or Gaussian derivatives up to order 4 evaluated at the interest point. These are scale normalized derivatives.
2. Histograms of either spatio-temporal gradients or optical flow computed over a windowed neighborhood or several smaller neighborhoods at different scales for each interest point. These are termed as position dependent histograms since the coordinates are measured relative to the interest point and used together with local image measurements.

Fig. 1 This figure shows the cuboidal features marked for actions boxing, hand clapping and hand waving. These are typical examples of simple activities. The figure is taken from [32]

3. A lower dimensional representation of either optical flow or spatio-temporal gradient vectors (L_x, L_y, L_z) computed over the spatio-temporal neighborhood of the interest point obtained using Principal Component Analysis (PCA).

There are also other descriptors defined such as position independent histograms and global histograms. The reader is recommended to look into [28] for details of these descriptors and a detailed description of the method.

Recognition is performed by examining the feature descriptors of each action. A classifier is trained on these descriptors to obtain the set of features which represent each activity. This method has been used in the recognition of simple actions like walk, run, wave, etc. The use of these features has also been extended to the recognition of multi-person activities [49].

2.1.2 Cuboidal Features

Spatio-temporal interest points are direct 3D counterparts to 2D interest points such as SIFT [34] which look at corners. They are sparse in nature and may not identify all interesting regions of the video. An alternative would be to use [14] which are more dense and represent any strong changes in the video. Here, we look at a response function defined over the image intensities directly. For a stack of images denoted by $I(x,y,t)$, the response function is given by

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{3}$$

where $g(x,y;\sigma)$ is the 2D Gaussian smoothing kernel applied along the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair [18] of 1D Gabor filters applied temporally. These are given by $h_{ev}(t;\tau,w) = -\cos(2\pi tw)\exp(-t^2/\tau^2)$ and $h_{ov}(t;\tau,w) = -\sin(2\pi tw)\exp(-t^2/\tau^2)$, where ω is taken to be $4/\tau$. The two parameters σ and τ correspond to the spatial and temporal scale of the detector. A stationary camera is assumed.

The interest points are detected as the local maxima of this response function. Any region with spatially distinguishing characteristics undergoing a complex motion is found to induce a strong response of the function. Therefore, in addition to

spatio-temporal corners, the detector picks points which exhibit periodicity in image intensities and other kinds of non-linear motion like rotations.

A spatio-temporal cuboid is extracted at each interest point, the size of which is chosen such that it contains most of the data which contributed to the response function. A normalization is applied to the cuboids to make them invariant to small changes in appearance, motion or translation. Three methods have been suggested to extract feature vectors, which are:

1. The normalized pixel values,
2. The brightness gradient calculated at each spatio-temporal location (x,y,t) giving rise to three channels (G_x, G_y, G_t) and
3. Windowed optical flow.

In each method, various kinds of descriptors were explored such as flattened vectors, 1-D histograms and N-D histograms. It was found that the flattened vectors gave the best performance.

This method has been used in behavior recognition of animals such as mice, detection of facial expression and also in the recognition of human activities. Examples of cuboidal features detected for single person activities are seen in Figure 1

2.1.3 Volumetric Features

The [25] are cuboidal volumes in the 3D spatio-temporal domain which represent regions of interest in the video. These features are capable of recognizing actions that may not generate sufficient interest points, such as smooth motion. Volumetric features are found to be robust to changes in scale, viewpoint and speed of action.

Volumetric features are computed on the optical flow of a video. The optical flow is separated into its horizontal and vertical components and volumetric features are computed on each component. For a stack of n frames, the horizontal and vertical components are computed as $v_x(x,y,t)$ and $v_y(x,y,t)$ at pixel locations (x,y) and time t. Two kinds of volumetric features are extracted, "one box features" which are the cumulative sum of the optical flow component and "two box features", which are the differences between cumulative sums over combinations of two boxes calculated over the stack of frames. These features are computed over a chosen number of combinations of windows placed over the volume. The size of the window is varied to extract features at different scales in space and time. The reduced set of volumetric features for each action is identified by a training procedure on these features. These have been effective in the recognition of activities which involve uniform motion such as drinking coffee, picking an object, etc.

2.2 Global Features

The idea behind global features is to represent the motion in an entire frame or a set of frames by a global descriptor. Activities are modeled as a sequence of these global features. Such features are also called sequential features [1]. Here, we will look at some examples of global features.

2.2.1 Motion Descriptors

Motion descriptors [15] are a set of descriptors used for the recognition of actions in low resolution sports videos where the bounding box of each person is available. Motion of the object in each frame is represented by a single . Actions are identified by matching the sequence of motion descriptors, which are based on optical flow.

First, a figure centric spatio-temporal volume is computed for each person. This is done by tracking the human figure and constructing a window at each frame centered at the figure. Optical flow is computed at each frame on the window. The optical flow vector field $\mathbf{V}$ is split into horizontal $\mathbf{V_x}$ and vertical $\mathbf{V_y}$ components, each of which is rectified into four non-negative channels $\mathbf{V_x^+}$,$\mathbf{V_x^-}$,$\mathbf{V_y^+}$ and $\mathbf{V_y^-}$. These are normalized and smoothed with a Gaussian to obtained blurred versions $\hat{\mathbf{Vb}}_\mathbf{x}^+$,$\hat{\mathbf{Vb}}_\mathbf{x}^-$,$\hat{\mathbf{Vb}}_\mathbf{y}^+$ and $\hat{\mathbf{Vb}}_\mathbf{y}^-$. The concatenation of these four vectors gives the motion descriptor of each frame.

Matching is performed by comparing the motion descriptor of each frame of one video with each frame of another video using normalized correlation to generate a frame-to-frame similarity matrix $\mathbf{S}$. The final motion-motion similarity is obtained by a weighted sum of the frame-to-frame similarities over a temporal window $\mathbf{T}$, assigning higher weights to near diagonal elements since a match would result in a stronger response close to the diagonal.

2.2.2 Histogram of Oriented Optical Flow (HOOF)

We have seen that optical flow has been used as a reliable representation of motion information in much of the work described above. It has been shown in [8], that histograms of optical flow are used as motion descriptors of each frame. An activity is modeled as a non-linear dynamic system of the time series of these features. This method has been used in the analysis of single person activities.

When a person moves through a scene, it induces a very characteristic optical flow profile. This profile is captured in the histogram of optical flow which can be considered as a distribution of the direction of flow. Optical flow is binned according to its primary angle after a normalization of magnitude which ensures scale invariance. This histogram is known as the . Histograms of different videos can be compared using various metrics like geodesic kernels, χ^2 distance and Euclidean

distance. The time series of these distances is then evaluated using Binet-Cauchy kernels.

2.2.3 Space-Time Shapes

Actions can be considered as three-dimensional shapes induced by silhouettes in the space-time volume. The authors in [60] have computed by extracting various shape properties using the Poisson's equation and used for representation and classification.

Silhouettes are extracted by background subtraction of the input frames. The concatenation of these silhouettes formes the space-time shape S of each video. A Poisson's equation is formed by assigning each space-time point within this shape the mean time required to reach the boundary. This is done through the equation

$$\nabla^2 \mathbf{F}(\mathbf{x},\mathbf{y},\mathbf{t}) = -1, \tag{4}$$

with $(x,y,t) \in S$ and the Laplacian of $\mathbf{F}$ being $F_{xx} + F_{yy} + F_{tt}$ subject to Dirichlet boundary conditions. $\mathbf{F}$ is obtained by solving the Poisson's equation. A 3×3 Hessian matrix $\mathbf{H}$ is computed at each point of $\mathbf{F}$. It is shown that the different eigenvalues of $\mathbf{H}$ are proportional to different properties of the space-time shape such as plateness (uniformity of depth), stickness (uniformity of hieght and width) and ballness (curvature of the surface). These properties put together form the global descriptor of the action. The Euclidean distance between these measures is taken to be the distance between two actions.

3 Recognition Strategies

In the previous section, we described a few examples of local and global features used in activity recognition systems. The in most of these approaches was either a distance measure computed over the feature descriptors or a shape matching strategy. These methods of recognition are known as non-parametric methods [57]. This section discusses some recognition strategies which attempt to design parametric models or reasoning based models for activities. We will look at a few examples which will describe the state of the art in model based recognition strategies.

3.1 Hidden Markov Models

(HMMs) are one of the most popular state space models used in activity recognition [57]. These models are effective in modeling temporal evolution of dynamical systems. In the activity recognition scenario, each activity is considered to be composed

of a finite set of states. The features which are extracted from a video are considered as the observed variables, whereas the states themselves are hidden. Given the observations, the temporal evolution of a video is modeled as a sequence of probabilistic changes from one state to another. Classification of activities is performed by a comparison of these models.

An example of the use of HMMs is found in [24] where it has been used for identification of persons using gait. The features used here are the outer contour of the binarized silhouette or the silhouette themselves. The features are fed to a Hidden Markov Model as the observations over a period of time and the transition across these features as the hidden variables. The modeling of gait of each person involves computation of the initial probability vector (π), the state transition probability matrix (**A**) and the output probability distribution (**B**). These parameters are estimated using the Baum-Welch algorithm.

3.2 Stochastic Context-Free Grammars

The use of grammars in activity recognition is one of the more recent approaches. These methods are suitable in the modeling of complex interactions between objects or in modeling the relations between sub-activities. These methods express the structure of a process using a set of production rules [57]. Context free grammars are a formalism similar to language grammars which looks at activities as being constructed from words (action primitives). Stochastic context-free grammars are a probabilistic extension of this concept. Here, the structural relation between primitives is learnt from training sequences.

A typical example of stochastic context-free grammars can be found in [23]. Here, stochastic context free grammars are used for recognizing patterns in normal events so as to detect abnormal events in a parking lot scene. An attribute grammar (AG) is defined as

$$AG = (G, SD, AD, R, C) \tag{5}$$

where G is the underlying context free grammar given by $G = (V_N, V_T, P, S)$ with V_N and V_T are the non-terminal and terminal nodes, P are the set of productions and S are the symbols; SD is the semantic domain consisting of coordinates and functions operating on the coordinates, AD are the attributes of each type associated with each symbol occurring in the productions in P, R is the set of attribute evaluation rules associated with each production $p \in P$ and C is the set of semantic conditions associated with P. The production rules here could be certain sub-events or features of the video. The attributes are the characteristics associated with each production, for example the location of the objects associated with the production.

The attribute grammar of an event is obtained by studying the relationships between the different attributes associated with the event. Any new event is parsed and the attributes are evaluated with the known attribute grammar. Any event which does not satisfy this grammar is termed as an abnormal event.

4 Complex Activity Recognition

There are four characteristics which can be used to define activities. These are: kinesics (movement of people or objects), proxemics (the proximity of people with each other or objects), haptics (people-object contact) and chronemics (change with time) [2]. Simple activities are those which consist of one or few periodic atomic actions. They typically span a short duration of time, not more than a few seconds. Some examples of simple activities are walking, running, jumping, bending, etc. Most work on simple activity recognition [49] has focused on the analysis of kinesics and chronemics. Although there is no formal definition of a , in this chapter we will describe a complex activity as one which could involve one or more persons interacting with each other or with some objects. Typical examples of complex activities are a soccer goal, people grouping together and two person activities such as handshake or punching. We see that these activities also involve proxemics and possibly haptics in addition to kinesics and chronemics.

In the field of activity recognition, focus has slowly been shifting from the analysis of simple activities to complex activities. This is because in a real world scenario, we often find that an atomic action does not occur by itself but occurs as an interaction between people and objects. We will now briefly discuss some of the work which has taken place in complex activity recognition.

As compared to a simple activity recognition system, the inherent structure and semantics of complex activities require higher-level representation and reasoning methods [57].There have been different approaches used to analyze complex activities. One common approach has been the use of . Graphical models encode the dependencies between random variables which in many cases are the features which represent the activity. These dependencies are studied with the help of training sequences. Some examples of graphical models commonly used are Belief networks (BNs), Hidden Markov Models (HMMs) and Petri nets. Belief networks and Dynamic Belief Networks (DBNs) are graphical models that encode complex conditional dependencies between a set of random variables which are encoded as local conditional probability densities. These have been used to model two person interactions like kicking, punching, etc by estimating the pose using Bayesian networks and the temporal evolution using Dynamic Bayesian networks [43][62]. A grid based belief propagation method was used for human pose estimation in [29]. Graphical models often model activities as a sequential set of atomic actions. A statistical model is created for each activity. The likelihood of each activity is given by the probability of the model generating the obtained observations [48].

A popular approach for modeling complex activities has been the use of stochastic and . It is often noticed that a complex large-scale activity often can be considered as a combination of several simple sub-activities that have explicit semantic meanings [63]. Constructing grammars can provide useful insights in such cases. These methods try to learn the rules describing the dynamics of the system. These often involve hierarchical approaches which parallel language grammars in terms of construction of sentences from words and alphabets. A typical example is when the activity recognition task is split into two steps. First, bottom-up statistical method can

be used to detect simple sub-activities. Then the prior structure knowledge is used to construct a composite activity model [20]. In another instance, context free grammars in [47] followed a hierarchical approach where the lower-levels are composed of HMMs and Bayesian Networks, whereas the higher level interactions are modeled by context free grammars [57]. More complex models like Dependent Dirichlet Process-Hidden Markov Models (DDP-HMMs) have the ability to jointly learn co-occurring activities and their time dependencies [27].

Knowledge and logic based approaches have also been used in complex activity recognition [57]. Logic based approaches construct logical rules to describe the presence of an activity. For instance, a hierarchical structure could be used by defining descriptors of actions extracted from low-level features through several mid-level layers. Next, a rule based method is used to approximate the probability of occurrence of a specific activity by matching the properties of the agent with the expected distributions for a particular action [37]. Recently, the use of visual cues to detect relations among persons have been explored in a social network model [13].

Description based methods try to identify relationships between different actions such as “before”, “after”, “along with”, etc. The algorithm described in [49] is one such method which uses spatio-temporal feature descriptors. The Bag of Words approach [21] disregards order and tries to model complex activities based on the occurrence probabilities of different features. Attempts have been made to improve on this idea by identifying neighborhoods which can help in recognition [26] and by accommodating pairwise relationships in the feature vector to consider local ordering of features [36]. Hierarchical methods have also been proposed which build complex models by starting from simpler ones and finding relationships between them [42].

Many of these approaches require either tracking body parts, or contextual object detection, or atomic action/primitive event recognition. Sometimes tracks and precise primitive action recognition may not be easily obtained for complex/interactive activities since such scenes frequently contain occlusions and clutter. Spatio-temporal feature based approaches, like [14], hold promise since no tracking is assumed. The statistics of these features are then used in recognition schemes [21]. Recently, spatial and long-term temporal correlations of these local features were considered and promising results shown. The work in [8] models the video as a time-series of frame-wide feature histograms and brings the temporal aspect into picture. A matching kernel using “correlograms” was presented in [52], which looked at the spatial relationships. A recent work [48] proposes a match function to compare spatio-temporal relationships in the feature by using temporal and spatial predicates, which we will describe in detail later.

Often, there are not enough training videos available for learning complex human activities; thus, recognizing activities based on just a single video example is of high interest. An approach of creating a large number of semi-artificial training videos from an original activity video was presented in [46]. A self-similarity descriptor that correlates local patches was proposed in [56]. A generalization of [56] was presented in [54], where spacetime local steering kernels were used.

4.1 Challenges in complex motion analysis

Activity recognition is a challenging task for several reasons. Any activity recognition system is efficient only if it can deal with changes in pose, lighting, viewpoint and scale. These variations increase the dimensionality of the problem. These problems are prevalent to a greater degree when it comes to complex activity analysis. There is a large amount of structural variation in a complex activity, therefore the dimension of the feature space is high. The feature-space also becomes sparser with the dimension, thus requiring a larger number of samples to build efficient class-conditional models thus bringing in the Curse of Dimensionality [57]. Issues of scale, viewpoint and lighting also get harder to deal with for this reason.

Most of the simple activity recognition systems in the past had been tested on sequences recorded in a noise free controlled environment. Although these systems might work reasonably well in such an environment, they may not work in a real world environment which contains noise and background clutter. This problem is more prominent in a complex recognition system since there are multiple motions in the scene and they can easily be confused with the clutter.

Another challenge in complex motion analysis is the presence of multiple activities occurring simultaneously. Although many approaches can deal with noise with sufficient training data, there are difficulties in recognizing hierarchical activities with complex temporal structures, such as an activity composed of concurrent subevents. Therefore many methods are more suited for modeling sequential activities rather than concurrent ones [48]. In addition, as stated in [48], as an activity gets more complex, many existing approaches need a greater amount of training data, preventing them from being applied to highly complex activities.

5 Some Recent Approaches in Complex Activity Recognition

As discussed in the previous section, there are several approaches which have been adopted to extend activity recognition to more complex scenarios. In this section, we will look into a few examples of activity recognition which involve activities of single or multiple objects in a natural setting.

5.1 Spatio-Temporal Relationship Match

The use of spatio-temporal features have been extended to the recognition of multi-person activities like handshake, push, kick and punch by the analysis of spatio-temporal relationships in [49]. Spatio-temporal interest points are often used in a Bag-Of Features framework where the combinations of interest points are learnt for classification by discarding its temporal ordering. Although this works fairly well in recognition of simple activities, ordering plays a key role in recognition of

more complex activities involving multiple tasks. Here, in addition to looking at the combination of interest points, we also need to study the between these interest points.

The method in [49] aims to compare the structure of interest points between two videos to determine their similarity. After computing the spatio-temporal interest points over the video and the time duration for which each point is detected (start time and end time), the spatial and temporal relationship between these points is calculated. The spatial relationships are quantified using predicates *near*, far, *xnear* (x-coordinate is near) and *ynear* (y-coordinate is near). These predicates are set based on a threshold. Similarly, some of the temporal predicates are *equals* (complete time overlap and equal durations), $before$ (one completes before the other starts), *meets* (one starts as the other ends), *overlaps* (partial time overlap), *during* (complete time overlap but unequal durations), *starts* (both start together) and $finishes$ (both end together). A 3D histogram whose dimensions are $featuretype \times featuretype \times relationship$ is formed for each video. Two videos are compared by computing the similarity between the bins of their corresponding histograms using a bin counting mechanism termed as a *spatio-temporal relationship match kernel*. This method has been shown to be effective in activity classification using a training database and in localization of events in a scene containing multiple unrelated activities using a partial matching of bins.

5.2 String of Feature Graphs

String of Feature Graphs are a generalization of the method in [49]. Here, a string representation which matches the spatio-temporal ordering is proposed as the feature model of a video. Each string consists of a set of , each of which contains the spatio-temporal relationship between feature points in a certain time interval. Similarities between activities are identified using a graph matching framework on these strings. This method has also been used to recognize activities involving multiple persons in the presence of other irrelevant persons [17].

Given a video, the first step is to extract the spatio-temporal interest points. The video is divided into time intervals of length t_I. A graph is constructed using the features in each time interval, the nodes of the graph being the feature points and edge weights being the pairwise spatio-temporal distance between them. The construction of a feature graph is illustrated in Figure 2a). The similarity between two such feature graphs can be calculated by finding the correspondence between these graphs. This can be done using the spectral technique given in [30].

The comparison of two video sequences can be performed by the comparison of their corresponding feature graphs. Since there could be a difference in speeds of actions even between two similar activities, the time series of feature graphs have to be normalized for comparison. A technique called dynamic time warping [51] uses a dynamic programming technique to match two time series in a flexible manner.

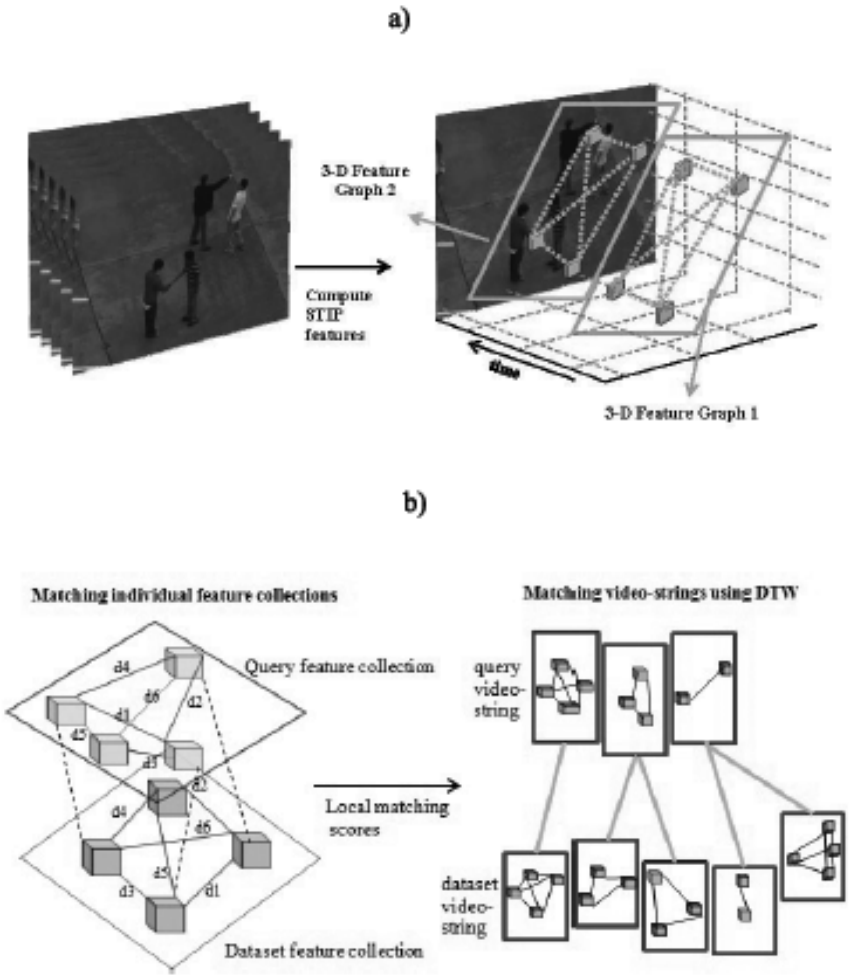

Fig. 2 This figure illustrates the construction and comparison of strings of feature graphs. Fig a) shows the construction of a graph using spatio-temporal relations between STIP featurs in a time window. Fig b) illustrates matching of two videos using dynamic time warping

This method is used here to compute the overall distance between two activities as illustrated in Figure 2b).

String of Feature Graphs can be used in a query based retrieval of activities with a single or very few example videos.

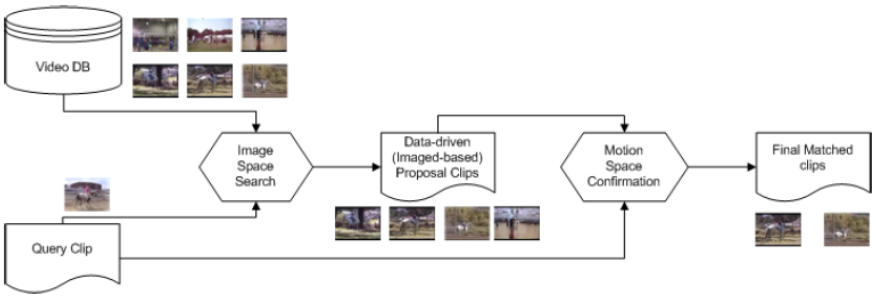

Fig. 3 This figure shows the overall algorithm for a hierarchical video search using stochastic integration of motion and image features.

5.3 Stochastic Integration of Motion and Image Features for Hierarchical Video Search

The problem of matching activity videos in a large database containing videos of several complex activities is a challenging problem due to the extremely large search space. The authors in [55] suggest a way of performing a search in such scenarios in an efficient manner. This method is a typical example of the application of stochastic search algorithms in activity recognition.

An activity is composed of two kinds of data - spatial data in the form of pixel values and motion data in the form of flow. The search algorithm proceeds in a acceptance-rejection manner alternating between the pixel and motion information. The search is based on the which is an improvised version of the traditional Markov Chain Monte Carlo technique. The Hamiltonian Monte Carlo search consists of the following basic steps.

1. Generate a random sample from the data distribution
2. Dynamic Transition Step - Perturbation via Hamiltonian dynamics, also known as the Leapfrog step
3. Metropolis-Hastings Step: acceptance/rejection of the proposed random sample

Given a video database, for each video, the image and motion features are computed. The image features used are shape of the silhouette or shape of the trajectory of an object. The motion features are the trajectory or spatio-temporal interest points. After computing these features on the entire database, a probability distribution is defined over both these features for each activity. The search for a particular activity would require finding the maxima of the joint distribution of both these features for that activity. To do this, the authors employ the Data Driven Hamiltonian Monte Carlo mechanism. The motion space is sampled randomly. The sample is then accepted or rejected based on the Metropolis-Hastings algorithm. The next sample is chosen after the dynamic transition step. The procedure repeats till the maxima is reached. The overall algorithm is illustrated in Figure 3.

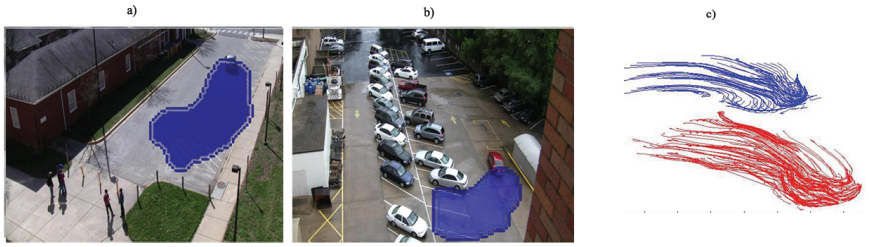

Fig. 4 The figure shows the utility of streaklines in activity recognition. a) and b) show the segmentation of streaklines for two videos of a car turning left. c) plots the streaklines in the two cases.

The method is found to be effective in search of natural videos, for example, videos from the Youtube database.

5.4 Dynamic Modeling of Streaklines for Motion Pattern Analysis

Natural videos consist of multiple objects interacting with each other in complex ways. The underlying patterns of motion contain valuable information about the activities and optical flow can be used to extract these patterns. A can be defined as locations of all particles in a vector field at a given time that passed through a particular point. Streaklines have been used in the analysis of crowded videos in [38]. This concept can be extended to the task of activity recognition. The streaklines representation can be combined with dynamical systems modeling approaches in order to analyze the motion patterns in a wide range of natural videos. This combination provides a powerful tool to identify similar segments in a video that exhibit similar motion patterns.

Given a video, we can compute streaklines over time windows of a particular size. These streaklines capture the motion information over the entire time window. Since similar motions result in similar streaklines, we can cluster the streaklines based on their similarity to identify segments of the video with similar motion. In a multi-object video, this segmentation helps to separate out different object motions. Next, these clusters are modeled as a linear dynamical system (eg: an Auto-Regressive Moving Average (ARMA) model). Here, the actual motion of the underlying pixel is taken to be the input to the model and the observed streaklines form the output. The distance between any two models representing motion patterns can be computed using the subspace angles as defined in [10]. Figure 4a) and b) show two videos containing cars turning left. The streakline cluster is overlaid on the image. We notice that the segmentation has extracted this motion and similar motions result in similar segmentation. Figure 4b) shows the segmented streaklines for the two videos. It can be seen that they look very similar.

This method is suitable to analyze videos which have multiple objects moving in the scene and exhibiting different motions. Since we are separately modeling each motion pattern, this allows for a partial matching of videos and also in the grouping of similar videos in an unsupervised manner.

6 Conclusion

In this chapter, we discussed a few techniques in activity recognition. An activity recognition system generally requires a set of features to represent the activity and a recognition strategy for classification. Features can either be local (describing a small region of the scene) or global (describing the entire scene). We have discussed different local features with different levels of sparsity. For example, spatio-temporal features are usually corners, cuboidal features are more dense and represent regions of strong changing motion, and volumetric features are capable of identifying regions of uniform motion. Global features represent the entire scene of a certain level of detail. For example, HOOF features represent an entire frame with the histogram of its flow field while space time shapes are a collection of silhouettes.

We have looked at a variety of recognition strategies. In some cases, a classifier is learnt on the set of features. In some other applications, activities are modeled using these features as observations. Graphical models and stochastic grammars can model co-occurrences of features. Logic based approaches classify based on relationships between these features. We have also defined what complex activities are, the challenges involved and the common recognition strategies used. We discussed some recent work in complex activity recognition which looked at different applications like modeling multi-person interactions, localizing motion patterns and searching methods for complex activities. Abnormal activity recognition is an application where activities in a scene are modeled so as to identify anomalies.

The main challenge for future work in this area will be to develop descriptors and recognition strategies that can work in natural videos under a wide variety of environmental conditions.

7 Further Reading

This chapter presents an overview the state of the art in activity recognition and briefly describes a few methods of activity recognition. The description in Section 1.1 is based on the survey presented in [57]. We recommend the reader to go through this reference for a comprehensive survey of work which has taken place in activity recognition. We also recommend the reader to [1] for an overview in sequential and hierarchical methods of complex activity recognition. There are several feature descriptors used in this field, a few of which we described. We also recommend the reader to the work in [39] for a comparison of the different local descriptors.

References

1. J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, pages 428–440, 1999.
2. P.A. Anderson. *Nonverbal Communication: Forms and Functions*. Waveland Press, Inc., Long Grove, IL, 2008, second edition, 2008.
3. J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1994.
4. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2001.
5. Y. Benezeth, P.M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. *Computer Vision and Pattern Recognition*, 0:2458–2465, 2009.
6. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision*, pages 1395–1402, Washington, DC, USA, 2005.
7. A.F. Bobick. and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
8. R. Chaudhary, A. Ravichandran, G.D. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition*, pages 1932–1939, 2009.
9. N.I. Cinbis and S. Sclaroff. Object, scene and actions: combining multiple features for human action recognition. In *European Conference on Computer Vision*, pages I: 494–507, 2010.
10. K.D. Cock and B.D. Moor. Subspace angles and distances between arma models. *Systems and Control Letters*, 46(4):265–270, 2002.
11. N.P. Cuntoor and R. Chellappa. Epitomic representation of human cctivities. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.
12. G. Denina, B. Bhanu, H. Nguyen, C. Ding, A. Kamal, C. Ravishanka, A. Roy-Chowdhury, A. Ivers, and B. Varda. Videoweb dataset for multi-camera activities and non-verbal communication. In *Distributed Video Sensor Networks*. Springer, 2010.
13. L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In *European Conference on Computer Vision*, pages IV: 410–423, 2010.
14. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 0:65–72, 2005.
15. A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference of Computer Vision*, pages 726–733, 2003.
16. W. Forstner and E. Gulch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, 1987.
17. U. Gaur. Complex activity recognition using string of feature graphs. Master's thesis, University of California, Riverside, CA, USA, 2010.
18. G.H. Granlund and H. Knutsson. *Signal processing for computer vision*. Kluwer, December 1995.
19. C. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151, 1988.
20. Y.A. Ivanov and A.F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.
21. H. Wang J.C. Niebles and L. Fei-fei. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference*, 2006.
22. F. Jiang, J. Yuan, S.A. Tsaftaris, and A.K. Katsaggelos. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115:323–333, March 2011.

23. S.W. Joo and R. Chellappa. Attribute grammar-based event recognition and anomaly detection. *Computer Vision and Pattern Recognition Workshop*, 0:107, 2006.
24. A. Kale, A. Sundaresan, A. N. Rajagopalan, N.P. Cuntoor, A.K. Roy-Chowdhury, V. Krueger, and R. Chellappa. Identification of humans using gait. *IEEE Transactions on Image Processing*, 13:1163–1173, 2004.
25. Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *International Conference on Computer Vision*, volume 1, pages 166 – 173, October 2005.
26. A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Computer Vision and Pattern Recognition*, pages 2046–2053, 2010.
27. D. Kuettel, M.D. Breitenstein, L.J. Van Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *Computer Vision and Pattern Recognition*, pages 1951–1958, 2010.
28. I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *First International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.
29. M.W. Lee and R. Nevatia. Human pose tracking in monocular sequence using multilevel structured models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):27–38, 2009.
30. M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *International Conference of Computer Vision*, volume 2, pages 1482 – 1489, October 2005.
31. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30:79–116, 1998.
32. H. Liu, R.S. Feris, V. Krueger, and M.T. Sun. Unsupervised action classification using space-time link analysis. *EURASIP Journal on Image and Video Processing*, 2010.
33. Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2006, 2006.
34. D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–, Washington, DC, USA, 1999.
35. D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man and Cybernetics*, 35(3):397–408, June 2005.
36. P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *European Conference on Computer Vision*, September 2010.
37. G. Medioni, R. Nevatia, and I. Cohen. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:873–889, 1998.
38. R. Mehran, B.E. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In *European Conference on Computer Vision*, pages III: 439–452, 2010.
39. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, October 2005.
40. P. Natarajan, V.K. Singh, and R. Nevatia. Learning 3d action models from a few 2d videos for view invariant action recognition. In *Computer Vision and Pattern Recognition*, pages 20006–2013, 2010.
41. B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.
42. S. Park. A hierarchical bayesian network for event recognition of human actions and interactions. In *Association For Computing Machinery Multimedia Systems Journal*, pages 164–179, 2004.
43. S. Park and J.K. Aggarwal. Recognition of two-person interactions using a hierarchical bayesian network. In *ACM SIGMM International Workshop on Video Surveillance*, pages 65–76, New York, NY, USA, 2003.
44. R. Polana and R.C. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, 23(3):261–282, June 1997.

45. S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.
46. M. S. Ryoo and W. Yu. One video is sufficient? human activity recognition using active video composition. In *IEEE Workshop on Motion and Video Computing*, 2011.
47. M.S. Ryoo and J.K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Computer Vision and Pattern Recognition*, pages II: 1709–1718, 2006.
48. M.S. Ryoo and J.K. Aggarwal. Semantic representation and recognition of rontinued and recursive human activities. *International Journal on Computer Vision*, 82(1):1–24, 2009.
49. M.S. Ryoo and J.K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *International Conference on Computer Vision*, pages 1593–1600, 2009.
50. M.S. Ryoo, C. Chen, J.K. Aggarwal, and A. Roy-Chowdhury. An overview of contest on semantic description of human activities (sdha) 2010. In *International Conference on Pattern Recognition*, pages 270–285, Berlin, Heidelberg, 2010.
51. H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 26(1):43–49, 1978.
52. S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, 2008.
53. C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Internation Conference on Pattern Recognition*, 2004.
54. H. J. Seo and P. Milanfar. Detection of human actions from a single example. In *International Conference on Computer Vision*, 2009.
55. R.J. Sethi, A.K. Roy-Chowdhury, and S. Ali. Activity recognition by integrating the physics of motion with a neuromorphic model of perception. In *IEEE Workshop on Motion and Video Computing*, 2009.
56. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition*, 2007.
57. P.K. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, November 2008.
58. N. Vaswani, A. Roy-Chowdhury, and R. Chellappa. "shape activity": A continuous state hmm for moving/deforming shapes with application to abnormal activity detection. *IEEE Transactions on Image Processing*, 2005.
59. I.S. Wersborg, T. Bautze, F. Born, and K. Diepold. A cognitive approach for a robotic welding system that can learn how to weld from acoustic data. In *Computational Intelligence in Robotics and Automation*, pages 108–113, Piscataway, NJ, USA, 2009.
60. A. Yilmaz and M. Shah. Actions sketch: A novel action representation. *Computer Vision and Pattern Recognition*, 1:984–989, 2005.
61. R.A. Young and R.M. Lesperance. The gaussian derivative model for spatial-temporal vision. *Spatial Vision*, 2001:3–4, 2001.
62. Z. Zeng and Ji. Qiang. Knowledge based activity recognition with dynamic bayesian network. In *European Conference in Computer Vision*, Crete, Greece, 2010.
63. Z. Zhang, K.Q. Huang, and T.N. Tan. Complex activity representation and recognition by extended stochastic grammar. In *Asian Conference on Computer Vision*, pages I:150–159, 2006.

Glossary

activity recognition The task of identification of actions or goals of objects in a video in an automated manner is known as activity recognition. 2

complex activity Complex activity is an activity involving multiple objects exhibiting motion in a natural setting . 25

features Features are descriptions of a video or a portion of the video. They are used for classification of activities. 3

motion patterns Motion patterns indicate the patterns in the motion field of a video. Objects which are moving in a similar manner create similar motion patterns. 17

recognition strategies Recognition strategies are the approaches used to classify activities given a set of features . 3

Index