Human Motion: Modeling and Recognition of Actions and Interactions

J.K. Aggarwal
Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712, USA
aggarwaljk@mail.utexas.edu

Sangho Park
Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712, USA
sangho@ece.utexas.edu

Abstract

Processing of image sequences has progressed from simple structure from motion paradigm to the recognition of actions / interactions as events. Understanding human activities in video has many potential applications including automated surveillance, video archival/retrieval, medical diagnosis, sports analysis, and human-computer interaction. Understanding human activities involves various steps of low-level vision processing such as segmentation, tracking, pose recovery, and trajectory estimation as well as high-level processing tasks such as body modeling and representation of action. While low-level processing has been actively studied, high-level processing is just beginning to receive attention. This is partly because high-level processing depends on the results of low-level processing. However, high-level processing also requires some independent and additional approaches and methodologies. In this paper, we focus on the following aspects of high-level processing: (1) human body modeling, (2) level of detail needed to understand human actions, (3) approaches to human action recognition, and (4) high-level recognition schemes with domain knowledge. The review is illustrated by examples of each of the areas discussed, including recent developments in our work on understanding human activities.

1. Introduction

Processing of image sequences has progressed from simple structure from motion paradigm to the recognition of actions / interactions as events. Understanding human activities in video has many potential applications including automated surveillance, video archival/retrieval, medical diagnosis, sports analysis, and human-computer interaction. Processing image and video data in real-time has become possible due to technological developments. With these developments, machine understanding of video data containing human activities is essential to the next generation of computer applications [13].

Computer vision-based recognition of human activity involves the understanding of human motion. Understanding human motion is a complex and challenging task in computer vision due to ambiguity caused by nonrigid body articulation, loose clothing, and mutual occlusion, as well as by image noise by shadow and illumination change. For example, recognition of outdoor activities is significantly influenced by weather and lighting.

The task of understanding human motion can be approached from various levels of detail according to the complexity involved in the behavior. Modeling and recognition of human behavior requires the characterization of motion understanding problems in terms of taxonomy of motion. In an early work on machine perception of motion, Nagel [36] used his taxonomy of "change, event, verb, episode, history" to reflect different dimensions of the problem. The different dimensions of the problem are related to the degree of domain knowledge required to achieve the task.

Bobick [6] used a different taxonomy of human motion: "movement, activity, action". In his taxonomy, *movements* are atomic primitives, requiring no contextual or sequence knowledge to be recognized. *Activity* refers to a sequence of movements or states, where the only real knowledge is the statistics of the sequence; much of the recent work in gesture understanding falls into this category of motion understanding. Finally, *actions* are larger scale events which typically include interaction with the environment and causal relations; action understanding straddles the gray division between perception and cognition, computer vision and artificial intelligence [6].

In this paper we give an overview of the *high-level* understanding of human motion: actions and interactions. High-level understanding of human motion requires various steps of low-level vision processing including segmentation, tracking, pose recovery, and trajectory estimation. These low-level processing tasks have been extensively studied. (See [2, 1, 17, 34] for review.) In this paper, we will concentrate on recent developments in one area

Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'04)
0-7695-2223-8/04 $ 20.00 IEEE

of high-level processing: recognition of actions / interactions as events. We will focus on the following aspects of high-level processing: (1) human body modeling, (2) level of detail in understanding human action, (3) approaches to human action recognition, and (4) high-level recognition schemes with domain knowledge. We will include examples of recent developments using our work in understanding human activities.

2. Human body modeling

Human motion analysis is a part of motion understanding in computer vision. Aggarwal and Cai [1] reviewed various studies in motion analysis and classified different types of motion as *rigid* or *nonrigid* motion, based on the degree of the nonrigidity of the objects. Following Kambhamettu et al.'s classification scheme [31], human motion is a kind of *articulated motion*, a subset of nonrigid motion. Articulated motion of a human body is composed of piecewise rigid motions of individual body parts, but the overall motion of the entire human body is not rigid.

We can classify the studies of articulated human motion into those methods which use *a priori* shape models, called 'model-based', and those that do not use shape models, called 'appearance-based' or 'view-based' methods. This classification is based on whether or not well-defined *a priori* knowledge of the object shape is employed in the motion analysis. Both approaches have advantages and disadvantages. Appearance-based approaches are applicable to more diverse situations because they don't require a specific object model. However, appearance-based approaches are sensitive to noise in general, because they lack any mechanism to distinguish noise from signal in visual input. On the other hand, model-based approaches can efficiently integrate shape knowledge and visual input, and are better for high-level understanding of complicated motions. However, model-based approaches usually require additional processing steps of model selection and parameter estimation to fit the model to a given visual input. Addition of a new activity or motion may require significant complexity to model-based techniques.

Appearance-based approaches build a body representation in a bottom-up fashion by first detecting appropriate features in an image, whereas model-based approaches build the body representation by fitting to the image data the predefined parameter values of a parametric body model. In model-based approaches, the fitting process involves either an optimization scheme such as the least square method [18] or a stochastic sampling scheme such as the particle filtering method [27].

In each approach, the human body can be represented at various levels of detail, involving either bounding boxes, stick figures, 2D contours, or 3D volumes, based on the complexity of model required in an application. The bounding box representation is one of the simplest models of the human body. Its representational ability is limited; the bounding box model is useful when the human body in the image sequence is so small that it occupies only a few pixels. The stick figure representation regards human body as a composition of sticks and the joints between them, based on the observation that human motion is essentially the movements of the supporting bones [54, 37, 12]. The 2D contour representation regards the human body as a projection from 3D space onto the 2D image plane, and approximates the human body by means of deformable contours, cardboards, or ribbons, e.g., the silhouette contour model [22, 5], 2D blob model [55], and cardboard model [23]. 3D volumetric models attempt to describe the detailed human body in 3D space by using polyhedrons such as elliptical cylinders [21, 41], generalized cones [16], or spheres [40]. As one moves from bounding boxes to stick figures, to 2D contours and 3D volumes, the model complexity increases along with the level of detail. More detailed models can represent more complex aspects of human activity, but they require more computational complexity. 3D models may require stereo information obtained from multiple cameras.

The degree of detail needed in the body representation depends on the application. For example, some applications may not need to represent the entire body, or may not need the details of body parts. In such cases, it is sufficient to use a simple representation. Intille et al. [26] represented each player in an American football game as a bounding box, and tracked each player by maintaining the bounding box across image frames. Another example of a simple representation of the human body is Wren et al.'s [55] use of 2D blobs, to represent the approximate locations of the body parts such as head, hands, torso, and feet. Color and intensity features were used to locate and track the body parts across image frames. Different applications require different degrees of detail of the human body. The degree of detail is related to limitations in the physical dimensions of the camera sensors; that is, the less the spatial resolution is, the wider area is covered, and vice versa. Therefore, a tradeoff exists between the spatial resolution and the viewing range. One way to overcome the tradeoff is to use multiple distributed cameras that cover different parts of the entire site. Cai and Aggarwal [11] presented a human tracking system that used distributed cameras mounted at various positions in an indoor environment to cover the wide area. Another way to overcome the limitation is to use an active camera with a panning/tilting head and zooming lens [58].

Various features other than color or intensity have also been used for the human body representation. For example, Bobick and Davis [7] used as the feature the binarized foreground portions of the image accumulated across the image sequence (called 'motion history image'), and classified

different action types of a person based on the motion history image. Sato and Aggarwal [51] used pixel velocity to track humans and recognize their interactions. In their system, moving pixels that have similar velocity are grouped to form moving human blobs. Each person is represented by a moving bounding box, and the relative translation patterns of the boxes are classified into different interaction classes.

Some research has used multiple features such as stereo, edge, sound, color, velocity in optical flow, and/or intensity to enhance performance. For example, Kakusho et al. [30] combined pixel intensity and beat information in music to classify the types of social dancing performed by a pair of persons. Zhou and Aggarwal [57] combined the motion, spatial position, shape and color of blobs to represent humans and vehicles. Their system represents each of the disjoint humans and vehicles as a single blob on a frame-by-frame basis and establishes the correspondence between consecutive frames to track the objects.

Park and Aggarwal [43, 44] exemplify a recent development in appearance-based human body modeling by processing the image at multiple levels, specifically the pixel, blob, body part, and sequence levels. Pixels are grouped into blobs according to color similarity, and multiple blobs are grouped to form body part regions such as head, upper body, lower body, face, hair, hands and legs. The body parts are tracked along the sequence.

3. Level of detail needed to understand actions

A significant amount of research on action recognition has been conducted on the analysis of single-person activities. Interaction recognition may involve the recognition of two-person interactions, group interactions among three or more persons, human-computer interactions, or the interaction between a human and objects. In general, each of these tasks requires a different level of image resolution and a different representation scheme. The more people that are included in the image, the fewer pixels will be occupied by each person, resulting in a low-resolution image. Therefore, different methods are needed for each case. Recognition of actions and interactions can be achieved at different levels of detail in the analysis: gross, intermediate, and detailed level.

At the gross level, individual persons are represented as distinct moving bounding boxes or ellipses. At this level, the recognition of human interaction is constrained to gross-level understanding about the moving patterns of the boxes/ellipses. Video surveillance applications often employ gross level recognition. Sato and Aggarwal [51] presented a system to recognize two-person pedestrian interactions such as *meet, depart, follow*, etc. Their system tracked individual persons as distinct moving boxes and classified the translation patterns of the two bounding boxes. Zhou and Aggarwal [57] presented a tracking system that discriminated human motion versus vehicle motion. Their system analyzed the average of the movements of individual pixels in the foreground area along the image sequence, but did not detect specific parts of the human bodies or the vehicles to classify their identity. Analyzing a sequence of people playing football, Intille and Bobick [24, 25] represented each person as a rectangular bounding box, and achieved recognition by interpreting the interaction types between the bounding boxes. Knowledge of the rules of American football was used to interpret the interactions between players, which required the user-involved construction of a rule-based network that represented the football rules and the interrelationships between the rules.

At the intermediate level, individual persons are represented by their major body parts such as head, torso, arms, and legs. Some video surveillance applications make use of intermediate-level recognition. Various methods have been proposed to segment the human body into the major body parts. Haritaoglu et al. [19] tracked multiple people using silhouette images. They applied a background subtraction method in order to segment the foreground regions of the image that contain a group of people, and projected a binarized foreground silhouette image onto the horizontal image axis to detect the head centroid of each person in the group. However, the goal of their system was to track the group of people, rather than to recognize interactions among the people in the group.

At the detailed level, several researchers have worked on recognition of human activities in terms of a single body part. This research domain mainly aims at developing the gesture-based human-computer interfaces (HCI). Recognition of hand gestures for human-computer interfaces has been studied by many researchers. Pavlovic et al. [46] surveyed the literature on visual interpretation of hand gestures in the context of its role in HCI. Hand gesture and arm motion have been considered to be promising candidates for indexing visual commands to control the computer. In this kind of application, the entire body model may not be necessary; a high-resolution image of the hand or arm is more crucial as input. Hand gesture is usually used to represent the vocabulary such as digits and alphabet letters, whereas arm motion is used to represent cursor movements and zooming-in/zooming-out action etc. in controlling the computer. The visual recognition of hand commands is closely related to the recognition of human activity in general, because both require the computer to 'classify and interpret' human motion types. Using a kinematic hand model for hand tracking, Rehg and Kanade [49] took a least squares approach to estimate and compare stored hand models to input hand-image sequences. Min et al. [33] proposed a combined approach that uses both a static representation for hand gesture and a dynamic representation for arm mo-

tion for human-computer interaction.

Recognition of interactions between a human and objects is also an important issue in visual surveillance, which uses knowledge about the objects or environment to understand 'what is going on in the scene'. Bobick and Pinhanez [8] proposed a system that combines knowledge of the environment and view-based vision algorithms. For a TV studio where a cooking show is broadcast, they developed a smart TV camera system that autonomously zooms in/out and selects an important scene. The system is an example of using contextual as well as visual information for high-level understanding of human-object interaction. Ayer and Shah [4] proposed a system that makes context-based decisions about the actions of people in a room. Their system recognizes behaviors such as entering a room, using a computer terminal, opening a cabinet, and picking up a phone. The system monitors pre-specified regions of interest (ROI) to detect the events. An accurate specification of the ROI is critical in this system.

Normally the interaction between a human and objects involves a single agent, while the interaction between two persons involves two independent agents. Therefore, a two-person interaction is defined by the relative relations between the two autonomous agents, i.e. the two humans. As an example of a coarse-level approach, Oliver et al. [39] proposed a system that models and recognizes human interactions in a visual surveillance task in a pedestrian plaza. The system classifies two-person interactions such as following another person, altering one's path to meet another, approaching and passing by, using Bayesian statistics to compare a test sequence with stored interaction models. Their system is a coarse-level recognition system in that it represents each person as a single, low-resolution blob in a birds-eye view image of a wide pedestrian plaza. Haritaoglu and Davis [23] developed a more detailed system that tracks each person's head, torso, hands, and feet, using a cardboard human body model and tracking each person's parts by template matching. Since the main goal of the system was to track individual persons in a scene, the recognition of interaction patterns between the persons was not attempted. Kakusho et al. [30] combined audio and visual information for recognition of social dancing of two persons. Auditory information from the beat of the dance music serves as break points that divide motion elements, called 'figures', that comprise different types of dance such as waltz, tango, and blues, etc. Park and Aggarwal [42] presented a system that recognizes two-person interactions at a detailed level. The system achieves the recognition by applying a K-nearest neighbor classifier to the parametric human-interaction model which describes the interpersonal configuration. The system independently classifies each frame by estimating the relative poses of the interacting persons, and provides a tool to detect the initiation and the termination of an interaction with no parsing procedure for video sequences.

4. Approaches to human action recognition

Human action recognition is carried out by classifying the video data as one of several types of actions. Traditionally two different paradigms exist; *direct recognition* and *recognition by reconstruction*. The paradigm of *direct recognition* recognizes human actions directly from image data without the reconstruction of body part poses. Polana and Nelson [47] proposed a system that recognizes pedestrian behavior with or without occlusion such as walking, running, and passing by another person. Their model-free approach uses periodicity information in cyclic motion and does not require a body model. Cyclic motion is characterized by repeated activity caused by arm swing and foot stepping, etc. They consider an image sequence as a spatio-temporal solid with two spatial dimensions and a time dimension. Repeated activity is indexed by periodic or semi-periodic bumps in the image solid that generate smoothing curves. They refer the curves as 'reference curves', and compare them the test curves in order to recognize activity types. The recognition of pedestrian motion is achieved by choosing the best matching between the reference curves and the test curves. Niyogi and Adelson [37] also used cyclic motion information for analyzing human gait patterns. They built a walking model with body translation and leg displacement in image sequence by using a simplified body model. In a spatio-temporal image solid, they extract pedestrians' different walking paths and track a person's walking. Baumberg and Hogg [5] track a pedestrian by tracking the pedestrian's contour image using an adaptive B-spline shape model. Their approach is view-based in that they do not attempt to locate body parts in the contour. They just fit an adaptive contour to the actual image in each frame, and track the pedestrian by keeping the adaptive contour model updated. The paradigm of *recognition by reconstruction* constructs the object poses from image and then recognize human actions. Park and Aggarwal [42] developed a system that estimates human body poses using a stick figure model and recognizes actions and interactions between two persons. Dever et al. [15] proposed a method to analyze silhouettes and recognize a classic holdup position of armed robbery. The recognition is achieved by first segmenting the skeleton of the silhouette into separate pieces of the body, then identifying the positions of the arms.

Some systems don't follow one of the two paradigms in a strict sense; they adopt a hybrid approach. Another useful distinction among recognition approaches is *static* vs. *dynamic* recognition. Human motion recognition entails the analysis of a series of images concatenated in time: the

video sequence. The video sequence can be analyzed either by using *static representation* of individual frames or by using *dynamic representation* of the entire sequence [52]. An approach using static representation analyzes the individual frames first and then combines the results into the sequence, whereas an approach using dynamic representation treats the entire sequence (or a fixed length of it) as its basic analysis unit; that is, the analysis unit is the trajectory information across the sequence. In an early research, Herman [20] used the static representation of stick-figures of a given frame to analyze different poses of a person. He inferred emotions and actions at a given frame based on the person's pose. His stick figure was built by manually locating body parts, and he analyzed individual frames separately, without considering the interrelations between the frames in a sequence. Akita [3] used static representation of silhouette images to recognize different motions of a person in tennis play. Most of static approaches have applied the method of template matching in recognition.

Most studies using dynamic representation have applied the methods of 'Dynamic Time Warping'(DTW) [50] or 'Hidden Markov Model'(HMM) [48]. DTW is a method of sequence comparison used in various applications such as DNA comparison in microbiology, comparison of strings of symbols in signal transmission, and analysis of bird songs and human speech. DTW deals with differences between sequences by operations of deletion-insertion, compression-expansion, and substitution, of subsequences. By defining a metric of how much the sequences differ before and after these operations, DTW classifies the sequences. DTW can also be applied to image sequences. DTW lacks, however, the consideration of interactions between nearby subsequences occurring in time. In many actual situations, a sequence has higher correlation between closer subsequences than between distant subsequences.

HMM considers this correlation between adjacent time instances by formulating a Markov process. HMM assumes that the observation sequence is stochastically determined by a hidden process which is composed of a fixed number of hidden states. HMM consists of a finite set of hidden states, a set of observation states, probabilities of state transition between hidden states, probability of state transition from hidden to observation states, and initial state probabilities. The success of HMM models in dealing with speech data motivated vision researchers to apply HMMs to visual recognition problems. Speech data is represented in a well-defined modeling unit (e.g. phonemes) of the spoken language. In contrast to speech recognition, computer vision lacks a general underlying modeling unit, i.e. how to map the images into symbols. Therefore, in order to recognize complex actions and interactions, researchers combine various HMM structures to build coupled HMM, abstract HMM, hierarchical HMM etc. [10]. Rabiner [48] presented a good tutorial for details of HMM. HMM has been more popular than DTW for dynamic representation of action because of its ability to handle uncertainty in its stochastic framework. For example, Bobick et al.[52] used HMM to classify hand motions. Min et al.[33] used both static representation and dynamic representation for hand gesture recognition. Yang et al.[56] applied HMM to classify human action intent and to learn human skills.

A significant limitation of the HMM is that it cannot handle three or more independent processes efficiently [39]. To alleviate this problem, researchers have developed dynamic Bayesian networks (DBNs) as generalization of HMMs [35]. DBNs are directed graphical models of a stochastic process, and can generalize HMMs by representing the hidden and observed states in terms of state variables, which can have complex interdependencies. The interdependencies among the state variables can be efficiently represented by the structure of the directed graphical models. Park and Aggarwal [44] presented a method for the recognition of two-person interactions using a hierarchical Bayesian network (BN). In their system the poses of simultaneously tracked body parts are estimated at the low level of the BN, and the overall body pose is estimated at the high level of the BN. The evolution of the poses of the multiple body parts are processed by a dynamic Bayesian network.

5. High-level recognition schemes with domain knowledge

In this section, we review interpretation schemes in high-level understanding of human actions and interactions. In order to *understand* 'what is going on in the scene', we require coherence in our interpretation and understanding of visual input and knowledge about the world. Various interpretation schemes have been proposed including the rule-based network, physics constraints, causal analysis, syntactic analysis, and finite automata method.

Intille and Bobick [25] developed a rule-based inference network to interpret American Football games. Their system was based on manual construction of the rules which are application-specific. Mann et al. [32] proposed a more universal scheme for understanding a scene using constraints based on Newtonian physics. They interpreted the event of moving a hand to manipulate objects in terms of physical laws such as gravity and friction etc. They analyzed a moving single arm that handles objects under the constraints of simple physical laws: contact, attach, body motor, linear motor, and angular motor, etc. Their system was to make inference about how a hand lifts a can, how a hand pushes an object, or how an object is supported by another object. They showed that physical laws can be used as effective causal constraints because every object in the world is situated under the control of some physical laws.

Their system infers a hypothesized hierarchy of physical laws that can most likely explain a given visual input. To solve conflicting explanations, they assumed a 'priority ordering' of those physics laws based on the hierarchy, and performed a breadth-first search[14] to effectively generate a consistent hypothesis. Their system automatically generates plausible interpretations about kinematic and dynamic properties of the scenes that contain the simple hand motions.

Ohno et al. [38] applied mechanical constraints to the tasks of tracking multiple players under occlusion and estimating the 3D position of a fast-moving ball in soccer games. Occluded players are successfully tracked and identified by using the position and velocity information of each player in recent frames. Estimation of the ball position is facilitated by limiting the search space by constraining the possible bouncing directions of the ball based on mechanics.

Human-involved scene understanding, however, needs more abstract and meaningful schemes than purely physical laws for interpreting 'what is happening in the scene'. Brand and Essa [9] interpreted body gestures of a single person in story telling by means of metaphor of actual motions in a real situation. They use the constraints of body kinematics and body dynamics to identify gestures that refer to actual motor plans such as 'lifting', 'pushing', 'opening', and 'resting' of arms. They formulate the knowledge about causal processes of the body kinematics/dynamics in terms of position, velocity, and acceleration of wrists, elbows, and shoulders. Therefore the 'physical constraints' of the human body parts change to more meaningful 'gestures' of a person. An example of knowledge about causal processes is that 'the greatest acceleration of hands occurs at the beginning of different actions.' Using these constraints, they designed filters to detect different motion types. The filters can be regarded as detectors for individual motion segments. The filters are combined to produce a preliminary segmentation of a video sequence into underlying motor plans. They analyzed the movements of the arms of a single person in front-view images. It seems that this approach can be generalized to more diverse situations.

Understanding very long image sequences requires another abstraction scheme: the 'event'. The event is regarded as a sort of summary of the whole sequence, and the summary is closely related to real world knowledge. The real world knowledge is efficiently represented in terms of the syntactic approaches to pattern recognition problem. Ivanov and Bobick [28] present an automatic surveillance system that labels events and interactions by using syntactic constraints. Their goal is to label person-vehicle interactions such as 'pick-up', 'drop-off', 'exit', and 'enter' in an open parking lot. In a similar way in Ayer and Shah[4], they built an environment map and selected ROI's. Their system is composed of a tracking module, an event generator, and a parser. The tracker tracks any moving objects, and the event generator maps the object tracks onto a set of predetermined discrete events. The event generator uses an environment map (i.e., the scene model of the parking lot) as a contextual information to assign visual changes to discrete events in the parking lot. The parser uses an activity grammar to parse the sequence of discrete events into meaningful labels of interactions between the person and the vehicle. This approach seems adequate for grouping into meaningful labels the discrete events distributed sparsely along a lengthy sequence of visual surveillance data. That is, the syntactic method makes it possible to extract meaningful interactions from the heterogeneous sequence that is composed of, and intermingled with, several different processes to which HMM methods can not be applied.

Iwai et al. [29] proposed a hybrid system that combines HMM and finite automaton to recognize continuous gestures of Japanese sign language. HMM is not efficient to recognize concatenated continuous gestures due to the transition of different gestures. Therefore an automaton is added and layered up on the HMMs to deal with contextual information of the gestures. That is, the HMMs deal with individual differences of gesture models, and the automaton changes the final output by dealing with the context information about what the previous gesture was. The context information contains as knowledge base the protocol gestures that a typical gesture often follows a certain gesture. Wada and Matsuyama [53] also presents a hybrid system composed of HMMs and finite automaton for recognizing the events of multiple persons' 'entering/exiting' a room. Their approach is a combination of bottom-up and top-down processes in that the image features detected in specific ROI's are fed up to the HMM modules in a bottom-up fashion and constraints on feasible hypotheses about events are controlled by the nondeterministic finite automaton in a top-down fashion.

Park and Aggarwal [45] proposed an event semantics to represent and recognize human-human interactions. The linguistic *verb argument structure* is used to represent human action in terms of *agent-motion-target* triplets. Spatial and temporal constraints are used for a decision tree to recognize specific interactions. In this framework, human action is automatically represented in terms of verbal description according to *subject* + *verb* + *object* syntax, and human interaction is represented in terms of *cause*-and-*effect* semantics between the human actions.

6. Conclusions

We have given an overview of past and current developments in the modeling and recognition of human motion. Our discussion has focused on (1) human body modeling,

(2) level of detail needed to understand human actions, (3) approaches to human action recognition, and (4) high-level recognition schemes with domain knowledge. The methods of human body modeling range from coarse representation of the body such as bounding box to fine representation such as 3D superellipsoids and mesh grids. There is a trade-off between the degree of fidelity and the computational cost of the system. The choice of a proper body model depends on the application. The level of detail in understanding human action is classified as gross-, intermediate-, and detailed-level understanding. Usually different levels are related to different application domains, and require different approaches and methodologies. Approaches to human action recognition can be classified as *recognition by reconstruction* vs. *direct recognition*, in which the former is proper for detailed recognition tasks and the latter is proper for gross-level recognition. An alternative classification of the approaches is *static* vs. *dynamic* recognition. An ultimate goal of understanding human motion is to know what is happening in the scene. The goal involves high-level knowledge about the scene and context. We have reviewed various high-level recognition schemes that use domain-specific knowledge such as TV show or universal knowledge such as language-based semantics. Research on human motion understanding is in its infancy, and new techniques are expected to solve related problems and to improve the performance of a system. Future directions of human motion understanding would include activity awareness in computer-supported environment such as unobtrusive monitoring of patients in hospitals. Multi-modal integration of video and audio data would be another promising research direction for better understanding of human activities.

References

[1] J.K. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):295–304, 1999.

[2] J.K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding*, pages 142–156, 1997.

[3] K. Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, 17 (1), 1984.

[4] D. Ayers and M. Shah. Recognizing human action in a static room. In *Proceedings of IEEE Computer Society Workshop on Interpretation of Visual Motion*, pages 42–46, 1998.

[5] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, 1994.

[6] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B*, 352:1257–1265, 1997.

[7] A. Bobick and J. W. Davis. An appearance-based representation of action. In *Proceedings of the International Conference on Pattern Recognition*, Vienna, Austria, August 1996.

[8] A. Bobick and C. Pinhanez. Controlling view-based algorithms using approximate world models and action information. In *Proceedings of IEEE Conference on Computer Vision and Pattern*, pages 955–961, Puerto Rico, 1997.

[9] M. Brand and I. Essa. Causal analysis for visual gesture understanding. *MIT Tech. Report*, 1995.

[10] H. Bui, S. Venkatesh, and G. West. Tracking and surveillance in wide-area spatial environments using the abstract hidden markov model. *Int'l Journal of Pattern Recognition and Artificial Intelligence*, 15(1):177–195, 2001.

[11] Q. Cai and J.K. Aggarwal. Tracking human motion in a structured environment using a distributed camera system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(11):1241–1247, 1999.

[12] Z. Chen and H. Lee. Knowledge-guided visual perception of 3d human gait from a single image sequence. *IEEE transactions on Systems, Man, and Cybernetics*, 22 (2):336–342, 1992.

[13] R. Cipolla and A. Pentland, editors. *Computer Vision of Human-Machine Interaction*. Cambridge University Press, 1998.

[14] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1998.

[15] J. Dever, N. da VitoriaLobo, and M. Shah. Automatic visual recognition of armed robbery. In *IEEE International Conference on Pattern Recognition*, Canada, 2002.

[16] D. Gavrila. *Vision-based 3-D Tracking of Humans in Action*. PhD thesis, Department of Computer Science, University of Maryland, 1996.

[17] D. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[18] D. Gavrila and L. Davis. 3-d model-based tracking of human upper body movement: a multi-view approach. In *Proceedings of Int'l Symposium on Computer Vision*, pages 253–258, 1995.

[19] I. Haritaoglu, D. Harwood, and L. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *Proceedings of Second IEEE Workshop on Visual Surveillance*, pages 6–13, Fort Collins, USA, 1999.

[20] M. Herman. *Understanding body postures of human stick figures*. PhD thesis, University of Maryland, 1979.

[21] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1 (1):5–20, 1983.

[22] D. H. I. Haritaoglu and L. Davis. Ghost: A human body part labeling system using silhouettes. In *Fourteenth International Conference on Pattern Recognition*, 1998.

[23] D. H. I. Haritaoglu and L. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. In *Third International Conference on Automatic Face and Gesture*, 1998.

[24] S. Intille. Tracking using a local closed-world assumption: Tracking in the football domain. Master's thesis, MIT, 1994.

[25] S. Intille and A. Bobick. Representation and visual recognition of complex, multi-agent actions using belief networks. Technical Report No. 454, MIT, 1998.

[26] S. S. Intille, J. Davis, and A. Bobick. Real time closed world tracking. In *Proceedings IEEE International Conference on Computer Vision*, pages 697–703, 1997.

[27] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[28] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[29] Y. Iwai, H. Shimizu, and M. Yachida. Real-time context-based gesture recognition using hmm and automaton. In *IEEE proc. on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 127 –134, 1999.

[30] K. Kakusho, N. Babaguchi, and T. Kitahashi. Recognition of social dancing from auditory and visual information. In *Proceedings of the Second Int't Conference on Automatic Face and Gesture Recognition*, pages 289–294, 1996.

[31] C. Kambhamettu, D. Goldgof, D. Terzopoulos, and T. Huang. Nonrigid motion analysis. *Handbook of PRIP: Computer Vision*, 2, 1994.

[32] R. Mann, A. Jepson, and J. Siskind. Computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65 (2):113–128, 1997.

[33] H. S. J. O. T. Min, B. Yoon and T. Ejima. Visual recognition of static/dynamic gesture: Gesture-driven editing system. *Journal of Visual Languages and Computing*, 10:291–309, 1999.

[34] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[35] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California at Berkeley, 2002.

[36] H. H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.

[37] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. In *Proceedings of Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[38] Y. Ohno, J. Miura, and Y. Shirai. Tracking players and estimation of the 3d position of a ball in soccer games. In *Proceedings on International Conference on Pattern Recognition*, volume 1, pages 145–148, Barcelona, Spain, September 2000.

[39] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.

[40] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2 (6):522–536, 1980.

[41] J. Park, S. Park, and J.K. Aggarwal. Model-based human motion capture from monocular video sequences. In *Lecture Notes in Computer Science: Computer and Information Sciences*, volume 2869, pages 405–412, 2003.

[42] S. Park and J.K. Aggarwal. Recognition of human interaction using multiple features in grayscale images. In *Int'l Conference on Pattern Recognition*, volume 1, pages 51–54, Barcelona, Spain, 2000.

[43] S. Park and J.K. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *IEEE Workshop on Motion and Video Computing*, pages 105–111, Orlando, FL, 2002.

[44] 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, Berkeley, CA, USA, 2003.

[45] S. Park and J.K. Aggarwal. Event semantics in two-person interactions. In *International Conference on Pattern Recognition*, Cambridge, UK, 2004.

[46] V. Pavlovic, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.

[47] R. Polana and R. Nelson. Detecting activities. In *Computer Vision and Pattern Recognition*, 1993.

[48] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 –286, 1989.

[49] J. Rehg and T. Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, pages 16–22, Austin, USA, 1994.

[50] D. Sankoff and J. Kruskal, editors. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley Publishing Company, Inc., 1983.

[51] K. Sato and J.K. Aggarwal. Temporal spatio-velocity transform and its application to tracking and interaction. *Computer Vision and Image Understanding*, 2004. to appear.

[52] M. Shah and R. Jain, editors. *Motion-Based Recognition*, chapter 9, pages 201–226. Kluwer Academic Publishers, 1997. State-Based Recognition of Gesture.

[53] T. Wada and T. Matsuyama. Multiobject behavior recognition by event driven selective attention method. *IEEE transaction on Pattern Analysis and Machine Intelligence*, 22(8):873–887, August 2000.

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

[55] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In *Proceedings of the second international conference on Automatic Face and Gesture Recognition*, pages 51–56, 1996.

[56] J. Yang, Y. Xu, and C. S. Chen. Human action learning via hidden markov model. *IEEE Transactions on Systems, Man and Cybernetics*, pages 34–44, 1997.

[57] Q. Zhou and J.K. Aggarwal. Tracking and classifying moving objects from video. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2001.

[58] X. Zhou, R. T. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *ACM International Workshop on Video Surveillance*, pages 113–120, Berkeley, CA, USA, November 2003.