Machine Vision and Applications (2003) 14: 199–201
Digital Object Identifier (DOI) 10.1007/s00138-002-0079-9

Machine Vision and Applications

Problems, ongoing research and future directions in motion research

J.K. Aggarwal

Computer and Vision Research Center, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084, USA

Published online: 8 August 2003 –

Key words: Interactions tracking – Human motion – Segmentation of motion – Motion of body parts

1 Introduction

Interest in human motion research has been spurred by the evolution of motion understanding and interpretation and a concomitant change in emphasis in sponsored vision research. Understanding human motion is one of the most compelling computer vision problems today. Motion understanding began with the study of planar rigid objects [1]. As successful algorithms were developed, the focus moved to 3D rigid objects, and then to piece-wise rigid/articulated objects [2]. Today we study non-rigid objects, of which the most important is the human body. A contributing factor to the growth of interest in human motion is the shifting emphasis of sponsored research from inanimate objects, such as planes and tanks, to human motion recognition. The problems we now address cross over into the realm of everyday human activities, such as modeling the movements of the human head in a car crash or developing automated surveillance systems to secure public facilities such as airports.

Motion understanding/interpretation is a tough problem. The words *understanding* and *interpretation* are similar in definition, meaning *to grasp the meaning* and *to explain or tell the meaning* respectively. To grasp the meaning of an image sequence involving human motion entails understanding everything about that sequence, including the tasks of monitoring, inferring intention, and ultimately interpreting the sequence. For human motion analysis to be useful, inferring the projected actions is an important component of understanding.

Consider a scene in which Person A walks down a hall and turns into a doorway. Shortly thereafter, Person B also walks down the hall and enters the same door. Is the similarity of their paths intentional or coincidental? Without the ability to infer projected actions, the usefulness of motion understanding/interpretation would be limited indeed. Inferring intention, knowing whether Person B is deliberately following Person A or just happens to have the same destination, is an important part of understanding the scene. We humans frequently draw mistaken inferences about what we observe. Automatic recognition, understanding and interpretation are even more challenging.

2 Problems

The study of human motion poses some interesting and challenging problems. *Occlusion* and *time-ordered correspondence* are two old problems that go hand in hand [3]. Given the flexibility of the human body, self-occlusion is a constant problem in human subjects. The occlusion problem can be overcome by using multiple cameras, but then one must establish correspondence among the camera models as well as time-ordered correspondence, both of which are difficult tasks.

Segmentation is key to recognition and must be addressed on multiple levels. It is used to distinguish individual figures from a crowd, to locate a subject's head among the other body parts, and to recognize the individual motions that make up a human activity. Certain body parts, such as the head, and their position and motion play a critical role in action recognition.

Recognizing interpersonal interactions is an emerging area that requires us first to locate individuals within a group (segmentation, again) and then determine how individual actions relate to one another. Interactions are hard to describe – they are so diverse! Even in choreography, a descriptive language is at best incomplete. We have begun to address this issue in a strictly constrained scenario, as discussed further below. Another new area is *recognizing continuous movement*. Although we now can recognize individual motions in highly constrained situations, we find that real-life human motion is a stream of continuous movement, one action followed by another.

3 Ongoing research

The University of Texas Computer and Vision Research Center has a long history of research in motion understanding. Current research involves the analysis, understanding and interpretation of human motion. Our goal is to achieve understanding of individual and interpersonal human activities.

In an early human motion project [4], we developed a prototype *multiple-camera tracking* system that follows a moving human in a wide-area, indoor environment. Sequences of synchronized, monocular grayscale images were captured from multiple fixed cameras. Bayesian classification schemes based on motion analysis of human features were used to track (spatially and temporally) a subject image of interest between consecutive frames. An automatic camera switching module predicted the position of the subject along a spatial-temporal domain and then selected the camera that would provide the best trajectory and require the least switching to continue tracking. Tracking was based upon the images of upper human bodies captured from various viewing angles, and non-human moving objects were excluded using Principal Component Analysis (PCA). The system tolerated limited degrees of occlusion and achieved a recognition rate of 96% using both location and intensity features.

More recent research on computer vision for human activity recognition addresses the complex task of *recognizing individual human motions*. Our system tracks the movement of the head over consecutive frames and uses the difference in the coordinates of the head to recognize 12 human motions from the front or side view. Human actions are inherently similar from person to person. People sit, stand, and walk in a more or less similar fashion. During each action, the head exhibits characteristic and unique movements. When a person stands up (arising from a sitting position), the head moves forward and then backward while moving continuously upward. By modeling the difference in the centroid of the head over successive frames, the system can match the test sequence to training models to classify the action. This technique recognizes actions despite varying physical traits of the subjects (e.g., height and weight), giving it an advantage over template-matching approaches that are sensitive to such variations. Our recognition algorithm depends on reliably tracking the centroids of the head, which may be found in various positions, including lower than the back, as happens in the bending down action. We cannot assume (as has some past research) that the head is the highest point in the body silhouette. To locate the head in each frame, we integrate the results of two motion-based segmentation techniques, background subtraction and frame differencing. Further research is needed to develop a methodology for motions that involve self-occlusion.

In a recent study, we developed a methodology for *automatic segmentation and recognition of continuous actions*. Our approach has its roots in the area of speech recognition, which parses speech into words and then phonemes. Phonemes are recognized and joined together to recognize words and complete sentences. In a somewhat similar manner, we break up a stream of continuous activity into discrete action primitives using the angles subtended by three major segments of the human body with the vertical. The angle of the torso, the angle of the upper leg, and the angle of the lower leg are used to detect the breakpoints or transitions from one action to another. We use image sequences taken from the lateral view in which human subjects have executed a series of actions with no pauses between the actions. The system has no prior knowledge of the commencement or termination point of the individual actions. Our aim is to segment the activity into its separate actions and correctly identify each action in the sequence. To achieve this, the system first generates a skeleton of the human subject from thresholded binary images. The skeleton is represented by three segments: the torso, upper leg and lower leg. We compute the angles subtended by these segments with the vertical axis and use this information to detect *breakpoint frames*, e.g., transitions from one action to another. Individual actions are then recognized by observing the three angles over a sequence of frames and using a nearest-neighbor algorithm.

In another ongoing project, we are developing a system to recognize *interactions between two persons*. Currently, we attempt to recognize four interactions patterns: shaking hands, pointing at the opposite person, standing hand-in-hand, and an intermediate or transitional situation in interaction. We integrate visual input and contextual knowledge in order to understand human interaction, by combining multiple image features in a bottom-up fashion. Intensity, silhouette and contour information extracted from the image are used to build a human body model, called a *functional stick figure*. The model emphasizes the body parts (e.g., head and hand) that are functionally significant in human interaction, and simplifies the less significant ones (e.g., leg and elbow). This produces a detailed upper-body and simplified lower-body representation. By emphasizing functionality, we utilize contextual knowledge about human interaction as well as visual input. Our system automatically detects the functionally significant feature points – head, shoulders, hands, and feet – without involving a human operator in the model-building process. These models are then combined into a *human interaction model* that represents the interpersonal configuration between the two persons in space and time. The human interaction model represents the two persons' shoulder widths, arm angles, head orientations, and the distances of the hands and heads between the two persons. Human interaction is treated as an event that involves systematic, significant changes in the interpersonal configuration. Contextual knowledge about the interaction, such as human intention, is inferred by detecting each person's visual attention pattern using head orientation as evidence of visual attention. Our system detects and segments the head from the image and classifies its orientation as one of eight classes that cover the entire range of head rotation (0–360 degrees) on the horizontal plane in 45-degree increments.

Our system recognizes human interaction by directly estimating individual image frames expressed in terms of the human interaction model. Our method is unlike approaches that represent temporal changes, such as velocity or motion history, across a fixed-length frame sequence. Those motion-based approaches use the fixed-length sequence as a single analysis unit; our approach applies a more atomic method: analysis of individual frames. The fact that humans can easily recognize an interaction pattern from a single glance shows us that sufficient information about an interaction is contained in one image frame. By estimating each frame's relative poses per se, our approach provides an efficient tool to detect the initiation and termination points of certain interactions without parsing the sequence. We use composition rather than parsing to interpret what is going on in the sequence of images. That is, we first obtain atomic analysis results from each frame and then combine the results into a coherent interpretation. This compositional approach allows easy integration of contextual knowledge and visual input in a coherent fashion. The system is hierarchically organized with modular processing units,

each of which detects different body features from the input image. Its open architecture can be easily extended by adding separate processing units to process additional features as the system requirements grow more sophisticated and elaborate. The system's hierarchical organization also enhances its overall performance by collaborating mutual constraints between the individual processing units.

4 Future directions

Despite the many advances in human motion research, it is clear that we are far from developing general-purpose computer vision systems for automatic recognition of human activities. At this time, it seems most productive to concentrate on task-oriented systems, such as tracking humans in structured environments. Such work could be the nucleus of a larger, modular system for monitoring public areas. A separate module could be added to the application to address issues such as recognizing a person's personal belongings. As progress is made in recognizing specific interactions, additional modules could expand the system's capability to monitor interpersonal interactions. This *modular, incremental approach* allows us to confront well-constrained problems and incorporate new results as issues are resolved.

Multisensor approaches need to be explored further. Thermal cameras, which are already widely used, can detect motion in the absence of light. Integrating information from audio sensors could assist in determining intentions in interpersonal interactions. The multisensor approach can overcome many of the limitations of relying on video alone.

Interdisciplinary studies are needed to deepen our knowledge of human motion. Ergonomics can give a better understanding of the mechanics of how humans carry out specific tasks. Kinesiology can provide a better understanding of the intricate ways in which the body moves and how the parts of the body interact. Psychology and sociology will provide valuable information for evaluating and predicting crowd behavior from observed actions and speech. Although we have come a long way, we are really just beginning the journey toward a full understanding of human motion.

References

1. Aggarwal JK, Duda RO (1975) Computer analysis of moving polygonal images. IEEE Trans Comput 24(10):966–976
2. Webb JA, Aggarwal JK (1982) Structure from motion of rigid and jointed objects. Artif Intell 19:107–130
3. Martin WN, Davis LS, Aggarwal JK (1981) Correspondence processes in dynamic scene analysis. Proc IEEE 69(5):562–572
4. Cai Q, Aggarwal JK (1999) Tracking human motion in a structured environment using a distributed camera system. IEEE Trans Pattern Anal Mach Intell 21(11):1241–1247