Machine Vision and Applications (2003) 14: 210–214
Digital Object Identifier (DOI) 10.1007/s00138-003-0124-3

Machine Vision and Applications

# Understanding human behavior from motion imagery

**Mubarak Shah**

Computer Vision Lab, Computer Science, School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

Published online: 28 August 2003

**Abstract.** Computer vision is gradually making the transition from *image understanding* to *video understanding*. This is due to the enormous success in analyzing sequences of images that has been achieved in recent years. The main shift in the paradigm has been from *recognition followed by reconstruction (shape from X)* to *motion-based recognition*. Since most videos are about people, this work has focused on the analysis of human motion. In this paper, I present my perspective on understanding human behavior.

Automatically understanding human behavior from motion imagery involves *extraction* of relevant visual information from a video sequence, *representation* of that information in a suitable form, and *interpretation* of visual information for the purpose of recognition and learning about human behavior.

Significant progress has been made in human tracking over the last few years. As compared with tracking, not much progress has been made in understanding human behavior, and the issue of representation has largely been ignored. I present my opinion on possible reasons and hurdles for slower progress in understanding human behavior, briefly present our work in tracking, representation, and recognition, and comment on the next steps in all three areas.

**Key words:** Video understanding – Human behavior – Representation – Human activity recognition

## 1 Introduction

Automatically understanding human behavior from video sequences is a very challenging problem. This involves *extraction* of relevant visual information from a video sequence, *representation* of that information in a suitable form, and *interpretation* of visual information for the purpose of recognition and learning human behavior. Video sequences contain a large amount of data; most of these data do not carry much information. Therefore, the first step in understanding human behavior is to extract relevant information that can be used for further processing. This can be achieved essentially through visual tracking. Tracking involves detection of regions of interest in image sequences that are changing over time. Tracking also involves finding frame-to-frame correspondence of each region so that location, shape, extent, etc. of each region can be reliably extracted.

Representation is very important and sometimes a difficult aspect of an intelligent system. The representation is an abstraction of the sensory data that should reflect a real-world situation, be view-invariant and compact, and be reliable for later processing. Once the representation has been defined, the first obvious thing to do is perform a comparison so that classification or recognition can take place. The methods usually involve some kind of distance calculation between a model and an unknown input; the model with the smallest distance is taken to be the class of motion to which the input belongs. The problem with this is that the system can only recognize a predefined set of behaviors. This kind of system needs a large number of training sequences, does not have the capability to explain what a particular behavior is, and cannot learn and infer new behaviors from already known behaviors.

Therefore, it is desirable to build a system that starts with no model and incrementally builds models of activities by watching people perform activities. Once these models are learned, the system should be able to recognize similar behaviors in the future. This is probably similar to how children learn actions by repeatedly watching adults perform different actions.

Human motion analysis is a very active area of research in computer Vision. (See [1,5,22] for an excellent survey of this work.)

In this paper, we present our perspective on understanding human behavior. Significant progress has been made in human tracking over the last few years. The next section deals with human motion tracking. We address issues such as: What is new in tracking? What is the impact of computing speed on tracking? What are the system-related issues in tracking? and What are the next steps in tracking? As compared to tracking, not much progress has been made in understanding human behavior. In Sect. 3, we express our opinion on possible reasons and hurdles for less progress in understanding human behavior and comment on the next steps in this area. Finally, conclusions are presented in Sect. 4.

*Correspondence to*: M. Shah (e-mail: shah@cs.ucf.edu)

## 2 Human motion tracking

Human motion tracking is a hard problem due to the non-rigid and articulated nature of the human body, large degrees of freedom, and clothing. However, the last few years have experienced an explosion of visual tracking techniques. This is due mainly to the availability of powerful and inexpensive computing and memory devices; high-resolution, low-power, and affordable sensing devices; and success in algorithm development. In this section, we address various issues related to human tracking.

### *2.1 What is new in tracking?*

For many years computer vision research was only limited to gray-level images; color was rarely used. The assumption was that color does not provide any additional information, but it contains three times as much data. However, recently color has been increasingly used in tracking, and its use has been quite successful. Since most tracking work has focused on tracking people, color is very useful for detecting different parts of the face (eyes, lips, eyebrows), hands, arms, elbows, etc. Therefore, many trackers explicitly employ a skin color predicate [14] to detect different parts of the body to be used in tracking. Difference pictures have been used in visual tracking and motion detection for a long time [10]. The difference picture can either be computed between the current image and the background image or between two consecutive images in a sequence. Earlier work was limited to gray-level images; therefore, only mean and standard deviation of a scalar gray value was used. Recently this same idea has been successfully extended to color images, and the mean vector and covariance matrix for the color vector have been employed to compute the Mahalanobis distance for motion detection purposes [2]. The use of an explicit human body model and human motion model in tracking is also new. Since model-based tracking can employ explicit constraints on human motion, tracking can be simplified as compared to tracking of an arbitrary object.

Three-dimensional tracking using a single camera by employing some constraints on the human body is also a new development. Humanoid bones have constant length. Thus, for example, since the distance from the elbow to the shoulder is fixed, the elbow can only move tangent to a certain sphere (sphere centered on shoulder). Ignoring the half of the sphere that lies behind the plane of the body, we are left with the surface of one hemisphere. If we look at the hemisphere along the polar axis, we can see the entire surface of the hemisphere. That is, along the polar axis we can collapse the hemisphere from 3D to 2D without loss of information, as well as recover the hemisphere from its 2D projection. Using this reasoning, the 3D trajectory of a finger can be computed using a video sequence captured by a single monocular camera [25]. The use of image synthesis in tracking is also new. Using a rough face model, for example, several possible views of face can be synthesized by considering different changes in face rotation, translation, scaling, etc. and compared with the input image for face tracking [4]. This approach helps us to deal with changes in images due to changes in 3D. In summary, several new ideas have been tried in human tracking during the last few years.

### *2.2 Impact of computing speed*

The earliest motion work was limited to only two frames, which was one of the reasons that *rigidity* assumption [23] became very popular, since two frames are sufficient for the rigidity assumption. After that researchers started to consider three frames so that *smoothness* of the motion constraint [19, 21] could be employed. The work in the structure from motion area on the *minimum number of points in the minimum number of frames* [9] was partly motivated by the limitation of computing power. One of the impacts of computing speed on current work is that we are now able to use hundreds and thousands of frames in motion analysis. In some cases, since computation speed is so fast, we are tempted to use a brute force method, which may require exhaustive search. Another impact of computation speed has been that we are currently able to solve real-time tracking problems in the context of video surveillance and monitoring, which would have been impossible a few years ago. Since the frame rate is 30 frames per second, processing speed must be comparable to that; otherwise some real-time events will be missed. However, computing speed still is not that high, which is why most operations we use are pretty simple, such as background differencing.

### *2.3 Systems issues in tracking*

There are several key practical "systems issues" that need to be addressed before human tracking can be performed from a live camera. In order to perform tracking from live cameras, we need to have access to uncompressed, noninterlaced, high-resolution, and high-frame-rate video. The compressed video contains artifacts, e.g., blockiness, mosquito noise, dirty window noise, wavy noise, etc. In the interlaced video, pixels in one field may be significantly displaced with respect to other fields depending on the motion of the camera or objects, which may create problems in tracking. It is always desirable to have high-resolution images so that in a scene containing multiple people, each person being tracked occupies a reasonable portion of the image. If the image size is too small, a person being tracked may consist of only a few pixels. Finally, the frame rate has to be comparable to the speed of motion. In order to capture rapid motion of some parts of the human body, for example the arm, a frame rate as high as 500 frames per second may be required at times. Also, since the field of view of a single camera is very limited, people performing activities will go out of the field of view. Therefore, it is necessary to use multiple cameras to track people for extended periods of time. The important problem in this context is to be able to seamlessly solve the camera hand-off problem from one field of view to another [11]. Another important issue relates to networking – how efficiently video can be sent over the network.

### *2.4 Next steps*

The next steps in tracking include tracking people in crowds, in complete occlusion, in scenes containing shadows and variable lighting, in sequences acquired with a moving camera, in a large field of view and tracking people who are moving slowly, who reverse the direction of motion, and who are handing over

objects to other people. Even though attempts have been made to address almost all of the above issues, much work remains to solve these problems completely.

For example, occlusion is a significant problem in human motion tracking [13]. People tend to walk and interact in groups with other people, thereby increasing the chances that persons will occlude each other completely or partially in images. The probability of observing occlusion can be decreased in general by placing the cameras at a higher angle of elevation from the plane of movement of people. That is, by placing the cameras looking vertically downwards, the chance of one person occluding another is minimized. Indeed most of the previous work in human tracking either uses this constraint on camera positioning ([7,3] or does not deal with occluding cases at all [2,18]. Limited solutions to the occlusion problem are presented in [6] and [8]. In [6], occlusion from static objects is dealt with using an occlusion reasoning framework, which maintains multiple hypotheses for occluded regions and keeps eliminating wrong ones as time progresses. However, this approach is demonstrated to be useful in simplistic cases and needs to be explored further in the case of more complicated scenarios. Moreover, it is limited to occlusion by static objects and may not generalize to the more complicated case of occlusion from nonrigid objects such as other persons. In [8], statistical features of the two persons before occlusion begins are used to resolve the labels after occlusion has ended, but the system cannot decide about which pixels belong to which person *during* the occlusion event. Therefore, future research in human tracking should address all of the above issues.

### *2.5 Tracking using nonvisual sensors*

Nonvisual sensors include infrared sensors, x-ray sensors, laser range finders, etc. Recently there has been considerable improvement in the quality and affordability of infrared sensors. The advantages of these sensors is that they can be used at night and in low lighting conditions, and they may not have much problem with clothing. Similarly, x-ray sensors can give us a complete image of a whole human body without any occlusions. One can easily get a stick figure model by analyzing bone structure in the x-ray images. Laser range finders are increasingly being used for scanning buildings, objects, and people. The laser range finder provides direct 3D information that with no ambiguity due to the projection process. Another way to simplify tracking is to attach some reflective markers on the human body or have people wear special body suits and gloves. Our view is that computer vision research should focus on developing methods for human tracking using only video cameras. The advantage of visual tracking is that it is nonintrusive. Also, video cameras are very inexpensive and very light. The use of other nonvisual sensors will make visual tracking easier and less challenging.

## 3 Understanding human behavior

### *3.1 Possible reasons and hurdles for slower progress*

One of the reasons for slower progress in understanding human behavior is that vision research has abandoned AI. Computer vision started as an AI problem, which is why vision has also been called *image understanding*. The original goal of vision was to understand a single image of a scene and locate and identify objects, their structures and spatial arrangements, and their relation to other objects. The MIT copy demo is a good example of this. The idea in the copy demo was roughly to have a computer vision program analyze the image of a scene containing several blocks stacked together, recover the structure of the blocks, generate a script, and have a robot build an exact copy of the block structure.[1] This, in fact, was a high-level vision problem. One of the motivations for the work in blocks, consistent line labeling, polyhedral junctions, etc. was in fact the copy demo. The researchers soon found out that low-level vision was not robust enough; they were not even able to extract lines from images to be used in this work. Therefore, it became necessary to first solve low-level vision problems before the high-level vision problems could be attacked. The research in low-level vision continued for some time. Then, during the 1970s, Marr [17], who popularized, among other things, shape-from-X methods, captured the attention of vision researchers. Since one dimension is lost during the projection of a 3D world onto 2D images, the aim of shape-from-X methods is to recover that lost dimension. The next two to three decades were spent developing algorithms for recovering 3D shapes from 2D images using stereo, motion (structure from motion), shading, texture, etc. The original AI problem was almost forgotten; not much work was done on high-level vision during those years. Currently we are living in era of vision research, when some shape-from-X problems, for example stereo, have been almost completely solved and are being used in industry. Other shape-from-X problems, such as shape from motion, has proved to be very difficult; and the remaining shape-from-X problems, such as shape from shading and texture, have become less interesting and applicable.

We feel the second reason that not much progress has been made in understanding human behavior is that too much emphasis has been placed on the structure-from-motion problem during the last three decades. That problem may be theoretically appealing, but it may not help us solve the problem of understanding human behavior. The shape-from-X methods compute intrinsic surface properties such as depth values. As correctly pointed out by Witkin and Tennenbaum [24], depth maps and other maps of the 2.5D sketch are still basically just images. They still must be segmented, interpreted, and so forth before they can be used for any more sophisticated task. Therefore, we feel 3D may not be necessary for recognition and interpretation. This is supported by one of two theories about the interpretation of motion by humans [5]. According to the first theory, people use motion information to recover the 3D structure and subsequently use that structure for recognition (structure from motion). In this case, the moving object would be identified first, then the motion it performs in the image sequence would be sought. According to the second theory, motion information is directly used to recognize a motion, without structure recovery. We believe the second theory is more suitable for understanding human behavior from video sequences. This is also obvious from recent success in human motion analysis work in computer vision, which has generated enormous interest in industry in this area.

[1] We feel it would have been easier to solve this problem if instead of a single image a sequence of images was analyzed.

The third reason for slower progress in understanding human behavior is the limitation of hidden Markov models (HMMs), which have been widely used. The standard approach to human behavior recognition is to extract a set of features from each frame of a sequence and use those features to train HMMs to perform recognition. In this research, most of the emphasis has been on discovering the appropriate features. Therefore, not much work has been done on HMMs; they have been treated as a black box. There are several important issues related to HMMs. First, since HMMs rely on probabilities, they require extensive training. Therefore, one needs to have a large number of training sequences for each activity to be recognized. Second, for each activity to be recognized, a separate HMM needs to be built. Therefore, this approach can only recognize some predefined set of activities. It does not have the capability to learn new activities. Third, since the HMM is treated as a black box, it does not explain what a particular activity is. It just outputs the probability that an unknown activity is recognized as the model activity.

The fourth reason for slower progress in understanding human behavior is that the issue of representation of features has largely been ignored. It is very important for a representation of action to be view invariant [20]. For example, since an action takes place in 3D and is projected onto a 2D image, depending on the viewpoint of the camera the projected 2D trajectory may vary. Therefore, trajectories of the same action may have very different projected trajectories, and trajectories of different actions may project to the same trajectory. This may create a problem in interpretation of trajectories at a higher level. One obvious thing to do is to recover the 3D trajectory, but that may be overkill, as stated above. However, if the 2D representation of an action captures characteristics that are view invariant, then the higher-level interpretation can proceed without any ambiguity.

In summary, abandoning of AI, excessive emphasis on structure from motion, the limitation of HMMs, and ignorance of view-invariant representations are some of the reasons for the slower progress in understanding human behavior.

### *3.2 Next steps*

We believe that one of the next steps in understanding human behavior is to work on dynamical perceptual organization. This work will extend the work on perceptual organization for a single image to sequences of images. The idea is to discover invariant relations that can be used to infer 3D information from 2D motion. The relations for a single image were originally proposed by Lowe[16] and include collinearity, curvilinearity, symmetry, parallelism, and vertices. Collinearity of points or lines in 2D implies collinearity in 3D, curvilinearity of points or arcs in 2D correspond to curvilinearity in 3D, skew symmetry in 2D implies symmetry in 3D, parallel curves over small visual angles correspond to parallel curves in 3D, and, finally, vertices in 2D represent curve terminations at a common point in 3D. Similar relations can be discovered between 2D motion and 3D motion. For instance, a 2D elliptical trajectory implies rotation motion in 3D. A set of elliptical trajectories with the same phase and the parallel major and minor axes correspond to the motion of points on a single rotating object in 3D. Two trajectories are parallel if they have equal speed and direction for all time instants. Similarly, parallel trajectories in 2D imply translational motion in 3D. The segments of trajectories with constant speed and direction correspond to constant motion in 3D.

Another possible next step is to build a system consisting of multiple uncalibrated and arbitrarily located cameras. The system should self-calibrate the cameras by watching people over extended periods of time. It should also establish correspondence among various camera views of the same object to maintain consistency in labeling those objects. Due to the availability of inexpensive video cameras and associated computer hardware, a large number of cameras can be used to completely cover entire regions of interest. In fact, we can assume to have almost one dedicated camera and a processor for each possible object of interest. This combination of video camera and processor is called an *agent*. Each agent can monitor its area, assimilate, learn, form concepts, and communicate with the controller (server). Each agent can learn about its position relative to other agents, just by observing human motion. The system can also learn about important places and frequent actions in the environment. Once the system has "matured," meaning that sufficient information is gathered from the environment, it can make high-level decisions. This monitoring can be done 24 hours a day, 7 days a week, 365 days a year. For indoor scenes the possible objects of interest in a room include people, doors, telephones, books, bookshelves, white boards, cabinets, keyboards, mouses, printers, cups, etc. Actions related to people are: enter, exit, walk, sit, stand, talk on the phone, work at a computer, fetch a book, read a book, write on white board, etc. In order to build such a system, we need to rely extensively on knowledge and context.

One more interesting problem for future work is motion-based recognition of people by habitual gestures: repeated motion of shoulders, wringing of hands, jerks of hands, twitching of head, blinking of eyes, etc. However, this is different from gesture and facial expression recognition work, which is essentially motion recognition. This deals with the recognition of people based on their motion and is relevant to the human ID problem.

The work on dynamic perceptual organization, the use of habitual gestures, and the work on a system consisting of uncalibarated, arbitrarily located cameras that can automatically learn human behavior by continuously watching people are a few of the possible next steps in understanding human behavior.

## 4 Conclusion

In this paper, we presented our perspective on understanding human behavior from video sequences. We identified three main steps in understanding human behavior: *extraction* of motion information (visual tracking) from video sequences, *representation* of this information, and *interpretation* of this information for recognition and inference purposes. We feel that much progress has been made in visual tracking. However, not much progress has been made in representation and interpretation. In order to make the proper progress, we need to solve high-level vision problems and use knowledge and context. It is our contention that 2D information is sufficient for understanding human behavior, which is supported by one

of the theories about the human visual system. Therefore, we support the direct approach to motion-based recognition, in contrast to the widely accepted view of reconstruction (3D shape and motion) followed by recognition.

## References


1. Aggarwal JK, Cai Q (1999) Human motion analysis: a review. CVIU 73(3):428–440
2. Azarbayejani A et al (1996) Real-time 3D tracking of the human body, MIT Media Lab, Perceptual Computing Section, TR No. 374
3. Bobick AF et al (1999) The KidsRoom: a perceptually-based interactive and immersive story environment, Teleoperators Virtual Environ 8(4):367–391
4. Cascia M, Sclaroff S, Athitsos V (2000) Fast reliable head tracking under varying illumination, IEEE PAMI, April 2000, pp 322–336
5. Cédras C, Shah M (1995) Motion-based recognition: a survey. Image Vis Comput 13(2):129–155
6. Fieguth P, Terzopoulos D (1997) Color-based tracking of heads and other mobile objects at video frame rates. CVPR-97, June 17 -19, 1997, Puerto Rico, pp 21— 27.
7. Grimson WEL et al (1998) Using adaptive tracking to classify and monitor activities in a site. CVPR-98, 23–25 June 1998, pp 22–29
8. Haritaoglu I, Harwood D, Davis L (1998) $W^4$ - who, where, when, what: a real-time system for detecting and tracking people. In: Proceedings of the international conference on automatic face and gesture recognition, April 14-16, 1998, Nara, Japan, pp 222-227.
9. Jain RC, Binford TO (1991) Dialogue: ignorance, myopia, and naivete in computer vision systems. Comput Vis Graph Image Process J 53(1):112–117
10. Jain RC, Militzer D, Nagel HH (1977) Separating non-stationary from stationary scene components in a sequence of real world TV-images., Proceedings of the international joint conference on artificial intelligence, NAGOYA, Aichi, Japan August 23-29, pp 612—618.
    1977, pp 612–618
11. Javed O, Khan S, Rasheed Z, Shah M (2000) Camera handoff: tracking in multiple uncalibrated stationary cameras, In: Proceedings of the IEEE workshop on human motion, Austin, Texas, December 2000, pp 113-118.
12. Kanade T et al (1998) Advances in cooperative multi-sensor video Surveillance. In: Proceedings of the DARPA image understanding workshop, Monterey, California, November 1998, pp 3—24.
13. Khan S, Shah M (2000) Tracking people in presence of Occlusion. In: Proceedings of the Asian conference on computer vision, Taipei, Taiwan, January 2000, pp 1132-1137. January 2000.
14. Kjeldsen R, Kender J (1996) Finding skin in color images.Proceedings of the international workshop on automatic face and gesture recognition, Oct 13-16, 1996, Killington, Vermont, USA , 1996, 312-317.
15. Lipton AJ, Fujiyoshi H, Patil RS (1998) Moving target classification and tracking from real-time video. In: Proceedings of the DARPA image understanding workshop, Monterey, California, pp 129–136 1998, pp 129–136
16. Lowe DG (1985) Perceptual organization and visual recognition, Kluwer, Amsterdam
17. Marr D (1982) Vision. W. H. Freeman Co., USA.
18. Olson TJ, Brill FZ (1997) Moving object detection and event recognition algorithm for smart cameras, Proceedings of the DARPA image understanding workshop, , New Orleans, May 1997, pp 159—175.
19. Rangarajan K, Shah M (1991) Establishing motion correspondence. CVGIP Image Understanding 54(1):56–73
20. Rao C, Shah M (2000) A view-invariant representation of human action. In: Proceedings of the international conference on control, automation, robotics and vision ICARCV2000, Singapore, 5–8 December 2000
21. Sethi IK, Jain R (1987) Finding trajectories of feature points in a monocular image sequence. IEEE PAMI 9(1):56–73
22. Shah M, Jain R (1997) Visual recognition of activities, gestures, facial expressions and speech: an introduction and a perspective, In: Motion-based recognition. Kluwer, Amsterdam, pp 1–12
23. Ullman S (1979) Interpretation of visual motion. MIT Press, Cambridge, MA
24. Witkin A, Tenenbaum M (1986) On perceptual organization. In: Pentland A (ed) From pixels to predicates, Ablex, Norwood, NJ, pp 149–169
25. Wu A, Shah M, da Vitoria Lobo N (2000) Virtual blackboard. In: Proceedings of the international conference on automatic face and gesture recognition, March 28-30, 2000 Grenoble, France 2000.