M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 413
Appears in the Royal Society Workshop on Knowledge-based Vision in Man and Machine, London, England — February, 1997
To appear: Proceedings of the Royal Socoiety

Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion

Aaron F. Bobick
MIT Media Laboratory
20 Ames Street, Cambridge, MA 02139
bobick@media.mit.edu
http://www.media.mit.edu/~bobick

Abstract

We present several approaches to the machine perception of motion and discuss the role and levels of knowledge in each. In particular we describe different techniques of motion understanding as focusing on one of movement, activity, or action. *Movements* are the most atomic primitives, requiring no contextual or sequence knowledge to be recognized; movement is often addressed using either view- invariant or view specific geometric techniques. *Activity* refers to sequences of movements or states, where the only real knowledge required is the statistics of the sequence; much of the recent work in gesture understanding falls within this category of motion perception. Finally, *actions* are larger scale events which typically include interaction with the environment and causal relationships; action understanding straddles the gray division between perception and cognition, computer vision and artificial intelligence. We illustrate these levels with examples drawn mostly from our work in understanding motion in video imagery and argue the utility of such a division is that it makes explicit the representational competencies and manipulations necessary for perception.

1 Introduction

Recently, there has been a shift in computer vision from the processing of static images to the consideration of video sequences. The majority of previous work on sequences of images has focused on recovering the geometry of either the scene — structure-from-motion, the camera motion — egomotion, or the motion of the pixels themselves — optic flow (see Cedras & Shah 1994 for a review). Current research, however, has begun to investigate the recognition of the action taking place in the scene. The fundamental question being addressed is no longer "How are things moving?" but "What is happening?" (Bobick 1996).

However, there has been much confusion about exactly which interpretation problems constitute understanding action. For example, Polana & Nelson (1994) and Shavitt & Jepson (1993) focus on direct motion properties of the image pixels to detect activities such as walking or running. There is no knowledge about time, sequence, or causality embedded in the interpretation process. The "action" is coded strictly in the statistics of image motion. Jumping, for example, has a particular signature in a local spatio-temporal region of the image sequence.

In sharp contrast is work such as that by Siskind (1995) and Mann, et al. (1996). In these approaches the interpretation of the motion of objects is accomplished by analyzing the action in terms of a qualitative physics description. Mann's system understands that the proposition of an attachment to an active — i.e. self-propelled — moving object is adequate to explain the movement of a passive entity. In these systems, understanding action implies producing a semantically rich description that includes primitives such as "pick-up" or "bounce." To produce such descriptions requires a representation of the causal relations in qualitative physics; often, an extended representation of time, as opposed to an instantaneous or signal-based view, is needed as well.

The goal of this paper is to analyze various approaches to understanding motion with respect to the nature and amount of the knowledge required. Drawing mostly from examples of our own work, we will propose three levels of motion understanding problems labeled, in increasing order of knowledge implied, as movement, activity, and action. One advantage of considering motion interpretation problems this way is that upon presentation of an algorithm or application task one can immediately compare the work to other approaches, and in particular consider the competence of the representation and knowledge employed.

Before continuing it is necessary to note the pioneering work of Hans Nagel in the general area of machine perception of motion (Nagel, 1977), and in the specific endeavor of attempting to characterize motion understanding problems (Nagel, 1988). His taxonomy of "change, event, verb, episode, history" reflect different dimensions of the problem than those discussed here, but it does provide an interesting alternative view. The 1988 paper begins with the sentence: "Today, the design of a program which 'understands' image sequences appears an ambitious but not totally unrealistic research goal." It still feels ambitious.

2 Perception of motion: movement, activity, and action

Suppose we wish to construct a system that recognizes different motions in a particular application domain: a baseball game. Let us consider three distinct "actions" one might wish to identify: swinging the bat, pitching the ball, and tagging out a runner. In this section we will argue that these three tasks are illustrative of three classes of motion understanding problems and that the techniques necessary to recognize them will differ in the type of knowledge required and how that knowledge is applied.

If one observes numerous players swinging a bat one would see little variation in the motion. While the exact stance and configuration of the static bat prior to the swing may vary, the motion itself is predictably similar from one instance to the next. We say "predictably" because the physical dynamics of the task — accelerating a stick to a speed sufficient to propel the ball (hopefully) 450 feet — and the kinematics of the human actuator constrain the motion to be performed in a particular manner

We term this type of motion a *movement:* a motion whose execution is consistent one instance to the next and easily characterized by a definite space-time trajectory in some configuration space (in this case the kinematics of the human body). For a given viewing condition execution consistency implies consistency of appearance: the appearance of the motion can be described reliably in terms of the motion of the pixels. The pixel-based description of the motion under different possible viewing conditions is the only knowledge required to see the movement.

Approaches to the perception of movements include the previously mentioned work by Polana & Nelson (1994) and Shavitt & Jepson (1993). These techniques are based on periodicity measurements of the pixels or blobs undergoing motion. In section 3 we will describe two techniques developed in our lab for recognizing human movements.

Pitching a baseball involves many more steps than hitting. Typically, but not always, a pitch involves 1) bringing the arms together in front of the body to achieve appropriate balance; 2) swinging the arms back; 3) kicking the front leg up while leaning back; 4) delivering the pitch.[1] Some instances diminish the effort or reduce the time put into one phase or another, or may even eliminate a stage entirely. The motion is no longer a single, primitive, consistent movement. Rather, it is an *activity,* a statistical sequence of movements. Recognition of such a motion requires knowledge about both the appearance of each constituent movement and the statistical properties of the temporal sequence.

An important domain area that requires addressing activity is that of recognizing gait. Rohr (1994) and Niyogi & Adelson (1994) make an explicit model of the sequence of movements or configurations that form the activity of walking. In both of these approaches the sequence is fixed and deterministic. The work by Black & Yacoob (1995) on understanding facial expression coded a qualitative variation over time of the shape of face features.

The recent surge in interest in hidden Markov models to process video sequences reflects the goal of explicitly representing statistical sequential information. One of the earlier efforts is Starner & Pentland (1995) where HMMs are used to understand American Sign Language; the success HMMs have attained in the speech recognition community was a strong motivation to apply them to the analogous ASL task. In section 4 we will describe some recent work in our lab that focuses on how activities may not be represented easily by a single feature set; as the activity progresses the underlying representation may need to vary.

Finally, what does it take to see a runner being tagged out? Semantically, the description is straightforward: a fielder with the ball causes his glove to come in contact with a base-runner who is not touching a base at the time. Visually, however, the appearance is difficult to define or describe because of the variability of how the movements may be made. The motion to be recognized needs to be understood in a context: the best explanation of the sequence of movements is that the fielder is intending to tag the runner which is why he is moving his arm down while the runner is trying to get to the nearest base. Tagging a runner is an *action* which we define to include semantic primitives relating to the context of the motion. For a system to recognize actions it must include a rich knowledge base about the domain and be able to hypothesize and evaluate possible semantic descriptions of the observed motion.

As defined, actions are at the boundary of where perception meets cognition. Indeed, artificial intelligence researchers proposing formal theories of semantics and inference of action (Schank 1975; Jackendorf 1990; Israel, et al. 1991) address motion at this level of analysis. Being primarily focused on computer vision our goal is to stay as connected to the visual signal as possible, where predefined actions (e.g. "tagging out a runner" or "mixing ingredients") and particular semantic labels (e.g. "base-runner" or "chef") have direct visual correlates. Mann, et al. (1996) accomplish this by postulating that certain behaviors suggest particular causal relationships, and that those relationships have visual consequences that can be verified. In section 5 we briefly outline a system we have developed that uses logical descriptions of actions to deduce visual correlates; these correlates are then used to control the selection of vision routines.

In the following three sections we describe some of our results in the machine perception and interpretation of motion that reflect the levels described here. Our main goal is to focus on the knowledge and the representation of time employed by each set of techniques. At the conclusion of the paper we will discuss the utility of the taxonomy of motion understanding problems presented.

3 Recognition of movement

Two recent efforts in our research group have focused on the direct recognition of movement. The first case, which we mention only briefly, is work on the recognition of ballet steps (Campbell & Bobick 1995). The approach we develop is based on the idea that different categorical movements, e.g. plié or relevé, each embody a different set of constraints on the motion of the body parts. These constraints are mostly easily observed in a phase-space that relates the independent variables of the body motion. This work presumes that the underlying 3-dimensional kinematics of the body are recovered from video (e.g. using a method such as that of Gavrilla and Davis, 1996). Our question is not how to recover the 3-dimensional structure; rather, given that structure, how do you see a plié?

Figure 1 illustrates an example. The phase plot on the

[1] Apologies to connoisseurs of the game; the descriptions here are simplified approximations.

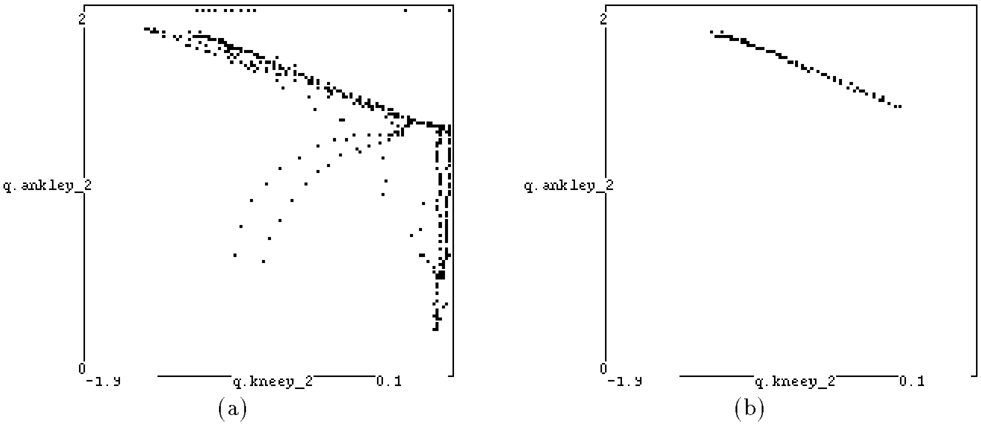

Figure 1: Two variable phase plots (ankle angle vs. knee angle) for (a) a wide variety of ballet moves, and (b) points during plié moves. The simple curve in the second plot represents a detectable constraint in force during the execution of a plié.

left displays shows the relation between the ankle angle and knee angle of one leg of a dancer performing a wide variety of ballet steps. The graph on the right contains only those points recorded during *plié* steps. Because the tight constraint in the second plot is not generally in force during other moves, detecting the presence of this relationship indicates the possibility of a plié being performed. By automatically learning from training data which sets of constraints are highly diagnostic of particular motions we can build constraint set detectors to recognize the movements. Note that this technique is only applicable to the recognition of atomic movements; in this approach sequences of steps can only be recognized if each individual movement is detected.

A more generic movement recognition method is developed in our recent work on temporal templates which aims for the direct recognition of movement from the motion in the imagery. Consider an extremely blurred sequence of motion; a few frames of one such example are shown in Figure 2. Even with almost no structure present in each frame, people can easily recognize the movement as someone sitting when the frames are displayed as a video sequence. Such capabilities argue for recognition from the motion itself, as opposed to first reconstructing a three-dimensional model of a person, and then recognizing the movement or action of the model.[2]

In (Bobick & Davis 1996a, Bobick & Davis 1996b, Davis & Bobick 1997) we propose a view-based representation and recognition theory that decomposes motion-based recognition into first describing *where* there is motion (the spatial pattern) and then describing *how* the motion is moving. The basic idea is that we project the temporal pattern of motion into a single, image-based representation — a *temporal template*.

The top row of Figure 3a contains key frames of a sitting sequence. The bottom row displays cumulative binary motion images — to be described momentarily — computed from the start frame to the corresponding frame above. As expected the sequence sweeps out a particular region of the image; our claim is that the shape of that region can be used to suggest both the movement occurring and the viewing condition, in this case horizontal viewing angle.

We refer to these binary cumulative motion images as *motion-energy* images (MEI). Let $I(x,y,t)$ be an image sequence, and let $D(x,y,t)$ be a binary image sequence indicating regions of motion; for many applications image-differencing is adequate to generate D. Then the MEI $E_\tau(x,y,t)$ is defined

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i)$$

We note that the duration τ is critical in defining the temporal extent of an action. During training we need to explicitly define τ. Fortunately, to perform real-time recognition we can exploit a backward-looking (in time) algorithm that can dynamically search over a range of τ yielding linear speed invariance in recognition.

In Figure 3b we display the MEIs of a sitting movement viewed over 90°. In (Bobick & Davis 1996a) we exploited the smooth variation of motion over angle to compress the entire view circle into a low order representation. Here we simply note that because of the slow variation across angle, we only need to sample the view sphere coarsely to recognize all directions.

To represent how motion is moving we enhance the MEI to form a *motion-history image* (MHI). In an MHI, pixel intensity is a function of the motion history at that point. For the results presented here we use a simple replacement and decay operator:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ \max\ (0, H(x,y,t-1) - 1) & \\ \quad \text{otherwise} & \end{cases}$$

The result is a scalar-valued image where more recently moving pixels are brighter. Examples of MHIs are presented in Figure 4. Note that unlike MEIs, the MHIs are sensitive to direction of motion. Also note that the MHI

[2] Example sequences are available on the Web at: http://vismod.www.media.mit.edu/vismod/archive

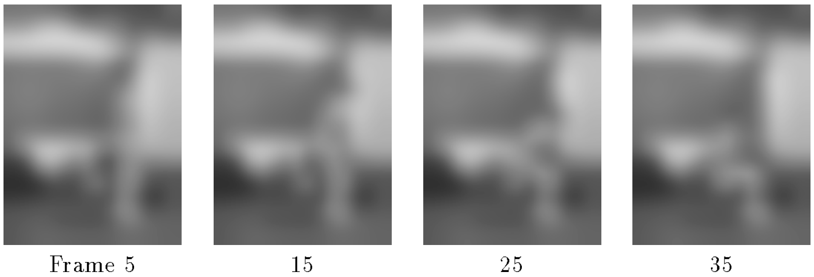

Figure 2: Selected frames from video of someone performing an action. Even with almost no structure present in each frame people can easily recognize the action as someone sitting.

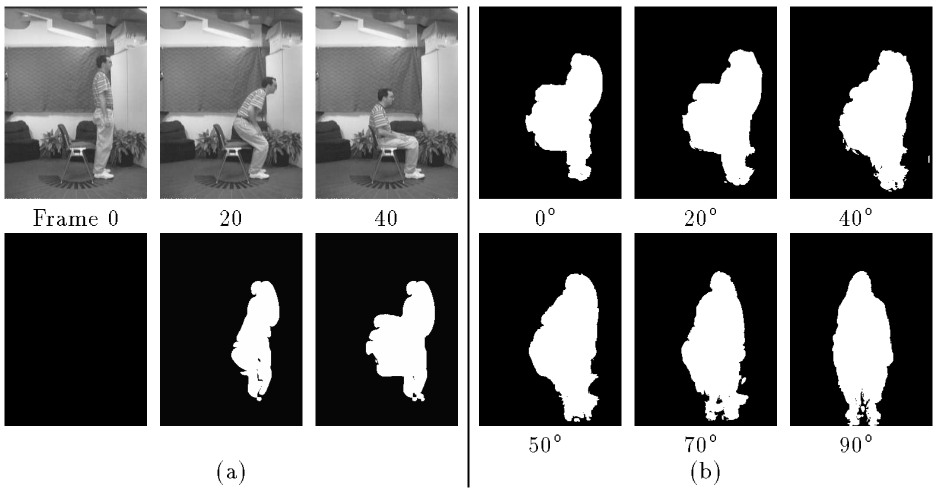

Figure 3: Example of someone sitting. (a) Top row contains key frames; bottom row is cumulative motion images starting from Frame 0. (b) MEIs for each of six viewing directions;the smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

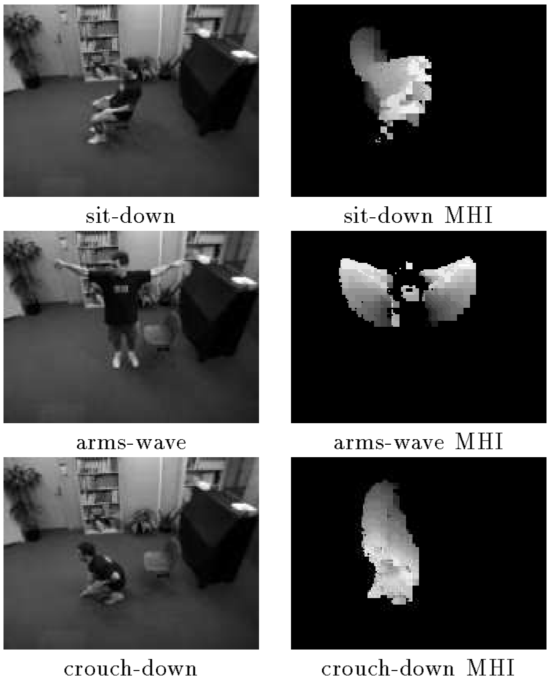

Figure 4: Action moves along with their MHIs used in a real-time recognition system.

can be generated by thresholding the MEI above zero. The MEI and MHI together form a temporal template of movement to be matched against unknown input motions.

To construct a recognition system, we need to define a matching algorithm for the MEI and the MHI. In (Bobick & Davis 1996b, Davis & Bobick 1997) we develop and test a scale and translation invariant technique using statistical moment descriptions of the MEI and MHI images. We first collect training examples of each movement from a variety of viewing angles. For each view of each movement a statistical model (mean and covariance matrix) is generated for shape moment parameters of both the MEI and MHI. The computation of the shape moments includes weighting by pixel intensities, giving different moments for the MEI and MHI. To recognize an input motion, a Mahalanobis distance is calculated between a shape moment description of the input and each of the known movements.

We have implemented a causal segmentation and recognition system that uses a backward looking variable time window to achieve speed invariance. The simple nature of the replacement operator allows the construction of a highly efficient algorithm capable of real time operation on a standard Unix workstation. For example, one implementation runs at approximately 13Hz using a color CCD camera connected to a Silicon Graphics Indy. The images are digitized to a size of 160x120 covering movements of duration from one to two seconds. The matching operation is virtually no cost once the input image statistics have been computed; adding more classes of movement does not affect the speed of the algorithm, only the accuracy of the recognition.

In summary, temporal templates are a representation suitable for the recognition of movements, designed for matching motion patterns between input and known models. The only statistics considered are the variability of appearance from one instance to the next. Time is handled implicitly by developing a matching method that is insensitive to linear scaling with respect to time, e.g. a simple change of speed.

Furthermore, there is no consideration of sequence. In fact, overly complicated motions give rise to temporal templates that overwrite themselves often (self-occlusion in space-time) and therefore are unreliable for matching. To recognize a series of atomic motions requires a more powerful representation of time and of the statistics of the temporal pattern.

4 Recognition of activity: gesture in communication

As defined, activities involve a sequence of motions. The components of the sequence can either be movements or static states. If explicitly based upon states, then the motion sequence is implicitly defined by the movements that are required to move from one state to the next.

The representation of the sequence defining an activity can either be explicit and deterministic, or implicit and statistical. An example of the former case is that of Rohr (1994) where the positions of silhouette edges of a walking person are encoded as a one degree of freedom function of the phase of the gait. These edges are matched against those of a person in each input image of a sequence. The gait phase "angle" is then estimated at each time instant yielding a description of the complete sequence in terms of a trajectory through the gait phases.

Examples of implicit and statistical representation of sequences are seen in the recent work on understanding human gesture (e.g. Starner & Pentland 1995, Wilson & Bobick 1995). Inspired by the successful application of hidden Markov models to speech recognition tasks, these methods represent activities — gestures — by probabilistic states where both the observed output of a given state and the transitions made between states are controlled by underlying probability distributions [Rabiner and Huang, 1993]. In the remainder of this section we will discuss the work in Wilson & Bobick (1995) because it not only maintains a Markovian model of the statistics of motion, but also learns the variations in representation of the imagery required to span the entire activity.

Two observations motivated the approach. First, human gestures are embedded within communication. As such, the gesturer typically orients the movements towards the recipient of the gesture (Darrell & Pentland 1993). Second, in the space of motions permitted by the degrees of freedom of the human body, there is a small subspace that we use in the making of a gesture. Taken together, these observations argue for a view-based approach in which only a small subspace of human motions is represented.

How should a system model human motion to capture the constraints present in the gestures? There may be no single set of features that makes explicit the relationships that hold for a given gesture. In the case of hand gestures, for example, the spatial configuration of the hand may be important (as in a point gesture, when the observer must notice a particular pose of the hand), or alternatively, the gross motion of the hand may be important (as in a friendly wave across the quad). Quek (1993) has observed that it is rare for both the pose and the position of the hand to simultaneously change in a meaningful way during a gesture.

We first presented an approach that represents gesture as a sequence of states in a particular observation space (Bobick & Wilson 1995). We then extended that work and developed a technique for learning visual behaviors that 1) incorporates the notion of multiple models — multiple ways of describing a set of sensor data; 2) makes explicit the idea that a given phase of a gesture is constrained to be within some small subspace of possible human motions; and 3) represents change over time as a probabilistic trajectory through states (Wilson & Bobick 1995). The basic idea is that the different models need to approximate the (small) subspace associated with a particular state, and membership in a state is determined by how well the state models can represent the current observation. The parsing of the entire gesture is accomplished by finding a likely sequence of states given the memberships and the learned transition probabilities between the states.

The details of the techniques are presented in Wilson & Bobick (1995). The approach is based upon state models that define a *residual* — how well a given model can represent the current sensor input. We then embed this residual-based technique within a Hidden Markov Model framework; the HMMs represent the temporal aspect of the gestures in a probabilistic manner and the associated parsing method provides a form of dynamic time warping for the recognition of gesture.

Here we illustrate the technique by way of two examples. Figure 5 — a wave gesture — consists of a single model example but shows the use of the HMM. The model in use is a principal component decomposition of the input image. The parameters describing each image are the coefficients of projecting the input image onto a basis set of images, where there is a different set for each state of the HMM. The basis set for a state consists of the most significant eigenvectors of the set of images determined to belong to that state. The residual between the input image and the best reconstruction using the basis set of a state determines the likelihhod that the given state could generate the input image. Because the basis set determines state membership but state membership is used to select the basis set, the entire estimation process is an iterative Expectation-Maximization algorithm; we add the basis set selection step to the traditional Baum-Welch technique of HMM parameter estimation [Rabiner and Huang, 1993]. The different basis sets are the varying representation of the activity to which we earlier referred.

In the first example, the input data consist of 32 image sequences of a waving hand, each about 25 frames (60 by 80 pixels, gray-scale) in length. The recovered Markov model, the mean image at each state, and plots of the memberships and residual for one sequence are shown in Figure 5. The recovered Markov model allows the symmetry of motion seen in the plot of membership over an observation sequence. Some of the observation sequences differ in the extent of the wave motion; in these cases the state representing the hand at its lowest or highest position in the frame may not be used. For a new instance of a wave gesture to be recognized, a high probability parse using the HMM must be possible.

Our second example describes the position and configuration of a waving, pointing hand (Figure 6). In each frame of the training sequences, a 50 by 50 pixel image window of the hand was tracked and clipped from a larger image with a cluttered background. Foreground segmentation was accomplished using the known background. The configuration of the hand is again modeled by the eigenvector decomposition of the image windows. The position of the hand is modeled by the location of the tracked hand within the larger image. The recovered Markov model is similar to that of the waving hand in the previous example except that now there are two components in the model of each state. As before, this gesture is recognized if a highly probable parse can be generated by the HMM.

The variance of each feature indicates the importance of the feature in describing the gesture. In this example both the position and configuration of the hand was relevant in describing the gesture. Had the location of the hand varied greatly in the training set, the high variance of the position representation would have indicated that position was not important in describing the gesture. The important point here is that each state defines the important models associated with that phase of the gesture.

The use of HMMs to encode the statistical sequence of movements or states associated with an activity has both advantages and disadvantages. The most important positive aspect of HMMs is their ability to learn the necessary states and transitions from training examples. Instead of a programmer explicitly coding the component movements or configurations, the learning algorithm will decompose the activity into natural phases. However, the disadvantage of HMMs is exactly the lack of control one has over

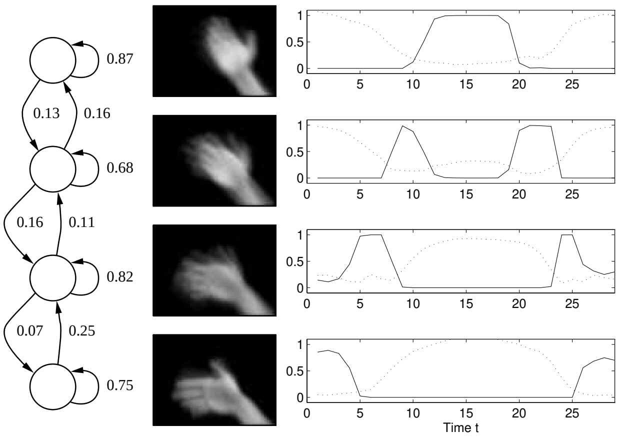

Figure 5: A wave gesture. The recovered Markov model for all training sequences at left shows the symmetry of the gesture. The mean image for each state is shown in the middle. On the right is a plot of membership (solid line) and residual (dotted line) for each state for one training sequence. The exact shape of the plots varies in response to the variance and length of the sequence.

(a)

(b)

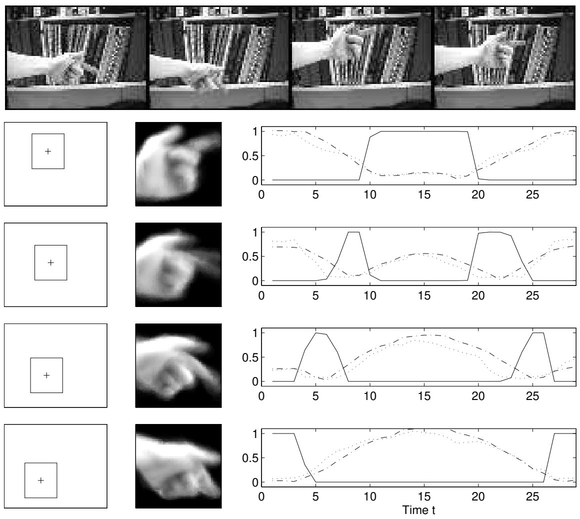

Figure 6: (a) Four representative frames (ordered left to right) are shown from one training sequence. (b) The mean location of the tracked hand in the larger image is shown on the left. The mean image for each state is shown in the middle. On the right is a plot of membership (solid line), configuration residual (dotted line), and the position residual (dash-dotted line) for each state for one training sequence.

the states recovered. Even if one has some a priori knowledge about the natural segments of an activity, it is difficult to incorporate them into the framework. Usually one can only affect the computation by specifying the topology of the state transition network to be learned.

An alternative to HMMs for the recovery of natural gesture was recently proposed in Wilson et al. (1996). In this approach, a fixed, finite state machine representation of a gesture sequence is employed. Each state is given a description in terms of the temporal properties of gesture. One such state would be described as being 1) similar in appearance to a "rest" state of the gesturer, 2) undergoing little motion, and 3) the hands typically remain in this state for a long duration. In this system only the duration parameters were learned from training data. The state transitions and descriptions were based on a priori understanding of the components of a sequence that make up the activity, in this case natural gesticulations. We demonstrated the ability to pick out semantically meaningful gestures comparable to an expert human observer.

Whether using HMMs or some other representation of activity, the requirements for recognition are the same: 1) a statistical or deterministic representation of a sequence of components that comprise the activity, and 2) a parsing mechanism that can temporally align the input signal with the known activity patterns. The knowledge encoded in these systems can be similarly considered as consisting of two elements. The first is the appearance or properties of the signal at different phases of the activity. While the gesture examples presented above are defined by individual static components, the baseball pitching example illustrates how some or all of the components may be atomic movements, relating the recognition of activity to the recognition of movements. The second element knowledge is the specificatoin of the quantitative temporal relationships between these components.

What is common between movements and activities is that neither refers to elements external to the actor performing them. That is, their occurrence can be "perceived" in the absence of knowledge of context or of the actor's interaction with other entities in the scene. Thus, the knowledge required to perceive movements or activties may considered strictly perceptual. In the next section we discuss motion understanding problems that are not so self-contained.

5 Recognition of actions: reasoning about space and time

The highest level of motion understanding is action recognition. Earlier we defined the action recognition task as requiring an interpretive context: a set of constraints on possible explanations for the observed motions. The discussion of Mann, et al.'s (1996) system for describing moving objects considered the use of the constraints of physics and measures of simplicity to derive likely explanations. The physical knowledge is exploited by providing consistency requirements and preference relations that any such explanation should satisfy.

A different focus is taken by Pinhanez and Bobick (Pinhanez & Bobick 1996, Bobick & Pinhanez 1997). In that work knowledge and the interpretive context is exploited to link perceptual signals to underlying actions. The application domain is SmartCams — cameraman-less cameras — that respond to a director's requests while filming a cooking show. Such cameras perform inverse video-annotation: given some symbolic description ("close-up chef") the system needs to generate the correct image.

From the perspective of this paper, the most important aspect of that work is that a script is available to the system; this script describes the actions that take place. These actions — e.g. *the chef is chopping the chicken* — are described using a logical formulation that allows perceptual inferences to be drawn. For example, the fact that the chef is chopping the chicken results in the assertion that the hands are moving and that they are likely near the chopping board. These inferences have visual implications and are exploited to select appropriate visual routines to perform the camera framing tasks.[3]

Fundamental to high-level action recognition is an explicit representation of time. One of the weaknesses of the SmartCam system as reported was a lack of a temporal reasoning mechanism that could consider temporal relationships between intervals; actions were strictly linear sequences. Recently, (Pinhanez & Bobick 1997) we have introduced the PNF constraint mechanism for temporal intervals that supports reasoning about time and the relationships between sensors and actions taking place at any given moment. To construct sophisticated action recognition mechanisms we need to be able to represent nontrivial temporal relationships such as partial ordering. The PNF formulation is real-time parsing mechanism based on Allen's temporal interval calculus that is designed to address such problems.

For example, using PNF it is simple to represent the action of *picking-up-a-bowl* as first the bowl is on the table, then the hands move towards and grasp the bowl, and then the bowl is off the table. Given a collection of such definitions and sensors capable of detecting events such as `hands-touching-bowl` or `bowl-off-table` the system can reason about which actions may be taking place currently. Using the PNF language it is also easy to say that A cannot take place at the same time as B but both must occur before C for some action to be considered as having taken place.

While the inferences supported by the system are not adequate to reason deeply about action (such as modeling any deep causality, [Israel, 1991; Brand, 1997]) many simple actions become possible to see. By simple we mean actions that have direct perceptual implications and can be recognized without extensive causal reasoning. Such reasoning is sometimes referred to as shallow (Jain and Binford, 1991) in that no explanatory theory is present. This is in contrast with systems such as that of Mann, et al. (1996) mentioned above where a qualitative physics theory is used to generate explanation based descriptions.

Whether the reasoning is shallow or deep, the knowledge required to perceive actions touches more than just the actor itself. Contextual or causal relations play a critical role. From the perspective of knowledge-based vision, the perception of action is the most knowledge intensive form of motion understanding.

[3] For more details and a demonstration visit the Web site: http://vismod.www.media.mit.edu/vismod/archive and search for SmartCams.

6 Conclusion: understanding motion

The problem of interpreting motion ("understanding action") has become a major thrust of computer vision research. Unlike object recognition, where generic classes were replaced with specific geometric models to make the problem tractable, the task of understanding actions will typically require representations of more than just geometry and appearance. The motion understanding taxonomy proposed — movements, activities, and actions — allows one to categorize particular approaches in terms of the representations and knowledge required to interpret the imagery.

Fundamental to the taxonomy discussed are the mechanisms necessary to manipulate time. The recognition of movements requires only simple linear (speed) invariance while the detection of activities employs more capable dynamic time warping methods. Finally, the perception of actions, even actions with direct visual correlates, necessitates reasoning about qualitative temporal relationships.

One of the utilities of this division of problems is the ability to immediately identify which techniques might be applicable to a given task. One cannot expect a movement-focused algorithm to extend trivially to the recognition of higher level action. For example, the temporal template method described will never perform generic shop-lifting detection unless the task can be formulated as a detecting a particular movement.

At the heart of the taxonomy are the knowledge and representations required to support the necessary inferences. Computer vision has developed numerous ways of representing a cup (Euclidean solids, superquadrics, spline surfaces, particles). How many ways do we have to represent throwing a baseball? Or even getting a wicket?

Acknowledgment: The taxonomy of action has its origins in discussions and collaboration with Stephen Intille and Claudio Pinhanez. Lee Campbell, Jim Davis, and Andy Wilson also contributed to the work presented here. The work presented here is supported in part by a research grant from LG Electronics and by ORD contract 94-F133400-000.

References

[1] Black, M. & Yacoob, Y. 1995 "Tracking and recognizing rigid and non-rigid facial motion using local parametric models of image motion," *Proc. Int'l Conf. Computer Vision*, Cambridge, MA.

[2] Bobick, A.F. 1996 "Computers Seeing Action," *British Machine Vision Conference,* Edinburgh, Scotland.

[3] Bobick, A.F. & Davis, J. 1996a "An appearance-based representation of action," *Proc. Int'l Conf. Pattern Recognition,* Vienna, Austria.

[4] Bobick, A.F. & Davis, J. 1996b "Real-time recognition of activity using temporal templates," *IEEE. Workshop on Applications of Computer Vision*, Sarasota, Florida.

[5] Bobick, A.F. & Pinhanez, C.S. 1997 "Controlling view-based algorithms using Approximate world models and action information," *Proc. IEEE Computer Vision and Pattern Recognition*, San Juan, Puerto Rico.

[6] Bobick, A.F. & Wilson, A.D. 1995 "A state-based technique for the summarization and recognition of gesture," *Proc. Int'l Conf. Computer Vision,* Cambridge, MA.

[7] Brand, M. 1997, "Physics-based visual understanding," *Computer Vision and Image Understanding,* 65(2).

[8] Campbell, L. & Bobick, A.F. 1995 "Recognition of human body motion using phase space constraints," *Proc. Int'l Conf. Computer Vision*, Cambridge, MA.

[9] Cedras, C. & Shah, M. 1995 "Motion-based recognition: A survey," *Image and Vision Computing*, 13(2), pp. 129-155.

[10] Darrell, T.J. and Pentland, A.P. 1993 "Space-time gestures," Proc. IEEE Computer Vision and Pattern Recognition, New York, New York.

[11] Davis, J.W. & Bobick, A.F. 1997 "The Representation and Recognition of Movement Using Temporal Templates," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Juan, Puerto Rico.

[12] Gavrilla, D.M., & Davis, L.S. 1996 "Tracking of humans in action:a 3-D model-based approach," *Proc. IEEE Computer Vision and Pattern Recognition,* San Francisco, CA.

[13] Israel, D., Perry, A., & Tutiya, S. 1991 "Action and Movements," *Proc. 12th Int'l Joint Conf. Artificial Intelligence*, Sydney, Australia.

[14] Jackendoff, R. 1990 *Semantic Structures,* MIT Press, Cambridge, Massachusetts.

[15] Jain, R.C. & Binford, T.O. 1991 "Ignorance, myopia, and naivete in computer vision systems," *CVGIP: Image Understanding,* 53(1), pp. 112-117.

[16] Mann, R., Jepson, A., & Siskind, J.M. 1996 "The Computational Perception of Scene Dynamics," *Proc. Fourth European Conf. Computer Vision*, Cambridge, England.

[17] Nagel, H.H. 1988 "From image sequences towards conceptual descriptions," *Image and Vision Computing,* 6(2), pp. 59-74.

[18] Nagel, H.H. 1977 "Analyzing sequences of TV-frames: system design considerations," *Proc. 2nd Int'l Joint Conf. Artificial Intelligence*, Cambridge, MA.

[19] Niyogi, S.A. & Adelson, E.H. 1994 "Analyzing gait with spatiotemporal surfaces," *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, Austin, Texas.

[20] Pinhanez, C.S. & Bobick, A.F. 1996 "Approximate world models: Incorporating qualitative and linguistic information into vision systems", *Proc. American Association of Artificial Intelligence*, Portland, Oregon.

[21] Pinhanez, C.S. & Bobick, A.F. 1997 "PNF Propagtion and the Detection of Actions Described by Temporal Intervals," *DARPA Image Understanding Workshop,* New Orleans, Louisiana.

[22] Polana, R. & Nelson, R. 1994 "Low level recognition of human motion," *IEEE Workshop on Non-rigid and Articulated Motion*, Austin, Texas.

[23] Quek, F. 1993 "Hand gesture interface for human-machine interaction," *Proc. of Virtual Reality Systems,* volume Fall.

[24] Rabiner, L. & Juang, B.H. 1993 *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ.

[25] Rohr, K. 1994 "Towards model-based recognition of human movements in image sequences," *CVGIP, Image Understanding*, 59(1), pp. 94-115.

[26] Schank, R.C. 1975 *Conceptual Information Processing*, North-Holland.

[27] Shavit, E. & Jepson, A. 1995 "Motion understanding using phase portraits," *IJCAI Workshop: Looking at People,* Chambery, France.

[28] Siskind, J.M. 1995 "Grounding Language in Perception," *Artificial Intelligence Review,* 8, pp. 371-391.

[29] Starner, T.E. & Pentland, A. 1995 "Visual recognition of American Sign Language using hidden Markov models," *Proc. of the Int'l. Workshop on Automatic Face- and Gesture- Recognition,* Zurich, Switzerland.

[30] Wilson, A.D. & Bobick, A.F. 1995 "Learning visual behavior for gesture analysis," *Proc. IEEE Int'l Symp. Computer Vision,* Coral Gables, Florida.

[31] Wilson, A.D., Bobick, A.F. & Cassell, J. 1996 "Recovering the Temporal Structure of Natural Gesture," *Proc. Second Int'l Conference on Automatic Face and Gesture Recognition,* Killington, Vermont.