M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 327
To Appear in the Proceedings of AAAI Fall '95 Symposium on Computational Models for Integrating Language and Vision

Causal Analysis for Visual Gesture Understanding

Matthew Brand
Learning and Common Sense Group
MIT Media Laboratory
Cambridge, MA 02139, U.S.A.
brand@media.mit.edu

Irfan Essa
Perceptual Computing Group
MIT Media Laboratory
Cambridge, MA 02139, U.S.A.
irfan@media.mit.edu

Abstract

We are exploring the use of high-level knowledge about bodies in the visual understanding of gesture. Our hypothesis is that many gestures are metaphorically derived from the motor programs of our everyday interactions with objects and people. For example, many dismissive gestures look like an imaginary object is being brushed or tossed away. At the discourse level, this implicit mass represents a referent in the conversation; at the scene-formation level, the dismissive gesture obeys many of the kinematic and dynamic constraints that would shape an actual tossing. Thus this metaphor provides us with constraints for both discourse annotation and visual processing. In this paper we present some preliminary results interpreting complex gesture sequences in video.

1 Introduction

As computers become visually competent and people are no longer fettered by keyboards, gesture will become an important channel of communication between humans and machines. In the foreseeable future, computers will called on to obey commands, mediate teleconferences, and process videos in which gestures convey significant content. Gestures as communicative behavior have been studied intensely by linguists in recent years (e.g., see [11, 5]) but little is known about how they are perceived or interpreted.

1.1 Gesture recognition

There is been a growing literature in body tracking and gesture recognition in computer vision (e.g., [7, 10, 15, 8, 13, 2]). Most of these efforts apply pattern recognition technologies to reduced-dimension codings of the image stream. These codings, while ingenious, shed no light on gesture understanding *per se*; many could just as well be directed to recognize flight-maneuvers of houseflies or car chases.

1.2 Recognition is not understanding

The language of gestures is body-specific, temporally variable, and subject to co-articulation effects. It is also generative, creative, and highly-context dependent for its actual meanings. In many circumstances "recognition" will be impractical or even ill-defined. At some point, a high-level system of meaning must be applied to the low-level data, both to resolve ambiguities in the input and to produce a useful interpretation of the data. In between, there are also mid-level constraints that are very useful. For example, the work of Essa [9] is a promising step in the direction of having a theory of gesture that has computational value: in order to track facial expressions, the dimension is reduced through a causal reconstruction of how the scene was produced, specifically, face images are coded in terms of the muscle contractions needed to produce each expression.

Similar attempts at "understanding" body movements in terms of model-based constraint propagation [12], and coding on the basis of joint-angles [4, 1] provide additional insight into the validity and a need for a higher level representation.

1.3 Gestures as metaphors of basic motor plans

We are exploring the use of high-level knowledge about bodies in the visual understanding of gesture. Our hypothesis is that many gestures are metaphorically derived from the motor programs for our everyday interactions with objects and people. For example, many dismissive gestures look like an imaginary object is being brushed or tossed away. At the discourse level, this implicit mass represents a referent in the conversation; at the scene-formation level, the dismissive gesture obeys many of the kinematic and dynamic constraints that would shape an actual tossing. Thus this metaphor provides us with constraints for both discourse annotation and visual processing.

Knowledge of how the body interacts with the physical world can be used to interpret several kinds of gestures, including:

- pushing, tossing, flicking, and brushing gestures, which metaphorically signal kinds of rejection (of ideas, lines of conversation, etc.);

- opening and closing gestures, physically controlling the access others have to one's body and metaphorically signaling kinds of receptivity;

- holding and blocking gestures, physically arresting moving objects and metaphorically used to slow down conversation;

- pointing and spanning gestures—controlled positioning or sweeps of the arm that spatially lay out a narrative and the positions of actors within it.

Our goal is to extract clues to arm positions from video, subject them to causal analysis, and produce construals like those above. This output will be useful for video editing and in virtual manipulation tasks, e.g. "virtual CAD." In this paper we present some preliminary results interpreting complex gesture sequences in video.

2 Causal analysis and high-level vision

The approach used here derives from work in scene analysis done by [3, 6]. The two central themes of this approach are *explanation-mediated vision* and *causal analysis*. In explanation-mediated vision, a high-level reasoning process directs and disambiguates low-level visual processes in the course of trying to assemble visual clues into a sensible explanation of what is going on in the scene. The scene representations used in this process are causal, meaning that they describe the underlying physics of the scene. Previous systems have used basic causal knowledge about substantiality, rigidity, mass, friction, and balance to intelligently understand and manipulate scenes of blocks towers, tinker-toy constructions, Lego Technics machines, and coffee mugs, solving a variety of visual problems along the way.

Generally, these systems have three layers. First, low-level visual processes extract clues to a scene's causal structure. From these clues, the scene elements receive a partial and tentative labeling. Finally, a high-level reasoning process tries to "explain" the labeling vis-a-vis causal knowledge of the domain. If domain constraints are violated, the reasoner proposes alternate labelings, along with justifications. If the labeling leads to an incomplete explanation, the reasoner speculates about objects, events, or relations that may not have been seen, also with justifications. Justifications incorporate predictions about the scene which are then used to guide or tune low-level sensing operations. The systems repeat this inspect-explain-predict cycle until the scene is understood.

To move to video, we "unroll" this cycle into time, so that events can be combined with domain knowledge to generate expectations for visual processing of later frames. In order to do this in the gesture of domain, we need to identify important clues to causal structure, how they relate to meaningful gesture labels, and what knowledge can be used to interpret gestures in a context of storytelling or conversation. The remainder of this paper is devoted to the extraction of these clues.

3 Causal analysis of motor plans

It is useful to contrast gesture and speech understanding to get a sense of these three kinds of knowledge available to constrain the interpretation process. They are:

- knowledge of the mechanics of the system that produces the signal, e.g., how the vocal tract produces speech or the body produces motion;

- knowledge of how this behavior maps to meaningful symbols, e.g., what sounds or motions are privileged as words and gestures;

- knowledge of why certain utterances make sense in sequence, e.g., utterances and gestures that are appropriate at any given point in a sentence or conversation.

At present, we are working on the first two—using knowledge about body kinematics to assist with the recovery of arm positions from video, and using knowledge about body dynamics to identify gestures that refer to actual motor plans. For example, consider a lifting gesture, which often follows on a shrug and metaphorically "raises a question." There are a number of interesting constraints which apply to the actual motor plan for lifting:

- the largest motion component at the wrists is upward;

- the greatest acceleration is at the beginning of the gesture, setting the arms in upward motion;

- there is work against gravity;

- the joints all keep in a vertical plane, because that is the most efficient way for human arms to brace against a weight.

We can design a filter for each of these, then combine them to produce a preliminary segmentation of a video sequence into underlying motor plans.

3.1 Some measures and their meaning

Here are some of the useful measures that we can extract from video that have both causal and metaphorical significance. These measures are constructed from shoulder, elbow, and wrist joint positions for the right arm, $S_r = (x_{rs}, y_{rs}, z_{rs})$; $E_r = (x_{re}, y_{re}, z_{re})$; $W_r = (x_{rw}, y_{rw}, z_{rw})$; $z = 0$ is the image plane.

- Wrist acceleration and deceleration. $|W|''$. This is useful for identifying gesture boundaries and for distinguishing between gestures that have similar trajectory, but different dynamics (e.g., repeated patting and lifting). We may further distinguish gestures that have inertial follow-through from those that halt in mid-motion, for example, a full-trajectory push versus a push which terminates mid-trajectory. One connotes rejection whereas the other connotes "stop" or "wait."
- Work done against gravity. $((W+2E+S)/4)'$. Useful for distinguishing gesture preps, especially if the acceleration is low, the hands are loose, or the hands are below the "gesture box." When the arms are lifted, it may also be a measure of "gesture loudness."
- Size of gesture. Perimeter of the rhombus connecting shoulders and wrists. An indicator of "gesture loudness."
- Area between arms. Calculated by fitting a plane to the wrists and shoulders and then measuring the area of the rhombus they project onto it. A good measure of openness. This measure goes negative when arms cross, typically signifying that one is closed to something.
- Angles between forearms. A sum of the angles each forearm makes to an imaginary line connecting the mid-forearms. Also a measure of openness and important for determining when arms are sufficiently coplanar to carry a mass.
- Nearness to body. $|W_z - S_z|$. Depending on whether arms cross, this could signal intimacy (holding something to one's self) or hostility (blocking something out).
- Verticality. Calculated by fitting a plane to the joints of each arm and measuring its vertical angle to the plane $x = 0$. Indicates degree of self-bracing under a load.
- Shaking and repetition. Extracted from phase plots of wrist position versus velocity. Signals emphasis or agitation.

Note that we do not use hand position in any of these. Even crude measures of hand position such as palm orientation would be very helpful in discriminating between potentially ambiguous gestures. Doing this visually is very difficult, since in typical full-body videos the hands only occupy a few pixels and even then can be difficult to segment from the rest of the moving arm and shirt. One possibility we intend to explore is the opportunistic use of limited information from existing hand-recognition systems: for those frames that they work, and to interpolate hand positions using constraints from the skeletal kinematics of the arm.

3.2 Some gesture filters

The measures above can be combined to form estimates of motor plans followed by a gesturer. G is a function that normalizes the measures and combines them as if they were confidence measures. Some of the filters we have constructed are:

- Opening. $G(\mathrm{forearm}_\theta, \mathrm{area})$. Widely angled forearms give others access to the body; the area measure ensures that there is actually space for someone to step in. Metaphor for receptivity.
- Lifting. $G(\mathrm{verticality}, \mathrm{acceleration_y}, \mathrm{work})$. The design of this filter is described above. Metaphorically holds up a discourse referent for discussion.
- Patting. $G(\mathrm{verticality}, \mathrm{deceleration_y})$. Also described above. Metaphor for regulating or holding down parts of the discourse.
- Pushing. $G(\mathrm{verticality}, \mathrm{acceleration_z}, -|(\mathrm{y_s}+\mathrm{y_e})/2 - \mathrm{y_w}|)$. Similar to patting in the depth plane. An additional constraint is that the wrists need to be nearly the same height as the mid-upper arms, so that the body's weight is centered behind the push and, kinematically, the palms can comfortable face forward. Metaphor for rejection.
- Stopping. $G(\mathrm{verticality}, \mathrm{deceleration_z}, -|(\mathrm{y_s}+\mathrm{y_e})/2 - \mathrm{y_w}|)$. Dynamical converse of pushing, in the same way that patting is to lifting.
- Clutching. $G(\mathrm{closeness}, 1 - |\mathrm{area}|)$. Degree to which something is being held to the torso, without arms being crossed. Metaphoric indication of significance to self.

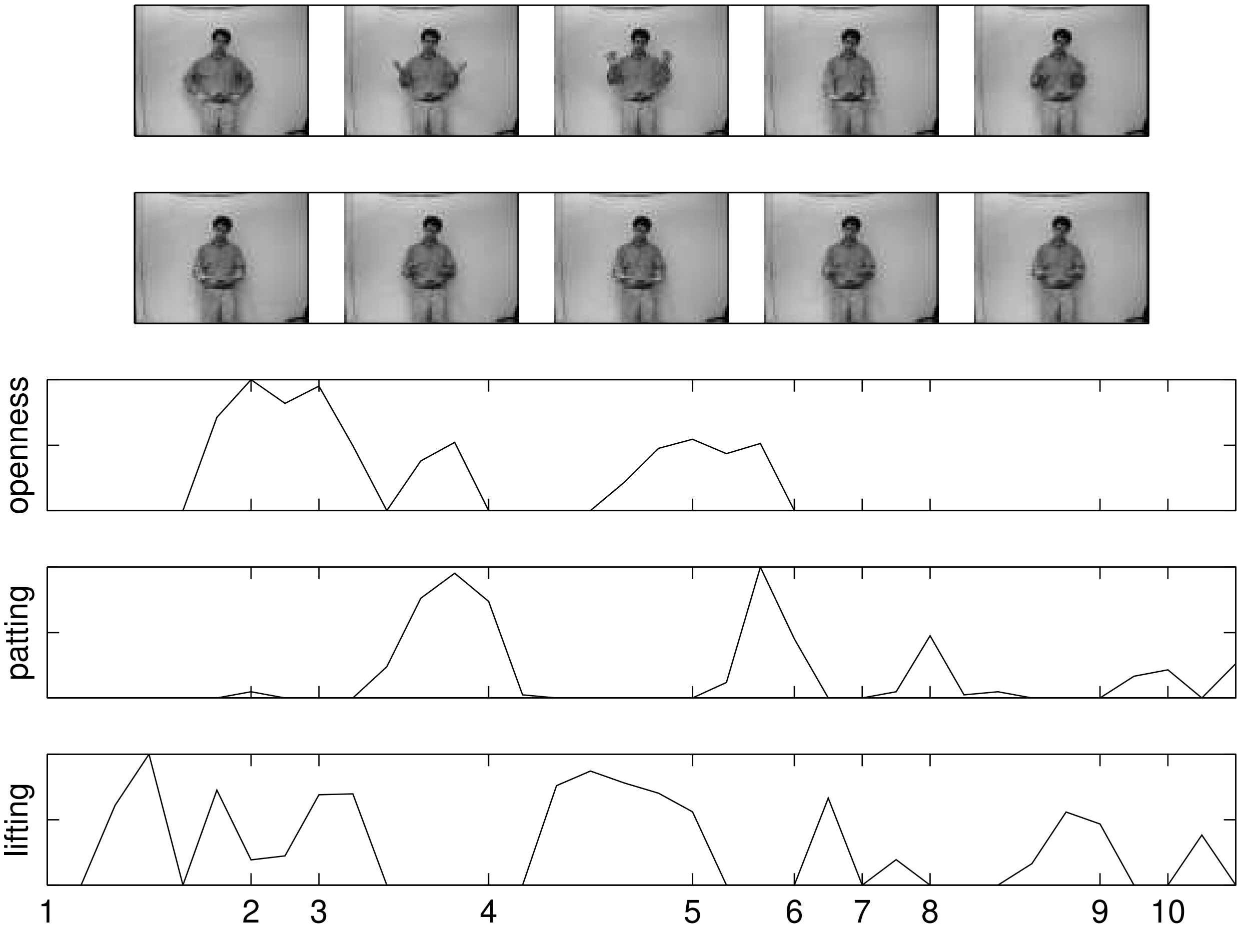

Figure 1: *Key frames and analysis from a shrug-then-repeatedly-pat sequence. The sequence was sampled at roughly 10 frames/second. Causally-constructed filters correctly label most of the sequence, even though some of the end pats are nearly undersampled.*

4 Examples

We have captured a number of storytelling and conversational video sequences. Two-dimensional joint positions are currently extracted by hand. From 2D data we invert the perspectival foreshortening to extract full 3D joint positions. At this point we identify causally significant arm motions by applying the gesture filters to the reconstructed motion data. We are working on a frame-subtraction system to identify and track shoulders and elbows using contours, and to track hands using color [14].

In the sequence in figure 1, a subject starts with hands on hips, makes an opening shrug, and then follows with a series of small patting gestures. In the audio track the subject expresses uncertainty and then talks about keeping things under control. The graphs show the response of each of the gesture filters. We use the area under the curve to determine the dominant filter for any segment of video, thus assigning extra confidence to filters whose responses persist in time. The opening shrug and all the pats are correctly labeled. The lifting filter also scores highly for appropriate segments, but we know that this is not a series of lifting gestures but a series of patting gestures with co-articulation phases that look like lifting. Since we don't having palm orientation to damp the "lifting" response, this is the sort of problem that we must use high-level inference to solve. This, and the co-articulation between opening and the first pat, will be suppressed when the top-level inference system is brought online.

Similarly, figure 2 shows a gesture sequence in which the subject makes a dismissive "pshaw" gesture–an exaggerated push, after which the arms follow through and fall to rest. The graphs show the responses of these two filters.

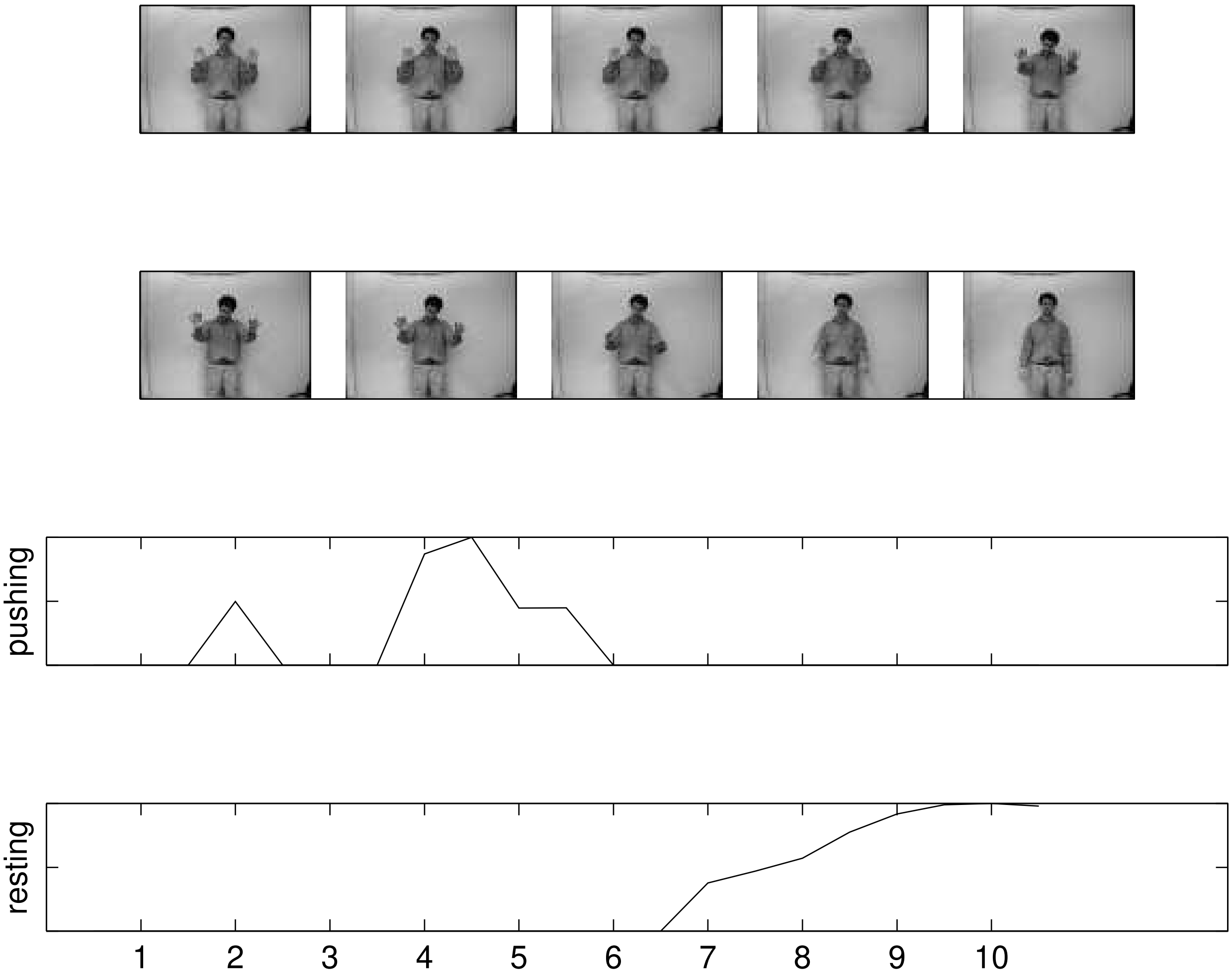

Figure 2: *Key frames and analysis from a dismissive push-then-rest sequence.*

5 Future Work

Our ultimate goal is to incorporate higher-level knowledge about discourse into this system. This knowledge ranges from rules predicting typical co-articulations (like those in the example) to expectations about how gestures follow upon each other. In the case of co-articulations, filter drop-outs, and visual tracking mistakes, we will need this knowledge to boost and suppress filters and reconstruct lost data, much in the way that reasoning was used to predict image structures and tune visual routines in [3]. In another avenue of exploration, we are looking at the use of knowledge to actually construct filters on the fly, employing the same reasoning that we demonstrated for the "lifting" filter. But the most important benefit of using knowledge is that the final output will be meaningful: A metaphor-driven interpretation system should provide a running commentary on the mood and possible meanings of the speaker. For example, in a conversation, one would expect an opening gesture signifying "I am receptive to your opinion" to be answered with a gesture signifying an evaluation, perhaps a dismissive toss or a standoffish push. One could exploit simple representations of discourse structure to produce outputs such as "person A is making a query; person B is signaling that they dislike the premise of the question."

6 Conclusion

We are exploring the use of high-level knowledge about bodies in the visual understanding of gesture. Our hypothesis is that many gestures are metaphorically derived from the motor programs of our everyday interactions with objects and people. For example, many dismissive gestures look like an imaginary object is being brushed or tossed away. At the discourse level, this implicit mass represents a referent in the conversation; at the scene-formation level, the dismissive gesture obeys many of the kinematic and dynamic constraints that would shape an actual tossing. Thus this metaphor provides us with constraints for both discourse annotation and visual processing. In this paper we use this insight to design a repertoire of salient filters for detecting the "prosaic" motor plans that underlie gesture, and present some preliminary results using these filters to interpret complex gesture sequences in video.

7 Acknowledgments

Many thanks to Alex Pentland and Justine Cassell for helpful discussions.

References

[1] N. I. Badler and S. W. Smoliar. Digital representation of human movement. *ACM Computing Surveys*, 11:19–38, March 1979.

[2] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proc. Fifth International Conference on Computer Vision*, pages 382–388, 1995.

[3] Matthew Brand, Lawrence Birnbaum, and Paul Cooper. Sensible scenes: Visual understanding of complex structures through causal analysis. In *Proc. National Conf. Artificial Intelligence*, 1993.

[4] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE Computer Society, June 1995. To Appear.

[5] J. Cassell and D. MacNeill. Gesture and ground. In K. Hall, J-P. Keonig, M. Meachman, S. Reinman, and L. Sutton, editors, *Proc. 16th Annual Meeting of the Berkeley Linguistics Society*, pages 57–68, 1990.

[6] Paul Cooper, Lawrence Birnbaum, and Matthew Brand. Causal scene understanding. *Image Understanding*, 1995. In press.

[7] Trevor Darrell and Alex Pentland. Space-time gestures. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 335–340, 1993.

[8] J. W. Davis and M. Shah. Gesture recognition. In *Proc. European Conf. Computer Vision*, pages 331–340, Stockholm, Sweden, 1994.

[9] Irfan A. Essa and Alex Pentland. A vision system for observing and extracting facial action parameters. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 76–83, 1994.

[10] Ioannis Kakadiaris, Dimitri Metaxas, and Ruzena Bajcsy. Active part-decomposition, shape, and motion estimation of articulated objects: A physics-based approach. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 980–984, 1994.

[11] David MacNeill. Iconic relationships between language and motor action revisited. In M.E. Landsberg, editor, *Syntactic iconicity and freezes: The human dimension*. Mouton de Gruyter, Berlin, 1993.

[12] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-2:522–536, 1980.

[13] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *Proc. European Conf. Computer Vision*, volume 2, pages 35–46, Stockholm, Sweden, 1994.

[14] Christopher R. Wren, Ali Azarbayejani, Trevor Darrell, Michael Johnson, Kenneth Russel, Thad Starner, and Alex Pentland. Hierarchical classification for human body tracking. In *Integration Issues in Large Commercial Media Delivery Systems*. SPIE, October 1995. to appear.

[15] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using a hidden markov model. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 379–385, 1992.