

VAMBAM: View and Motion -based Aspect Models for Distributed Omnidirectional Vision Systems

Hiroshi Ishiguro* and Takuichi Nishimura**

*Department of Computer & Communication Sciences, Wakayama University, Japan

** Cyber Assist Research Center,

National Institute for Advanced Industrial Science and Technology (AIST)

*ishiguro@sys.wakayama-u.ac.jp, ** t_nishimura@aist.go.jp

Abstract

This paper proposes a new model for gesture recognition. The model, called view and motion-based aspect models (VAMBAM), is an omnidirectional view-based aspect model based on motion-based segmentation. This model realizes location-free and rotation-free gesture recognition with a distributed omnidirectional vision system (DOVS). The distributed vision system consisting of multiple omnidirectional cameras is a prototype of a perceptual information infrastructure for monitoring and recognizing the real world. In addition to the concept of VABAM, this paper shows how the model realizes robust and real-time visual recognition of the DOVS.

1 Introduction

An open problem of visual recognition is “segmentation” of sensory data. Based on Marr’s paradigm [Marr, 1982], many researchers have tried to build recognition systems using 3-D models as the intermediate representations. There are mainly two methods. One is to detect visual features such as line segments and surfaces bounded by the lines and then to directly fit wire-frame models to the detected features or by estimating 3-D positions of the features by using multiple views. Another more powerful method is to detect targets by multiple-camera stereo [Kanade, xxxx]. The target segmentation by cameras surrounding the target is robust against lighting conditions. A problem of the former method is in the feature segmentation in the image. It has a serious problem for deformable targets, such as humans. A problem of the later method is in the computational const. Even if we use the most powerful computer, it is difficult to compute in real-time.



Fig. 1: Omnidirectional camera

The purpose of this paper is to develop a real-time vision system that tracks and recognizes human behaviors

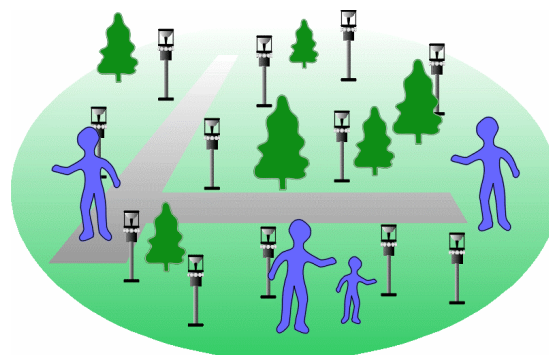


Fig. 2: Distributed omnidirectional vision system

in a wide area. Therefore, the methods based on the 3-D models are not suitable for our purpose. We, rather, take a new but well-known memory-based approach for the segmentation problem. Recent progress of computer hardware enabled us to use a large volume of memory and to access in real time. This paper follows the basic idea and proposes a general framework of visual recognition in a wide environment.

The system introduced in this paper consists of omnidirectional cameras. Each camera has an omnidirectional visual field and observes a wide range. Fig. 1 shows a compact omnidirectional camera [Ishiguro, 1998]. By using multiple omnidirectional cameras, the system redundantly covers the wide environment as shown in Fig. 2. We call it *Distributed Omnidirectional Vision System* (DOVS). The goal of our research is to develop a *perceptual information infrastructure* [Ishiguro, 1997] by using the distributed omnidirectional vision system as an extension of computer and sensor networking.

The key function of the DOVS is to track walking humans and to recognize the gestures in real time. That is, it is *location-free and rotation-free gesture recognition*. This paper proposes a new model of visual recognition for realizing the function. We call the model *View And Motion -Based Aspect Model* (VAMBAM). The model has two features as follows:

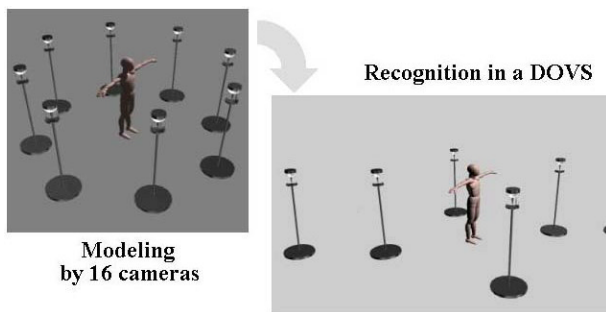


Fig. 3: Modeling and recognition

1. Omnidirectional model of a gesture
2. Motion-based segmentation of the gesture

As shown in Fig. 3, the multiple cameras observe a human from various viewing point to build the VAMBAM. Then, the system recognizes human gestures in the distributed vision system that consists of multiple cameras distributed in the environment. A VAMBAM maintains the visual information in a form of view-based aspect model. In the modeling, a gesture is observed as image sequences. Then, the image sequence is transformed to a feature vector that represents an aspect of the gesture based on the motion information. Thus, VAMBAM maintains omnidirectional and motion-based gesture information. This gesture recognition not using 3-D models has not been developed so far [Pavlovic, 1997].

In the following sections, Section 2 explains the design policies of VAMBAM. To utilize VAMBAM, we have developed a 3-D dynamic programming matching. Section 3 describes the detail. Finally, Section 4 reports experimental results and effectiveness of VAMBAM.

2 VAMBAM

2.1 Omnidirectional Aspect Model

The system has two phases for recognizing human gestures: modeling phase and recognition phase (see Fig. 3). In the modeling phase, the system acquires VAMBAMs by using 16 cameras surrounding a human. The human makes several gestures and the system memorizes each gesture as an individual VAMBAM. In the recognition phase, the system consisting of multiple omnidirectional cameras finds a human, tracks, and then matches acquired images from several cameras with the stored VAMBAMs. This section explains the modeling phase.

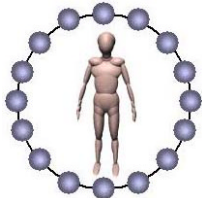


Fig. 4: 16 feature vectors

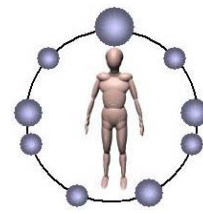


Fig. 5: Weighted 16 feature vectors

An ideal model for object recognition is to have all visual information from all directions like a complete 3-D model. If there are an infinite number of cameras surrounding an object, we can directly acquire such an ideal view-based model. VAMBAM is an approximation of such an ideal model.

A problem is how to reduce the number of cameras surrounding a modeling target in the modeling phase. It depends on complexity of the target. Since the purpose of the DOVS is to recognize coarse human gestures, we have decided the number as 16 based on a preparatory experiment.

By using 16 cameras, the system acquires 16 view sequences and transforms them to 16 feature vectors. In Fig. 4, the sphere shows a feature vector. Here, we suppose all of the cameras are located at the same height. Of course, we can extend the camera arrangement, but it is not needed for our current gesture recognition.

All 16 cameras do not have the same priority. For recognizing a particular gesture, for example breathing gesture, the front camera is more important than the side cameras. Therefore, each feature vector has a weight estimated from the motion information. If the camera observes more moving pixels by background subtraction than the others, it has a bigger weight. Fig. 5 shows the conceptual figure of the weight vectors. *The weights define aspect changes* in VAMBAM. Fig. 5 shows 9 feature vectors that have positive weights, and it represents 9 aspects. The system efficiently refers to the weight for searching the target in the recognition phase.

The VAMBAM shown in Fig. 5 is obtained by observing the target with a constant distance. If the camera is distant from the object, the aspect also changes. Bobick and Bolles [Bobick, 1992] reported an important idea that the model structure changes according to the resolution of a vision sensor. We follow the idea and

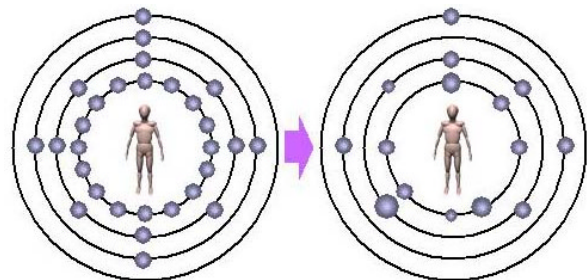


Fig. 6: Full VAMBAM

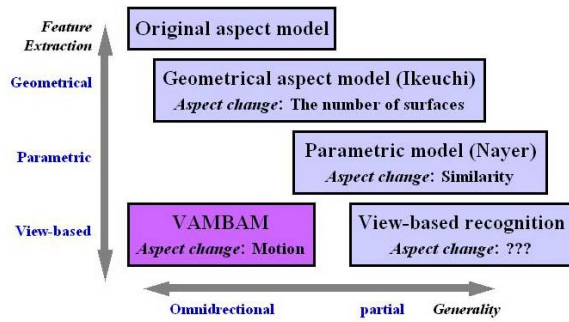


Fig. 7: Several approaches for visual modeling

extend a VAMBAM to a full VAMBAM as shown in Fig. 6. The full VAMBAM has several layers to handle different resolutions. In Fig. 6, the lateral circle represents the coarsest resolution (or distant observation). Although we have not developed the system using the Full VAMBAM, this concept is very important for our future extension of the basic model.

Here, let us state VAMBAM as an aspect model. Aspect models represent omnidirectional visual information surrounding the target. Koenderink [Koenderink, 1979] proposed the original concept based on 3-D structure of the object. Ikeuchi [Ikeuchi, 1988] modified for passive vision systems and defined the aspect change by the number of observable surfaces of the object. In their works, the important point is that the aspect model maintains omnidirectional visual information for the generality. On the other hand, view/appearance-based methods and parametric methods are also a kind of aspect model. The view-based method defines the aspect change as a difference among views by template matching or feature-based comparison. The parametric method defines the aspect change as a distance in the principle component parameter space. They, however, do not maintain the omnidirectional visual information in the previous works. VAMBAM is a view-based approach as described in the next subsection, but it is also an aspect model that maintains omnidirectional visual information. Fig. 7 shows relationships among the existing methods and our method.

2.2 Motion-based Segmentation

Another important feature of VAMBAM is motion-based segmentation of visual information. As mentioned before, the motion-based segmentation reduces the amount of memory and enables us to utilize omnidirectional view-based aspect models for practical systems.

In Fig. 4, each feature vector is given as shown in Fig. 8. The camera finds a moving object by background subtraction and tracks it as a candidate of a human. Then, it detects the center of gravity by making histograms along vertical and horizontal axes. The histograms give the height and width of the human. Based on the parameters, the system defines 9 regions that cover the

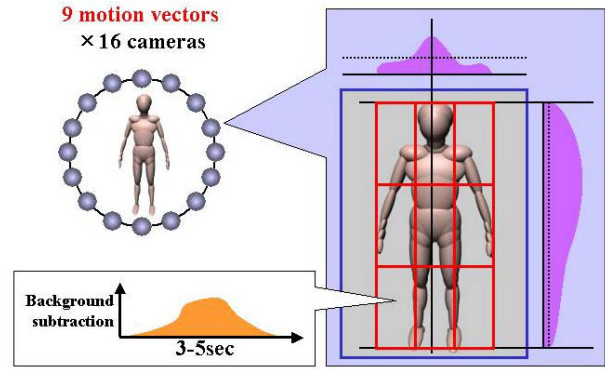


Fig. 8: Acquisition of a feature vector

human as shown in the left of Fig. 8, tracks the human fitting the regions. In each region, the system recodes the moving pixels detected by background subtraction and acquires 9 motion vectors in 3-5 sec. observation. Here, the observation length is adjusted according to the gesture. The acquired 9 motion vectors define a feature vector as represented with a sphere in Fig. 8. As mentioned before, several feature vectors do not contain motion information because of the viewing direction. By referring to the average and variance of the 9 motion vectors, the system assigns a weight to the feature vector.

The motion-based segmentation by background subtraction is one of the most robust and stable feature extractions and many practical systems use background subtraction. If we consider that “motion” represents relationships between static objects and events and gives meaning to the static objects, it is natural to build a recognition system based on the motion-based feature extraction.

Further, this representation of a gesture is compact. Suppose to record a 5 sec. gesture. A motion vector consists of 5×30 integers and 16 feature vectors have $5 \times 30 \times 9 \times 16 = 21600$ integers and the quarter is selectively memorized by the weights. If we represent an integer with 8bit, the necessary memory size will be 4000 bit and it is not large for recognizing a dozen of gestures in real time.

3 Two Techniques to Support VAMBAM

3.1 Real-time Tracking of Walking Humans

For realizing the location-free and direction-free gesture recognition, the system needs two more functions: one is tracking of walking humans in real time, the other is a robust matching method with gesture models.

By utilizing redundant visual information of the DOVS, we can build a robust and real-time multiple camera stereo system. This stereo, called *N-ocular stereo*, is an extension of trinocular stereo [Kitamura, 1990]. The basic process of N-ocular stereo is as follows:

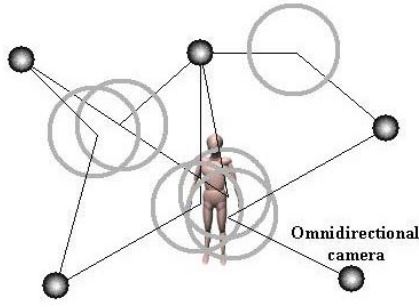


Fig. 9: N-ocular stereo by a DOVS

1. Detect moving regions on the images by background subtraction.
2. For all combination of two moving regions on different omnidirectional cameras, estimate the distance to the region and apply a circle as a human model (the circles in Fig. 9).
3. Check the overlapping of the circles and determine human locations.

As described above, we implement the N-ocular stereo as a model-based stereo method. Generally, the model-based stereo is more stable than the feature-based stereo in the case where the target is known. However, appearance of a human body frequently deforms on the image. Therefore, we have approximated the human body with a circle (the diameter is 60 cm) on the horizontal plane that is parallel to the floor. When three or more circles overlap each other, the system decides a human exists in the center of gravity of the circles. [Sogo, 2000, Sogo, 2001].

This N-ocular stereo for tracking humans in a DOVS is executed in real-time with a standard computer and it robustly tracks up to 3 humans simultaneously with four omnidirectional cameras. The reason of the robustness is in the redundant observation by multiple omnidirectional vision sensors. The method based on background subtraction is, generally speaking, noisy against change of lighting condition. However, the multiple observations from different viewing angles suppress the noise.

Another issue in the N-ocular stereo is precise localization and identification of the omnidirectional cameras. For this issue, we have already developed a method that automatically performs the localization and identification [Kato, 1999].

3.2 3-D Dynamic Programming for Matching

The recognition system needs to handle multiple gesture models simultaneously. Therefore, a robust and real-time matching technique is required.

The basic idea is to use conventional continuous dynamic programming (CDP). We have extended this CDP to interpolate discrete observation by 16 cameras, which is called *Spatio-Temporal Continuous Dynamic*

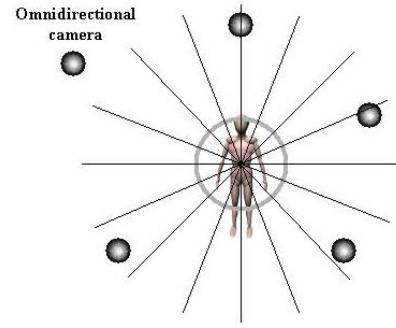


Fig. 10: Matching with a VAMBAM

Programming (STCDP, see Fig. 11). In the recognition phase, a stored model does not exactly match to the observations by the DOVS as shown in Fig. 10. The STCDP searches the best match by swiveling the model within 360/16 degrees. In addition to this, the STCDP finds the direction of target by rotating the models 360 degrees. Thus, this 3-D matching realizes the rotation-free gesture recognition by using the discrete model.

By taking account of all conditions in the matching, we formalize the STCDP as follows. Let us denote the cumulative distance of a point (i_q, t, t) for evaluation of the DP matching as $S(i_q, t, t)$. Suppose the q th feature vector is rotated i_q up to N_q and the model length is T . The STCDP calculates $S(i_q, t, t)$ by using the following recursive equations.

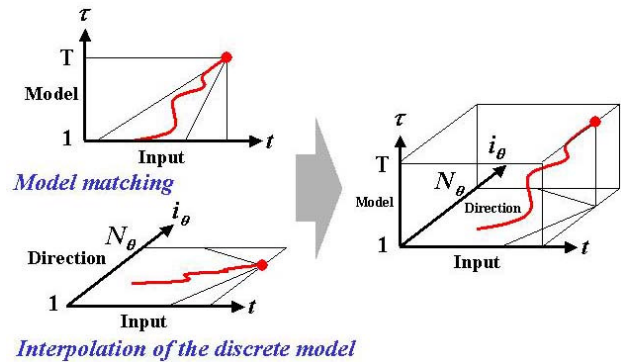


Fig. 11: STCDP

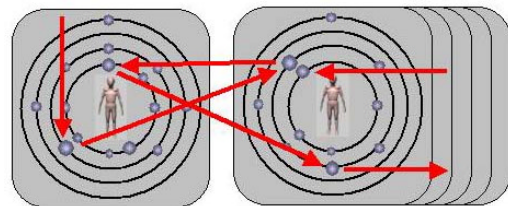


Fig. 12: Attention control in the matching

Boundary conditions ($1 \leq i_q \leq N_q, 1 \leq t \leq T, 0 \leq t$):

$$S(i_q, 0, t) = \infty$$

$$S(0, t, t) = S(N_q + 1, t, t) = \infty$$

$$S(i_q, t, -1) = S(i_q, t, 0) = \infty$$

Recursive equations ($1 \leq t$):

$$S(i_q, 1, t) = 3 \cdot d(i_q, 1, t)$$

$$S(i_q, t, t) =$$

$$\min \begin{cases} S(i_q, t-1, t-2) + 2 \cdot d(i_q, t, t-1) + d(i_q, t, t) \\ S(i_q, t-1, t-1) + 3 \cdot d(i_q, t, t) \\ S(i_q, t-2, t-1) + 3 \cdot d(i_q, t-1, t) + 3 \cdot d(i_q, t, t) \\ S(i_q-1, t-1, t-2) + 2 \cdot d(i_q-1, t, t-1) + d(i_q, t, t) \\ S(i_q-1, t-1, t-1) + 3 \cdot d(i_q, t, t) \\ S(i_q-1, t-2, t-1) + 3 \cdot d(i_q-1, t-1, t) + 3 \cdot d(i_q, t, t) \\ S(i_q+1, t-1, t-2) + 2 \cdot d(i_q+1, t, t-1) + d(i_q, t, t) \\ S(i_q+1, t-1, t-1) + 3 \cdot d(i_q, t, t) \\ S(i_q+1, t-2, t-1) + 3 \cdot d(i_q+1, t-1, t) + 3 \cdot d(i_q, t, t) \end{cases} \quad (2 \leq t \leq T)$$

In the current implementation, the STCDP sequentially matches the input with 10 stored models. However, the STCDP needs to be modified to handle a large number of models. The weights assigned to the VAMBAM reduce the computational time drastically. By referring to the weights, the STCDP can dynamically select possible feature vectors and ignore redundant search for impossible models in the early stage as shown in Fig. 12. In other words, we can see attention control in the recognition phase using VAMBAM.

3 Experimental Results

For the modeling, we have arranged 16 cameras along a circle of 3m diameter and acquired ten model gestures: (a) deep-breathing, (b) turning around the upper body, (c) kicking, (d) throwing, (e) jumping, (f) opening two legs, (g) picking up, (h) sitting down, (i) bowing, (j) Sumo-stomping. Fig. 13 shows the modeling system.

Then, we have verified the performance of the proposed method by using 4 omnidirectional cameras. Fig. 14 shows the top view of 4 cameras tracking a walking human and Fig. 15 shows a part of the trajectory. Fig. 16 shows an image sequence while the system recognizes “kick” gesture. Fig. 17 shows the monitor output of the system.

As shown in Table 1, we could verify good performance of the system. In the recognition phase, we have prepared different humans with the modeling phase and they have randomly iterated the 10 gestures 10 times.



Fig. 13: Modeling system

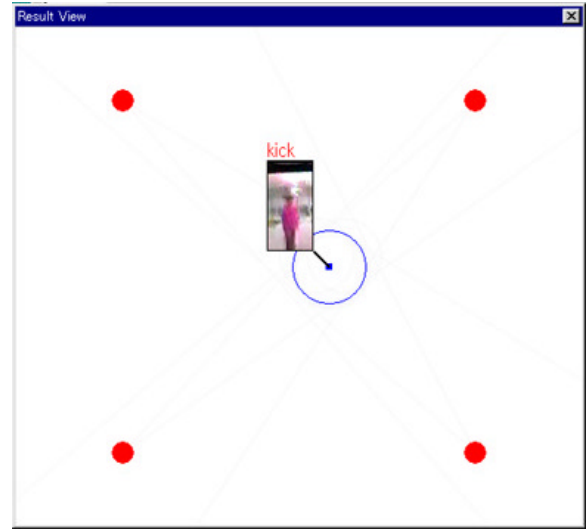


Fig. 14: Tracking of a walking human

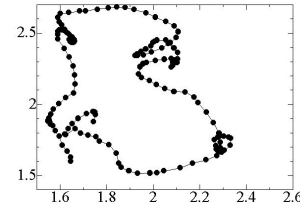


Fig. 15: Trajectory of the walking human

When the human behaves a gesture at an arbitrary location with an arbitrary orientation, the recognition results using STCDP was 100% with 4 cameras as shown in Table 1(a). Further, a human can behave similar behavior even if he/she is moving. In the case, the performance was 87%. In this experimentation, the system has recognized in real time.

The purpose of this system is to realize gesture recognition in a wide area. Although the experimental results are not sufficient, we could verify the basic function.

(a) Recognition results when the human stops

	CDP	STCDP
1 camera	65%	83%
4 cameras	100%	100%

(b) Recognition results when the human moves

	CDP	TSCDP
1 camera	31%	62%
4 cameras	64%	87%

Table 1: Recognition results

4 Conclusions

This paper has proposed a new model for gesture recognition, called VAMBAM. By using the model for the distributed omnidirectional vision system, we have developed a location-free and rotation-free gesture recognition system.

There are several remained problems yet. One is resolution of the gesture recognition. The current system handles only large gestures using arms and legs. However, we consider high-resolution cameras for the omnidirectional cameras can solve this problem. Another problem is the discrete omnidirectional models using 16 cameras. A better method is to acquire continuous models from the discrete observation and search by 2-D dynamic programming. We are now improving the algorithm.

The model, VAMBAM, is simple but general. Therefore, we consider this can be extended and realize more sophisticated recognition systems that can recognize more complex scenes, such as interactions among humans. Our final goal is to develop a *perceptual information infrastructure* to support human activities based on the proposed method.

References

- [Bobick, 1992] A. Bobick and R. C. Bolles, The representation space paradigm of concurrent evolving scene descriptions, IEEE Trans. PAMI, Vol. 14, No. 2, pp. 146-156, 1992.
- [Ikeuchi, 1988] K. Ikeuchi, Automatic generation of object recognition program, Proc of IEEE, Vol. 76, No. 8, 1988.
- [Ishiguro, 1997] H. Ishiguro, Distributed vision system: A perceptual information infrastructure for robot navigation, Proc. IJCAI, pp. 36-41, 1997.
- [Ishiguro, 1998] H. Ishiguro, Development of low-cost and compact omnidirectional vision sensors and their applications, Proc. Int. Conf. Information systems, analysis and synthesis, pp. 433-439, 1998.
- [Kato, 1999] K. Kato, H. Ishiguro and M. Barth, Identifying and localizing robots in a multi-robot system environment, Proc. IROS, 1999.
- [Kitamura, 1990] Y. Kitamura and M. Yachida, Three-dimensional data acquisition by trinocular vision, Advanced Robotics, Vol.4, No.1, pp. 29-42, 1990.

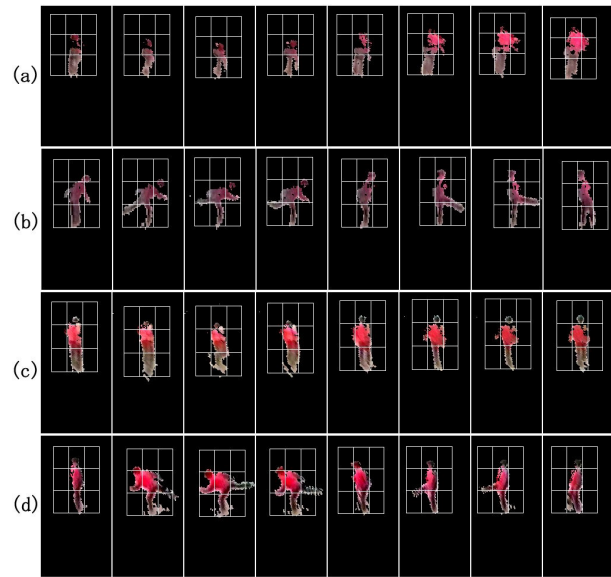


Fig. 16: An image sequence of "kick" gesture

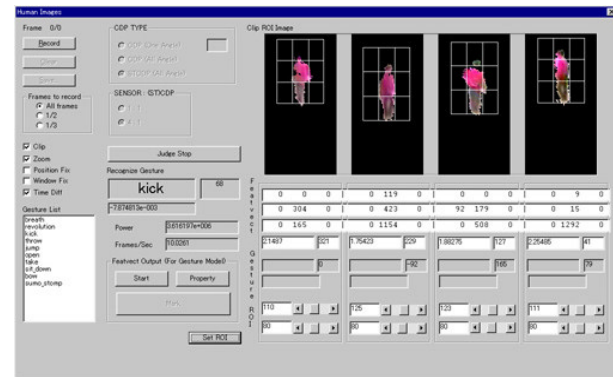


Fig. 17: Monitor output of the system

- [Koenderink, 1979] J.J. Koenderink and A.J. van Doorn, The internal representation of solid shape with respect to vision, Biological Cybernetics, 32:211-216, 1979
- [Marr, 1982] D. Marr, Vision: A computational investigation into the human representation and processing of visual information, Freeman, W. H. and Company, 1982.
- [Pavlovic, 1997] V.I. Pavlovic, R. Sharma and T.S. Huang, Visual Interpretation of hand gestures for human-computer interaction: A review, IEEE Trans. PAMI, vol. 19, no. 7, pp. 677-695, 1997.
- [Sogo, 2000] T. Sogo, H. Ishiguro and M. Trivedi, Real-time target localization and tracking by N-ocular stereo, IEEE Workshop on Omni. Vision, pp. 153-160, 2000.
- [Sogo, 2001] T. Sogo, H. Ishiguro and M. Trivedi, N-ocular stereo for real-time human tracking, In R. Benosman and S. B. Kang eds., Panoramic Vision: Sensors, Theory and Applications, Springer-Verlag, Berlin, 2001. (to appear).