13—15

Human Tracking Based on Hierarchical Attention Control

Junji SATAKE Takeshi SHAKUNAGA*

Department of Information Technology
Okayama University

Abstract

A hierarchical attention control model is introduced for implementing a human-like attention control in computer vision. The lowest layer of the hierarchy is called a window layer, in which each attention window is controlled by signal processing in the same manner as in the conventional attention control. The other four layers, however, are designed for implementing an attention control with high-level information processings. In our model, the attention control is accomplished using a tree structure across the five layers and both the bottom-up and top-down processings are accomplished on it. This paper provides its application to human tracking in complex situations with successful experimental results.

1 Introduction

An attention control is one of the most important problems in both human and computer visions[1]. In human vision, two different mechanisms are implemented. In the neighborhood of fovea, a focussing point is controlled by eye movement and intensive processing is made using fine visual information such as color and shape. In the other wide regions, an extensive processing is made using coarse visual information such as motion. The control of focussing point is often called an attention control, and it has been one of key subjects in psychology for a long time. Although most of work has discussed the attention control only on the image plane, Ken Nakayama of Harvard University recently proposed a new interpretation of human attention in a close relation with high-level concepts. This idea sounds reasonable because attention control has more logical aspect in human brain.

The motivation of our research is based on his interpretation, and our goal is to implement such a humanlike mechanism in computer vision. In this paper, a hierarchical attention control model is introduced for the purpose. This approach is different from the conventional computer vision [2, 3].

2 Attention Control

2.1 Hierarchical Attention Control Model

We propose a hierarchical model of attention control which consists of five layers, as shown in Fig.1. The lowest layer is called a window layer, in which each attention window is controlled by signal processing in the same manner as in the conventional attention control. The other four layers, however, are designed for high-level control, of which the top layer is used only for the mission control. In our model, the attention control is accomplished using the tree structure across the five layers.

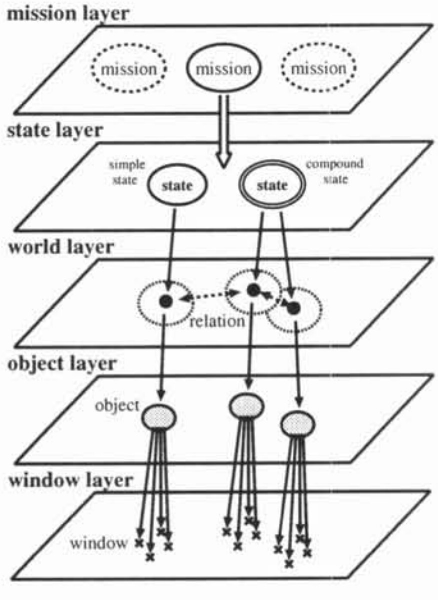


Figure 1: Hierarchical attention control model

2.2 Consitution and Role of Each Layer

Five layers in the hierarchical model process different information to realize the attention control. The constitutions and roles of five layers are summarized as follows:

1) window layer: The lowest layer is called a window layer, in which each attention window is controlled by signal processing in the same manner as in the conventional attention control. Although human

*Address: Tsushima naka 3-1-1, Okayama 700-8530, Japan. E-Mail:{satake,shaku}@chino.it.okayama-u.ac.jp vision treats only one attention due to its hardware implementation, our model treats multiple windows in the same time because no such restriction exists for the implementation. A priority is attached to each window in order to realize a flexible attention control.

2) object layer: The object layer deals with information on each object. An object node is generated when a remarkable translation is observed in the window layer. An object has a set of pointers to multiple windows. The object layer controls a process of each window according to the state of the object.

3) world layer: The world layer deals with information of a relationship among objects in the physical space and in the feature space. For example, when an object goes into the neighborhood of another, a signal is generated in the world layer and a state transition is invoked in the state layer.

4) state layer: In the state layer, a state of each object is controlled. There are two kind of states, called simple states and compound states. A simple state is bound with an object and shows its property. On the other hand, a compound state is bound with multiple objects and show its relationship.

5) mission layer: The mission layer controls the whole system. When the input image and the purpose of process are given, the mission layer controls the other four layers. In this paper, the mission is fixed to the human tracking from now on.

3 Human Tracking Based on Hierarchical Attention Control

3.1 Outline of Human Tracking

Let us pick up the human tracking as an example problem of conputer vision. This section shows how to implement the human tracking in the frame work of the hierarchical attention control.

Multiple windows are generated on the image plane and processed in the window layer. A priority is attached to each window, which shows the importance for the tracking. The priority is controlled in the object node based on the bottom-up and top-down information.

3.2 Human Tracking as a Set of Multiple Windows

Because the tracking with each window is not enough reliable, a human is tracked using a set of multiple windows. When some of the windows fail in tracking, they are detected and corrected in the object layer.

Let us calculate the average location and the variances σ_x^2 and σ_y^2 over the set of windows. A human region is defineded by a rectangle, of which the center locates at the average, and the horizontal and vertical lengthes are $2\lambda\sigma_x$ and $2\lambda\sigma_y$, respectively. When a window is detected out of the human region, the window is considered to be in failure and removed from the set.

When a distance between two windows is smaller than a threshold, one of them with a lower priority is deleted. In the current implementation, the threshold is a half of the side length of window.

3.3 States and State Transition

The states of each human are classified into four states, Solitude, Approaching, Distinguishable and Indistinguishable. The classification is based on distances in the physical and feature spaces, in the world layer. The states are first classified into three based on the minimum distance from the other objects in the physical space. Then, the last state is classified into two states by the minimum distance in the feature space. The state transition diagram is defined as shown in Fig.2.

1) Solitude: No object exists within 5σ in the physical space.

2) Approaching: No object exists within 3σ in the physical space and one or more objects exist within 5σ.

3) Rendzvous: One or more objects exist within 3σ in the physical space.

3a) Distinguishable: A Fisher's criterion is larger than a threshold.

3b) Indistinguishable: The Fisher's criterion is smaller than a threshold.

3.4 Attention Control in Each State

1) Solitude: The priority of the window is set according to the distance between the human region and the background in the feature space. When a window is detected out of the human region, the window is considered to be in failure and removed from the set. The number of windows is controlled to be nearly constant during the tracking. To keep the number, new windows are often generated using

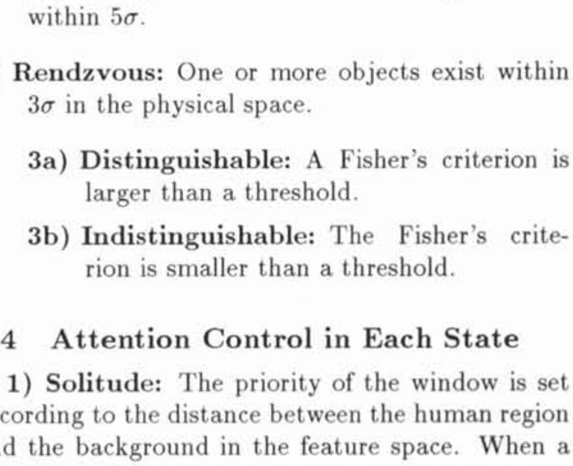


Figure 2: State transition diagram on human tracking

the difference between the input and background images.

2) Approaching: The priority of the window is set according to the distance between the target human region and the other human region in the feature space. New windows are not generated in the overlapped region, because they often cause a wiondow generation in the wrong person.

3) Distinguishable: The priority of the window is set according to the distance between the target human region and the other human region in the feature space. Windows with a low priority are not used for tracking.

4) Indistinguishable: The positions of the target and the other are expected using the temporal velocities. If the velocities can not be estimated, the target and the other are tracked together.

3.5 Human Tracking Based on Hierarchical Attention Control

The state of each object is decided in the bottom-up process as shown in Fig.3(a). Each window is tracked in the window layer. In the object layer, the position of the object is calculated from the positions of multiple windows. In the world layer, relationships are managed among objects both in the physical and feature spaces. In the state layer, the state of each object is decided using the distance among objects.

The attention is controlled accoding to the state of the object in the top-down process as shown in Fig.3(b). The state is informed to the object layer from the state layer. In the object layer, the tracking process is controlled according to the state as discussed in 3.4. Thus, the flexible attention control is realized by the combination of the top-down and bottom-up processes.

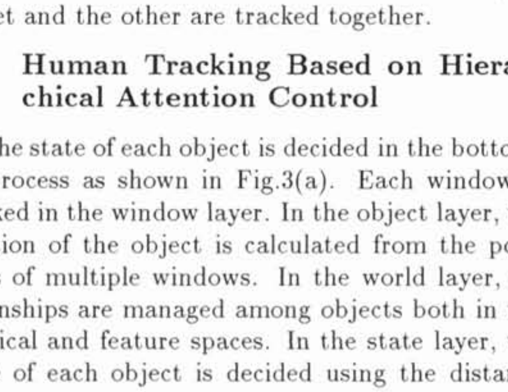


Figure 3: (a)Bottom-up and (b)Top-down processes

4 Experimental Results

This section shows results of human tracking in the real environments. Each human region is drawn in a solid line or a broken line according to the state as shown in Tab.1.

Table 1: Human regin

State	5σ	3σ
Solitude	solid line	broken line
Approaching	solid line	solid line
Distinguishable	broken line	solid line
Indistinguishable	broken line	broken line

4.1 Outdoor Scene

Figure 5 shows a tracking result for an outdoor scene. Only two human region are drawn in the figure. First, when the target is far from the other, the state is Solitude(Fig.5(a)). The priorities of windows are set according to the distances between the human region and the background in the feature space. When the target approaches to another, the state is transferred to Approaching(Fig.5(b)). The priorities of the windowes are set according to the distances between the target human region and the other human region in the feature space. The new windows are not generated in the overlapped region. When target object cross the other, the state is transferred to Distinguishable(Fig.5(c)). Windows with a low priority are not used for tracking. Finally, the distance between objects becomes larger, the state is transferred to Approaching and Solitude again (Fig.5(e),(f)).

Figure 4 shows a tracking result when only the window and object layers are effective. This resulted in failure because no flexible control is realized without using the world and state layers.

4.2 Indoor Scene

Figure 6 shows a tracking result for an indoor scene. When the state is Approaching or Distinguishable, the priorities of the window is set according to the distances between the target human region and the other human region in the feature space using the Fisher's method. When the distinguishable regions are hidden, the state is transferred to Indistinguishable(Fig.6(c)). When the distinguishable regions appear again, the state is transferred in Distinguishable(Fig.6(d)).

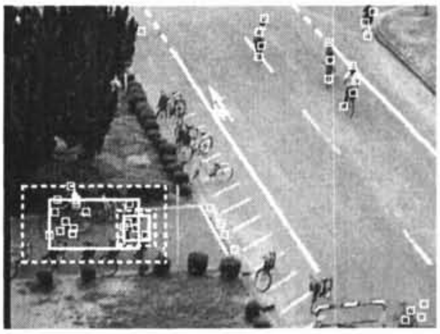


Figure 4: Failure of tracking

5 Conclusions

A hierarchical attention control model is introduced for implementing a human-like attention control in computer vision. This model effectively works for the human tracking as shown 4. Future work includes applications to other problems in computer vision.

Acknowledgment

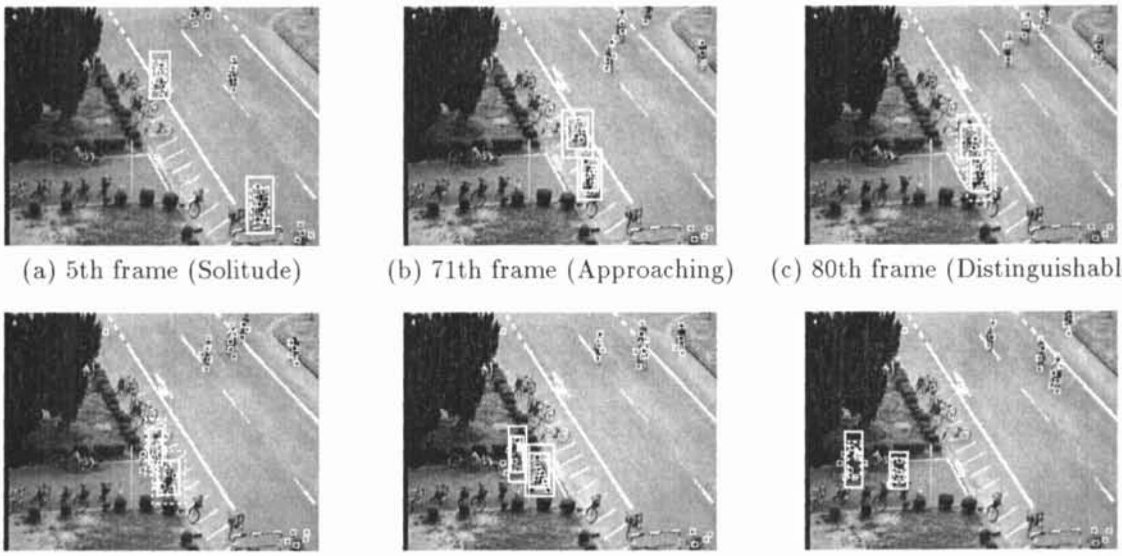
A part of this work was supported by Grant-In-Aid for Science Research under Grant No.10480074 from the Ministry of Education, Science, Sports, and Culture, Japanese Government.

References

[1] David Marr: Vision: A computational Investigation into the Human Representation and Processing of Visual Information, Freeman, W.H. and Company, New York, 1982.

[2] B.K.P.Horn: Robot Vision, The MIT Press, McGraw-Hill Book Company, 1986.

[3] Rosenfeld,A., A.C.Kak: Digital Picture Processing, Vols. 1&2, Second Edition, Academic Press, New York, 1982.



(a) 5th frame (Solitude)

(b) 71th frame (Approaching)

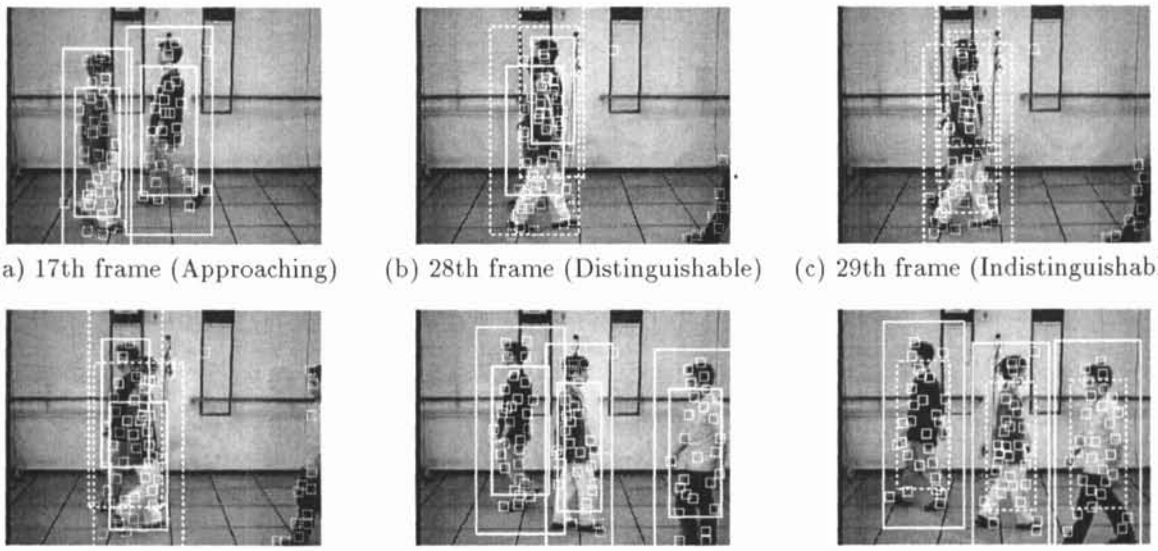
(c) 80th frame (Distinguishable)

(d) 91th frame (Distinguishable)

(e) 114th frame (Approaching)

(f) 150th frame (Solitude)

Figure 5: Results of tracking (outdoor scene)



(a) 17th frame (Approaching)

(b) 28th frame (Distinguishable)

(c) 29th frame (Indistinguishable)

(d) 32th frame (Distinguishable)

(e) 40th frame (Approaching)

(f) 46th frame (Solitude)

Figure 6: Results of tracking (indoor scene)