MVA'94 IAPR Workshop on Machine Vision Applications Dec. 13-15, 1994, Kawasaki

SHAPE IDENTIFICATION USING PROBABILISTIC MODELS OF ATTENTIONAL SEQUENCES [1]

H. Işıl Bozma[†] and Çağatay Soyer[‡]

[†] Department of Electrical and Electronic Engineering
Boğaziçi University
Bebek, Istanbul 80815, Turkey

[‡] Institute of Biomedical Engineering
Boğaziçi University
Istanbul 80815, Turkey

ABSTRACT

This paper presents a system which demonstrates a biologically motivated approach to artificial vision. As a typical application 2-D shape recognition is chosen. Traditionally recognition problem is divided into two phases: sophisticated preprocessing of the whole image so that regions with similar features can be grouped together and then the recovery of shape parameters of each region using one among the plethora of existing methods. In contrast an approach based on selective attention makes use of simple methods to guide visual processors to the interesting regions of the image. Previous work on selective attention has focused on pre-attentive and attentive stages while almost no work has been done on the final stage - cognition. This paper presents the beginnings of how to complete the pre-attention, attention, cognition sequel of selective attention. An approach to shape recognition based on probabilistic models of attentional sequences is described and the capabilities of the system are investigated. Experiments performed on rotated shapes demonstrate the performance of the algorithm.

1. INTRODUCTION

Studies on the vertebrate visual system provide us with many clues about how recognition can be so fast when the neurons - the major processing elements - are rather slow [1, 2, 3, 4, 5]. It has been determined that in addition to being massively parallel, biological vision systems also selectively process the incoming data in order to reduce the amount of processing. In particular, selective attention, defined as the intelligent use of limited resources, has been advocated to be a biologically verified means of alleviating the computational requirements of visual processing [3, 4, 5]. In this perspective, visual processing consists of three stages: pre-attention, attention and cognition. Previous work on eye movements has motivated the representation of an object as a feature ring - a temporally ordered sequence of fixations with their corresponding features and spatial relations. It is concluded that recognition involves temporal matching of a feature ring elicited from the scene to those stored in memory [3, 4, 5]. However, most of the studies in selective attention have primarily focused on the first two stages while recognition based on attentional sequences has remained relatively unexamined [8, 6, 7, 9]. In this paper,we examine the performance of an approach based on probabilistic models of attentional sequences, within the context of shape identification. In this work, recognition is achieved by modelling the attentional sequences generated on library of shapes using hidden Markov models (HMMs), and then finding the observation probabilities of a generated sequence by these models. The sequence is classified to be the library shape whose HMM generates the maximum observation probability (MOP).

The organization of this paper is as follows: Section 2 presents the generation of attentional sequences. The modelling of these sequences using HMMs are explained in the next section including a brief overview of hidden Markov models. In Section 4 shape identification using attentional sequences is explained. The performance of the system on real images is evaluated in the Experiments section. We conclude with a brief summary.

2. SEQUENCE GENERATION: PRE-ATTENTION & ATTENTION

An emergent attentional sequence consists of a set of ordered foveas which trace, via attention and fixation, the most informative parts of a scene. The visual element which is a part of selective attention hardware is modelled as a spatially variant sensor like the vertebrate retina [1]. The visual acuity of the sensor is determined by a 2-D gaussian centered at the optical axis or the fixation point. For simplicity we approximate the gaussian by a square function. During each fixation, the attentive stage consists of detailed processing of the fovea to find features which will be added to the attentional sequence. In the pre-attentive stage simpler analysis is performed on the candidate foveas in the visual field, to find the next fixation point. Candidate foveas are allowed to overlap by a certain amount so that, salient image points can be traced with controllable accuracy. In order not to examine the same area twice, an region of pre-defined size around the fixation point is inhibited during pre-

[1]This work was supported by Boğaziçi University Research Fund grant #93HA0230 and partially by TUBİTAK MISAG grant #24-1992.

attentive processing. As a result pre-attentive vision is performed on a region around the fovea, which is called the periphery. Note that fovea and periphery together make up the visual field.

Let us suppose that a moving camera starts with an arbitrary orientation and consequently an arbitrary optical axis. The fovea is a small region around the fixation point, and the visual field contains the fovea at its center. The camera then fixates to the center of the most salient of overlapping foveal regions inthe periphery. This fixation is done by moving the camera's optical axis to the new fixation point which determines the new fovea and the new periphery. Saliency is a measure of interest based on presence of simple features and should be computationally very cheap. For example if the fovea image is represented by $I_1(x, y)$ and the intensity gradient is $\nabla I_1 = [I_{1_x} I_{1_y}]$, then saliency may be computed as the sum of linearly weighted combination of the responses $c_1 \mid I_{1_x} \mid + c_2 \mid I_{1_y} \mid$ of each point in the fovea. The sequence of fixations with the given measure of saliency and corresponding visual fields are shown in fig. 1. Using different measures of saliency the fixation

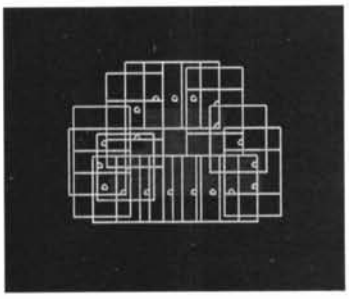

Fig. 1. Top: Image and fixations; Bottom: Feature sequence.

sequence can be biased to exhibit desired properties depending on the task or type of image.

In the attentive stage, the fixation fovea is further processed to extract features which characterize this fovea. The feature vector obtained in this way is assigned a state number and thenadded to the attentional state sequence. Thus, each feature vector corresponds to a state. Using the measure of saliency presented above and defining edgetype as a feature, the state sequence consisting of 1-D vectors is obtained as in fig. 1. This sequence comprises of four different types of states representing four directional edges.

The pre-attention and attention process recursively continues tracing all the interesting points in the image. The termination of fixations may or may not be considered, since recognition can be simultaneously occuring during sequence generation. However, when no salient fovea can be found in the current periphery, the system starts fixating on arbitrary points in the periphery. This

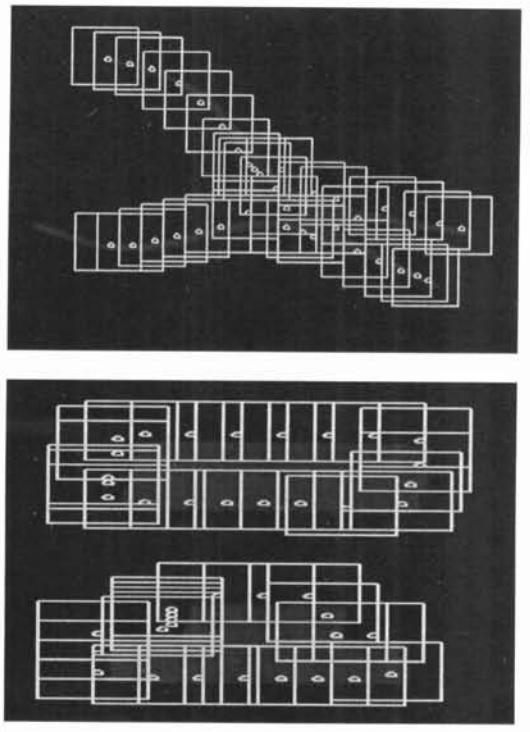

Fig. 2. Left: Fixations on tools and two object image.

can be prevented by using a saliency threshold forsequence termination. Alternatively, by keeping track of foveas from previous peripheries, fixations can be made on unexamined regions. Example fixations on relatively complex scenes using such memory and saliency threshold are shown in fig. 2.

3. ATTENTIONAL SEQUENCE MODELS

In order to model an attentional sequence probabilistically so that it can be recalled at a later time, a hidden Markov model is used [10, 11, 6]. In the next part, we will briefly review HMMs.

If k is an index to the 2d shape, its HMM is characterized by the 3-tuple (A_k, B_k, π_k) as:

1. N, the number of states in the model. Individual states are denoted by $S = \{S_1, S_2, ..., S_N\}$, and the state at time t by q_t.

2. M, the number of distinct observation symbols per state, corresponding to the observed physical output of the process at each state. Output symbols are denoted by $O = \{O_1, O_2, ..., O_M\}$ and the observed output at time t by o_t.

3. State transition probability matrix $A_k = \{a_{ij}^k\}$ where

$$a_{ij}^k = P(q_{t+1} = S_j | q_t = S_i, k), \qquad 1 <= i, j <= N$$

4. The observation symbol probability distribution in state j,

$$B_k = \{b_j^k(l)\}, \qquad \text{where}$$

$$b_j(l) = P(o_t = O_l | q_t = S_j), \ 1 \leq j \leq N, \ 1 \leq l \leq M, \ \forall t$$

5. The initial state distribution $\pi_k = \{\pi_i^k\}$ where

$$\pi_i^k = P(q_1 = S_i), \qquad 1 \leq i \leq N$$

In our approach, letting N_F be the number of foveas in an attentional sequence, each fovea in the sequence corresponds to an observation o_t, $t = 1, \ldots, N_F$. The value of each observation corresponds to the extracted features O_i, $i = 1, \ldots, M$. Each observation has a corresponding state s_t, where the value of the state is S_j, $j = 1, \ldots, N$ also corresponds to the extracted feature.

4. CLASSIFICATION USING HMMs

In order to use HMMs in the recognition of 2-D shapes, a library model is obtained for the attentional sequence generated on each library shape. These models are calculated by considering the number of transitions between states during fixations on the library shape. Thus, L library objects are represented by the transition probability matrices A_l, $(l = 1, \ldots, L)$.

During classification, an emergent attentional sequence $O = \{o_1, \ldots, o_N\}$ is generated on the object to be classified. The observation probability of this sequence by the model l is given by,

$$\begin{aligned} P(O|l) &= P(o_1)P(o_2|o_1, l).....P(o_n|o_{n-1}, l) \\ &= P(o_1)\prod_{i=1}^{N-1} P(o_{i+1}|o_i, l) \end{aligned}$$

To find the maximum observation probability, $P(0|l)$ is maximized over all library models,

$$MOP = \max_{l=1,\ldots,L} P(O|l)$$

The output class is the one whose library model gives the MOP for the input sequence.

5. EXPERIMENTS

The above approach is used in a system implemented for recognition of 2-D shapes. Experiments are performed on 640x480 pixels 8 bit gray scale images of various shapes. Library models are generated from images of shapes without any rotation and rotated shapes are used to test the performance of the system. These images are shown in fig. 3 through fig. 8.

Fig. 3. Library images: rec1,t1,el1.

Fig. 4. Experiment set: rec2,t2,el2.

Fig. 5. Experiment set: rec3,t3,el3.

Fig. 6. Experiment set: rec4,t4,el4.

Fig. 7. Experiment set: rec5,t5,el5.

Fig. 8. Experiment set: rec6,t6,el6.

Experimental configuration consists of a 60x60 visual field and 10x10 fovea. Note that, in hardware implementation the size of the visual field is physically defined by the focal length of the camera lens. Even with a small angle lens, the visual field is too big for many applications. Therefore, we propose using a subset of the image seen by the camera. In this simulation, the camera is assumed to see only the 60x60 region of the world which is 640x480. Saliency measure is a simple gradient which returns the edge contents of the overlapping foveas in the periphery. The amount of overlap is 50%. The size of the attentional sequence is limited to 20 fixations, which is a reasonable value for these images. In the attentive stage each fovea is processed to obtain its features and find its corresponding states. We experimented with different features and feature vectors like edge strength, edge type, corner type and average intensity. Among these the best results are obtained by using only the edge type as the state determining feature. Therefore we had four different types of states in attentional sequences.

From no-rotation images of three shapes, library models are obtained. Transition probability matrices for these HMMs are shown in fig. 9, where S_0 through S_1 represent four edgetypes.

Using these models and the above parameters, rotated shapes are classified correctly with satisfactory values except for t6, where a fixation size of 20 resulted in more than few empty fixations. Observations probabilities for these experiments are tabulated below. Here L_0, L_1, and L_2 represent three library models and image suffixes indicate different amounts of rotation. Values of observation probabilities are scaled by 10^{10} for convenience.

	S_0	S_1	S_2	S_3
S_0	.636	.090	.181	.090
S_1	.100	.600	.100	.200
S_2	.142	.285	.428	.142
S_3	.285	.142	.142	.428

	S_0	S_1	S_2	S_3
S_0	.400	.100	.100	.400
S_1	.100	.600	.200	.100
S_2	.418	.142	.285	.142
S_3	.250	.250	.250	.250

	S_0	S_1	S_2	S_3
S_0	.250	.125	.375	.250
S_1	.111	.555	.111	.222
S_2	.400	.100	.400	.100
S_3	.125	.250	.125	.500

Fig. 9. Transition probability matrices A_0, A_1 and A_2 for library objects rec1,t1, and el1.

	$A_0(rec1)$	$A_1(t1)$	$A_2(el1)$
t2	7	75	3
el2	5	40	55
rec2	5171	6	5
t3	3	19	6
el3	0	2	37
rec3	829	1	2
t4	1	82	14
el4	0	15	15
rec4	435	38	72
t5	3	26	3
el5	0	8	186
rec5	0	8	186
t6	4	16	22
el6	34	1	215
rec6	1161	2	2

Fig. 10. OP values for various objects.

From these results we conclude the following: The measure of saliency determines the fixation sequence. This property can be used to guide the camera or robot to the desired points of the world like edges in our experiments. The features used in formulating the states are another key component in the algorithm. These features must have the necessary discrimination power for the given task. For example, when we perform the above experiments using edge strength as our state determining feature, the results are not as satisfactory due to the almost equal values of edge strength in fixation foveas.

Other parameters affecting the system performance are the sizes of visual field and fovea, the amount of overlap, and inhibition area around fixation point. These all primarily determine the resolution and accuracy of fixations, and need to be adjusted appropriately for a good sequence with minumum redundancy. For example, a small fovea in the above experiments may lead to fixations concentrating on a corner and finally locking themselves among already fixated foveas. Unless we use some kind of memory to remember saliencies in previous peripheries, this situation is unrecoverable.

6. CONCLUSION

We presented preliminary results on the use of selective attention in shape identification. In this work, recognition is achieved by modelling the attentional sequences generated on library shapes using HMMs, and then finding the observation probabilities of a generated sequence by these models. Considering the experimental results and the strong biological evidence behind it, we find this approach promising. The contribution of our work has been the investigation of the use of selective attention in visual recognition. Our future work in this field, will concentrate on different models and recognition schemes of attentional sequences, as well as fixation control and feature selection. A hardware implementation is also being set up to make use of this approach in mobile robot guidance.

REFERENCES

[1] P. Gouras and C.H.Bailey. The retina and phototransduction. In J.H. Schwartz and E.R.Kandel, editors, *Principles of Neural Science*. Elsevier, 1986.

[2] J.P. Kelly. Anatomy of the central visual pathways. In J.H. Schwartz and E.R.Kandel, editors, *Principles of Neural Science*. Elsevier, 1986.

[3] D.H. Ballard. Animate vision. *Artificial Intelligence*, (48):57–86, 1991.

[4] D. Noton and L.Stark. Eye movements and visual perception. *Scientific American*, 224(6), 1971.

[5] L. Stark and S.R.Ellis. Scanpaths revisited: Cognitive models direct active looking. In Monty Fisher and Senders, editors, *Eye Movements: Cognition and Visual Perception*, pages 193–226. Erlbaum, NJ, 1981.

[6] R.D. Rimey and C.M.Brown. Selective attention as sequential behavior: Modelling eye movements with an augmented hidden markov model. Technical report, The University of Rochester, Computer Science Department, February 1990.

[7] D.H. Ballard and C.M.Brown. Principles of animate vision. *CVIP: Image Understanding*, 56(1), July 1992.

[8] A.L. Abbott. A survey of selective fixation control for machine vision. *IEEE Control Systems*, pages 25–31, August 1992.

[9] J.J. Clark and N.J.Ferrier. Modal control of an attentive vision system. In *Proceedings of 2nd International Conference on Computer Vision*. IEEE, 1988.

[10] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), February 1989.

[11] Y. He and A.Kundu. 2-d shape classification using hidden markov model. *IEEE Tansactions On Pattern Analysis and Machine Intelligence*, November 1991.