

Visual Perception and Sequences of Eye Movement Fixations: A Stochastic Modeling Approach

Selim S. Hacisalihzade, *Senior Member, IEEE*, Lawrence W. Stark, *Fellow, IEEE*, and John S. Allen

Abstract—Sequences of visual fixations, while looking at an object, are modeled as Markov processes and statistical properties of such processes are derived by means of simulations. The sequences are also abstracted as character strings and a quantitative method of measuring their similarity, based on minimum string editing cost (actually dissimilarity distance) is introduced. Interrelationships between the structure and size of the generating Markov matrices and the string editing distance shed light on the relative roles of deterministic and probabilistic processes in producing human visual scanpaths.

I. INTRODUCTION

EYE MOVEMENTS are necessary for vision while looking at an object that spans more than several degrees in the subject's field of view, because detailed visual information can only be obtained through the fovea, the small (about one degree) central area of the retina that has the highest photoreceptor concentration. Therefore, the brain directs the eye to move in such a way as to foveate successively onto the points of interest [15]. While viewing a stationary object, the eyes alternate between fixations and saccades, very rapid eye movements. Each saccade leads to a new fixation. Typically there are about three saccades per second, but since saccades are so fast, they occupy only about 10% of the total viewing time. Vision is suppressed during the saccades and thus almost all the visual information is collected during the fixations [10]. Clearly, eye movements play a very important role in visual perception. It has been found about two decades ago that people have repetitive and idiosyncratic ways of inspecting and recognizing a particular familiar object; these patterns were named scanpaths [7].

Storing and retrieving memories are important components of visual learning and recognition. Therefore, the memory system of the brain must contain an internal representation of every object that is to be recognized. Thus, familiarizing oneself with an object may be considered as the process of constructing this representation. Similarly, recognition of an object may be viewed as the process of matching it with its stored internal representation. A non-Gestalt view suggests that the internal representation is made of components and that during recognition the features of the model are matched step

by step with the object. In support of a serial process, it is known that the eyes seem to visit the features of the object under study cyclically, following somewhat regular scanpaths rather than crisscrossing it at random. A serial model of internal representations of objects based on this evidence is the so called "scanpath feature ring" [8]. This model maintains that the representation of objects are composed of sensory memory traces recording the features and motor memory traces of the eye movements from one feature to the other. A modified and more realistic version of this model introduces randomness into the generation of scanpaths.

Markov matrices were used by Stark and Ellis [11] in earlier studies in an attempt to go beyond visual inspection of the eye movement traces and application of a subjective test for similarity of such traces. They used Markov matrices identified from experimental sequences and showed the existence of a few structured processes. They also looked at structures beyond the first order Markov matrix, i.e., do the states $n-1$, $n-2, \dots$ previous to the present state n affect the probability of transition to the next state $n+1$? Unfortunately, the size of the higher order Markov matrices increases geometrically with the order. Thus very large experimental sequences would have been necessary for the identification of these matrices. For a nonstationary generator, as the actively looking human, this poses obvious experimental difficulties. Can one further quantify the similarity of sequences of visual fixations?

String editing is also a possible way to study the similarity of sequences by looking at the similarity of the corresponding strings. A reasonable way of doing this is to define a distance between strings that characterize sequences of visual fixations and to set a threshold below which strings, thus sequences of visual fixations, are similar. The question about the similarity of strings is a problem that has been occupying computer scientists [13] as well as biologists studying RNA and DNA sequences [3].

Aim of this paper is to demonstrate the feasibility of quantifying the similarity of visual fixation sequences while looking at familiar objects.

II. A MARKOVIAN MODEL OF SEQUENCES OF VISUAL FIXATIONS

Let us now divide the image under study (Fig. 1) into several regions of interest and label them with letters like the hand (*A*), the mouth (*B*), the nose (*C*), the left eye (*D*), the right eye (*E*), the neck (*F*), and the ear (*G*). If we call these regions states into which the fixations must be located and postulate that the transitions from one state to another have

Manuscript received November 16, 1990; revised August 22, 1991. This work was supported in part by the Swiss Academy of Medical Sciences and in part by the Swiss National Science Foundation grant 5.521.330.615/7.

S. S. Hacisalihzade is with Landis & Gyr, Corporate Research and Development, CH-6301 Zug, Switzerland.

L. W. Stark is with the University of California, Berkeley, CA 94720.

J. S. Allen is with the Department of Anthropology, University of Auckland, New Zealand.

IEEE Log Number 9104550.

0018-9472/92\$03.00 © 1992 IEEE

certain probabilities, we can effectively describe the generating process for these sequences of fixations as Markov processes [5]. In particular, the sequence σ_A of fixations in Fig. 1 are BBFFAAAABCEDCG. We can see that this sequence could have been generated by the Markov matrix M_G (note the terminal fixation on the ear):

$$M_G = \begin{bmatrix} 0.75 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0 & 0 & 0.33 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The same matrix could generate many other sequences as other realizations of the generating process, for instance $\sigma_B = \text{BFFAAAABFABBCEDCEDCGGGG}$; note that lengths of the sequences do not need to be equal; also note that in this case a terminal string of G's is produced. However, when we calculate the transition probabilities that occurred in this second sequence σ_B and summarize these probabilities in another Markov matrix $\tilde{M}_B$ resulting from the second sequence, we get

$$\tilde{M}_B = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.67 & 0 & 0.33 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0.67 & 0 & 0 & 0 & 0 & 0.33 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\tilde{M}_B$ is an identification of the generating process characterized by M_G ; the sequential data in σ_B is summarized in $\tilde{M}_B$.

We now define

$$E = M_G - \tilde{M}_B$$

as the error or statistical discordance matrix between the idealized generating matrix M_G and the estimated or observed matrix $\tilde{M}_B$. Note: We have assumed M_G as the generating matrix and thus denied it is an $\tilde{M}_A$ even though we in fact obtained this illustrative case backward. E is clearly a function of the number of states and the lengths of the sequences as well as a function of the elements of the generating matrix, that is, the structure of the matrix. A possible scalar measure of the statistical discordance matrix E is the typical error of each element defined as

$$\bar{e} = \frac{1}{n^2} \sum_{i,j=1}^n |e_{ij}|$$

with e_{ij} as the elements of the error matrix E and n its dimension.

Let us now look at the typical errors resulting from several processes with different matrix sizes, sequences of different lengths and structures. The first process in consideration is a process in which a transition occurs from one state to the next with 90% and to the one after the next with 10% probability;

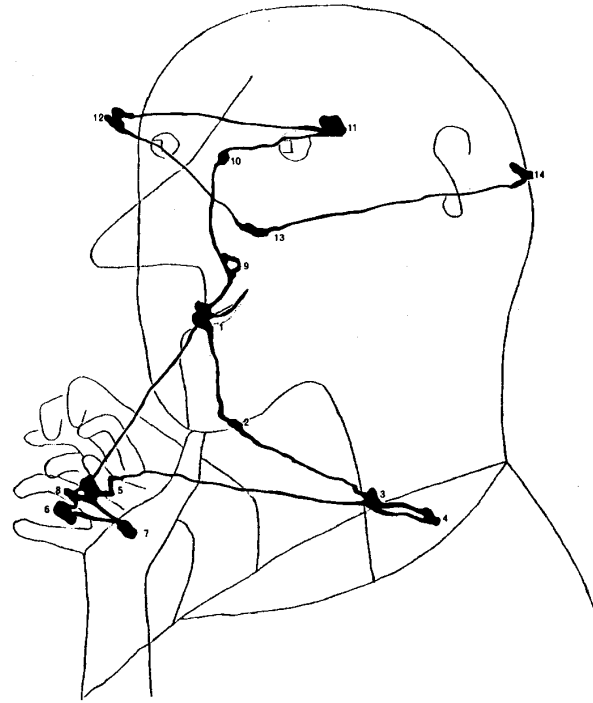


Fig. 1. Eye movements made by a subject viewing for the first time a drawing adapted from the Swiss artist Klee. Numbers show the order of the subject's visual fixations on the picture during free viewing. Lines between fixations represent rapid saccades from one fixation to the other.

no other transitions are possible. For $n = 4$ the corresponding Markov matrix is

$$M_1 = \begin{bmatrix} 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.9 & 0.1 \\ 0.1 & 0 & 0 & 0.9 \\ 0.9 & 0.1 & 0 & 0 \end{bmatrix}.$$

In the second process a transition occurs from one state to the next with 70% and to the one after the next with 30% probability; no other transitions are possible. For $n = 4$ the corresponding Markov matrix is

$$M_2 = \begin{bmatrix} 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 0.3 & 0.0 & 0.0 & 0.7 \\ 0.7 & 0.3 & 0.0 & 0.0 \end{bmatrix}.$$

In the third process a transition occurs from one state to the next with 50% and to the one after the next with 10% probability; all other transitions are equiprobable. For $n = 4$ the corresponding Markov matrix is

$$M_3 = \begin{bmatrix} 0.2 & 0.5 & 0.1 & 0.2 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.2 & 0.2 & 0.5 \\ 0.5 & 0.1 & 0.2 & 0.2 \end{bmatrix}.$$

In the fourth process all transitions occur equiprobably from one state to any other. For $n = 4$ the corresponding Markov

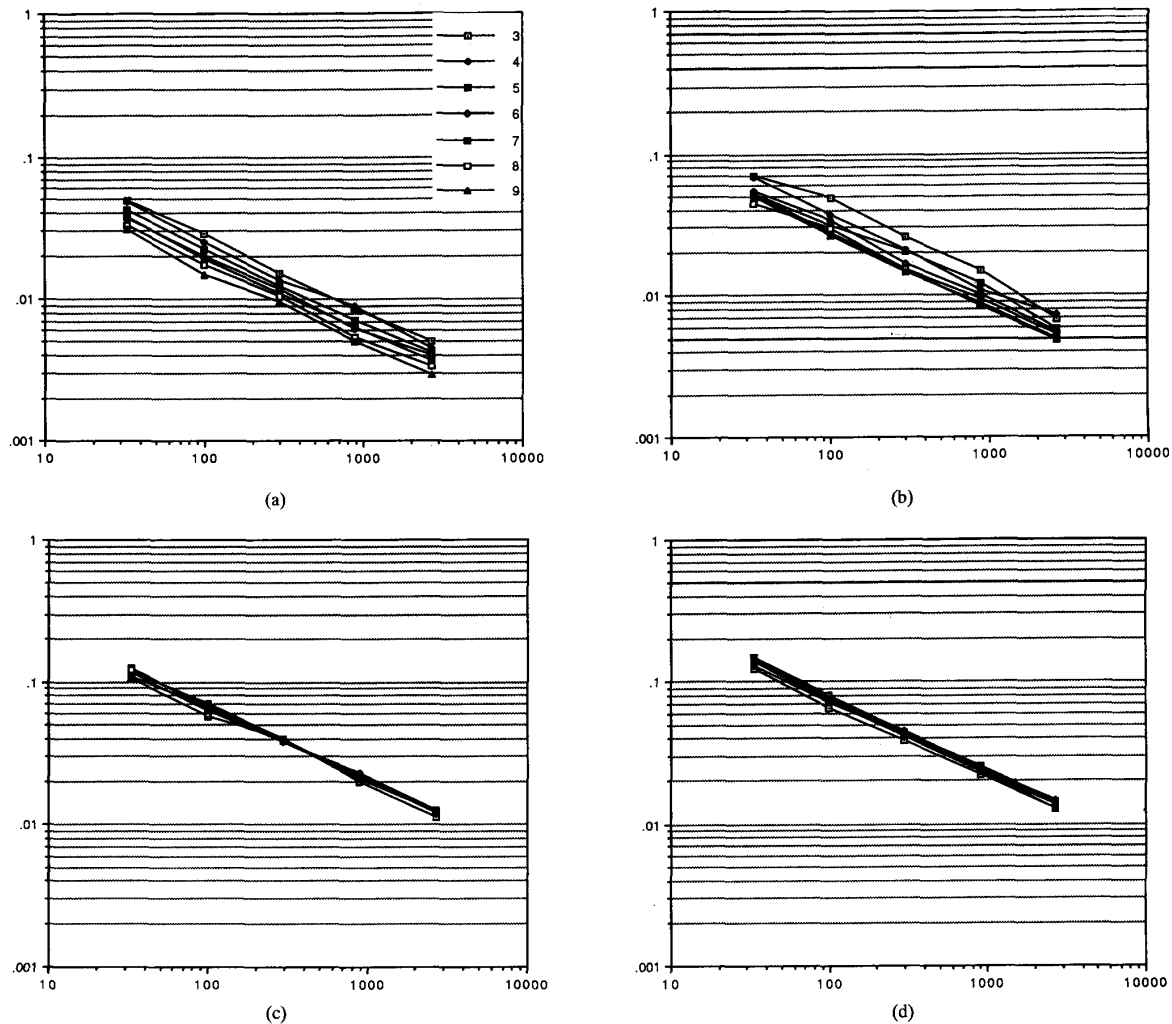


Fig. 2. Charts (a)–(d) show the typical size of the elements of the error matrices $E_1 \dots E_4$, which are the difference between the Markov matrices $M_1 \dots M_4$ (generating sequences) as defined in the text and the Markov matrices $\bar{M}_1 \dots \bar{M}_4$ (computed from the generated sequences) for $3 \dots 9$ states and $33 \dots 2673$ long sequences.

matrix is

$$M_4 = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

Typical errors for $M_1 \dots M_4$ with $n = 3 \dots 9$ and sequence lengths of $33 \dots 2763$ averaged over 30 simulations for each combination document a number of interesting relationships (see Fig. 2): The typical error gets smaller as a linear function of the string length in a double logarithmic scale. Typical error is also about ten times less for the quasideterministic processes characterized by M_1 and M_2 . Also, the number of states almost does not affect the typical error for a given string length. This is more true for the more random processes characterized by M_3 and M_4 .

Fig. 3 shows some simulated scanpaths generated by the matrices $M_1 \dots M_4$ superimposed on a drawing.

III. A STRING EDITING MODEL OF SEQUENCES OF VISUAL FIXATIONS

It is possible to define the distance between two strings of not necessarily the same length as the cost of editing one to get the other. Editing a string has three basic operations: substitution, deletion and insertion. One must first define the cost for each such operation.

For example, let us say that all substitutions are assigned a cost of 2 and both deletions and insertions a cost of 1. To transform the string ACA to CADAC one has to insert a C at the beginning (1) and a C at the end (1) and substitute the C in the middle with a D (2), resulting in a total cost, thus a distance, of 4. It is, of course, also possible to have more complicated cost assignments like setting the cost of substituting a letter with a letter following or preceding it in the alphabet as 1, the cost of substituting a letter with a letter following or preceding it in the alphabet by two letters as 2

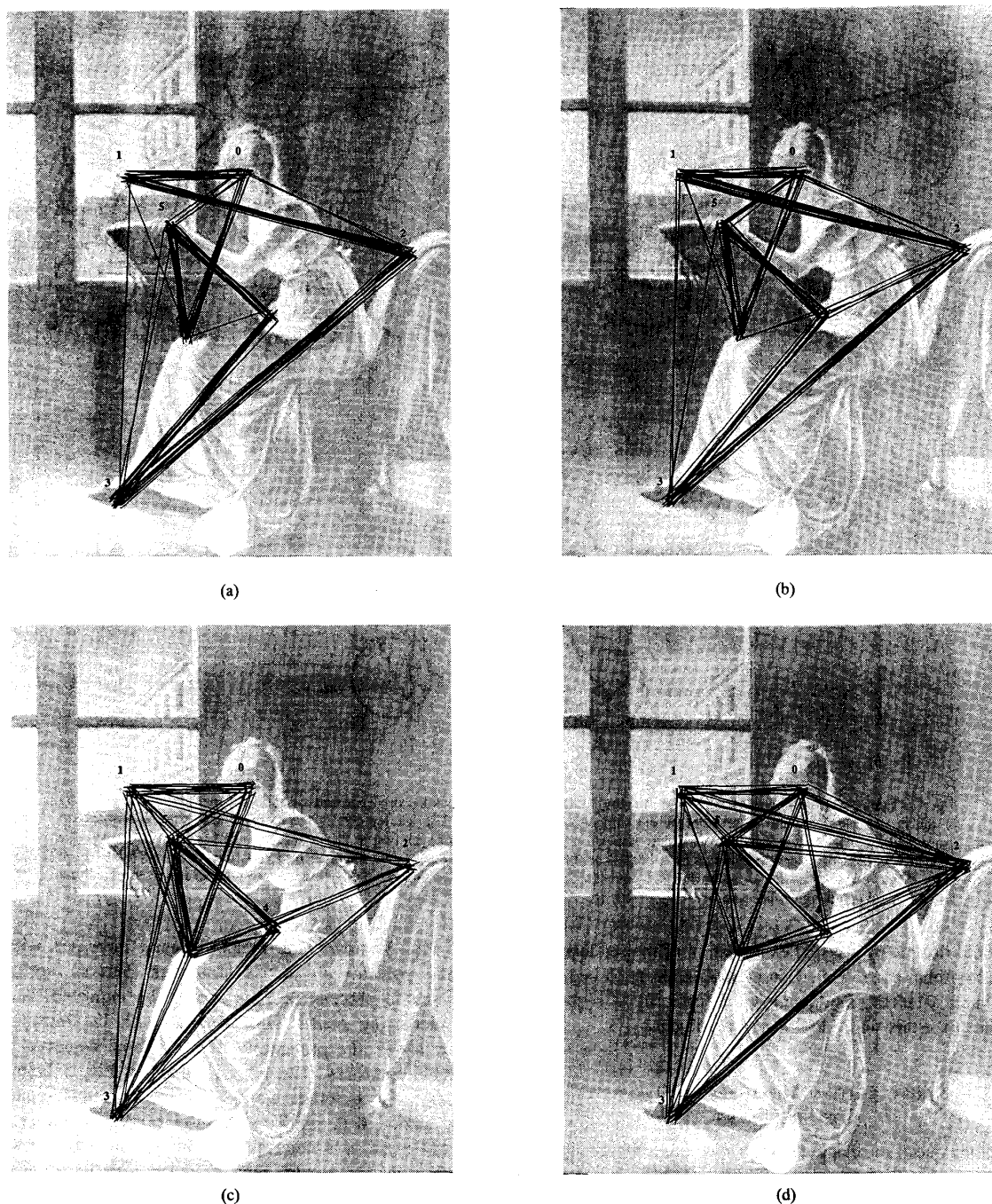


Fig. 3. (a)–(d) Simulated scanpaths on seven likely fixation points generated by matrices $M_1 \dots M_4$ superimposed on a line drawing adapted from a painting by Charpentier (figure provided by G. Tharp.)

and so on. As the strings get longer, the ways of transforming them increase very fast and it becomes no longer trivial to find the transformation that costs least. Therefore, an algorithm based on a modified dynamic programming that guarantees to find the minimum distance between two strings was developed exactly for that purpose [14].

When we apply this algorithm with the cost of substitution

as 1, of insertion as 2 and of deletion as 3 (these costs were found empirically) on sequences of visual fixations depicted in Fig. 4 we get the distance d between Figs. 4(a) and (b) as 10, between Figs. 4(a) and (c) as 15 and between 4(b) and (c) as 25. Therefore, the sequences in Figs. 4(a) and (b) are more similar to each other than the ones in Figs. 4(a) and (c) or the ones in Figs. 4(b) and 4(c). This result confirms what

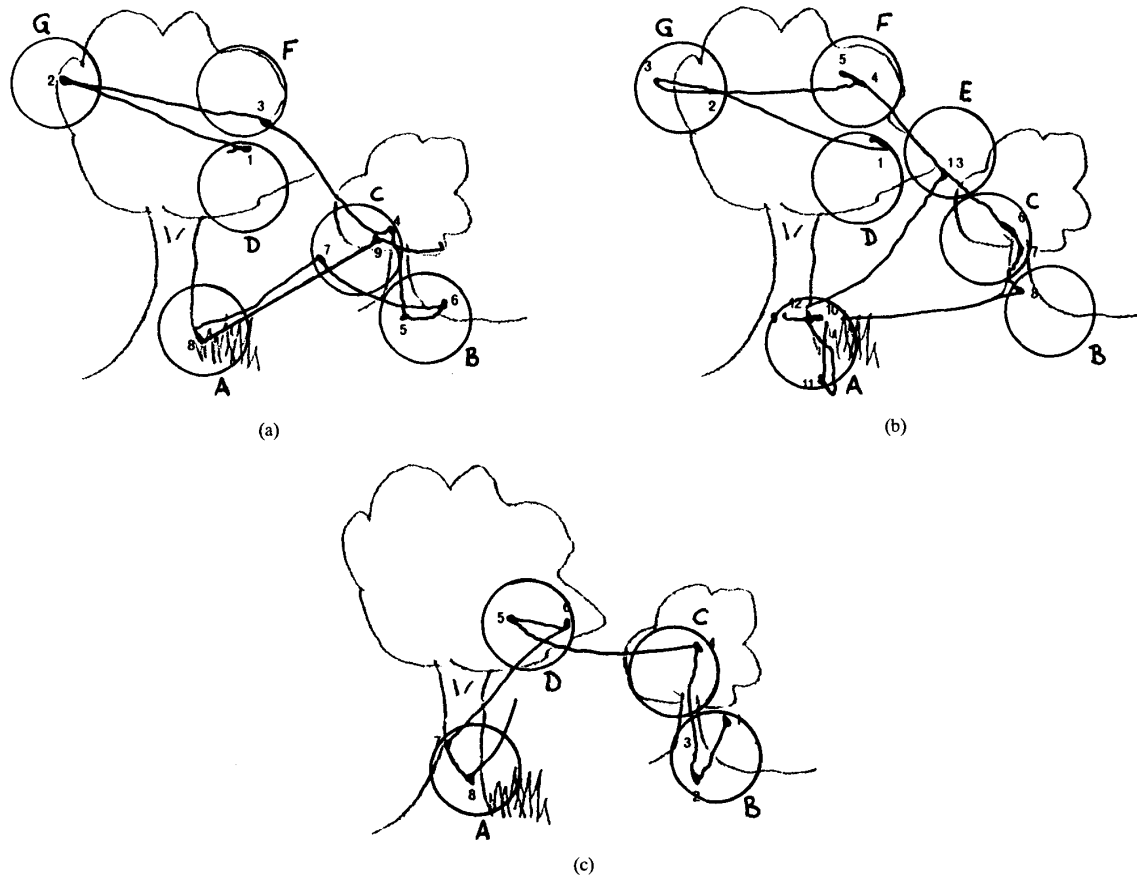


Fig. 4. Eye movements of a subject while viewing a simple drawing shows the presence and absence of repetitive cyclic scanning of the image. Numbers show the order of the subject's visual fixations. The labeled circles were drawn after the experiments to group and label the visual fixations. A visual inspection shows the paths of scanning in figures (a) and (b) to be similar, while the path in (c) does not resemble the ones in either (a) or (b). The fixation sequences are characterized as strings of letters where each letter denotes a fixation in the region labeled with that letter. The distances between the strings are computed as a cost of editing one string to attain the other and are (a)-(b) 10, (a)-(c) 15, and (b)-(c) 25.

one would deduce by a visual examination of the sequences of visual fixations alone. Thus this method appears to be useful to automate, objectify and quantify the similarity of sequences of visual fixations while looking at an object.

IV. STRING EDITING MEASURES OF SIMILARITY OF SEQUENCES

Another question of interest that now arises in the light of the past two sections is the following: What are the string editing measures between different realizations of the same Markov process, or in other words, what can we say about d , the distances between sequences generated by the same Markov matrix?

A simulation study was conducted, where for each of the Markov matrix structures $M_1 \dots M_4$ and each number of states $3 \dots 9$, 300 sequences of the length 33 (typical number of fixations during a viewing period of 10 s) were generated. Subsequently, the distances between these 300 sequences were measured with costs of substitution, deletion and insertion being chosen as unity.

As Fig. 5 shows, the mean distance increases with increasing size for all matrix structures (less so for quasi-deterministic processes). This was to be expected, because with increasing matrix size, that is, with the introduction of new states, the probability of a different sequence being generated increases. The lowest curves in Fig. 5 belong to M_1 and M_2 . In other words, the most similar sequences are generated with rigidly structured Markov processes that are very close to a deterministic process in which case the transition matrix would have only ones in the superdiagonal (Jordan form) and the sequences generated by such matrices would have a distance equal to zero from each other. Clearly, M_1 is closer to a deterministic process than M_2 . Therefore, the mean distance between the sequences generated by M_1 is smaller than the ones generated by M_2 . In the same vein, the process that is "most random," that is, with equal transition probabilities between each state, results in the largest distances between the resulting strings. Fig. 6 shows how the distribution of distances vary with the number of states for the four generating matrix structures.

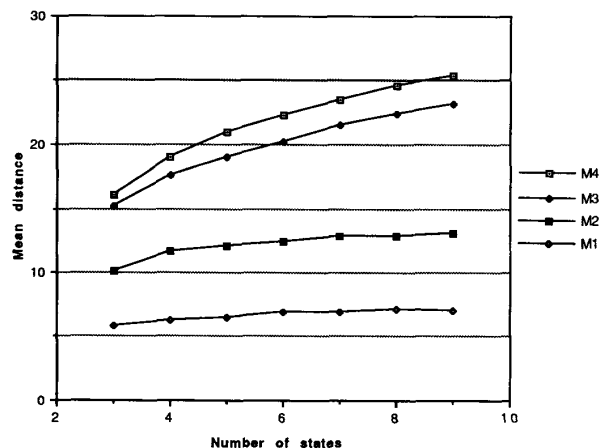


Fig. 5. A Markov process generates sequences of a given length. The mean distance between such sequences of the length 33 as obtained from 300 simulations for each combination is shown as a function of the number of states and the corresponding matrix structure.

V. DISCUSSION

Scanpath theory predicts similar sequences of visual fixations for a subject looking at a particular image. The presence of similarity is mostly determined by a visual inspection of the fixation sequences. The use of string editing method can help with the automation of determining this similarity by reducing a sequence of visual fixations to strings and by determining the distance between the strings. But in order to increase the statement power of this tool, we also had to find out about the statistics of randomly generated strings. This way, it can be stated with a certain confidence that the distance between two given sequences of visual fixations is below a threshold due to their similarity and not simply by coincidence.

An important problem inherent to this approach is the arbitrariness involved in the *clustering and regionalization of fixations*. For instance, why did we choose to include the fixation number 13 in Fig. 1 in the nose region? We could just as well have chosen to include it in the mouth region by changing the threshold between the mouth and the nose. Also, why didn't we subdivide the hand to fingers and palm and include fixation 7 in the palm area and fixations 5, 6, and 8 in the fingers area? It might be interesting to apply clustering algorithms as they are used in cosmology or elsewhere in image processing to decide to which group of points a certain fixation belongs [2]. But even such algorithms have several free parameters that the user must set. It is, of course, possible to divide the image into a regular grid but this loses any reference to the contents of the image.

Another problem involves the arbitrariness in *choosing the costs* of different editing operations in determining the distance between two strings. That is, while studying the similarity of sequence of fixations in Fig. 4, why did we choose the costs of substitution, insertion and deletion as 1, 2, and 3 respectively? We could have assigned to them equal values as in the simulation study of the previous section. Also, assuming in a further, hypothetical example that a certain region *D* is close to another region *E* but far away from region *F*, we

could make the costs string element dependent, such that the distance between the sequences *ABCD* and *ABCE* is less than the distance between *ABCD* and *ABCF*.

Features that are frequently utilized in active looking may have resulted in the evolution of internal mechanisms or *feature detecting processes* that have been embedded into "firmware" in the visual brain. When people look at an image, they look mainly at the parts of the picture that are regarded as being its features; they are the parts that hold the most information about the image. When subjects view simple pictures, their fixations tend to cluster around the points of the image where line directions change most abruptly [1]. Therefore, one can hypothesize that the angles are the principal features the brain employs to store and recognize images. There is further evidence encouraging that conclusion, namely the existence of angle-detecting neurons in the frog's retina [6], and complex cells in cats and monkeys [4]. This also make sense from a memory space optimization point of view (data compression): if the object is divided into straight segments connected with each other, it is more reasonable to store the length of a segment and the angle that connects it to the next segment rather than storing the continuation of the segment at predetermined intervals. This is analogous to the storage of large sparse matrices (as encountered, for example, in power network problems) where the positions of the elements different from zero and their values are stored instead of storing, say, 10 000 elements of which only about one percent are nonzero.

Understanding how humans recognize objects can also be transferred to machine vision to result in top-down image processing methods [12]. Furthermore, a simplistic analogy of the feature ring theory can be applied for the recognition of (convex or concave) polygons [9]. First, the image must be preprocessed to enhance contrast and to detect edges. Then, the image is recoded in terms of corners and lengths of straight edges between corners (for curved edges, discretization can be employed to introduce corners). Corners can, for instance, be labeled with upper case letters and lengths with lower case letters or other special characters. It makes sense to normalize the lengths by the first length. This way, the string will be characteristic of the object independent of its translational or rotational position or distance from the camera. (Of course, this string will initially be dependent on the first corner that the system recognizes, but this can be taken care of by a shift operator.) Once the image is compressed to a string representation, it can be compared with objects known to the system as described previously. The system will recognize the object under study as the object that has the shortest distance in the string editing sense from known objects. If the minimal distance is larger than a predefined threshold, the system can be programmed to learn and store it as a new object in its library. This approach of minimizing distances makes the method robust with respect to noise and errors in contrast enhancement and edge detection.

In summary, sequences of visual fixations while looking at an object were modeled as Markov processes. A method of quantifying the similarity of eye movements while looking at an object was introduced. This method is based on reducing

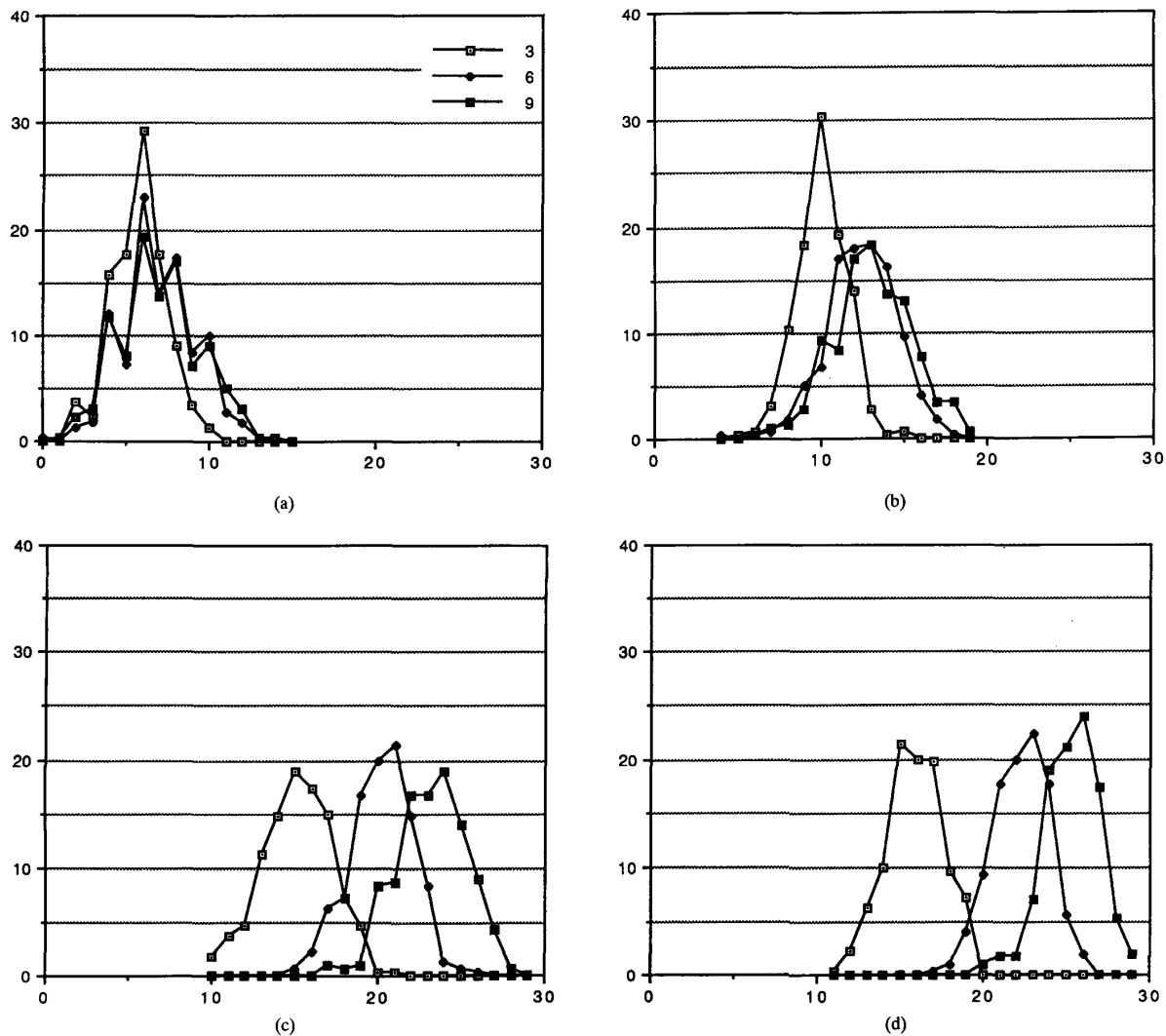


Fig. 6. The distribution of distances between sequences generated by the same Markov matrix gets more evenly and widely spread around the mean distance as the randomness of the process increases from (a) M_1 –(d) M_4 .

the sequence of visual fixations to a sequence of letters and defining an editing cost as the distance between such strings. Advantages and shortcomings of this method were discussed together with the possibility of its application in machine vision for the robust recognition of objects. Results of an experimental study of visual fixations with the string editing algorithm will be presented in a future paper.

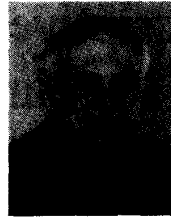
ACKNOWLEDGMENT

The authors thank Greg Tharp for providing Fig. 3.

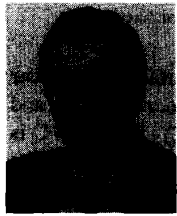
REFERENCES

- [1] F. Attneave, "Some informational aspects of visual perception," *Psychol. Rev.*, vol. 61, pp. 183–193, 1954.
- [2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
- [3] M. Eigen, R. Winkler-Oswatitsch, and A. Dress, "Statistical geometry in sequence space: A method of quantitative comparative sequence analysis," *Proc. Nat. Acad. Sci.*, vol. 85, 1988, pp. 5913–5917.
- [4] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.
- [5] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. New York: Springer, 1983.
- [6] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts, "What the frog's eye tells the frog's brain," in *Proc. IRE*, vol. 47, 1959, pp. 1940–1951.
- [7] D. Noton, L. W. Stark, "Scanpaths in eye movements during pattern perception," *Science*, vol. 171, pp. 308–311, 1971.
- [8] ———, "Eye movements and visual perception," *Scientific Amer.*, vol. 221, no. 6, pp. 34–43, 1971.
- [9] R. Schubiger, J. Moser, S. S. Hacisalihzade, and M. A. Müller, "Machine vision based on human perception and eye movements," to be presented at the IEEE Eng. Medicine Biology Soc. 13th Annu. Int. Conf., Orlando, FL, Nov. 1991.
- [10] L. W. Stark, J. A. Michael, and B. L. Zuber, "Saccadic suppression: A product of the saccadic anticipatory signal," in *Attention in Neurophysiology*. C. R. Evans, and T. B. Mulholland, Eds. London: Butterworths, 1969.

- [11] L. W. Stark and S. R. Ellis, "Scanpaths revisited: Cognitive models, direct active looking," in *Eye Movements: Cognition and Visual Perception*, D. F. Fisher, R. A. Monty, and J. W. Senders, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1981.
- [12] L. W. Stark, B. Mills, A. H. Nguyen, and H. X. Ngo, "Instrumentation and robotic image processing using top-down model control," in *Robotics and Manufacturing*, M. Jamshidi, *et al.*, Eds. New York: ASME Press, 1988.
- [13] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, pp. 168–173, 1974.
- [14] R. A. Wagner, "On the complexity of the extended string-to-string correction problem," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff, J. B. Kruskal, Eds. Reading, MA: Addison-Wesley, 1983.
- [15] B. L. Zuber, *Models of Oculomotor Behavior and Control*. Boca Raton, FL: CRC Press, 1981.



Lawrence Stark (SM'61–F'70) is a Professor at the University of California, Berkeley, where he divides his teaching efforts between the EECS and ME Departments in engineering and between the Physiological Optics and Neurology Departments in biology and medicine. His research interests are in bioengineering, with emphasis on human and robotic control of movement and vision. He pioneered the application of control and information theory to neurological systems.



Selim Hacisalihzade (S'81–M'81–SM'90) was born in Istanbul, Turkey, in 1957. He received the diploma of electrical engineering, the postdiploma in automatic control, and the doctorate in electrical engineering, all from the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, in 1980, 1983, and 1986 respectively.

He was a Research Associate at University of California, Berkeley (1987–1989), and an NRC Fellow at NASA Ames Research Center, California (1988–1989). He is currently head of Technology

Observation and European Community (EC) R&D Projects at Landis & Gyr and a lecturer at the Swiss Federal Institute of Technology (ETH).

Dr. Hacisalihzade is the founder and the current Chairman of the IEEE Engineering in Medicine and Biology Society Chapter in Austria/Germany/Switzerland. He is the author or coauthor of more than 50 papers in the general field of automatic control applications in life sciences.



John Allen was born in 1961 in Iowa City, IA. He received the B.A. degree in molecular biology and anthropology in 1983, and the Ph.D. degree in biological anthropology in 1989, from the University of California, Berkeley.

His research interests include human behavioral evolution, the use of biological markers in the cross-cultural study of behavioral diseases (e.g., eye movement dysfunction and schizophrenia), and the history of anthropology. He has published articles in *Perspectives in Biology and Medicine*, *Ergonomics*,

Human Biology, *Current Anthropology*, and *Biological Psychiatry*. After completing a postdoctoral fellowship in the Stanford University School of Medicine, Department of Psychiatry and Behavioral Sciences, he has recently been appointed a Lecturer in Biological Anthropology at the University of Auckland, New Zealand.