Copyright © 1998, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

ALGORITHMS FOR DEFINING VISUAL REGIONS-OF-INTEREST

by

Claudio Privitera and Lawrence W. Stark

Memorandum No. UCB/ERL M98/56

10 October 1998

COVER

ALGORITHMS FOR DEFINING VISUAL REGIONS-OF-INTEREST

by

Claudio Privitera and Lawrence W. Stark

Memorandum No. UCB/ERL M98/56

5 October 1998

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Algorithms for Defining Visual Regions-of-Interest:

Comparison with Eye Fixations

Claudio Privitera and Lawrence W. Stark

Neurology and Telerobotics Units

486 Minor Hall, University of California, Berkeley 94720-2020

Abstract

Many machine vision applications, such as compression, pictorial database querying and image understanding, often need to analyze in detail only a representative subset of the image that may be arranged into sequence of loci called regions-of-interest, ROIs.

We have investigated and developed a methodology that serves to automatically identify such a subset of aROIs (*a*lgorithmically detected ROIs) using different image processing algorithms and appropriate clustering procedures. In human perception, an internal representation directs top-down, context-dependent sequences of eye movements to fixate on similar sequences of hROIs (*h*uman identified ROIs). In this paper we introduce our methodology and we compare aROIs with hROIs as a criteria for evaluating and selecting optimal bottom-up, context-free algorithms. Some applications are discussed and defined.

1 Introduction

Eye movements, are an essential part of human vision because they must carry the fovea, and consequently the visual attention, to each part of an image to be fixated on and processed with high resolution. An average of three eye fixations per second generally occurs during active looking; they are intercalated by rapid eye jumps, called saccades, during which vision is suppressed. Only a small set of eye fixations, hROIs, *h*uman detected Regions-Of-Interest, is usually required by the brain in order to recognize a complex visual input (Figure 1, upper panels).

We have been studying and defining a computational model of this complex cognitive mechanism based on intelligent processing of digital images.

Image processing algorithms, IPAs, are usually intended to detect and localize specific features in a digital image, analyzing for example, spatial frequency, texture conformation or other informative values of loci of the visual stimulus. Many algorithms have been proposed in the literature and they might be classified into three principal approaches; for a survey, see Haralick [8] and Reed and Hans Du Buf [15]. Firstly, structural approaches, based on an assumption that images have detectable and recognizable primitives distributed according to some placement rules; examples are matched filters. Secondly, statistical approaches, based on statistical characteristics of the texture of the image; examples are co-occurrence matrices and entropy functions. Thirdly, model approaches that hypothesize underlying processes for generation of local regions: the images are analyzed on the basis of specific parameters governing these generators; examples are fractal descriptors.

For the purpose of our study, we have selected and adapted elements from this taxonomy in an attempt to simulate certain aspects of human perception.

Applying an IPA to an image means to transform that image into a new range of values

defining the corresponding algorithm parameter for each pixel. Local maxima in the transformed image represent loci wherein that particular parameter is particularly accentuated and they can, consequently, be used as a basis for identifying aROIs, algorithmic detected Regions-Of-Interest. Many local maxima may be generated by an image transformation: therefore a clustering procedure is required to reduce the initial large set of local maxima into a final small subset of aROIs (Figure 1, lower panels).

aROIs and hROIs can be compared to each other, analyzing their spatial locations or structural binding, and temporal order or sequential binding. The result of these comparisons measures the hROIs prediction capability of an IPA together with its clustering procedure. Thus our aim is explicit and our measures quantitative. The over-riding question is whether IPAs can treat a image in a fashion similar to human sequential glimpses.

In the following section the experimental protocol to acquire eye movement data is discussed in detail. Section 3 will be devoted to defining a list of IPAs from the above-mentioned taxonomy; the clustering and sequencing issue is considered and explained in section 4. The computational and statistical platform used to compare hROIs and aROIs is introduced in section 5 with a particular emphasis on scanpath human characteristics; in section 6 the results of the comparisons are discussed, and finally some applications are presented and defined in section 7.

2 Stimulus presentation and eye movement measurement

Computer controlled experiments present pictures and carefully measure eye movements using video cameras [20, 21]. An infrared source light is projected towards the eye of the subject generating a bright Purkinje reflection which is easy to track (Figure 2). An infrared video camera is focused on one of the subject's eye and the video image of the eye is digitized by means of a framegrabber into a PC-Pentium 166, the eye tracking server, which tracks the

Purkinje reflection at a sampling frequency of 60 Hz. The subject is instructed to watch the sequence of images, the visual stimuli, on a Silicon Graphics Indigo2 screen, the client stimulus controller, which is socket-connected to the eye tracking server. The subject is seated in front of the screen with his head secured onto an optometric chin-rest structure.

Prior to display of a visual stimulus, the subject is instructed to fixate on a sequence of nine different calibration points shown in random order on the SGI screen: this calibration procedure establishes a mapping between the location of the Purkinje reflection in the video image and the direction of the subject's gaze on the stimulus screen.

A fixation analysis algorithm is then applied to the eye movement data to distinguish rapid saccades jumps (Figure 1, upper right panel, arrows) from location of eye fixations (Figure 1, upper right panel, squares). The fixation algorithm analyzes the time derivative of eye movement data in order to localize regions in the screen where the eyes slow down below a pre-defined speed threshold (note eye movement sampling, Figure 1, upper left panel).

Seven different subjects were used during eye movement experiments. Fifteen different images were utilized: two Mars Terrain photographs; a Chilean Desert overview; a Country Landscape and Woman with Boy paintings; Leonardo's Mona Lisa; a Cave Interior with ancient horse-hunting art; After the Shower, a painting representing different figures in an interior country house and Madame, a female figure with a window in the background. We also used image modifications of some of these stimuli, such as embossed effect or binary thresholding.

Visual stimuli were not only presented (for a duration of 4 seconds, plus a double calibration time of 18 seconds for each stimulus, before and after data acquisition) to the group of subjects with several repetitions per person, but were also processed by all the IPAs described in the next section.

Repetitions were achieved by asking subjects to repeat the experiments within a few days for

a total of at least three sessions: in this way, consistency in the way each subject looks at specific visual stimulus could be studied and compared with algorithmic performance. During one experimental run, the complete sequence of images (each time in different order) was displayed to the subjects.

3 Image processing algorithms, IPAs, used for identifying aROIs

The information content of a generic image can be abstracted by different image parameters that in turn can be identified by relevant IPAs. In this sense, applying algorithms to an image means mapping that image into different domains, where for each domain, a specific set of parameters is extracted. After the image has been processed, only the loci of the local maxima from each domain are retained; these maxima are then clustered in order to yield a limited number of aROIs. The algorithms we studied are:

$1 - \mathcal{X}$, an x-like mask of 7×7 pixels, positive along the two diagonals and negative elsewhere, was convoluted with the image. We have also used different high-curvature mask convolutions, for example the " < " -like mask whose definition is intuitive. The block size of 7×7 (0.3×0.3 degree block) depends on image scaling.

$2 - \mathcal{S}$, symmetry, a structural approach, appears to be a very prominent spatial relation. For each pixel x, y of the image, we define a local symmetry magnitude $\mathcal{S}(x, y)$ as follows:

$$\mathcal{S}(x,y) = \sum_{(i_1,j_1),(i_2,j_2)\in\Gamma(x,y)} s((i_1,j_1),(i_2,j_2)) \tag{1}$$

where $\Gamma(x, y)$ is the neighborhood of radius 7 of point x, y defined along the horizontal and vertical axis ($\Gamma(x,y) = (x-r,y),...,(x,y),...(x+r,y),(x,y-r),...,(x,y+r)$) and $s((i_1,j_1),(i_2,j_2))$ is defined by the following equation:

$$s((i_1,j_1),(i_2,j_2)) = G_\sigma\ (d((i_1,j_1),(i_2,j_2)))\ |cos(\theta_1 - \theta_2)| \tag{2}$$

The first factor G_σ is a gaussian of fixed variance, $\sigma = 3$ pixels and $d(\cdot)$ represents the distance function. The second factor represents a simplified notion of symmetry: θ_1 and θ_2 correspond to the angles of the gray level intensity gradient of the two pixels (i_1, j_1) and (i_2, j_2). The factor achieves the maximum value when the gradients of the two points are oriented in the same direction. The gaussian represents a distance weight function which introduces localization in the symmetry evaluation. Our definition of symmetry was consequently based on the orientation correspondences of gradients around the centered point [16]. Alternatively, a normalization of the axial quadratic moment could be used instead to compute the symmetry transform [7].

$3 - \mathcal{W}$, a discrete wavelet transform is based on a pyramidal algorithm which splits the image spectrum into four spatial frequency bands containing horizontal lows/vertical lows (ll), horizontal lows/vertical highs (lh), horizontal highs/vertical lows (ll) and horizontal highs/vertical highs (hh). The procedure is repeatedly applied to each resulting low frequency band resulting in a multiresolution decomposition into octave bands. The process of image wavelet decomposition is achieved using pair of conjugate quadrature filters (CQFs) [25] which acts as a smoothing filter (i.e. a moving average) and a detailing filter respectively (see for example [18]). We have used different orders from the Daubechies $\mathcal{W}_{db}$ and Symlet $\mathcal{W}_{sy}$ family bases [5, 6] to define CQF filters. For each resolution i, only the wavelet coefficients of the highs/highs hh_i matrix were retained and finally relocated into a final matrix HH (with the same dimension as the original image) by the following combination:

$$HH = \sum_{i=1}^{n} \zeta^i(hh_i) \tag{3}$$

where n is the maximum depth of the pyramidal algorithm ($n = 3$ in our case) and where $\zeta(\cdot)$ is a matrix operation which returns a copy of the input matrix hh by inserting alternatively rows and columns of zeros.

$4 - \mathcal{F}$, a center-surround 7×7 quasi-receptive field mask, positive in the center and negative

in the periphery, was convoluted with the image.

5 $-$ $\mathcal{O}$, difference in the gray-level orientation, is possibly also analyzed in early visual cortices (see also [10]). Center-surround orientation difference is determined first convoluting the image with four Gabor masks of angles 0°,45°,90° and 135° respectively. For each pixels x, y, the scalar result of the four convolutions are then associated with four unit vectors corresponding to the four different orientations. The orientation vector $\bar{o}(x,y)$ is represented by the vectorial sum of these four weighted unit vectors. We define the center-surround orientation difference transform as follows:

$$\mathcal{O}(x,y) = (1 - \bar{o}(x,y) \cdot \bar{m}(x,y)) \ \| \bar{o}(x,y) \| \| \bar{m}(x,y) \| \tag{4}$$

where $\bar{m}(x,y)$ is the average orientation vector evaluated within the neighborhood of 7×7 pixels. The first factor of the equation achieves high values for big differences in orientation between the center pixel and the surroundings. The second factor acts as a low-pass filter for the orientation feature.

6 $-$ $\mathcal{E}$, edges per unit area, is determined by detecting edges in an image, using the Canny extension of the sobel operator [3] and then congregating the edges detected with a gaussian of $\sigma = 3$ pixels.

7 $-$ $\mathcal{N}$, entropy is locally calculated as $\sum_{i \in G} f_i \ log \ f_i$ where f_i is the frequency of the $i-th$ gray level within the 7×7 surrounding region of the center pixel and G is the local set of gray levels. Local maxima defined by this factor emphasize texture variance.

8 $-$ $\mathcal{C}$, Michaelson contrast, is most useful in identifying high contrast elements, generally considered to be an important choice feature for human vision. Michaelson contrast is calculated as $\|(\mathcal{L}_m - L_M) \ / \ (\mathcal{L}_m + L_M)\|$, where $\mathcal{L}_m$ is the mean luminance within a 7×7 surrounding of the center pixel and L_M is the overall mean luminance of the image. $\mathcal{L}_m$ was also used in our study.

9 – $\mathcal{H}$, the discrete cosine transform, DCT, introduced by [1], is used in several coding standards as, for example, in the JPEG-DCT compression algorithm (see section 7). The image is first subdivided into square blocks (i.e. 8×8); each block is then transformed into a new set of coefficients using the DCT; finally, only the high frequency coefficients are retained to quantify the corresponding block.

10 – $\mathcal{L}$, the laplacian of the gaussian, is convoluted with the image.

4 Clustering and sequencing

The IPAs result in defining local maxima widely over the image; a clustering procedure is then applied to reduce this large set of local maxima into the final small ($n \approx 7$) subset of aROIs. Thus, the resulting string of aROIs were similar in number to human eye movement fixation glances looking at similar images.

The initial set of local maxima is clustered by connecting local maxima and gradually increasing the acceptance radius for joining them. During each step of the clustering process, all local maxima less than a specific radius apart are clustered together (Figure 3). Each cluster inherits the maximum value of its component points (local maxima): the locus of this highest valued maximum for each cluster then also determines the locus of that cluster. Only that maximum point is retained; all the other composing local maxima are deleted. The procedure is repeated while increasing the acceptance radius at each step. The decision to end the clustering process is set when only a pre-defined number n of clusters remain. The values of the remaining clusters, ordered from highest to lowest, permits us to relate the sequence of clusters, aROIs, to sequences of human fixations.

Algorithm $\mathcal{N}$ was applied for example to a Chilean desert photo (Figure 4, upper left panel) and the initial set of local maxima (Figure 4, upper right panel) was then clustered (Figure 4,

lower left panel; partway through the clustering process). The final ordering is indicated by the arrows connecting the cluster loci and superimposed on the original image(Figure 4. lower right panel). Note the maximum valued locus for each cluster. No initial conditions are required for the clustering. The overall procedure, implemented using sparse matrix representation, is fast in execution, even for large images.

Other clustering procedures have been investigated. Changing the criteria to detect the locus and the value of the clusters during each iteration can modify the previous procedure: for example, the number of included local maxima could be used to affect the value of a specific cluster. However, no significant disparities in the overall performance of our system have been noted when different clustering procedures were compared to each other. Each of our IPAs, of course, contributes the intensity of its selected parameter in finding the local maxima and thus the values of resulting clustered aROI domains. This may be quite intuitive; it is the nature of the processed image (i.e. the IPA used), more than the clustering procedure used for the identification of the final aROIs, that most influences the final distribution of aROIs (Privitera et al., in preparation).

If we had used only IPAs and not the clustering procedure, we could have selected, say, the seven highest local maxima directly and defined them to be the aROIs. Those selected aROIs however, might be much more closely spaced. Thus the clustering procedure is actually an eccentricity-weighting procedure, where even lower local maxima that are eccentrically located may finally be selected to form an aROI.

5 Comparing and sorting procedures

The aROI loci selected by our different IPAs and those loci defined by human eyes movement fixations, hROIs, can be compared. In this section we describe the statistical and computational

platform we have been using for these comparisons (see also [14]). We also introduce the scanpath theory.

5.1 Comparison of two set of ROIs

Comparison of final clusters of ROIs began with taking two sets of ROIs (Figure 5, middle, upper and lower panels) and clustering these two sets using a distance measure derived from a k-means pre-evaluation. This evaluation determined a region for calling coincident any ROIs that were closer than this distance and non-coincident for ROIs that were further apart than this distance; the distance was about two degrees and similar in size to human foveal spans for moderate visual acuity. All the coincident ROIs (named *joined-ROIs*) were labeled with the same alphabetic character (Figure 5, right panel) and they then enabled a similarity metric, S_p, to determine how many ROIs two algorithms (as in the example shown in Figure 5, see also the processed image in the left panels), or two humans, or an algorithm and a human have in common; the final value was normalized based upon string length. The individual sources of the elements, that is the original ROIs, used in these final interactive steps are preserved as circles and squares (Figure 5, right panel) to illustrate the procedure.

As mentioned above, ROIs are ordered by the value assigned by the IPA or by the temporal ordering of human eye fixations in a scanpath. Then, the joined-ROIs can finally be ordered into strings of ordered points. Here, (Figure 5), we have for example: $string_{\mathcal{E}} = abcfeffgdc$ and $string_{\mathcal{S}} = afbffdcdf$. The string editing similarity index S_s was defined by an optimization algorithm [20] with unit cost assigned to the three different operations *deletion*, *insertion* and *substitution*.

Our comparisons yielded two different indices of similarity which tells how closely two set of ROIs resemble each other in locus, S_p, and in sequence, S_s (see the "toy" diagrams on Figure

6). For the example illustrated above (Figure 5) .we have: $S_p = 1$ and $S_s = 0.34$.

5.2 Y-matrices and parsing diagrams

Similarity coefficients can be sorted and represented for the two measures, S_p and S_s in Y-matrices (two human subjects are for example compared and averaged for two different pictures, Figure 7, upper panels; again in Figure 8 where we compare not different subjects but different IPAs) and in parsing diagrams, (Figure 7, lower panels, for all the subjects). The parsing diagram collects averages of these similarity coefficients: R, for repetitive scanpaths, same subject looking at the same picture at different times; Local = L, different subjects same picture; Idiosyncratic = I, same subjects different pictures; Global = G, different subjects different pictures.

The most important distinction is that between Repetitive similarity, R, upper left box (Figure 7, lower panels), and Global similarity, G, lower right box: the R value for human with the S_p measure, equals 0.64. This means that the string for repetitive viewing of the same stimulus for the same subject have loci that were 64% within fixational or foveal range — this represents continuing support for the scanpath theory, (see the following section).

For Global, all different subjects looking all different stimuli had an S_p value of only 0.28. This number was somewhat different from the expected S_p value of 0.21 based up on consideration of a random model, Ra, (bottom box).

5.3 The scanpath theory

The scanpath has been defined on the basis of experimental findings. It consists of sequences of alternating saccades and fixations that repeat themselves when a subject is viewing a picture. Only ten percent of the duration of the scanpath is taken up by the collective duration of the saccadic eye movements providing an efficient mechanism for traveling over the scene or regions of interest; thus the intervening fixations or foveations onto hROIs, have at hand ninety percent

of the total viewing period (see [2, 24] and also section 2).

Scanpath sequences appeared spontaneously without special instructions to subjects and were discovered to be repetitive: note the high R index in the parsing diagrams (Figure 7). This repetitiveness suggested to Noton and Stark [11] that a top-down internal cognitive model[1] controls perception and active looking of eye movements in a repetitive sequential set of saccades and fixations, or glances, over features of a scene so as to check out and confirm the model [20].

Of course, the objective or task can affect the active looking of eye movements [27, 26]. Nevertheless, without any specific task instruction, for general viewing conditions, the high R values in the parsing diagrams (Figure 7: 0.64 for S_p and 0.42 for S_s) suggest that a very similar set of representative regions of interest are sequenced and searched by the brain each time. A considerable consistency is also reported by the high L values when different subjects look at the same picture (Figure 7: 0.54 for S_p and 0.28 for S_s).

The strong consistency reported in human experiments, when no specific objective is given to the subjects, means that only a specific restricted set of representative regions of the picture is essential for the brain to perceive and recognize that picture. This representative set is similar for different subjects, and this important characteristic brings us to the main scientific objective of our work: whether it is possible to automatically identify this set by using IPAs.

Comparing aROIs with hROIs is the standard utilized to study and select which IPAs are more successful in this objective; if a specific task is given, different hROIs may result, and consequently, different algorithms may be selected from our collection [13].

[1]This internal cognitive model must, of course, approximate the external world; otherwise, perception would not be possible. Thus, what we call an internal model and what is generally understood to be memory are two very interrelated cognitive entities. It is beyond the scope of the present paper to go into detail about how this model is created and organized in the brain; what is perhaps worth emphasizing here, is that internal cognitive models often must match closely the visual stimuli inputs.

5.4 Study of the ANOVA: ANalysis-Of-VAriance

In the following section, results from the aROIs vs. hROIs comparisons are presented and discussed.

In order to better evaluate and interpret the final results, we also used an Anova, analysis-of-variance. The Anova value is compared with a critical value F of Fisher distribution with $k-1$ degree of freedom in the numerator (where k is the number of a distribution that we are comparing) and $n-k$ in the denominator (where n is the total number of observation in the k distributions). If the Anova test value is less than the F-Fisher critical value for an α level of significance (for example, in this paper, α was set equals to 0.01), then it is possible to infer that the two means are not different enough to come from different distributions; on the other hand, if the Anova test value is greater than the F-Fisher critical value, this signifies that the means likely come from different distributions.

Our standard format for presenting our data in the algorithms parsing diagram was as a triplet; for example 0.33 (0.04, 18.7) with 0.33 equal to the mean value, 0.04 equal to +/- the standard deviation, and 18.7 equal to the Anova test value. Our quantitative conclusions presented in the result section below were strongly sustained by the relationship between Anova test values and F-Fisher critical value (for $\alpha = 0.01$) of 7.5.

6 hROIs vs. aROIs comparisons

We were, of course, most interested in the several uses of our methodology on our data — to analyze not only the capability of IPAs and clustering procedure to predict eye fixations, but also the inter-relationships among algorithms. In this section, we present and discuss these different results which derive from the comparison results.

6.1 Relationships among IPAs

We wished to obtain as wide a variety of image processing algorithms as possible and to keep small the coherence between pairs of image processing algorithms. Thus, our wide variety of image processing algorithms would have independent actions on the images and they could serve to identify aROIs for a variety of picture types, and for a variety of visual identification tasks [13].

The coefficients of the Y-matrix (Figure 8, see also [14] for preliminary results) indicated the coherence between each pair of a selected group of algorithms as explained above. Enclosed within a dashed box are two different group of algorithms: each group is internally characterized by high S_p similarity but cross-similarity between the two groups is very low. For example, a coefficient value of 0.69 (Figure 8, left panel) between algorithms $\mathcal{N}$ and $\mathcal{F}$ demonstrated a strong coherence between those two algorithms while the value of 0.15 (Figure 8, left panel) between algorithms $\mathcal{C}$ and $\mathcal{W}_{db}$ demonstrated a high independence. A string-editing similarity coefficient of zero (Figure 8, right panel) between algorithms $\mathcal{N}$ and $\mathcal{W}_{sy}$ represents complete independence of two compared sequences. Note, that the coefficients for S_s (aROIs string-editing similarity) were much lower than the coefficients for S_p (aROI-loci similarity).

Our collection of algorithms could thus be sorted for similarities or for differences in generating aROIs. This is of value in selecting algorithms for different tasks [13].

6.2 Parsing diagrams

Again, we gathered the crucial comparisons between algorithms and eye fixations together into a parsing diagram (Figure 9). The ability of the algorithms (labeled A in Figure 9) to predict eye fixations was demonstrated by the number in the upper right box, L, of the left panel, S_p. The average for all the algorithms was 0.33.

On the basis of this large number of measures between algorithms, images and subjects, we may select a sub-group of algorithms ($A* = \mathcal{W}_{db}$, $\mathcal{L}$, $\mathcal{O}$ and $\mathcal{S}$). For this selection, the S_p similarity rose to 0.36 and the Anova test showed a considerable significance (27.0 related to the F-Fisher critical value of 7.5). Two different examples of high S_p values are shown (Figure 10: hROIs left panels and aROIs right panels, algorithms $\mathcal{L}_m$, upper, and $\mathcal{W}_{sy}$, lower).

The global, G, S_p value represented the average S_p for algorithms applied to different images and it could be considered a bottom anchor for S_p; a further bottom anchor was Ra, the random S_p, calculated for coincidence among randomly identified loci.

The S_s parsing diagram shows little coherence even among all the algorithms providing support of an earlier preliminary study, [23], that the IPAs and the clustering procedures we used cannot predict sequencing of human eye movements.

We also selected four of the algorithms that seem to cohere (average $S_p = 0.56$) with the human subjects for two particular images (After the Shower and Madame). This overall S_p was further segregated to show each algorithm and each subject separately (Figure 11, upper right box; subjects A,C,H,T and algorithms $\mathcal{X}$, $\mathcal{S}$, $\mathcal{W}_{db}$ and $\mathcal{F}$) so the variability can be judged. Another coherence strengthens our overall result by documenting the strong S_p indices among these four selected algorithms; note the high average S_p, that equals 0.75 among these algorithms (lower triangle). A third coherence (not shown) was achieved for the set of Mars terrain images alone, and algorithms $\mathcal{C}$, $\mathcal{W}_{sy}$ and $\mathcal{E}$(average coherence between aROIs and hROIs was 0.43).

We might further improve eye fixation loci prediction by choosing other sets of structurally different algorithms and combining these algorithms and the others used in this paper in some optimal fashion. The overall result here is that IPAs in conjunction with the clustering procedure can predict hROIs with appreciable average S_p and with statistical significance. If we select some of the algorithms on the basis of specific type of images, then this prediction is as good as the

ability of one human to predict the fixation locations of another subject; the local, L, relationship in our S_p parsing diagrams equaled (Figure 7). Thus, our large collection of algorithms can provide different selection policies both for different images and for different tasks [13].

Several subjects from our lab, extraneous to this project, have been asked to qualitatively judge the distribution of aROIs over the pictures for all the algorithms and the pictures that have been used in this study. In this way, an evaluation different and independent from the S_p metric, could be taken into account in order to validate our results and better interpret the meaning of the coefficients reported in the tables. The subjects participating in the evaluation were asked to analyze the aROIs for each algorithm and each picture and express a personal valuation using three different grades: good, medium and bad. First, a total of 64% of aROIs were considered acceptable or good. Then, ordering these results for each algorithms and for each image, we computed the correlation with the ordering generated by our S_p metric. The average correlation was quite high, around 0.7, and Anova analysis confirmed the relationship between these human qualitative evaluations of aROIs and the S_p coefficients.

7 Some applications of aROIs

Certain computer vision applications might benefit from an apparatus that automatically identifies regions of visual interest in a digital image. Some of these applications have been investigated within the last few years in our lab.

Compensating for a variety of visual defects might be arranged by using idiosyncratic anatomical information about the localized lesion for an individual patient. We are studying a head mounted vision apparatus wherein important aROIs can be shifted so as to avoid the patient scotoma or blinded area.

Internet communications usually have a bottleneck in image retrieval: pictorial databases

occupy large amounts of disk memory, and image searching through the internet is usually very slow when the entire image has to be analyzed for all the images in the database. An intelligent pictorial database querying system could be based on aROIs in the digital image. The query can be formulated in terms of these regions, and the subsequent search implemented only analyzing the aROIs associated to each image in the database. The smaller set of images that match aROIs present in the query would eventually be sent back through the net (many interesting alternatives are present in the literature, see for example [17, 4] and the proceedings corresponding to the second reference).

A suggestion was put forward by Stark and Ellis in 1981 [22] for video transmission to send high resolution local ROIs alternately with the ordinary resolution of the entire image. In this way, subjects looking at the transmitted images would likely obtain an impression that the entire image had been seen with very high resolution. Algorithms for detecting aROIs can be utilized for automatically identifying these high resolution ROIs.

An extended version of the JPEG encoder, the *Selective JPEG* encoder, based upon on aROIs, has been implemented in our lab. The JPEG image compression standard [12], uses the Discrete Cosine Transform, DCT, of 8×8 image blocks followed by a lossy quantization and loss-less entropy (Huffman) coding for each block. Quantization is usually performed by the following division and rounding operation:

$$C_i = \lfloor f_i \pm Q_i/2 \rfloor / Q_i \tag{5}$$

where i is one element of the 8×8 block; C_i is the quantized coefficient, f_i is the $i-th$ DCT frequency coefficient, and Q_i is the corresponding quantizer factor. The $\pm$ sign is the same as the f_i coefficient.

The quantizer factors are grouped into a quantizer factors matrix, $Q = \{q_i : i = 1, ..., 64\}$, each factor corresponding to a specific DCT frequency, and they control the lossy compression

level of the image: the bigger the factors in the matrix are, the more the blocks are compressed. Hence, the more frequency information is lost: a quantization that is too coarse may eventually produce the classical *blocky* effect.

In the standard JPEG baseline, the quantization matrix is usually standardized, with higher coefficients corresponding to higher frequencies, and it is uniquely defined for all the blocks of the image. Consequently the quantization cannot be differentiated over different regions of the image. In our *Selective JPEG* baseline, the magnitude of the quantizer factors are adaptively related to the distance from the set of aROIs by means of the following rule:

$$Q(x,y)_i = Q_i S(d_{min}(x,y)) \tag{6}$$

where $d_{min}(x,y)$ is the minimum distance between the block x,y in the image and the set of aROIs. $S(\cdot)$ is a stepwise monotonic function equal to the unity for distance $d_{min}(x,y)$ that is appropriately small and then increasing with the distance; Q_i is the original standard quantizer coefficient.

The stepwise function $S(\cdot)$ can be tuned appropriately both in size and in slope: we can choose for example to have small high resolution regions and then gradually increase the quantization in equation 6) in the periphery or define large high resolution regions and then strongly increase the quantization in the periphery.

Algorithm $\mathcal{O}$ has been applied, for example, to a countryside photo and five aROIs have been identified (Figure 12, upper panel). The *Selective JPEG* compression was then applied to the image based on those identified aROIs where $S(d(x,y)) = 2$ for $0° \leq d(x,y) < 1°$, degree of the visual angle (note that also aROIs are slightly compressed); $S(d(x,y)) = 250$ for $1° \leq d(x,y)$. The *Selective JPEG* compressed image (Figure 12, middle panel) can be compared with the standard JPEG compression (Figure 12, lower panel) with the same amount of compression (100:4): the visual appearance is much better with the *Selective JPEG* compression.

Alternative solutions have been proposed in the literature (see for example [19, 9]): aROIs are usually identified at the block level and severe discontinuities can appear for strong compression rate even with selective quantization. Moreover, some of the proposed methods may not be really JPEG-compatible. An interesting proposal is discussed in [28] where a fuzzy scheme is used to determine important regions. Our method is based on a number of compact regions of interest whose size and distribution over the image is based on and strongly inspired by human eye movement studies. Indeed, the biological plausibility of our model results in an evidently better qualitative impression of the compression.

8 Discussion

Our method provides a precise task for the IPAs we have studied — to predict human scanpaths, both loci and sequences of eye movement fixations or foveations. The method also provides for quantitative measurements of prediction accuracy.

In this paper, we have validated that a constellation of IPAs used in conjunction with a clustering procedure can predict, for S_p, the loci of human fixations. Our results indicate, however, that the algorithms can not predict the sequential ordering, S_s, of the sub-features used by a person.

The wide selection of algorithms gave us an opportunity to study the differences and similarities in terms of the precise task we consider. These algorithm characteristics are of great interest to us as indicators of the general nature of an picture and how either algorithms or humans process it. We might need to provide weighting coefficients for the different algorithms in order to optimize the prediction capabilities of the ensemble.

Our scale of similarity indices is anchored at the bottom both by the random, Ra, values and by the global, G, values, that is for all subjects and algorithms and pictures. The top of

the scale is anchored for human studies by the repetitive, R, value, the closeness of fit of a single subject's scanpaths to her scanpaths with the same picture at another time with the same task instructions. Can we similarly use trivial modifications of the pictures to obtain repetitive indices for the algorithm studies?

The clustering procedures we used require a good deal of thought and preliminary studies have been reported in section 4. As we indicated above, the clustering procedure distributes strings of aROIs in more eccentric locations than they would be in without the clustering procedure. This eccentricity asserted a positive effect on the similarity between aROIs and hROIs.

In summary, the methodology defined in this paper has been tested on a varied set of digital images that ranges from portraits to landscapes and terrain images. A number of subjects were used for the eye movement experiments. Finally independent subjective evaluations by naive subjects in order further validated the results.

The overall results are very encouraging and we have started to define and implement different applications such as image retrieval from pictorial databases. For use in image compression, a *Selective JPEG* image compression encoder has been defined in more detail in the last section of the paper.

Acknowledgements

We thank our sponsors for partial support:- NASA-Ames Research Center (Dr. Stephen Ellis), Fujita Research (Dr. Ken Kawamura) and Neuroptics Corporations (Dr. Kamran Siminou). Our colleagues in the laboratory, Michela Azzariti for her indispensable statistical analysis; Ted Blackmon, Yeuk Fai Ho, Veit Hagenmeyer, Yong Yu. Others have been generous in their advice — Irwin Sobel at Hewlett-Packard, and at UCB, Jerry Feldman, CS, and David Brillinger, Statistics; also the anonymous referees for their incisive, constructive and motivating comments.

References

[1] N. Ahmed, T. Natarajan, and K. Rao. Discrete cosine tranform. *IEEE Trans. Computer*, 23:90–3, 1974.

[2] T.A. Bahill and L.W. Stark. Trajectories of saccadic eye movements. *Scientific American*, 240:84–93, 1979.

[3] J. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 42–9. IEEE Comput. Soc. Press., Los Alamitos, CA, 1997.

[5] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:906–66, 1988.

[6] I. Daubechies. *Ten Lectures on Wavelets.* Philadelphia: SIAM, 1992.

[7] V. Di Gesú and C Valenti. The discrete symmetry transform in computer vision. Technical Report 011-95, Laboratory for Computer Science (D.M.A.), University of Palermo, 1995.

[8] R.M. Haralick. Statistical and structural approaches to texture. *Proc. IEEE*, 67:786–804, 1979.

[9] A. Kundu. Enhancement of JPEG coded images by adaptive spatial filtering. In *Proc. International Conference on Image Processing*, volume 1, pages 23–26. IEEE Comput. Soc. Press., Washington, DC, 1995.

[10] E. Niebur and C. Koch. Control of selective visual attention: Modeling the "where" pathway. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–8. The MIT Press, 1996.

[11] D. Noton and L.W. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171:308–11, 1971.

[12] W.B. Pennebaker and J.L. Mitchell. *JPEG Still Image Data Compression Standard.* Van Nostrand, 1993.

[13] C.M. Privitera, M. Azzariti, and W.L. Stark. Autonomous image processing algorithms locate regions-of-interest: the Mars Rover application. *Tech. Report, Electronics Research Laboratory, University of California, Berkeley*, (UCB/ERL M98/8), March 1998.

[14] C.M. Privitera and L.W. Stark. Evaluating image processing algorithms that predict regions of interest. *Pattern Recognition Letters*, 1998. (forthcoming to appear).

[15] R. T. Reed and J. M. Hans Du Buf. A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image Processing*, 57(3):359–72, 1993.

[16] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*, 14:119–130, 1995.

[17] E. Remias, G. Sheikholeslami, Z. Aidong, T. Fathima, and F. Syeda-Mahmood. Supporting content-based retrieval in large image database systems. *Multimedia Tools and Applications*, 4(2):153–70, 1997.

[18] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Trans. on Information Theory*, 38(2):569–86, 1992.

[19] R. Rosenholtz and A.B. Watson. Perceptual adaptive jpeg coding. In *Proc. International Conference on Image Processing*, volume 1, pages 901–4. IEEE Comput. Soc. Press., Lausanne, Switzerland, 1996.

[20] L. Stark and Y. Choi. Experimental metaphysics: The scanpath as an epistemological mechanism. In W. H. Zangemeister, H. S. Stiehl, and C. Freksa, editors, *Visual Attention and Cognition*, pages 3–69. Elsevier, Amsterdam, 1996.

[21] L. Stark, Y. Choi, and Y. Yu. Visual imagery and virtual reality. In *Visual Science.* Dordrecht, the Netherlands: Elsevier, 1996.

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

[23] L. Stark and C.M. Privitera. Top-down and bottom-up image processing. In *Proc. of IEEE International Conference on Neural Networks*, volume 4, pages 2294–9. Houston, TX, June 1997.

[24] L.W. Stark, W. Hoyt, K. Cuiffreda, R. Kenyon, and F. Hsu. Time optimal saccadic trajectory model and voluntary nystagmus. *Models of Oculomotor Behavior and Control*, 1980.

[25] P.P. Vaidyanathan. *Multirate System and Filter Banks.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[26] A.L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967.

[27] W.H. Zangemeister, K. Sherman, and L.W. Stark. Evidence for global scanpath strategy in viewing abstract compared with realistic images. *Neuropsychologia*, 33(8):1009–25, 1995.

[28] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsushita. A JPEG codec adaptive to the relative importance of regions in an image. *Transactions of Information Processing Society of Japan*, 38(8):1531–42, 1997.

FIGURE LEGENDS

Figure 1 - Computer and Human Processing

Comparing *h*uman identified Regions of Interest, hROIs, (upper right) with *a*lgorithmic identified Regions of Interest, aROIs, (lower right). Note eye movement sampling, (upper left) and local maxima in the processed image, (lower left).

Figure 2 - Measuring Eye Movements

The video image of the subject's eye is digitized by means of a frame-grabber and the bright Purkinje reflection is tracked in x-y coordinates.

Figure 3 - Clustering Procedure: a Single Step

During each step of the clustering process, all local maxima less than a specific radius apart, D_i, are clustered together. Then the highest valued maximum for each cluster determines the locus of that cluster and all the remaining maxima are removed. The process continues while increasing the acceptance radius D_i at each step.

Figure 4 - Clustering Procedure: to Completion

Algorithm $\mathcal{N}$ was applied to a Chilean desert photo (upper left). The initial set of local maxima (upper right) was then clustered using the defined iterative procedure (lower left: partway through the process). The final ordering, superimposed on the original image, is shown in the lower right panel; the maximum-valued locus for each cluster is inserted in the figure.

Figure 5 - ROIs Comparisons Procedures

Actions of each IPA yields a transformed image (left column) for two examples, $\mathcal{E}$ (upper), and $\mathcal{S}$ (lower). Final aROIs in each image are ordered by value and connected by arrows in analogy to eye movement sequences of fixations (central column). The two set of aROIs are finally combined (right panel) into a number of *joined*-ROIs, further used to define distance measures between the two sets.

Figure 6 - Similarity Measures

Two sets of ordered ROIs (left) whose loci are widely separated: S_p low and S_s low. Two sets of ROIs (middle) with closely matched loci, but whose ordered sequences are different: S_p high and S_s low. Two sets of ROIs (right) whose loci and ordered sequence are similar: S_p high, S_s high.

Figure 7 – Y-matrices and Parsing Diagrams

S_p and S_s similarity indices for different subjects (or different algorithms) and for different pictures can be arranged in a Y-matrix (upper panels) with each value being the average of several repetitions. Parsing diagrams, (lower panels), represent averages of these similarity indices in a more collected and intuitive fashion.

Figure 8 - Coherence and Independence among IPAs

Cross-comparison values of six algorithms for two indices, S_p and S_s. Enclosed within the dashed boxes are two different group of algorithms: each group is internally characterized by high S_p similarity, but cross-similarity in S_p between groups is very low. Note that S_s values are very low.

Figure 9 - Parsing Diagrams for Comparing aROIs and hROIs

Crucial comparisons between algorithms and eye fixations are gathered in the parsing diagrams (see text). Test of significance are made by comparing the Anova test value with the F-Fisher critical value which is 7.5.

Figure 10 - Comparisons of hROIs and aROIs for Different Pictures

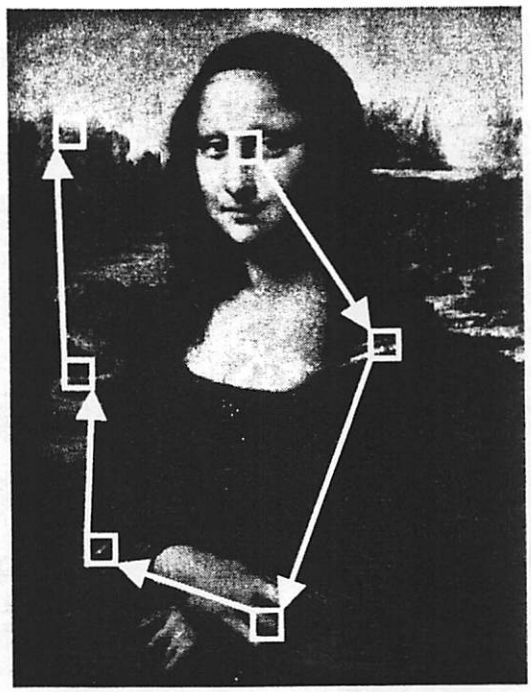
Two examples of good S_p-similarity between aROIs (right column: $\mathcal{L}_m$ upper, $\mathcal{W}_{sy}$ lower) and hROIs (left column). $S_p = 0.62$ and $S_s = 0.13$, upper panel; $S_p = 0.87$ and $S_s = 0.13$, lower panel. Note low values of S_s indicating that string sequences could not be identified by the IPAs.

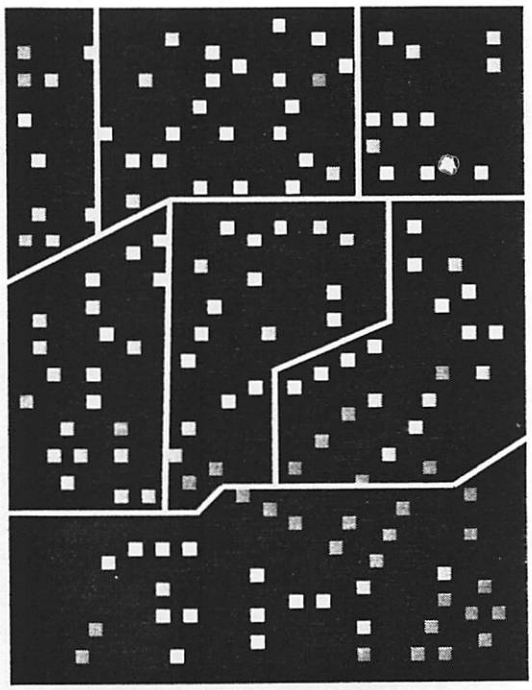
Figure 11 – Coherence and Independence among Selected Algorithms

Y-matrixes for two particular images (After the Shower and Madame), a selected group of algorithms, $\mathcal{X}$, $\mathcal{S}$, $\mathcal{W}_{db}$ and $\mathcal{F}$ and subjects, A,C,H,T.

Figure 12 – *Selective JPEG* compression based on aROIs

Five aROIs (arrows) were identified using $\mathcal{O}$; aROIs were maintained at higher resolution by the *Selective JPEG* compression. The *Selective JPEG* compressed image is shown in the middle panel and a standard JPEG compression in the lower panel. Total compression was the same in both examples; the visual fidelity is much higher with *Selective JPEG* compression.





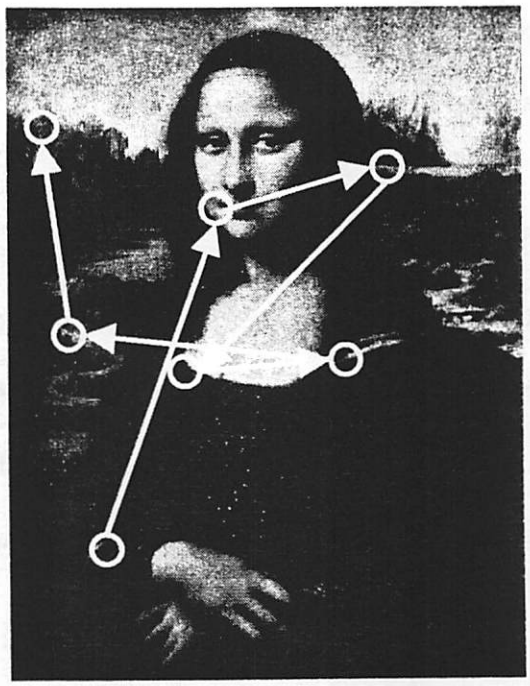


Figure 1

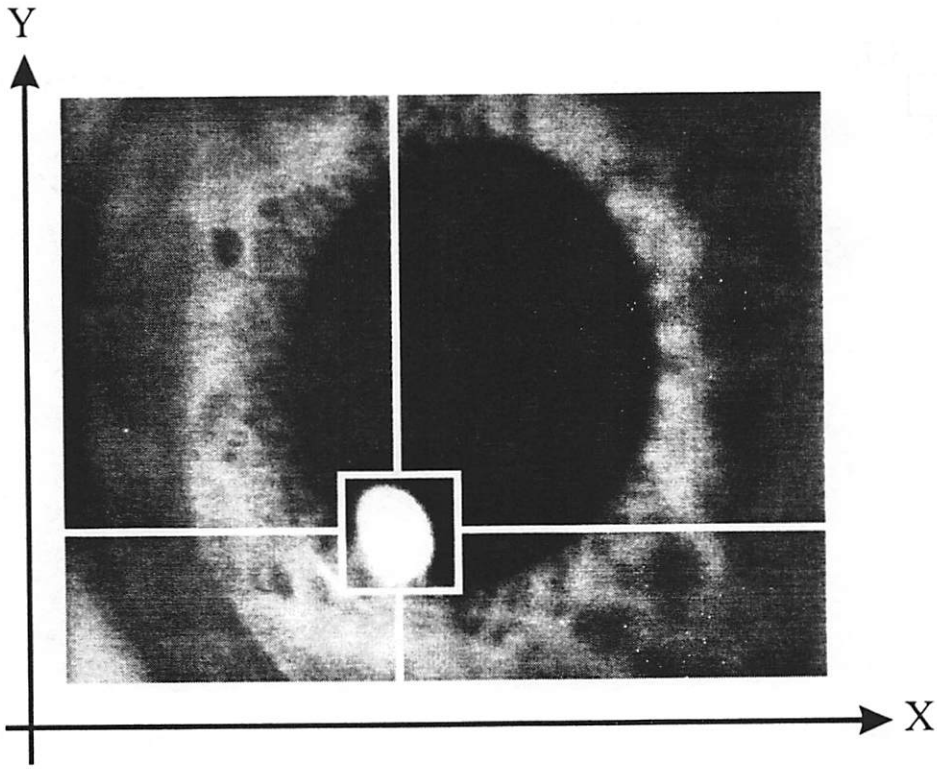


Figure 2

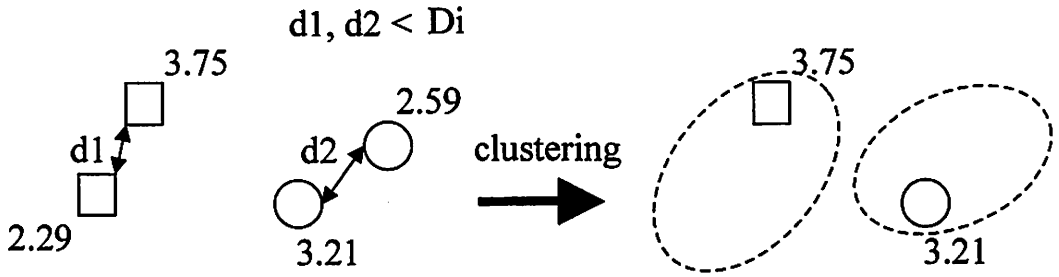
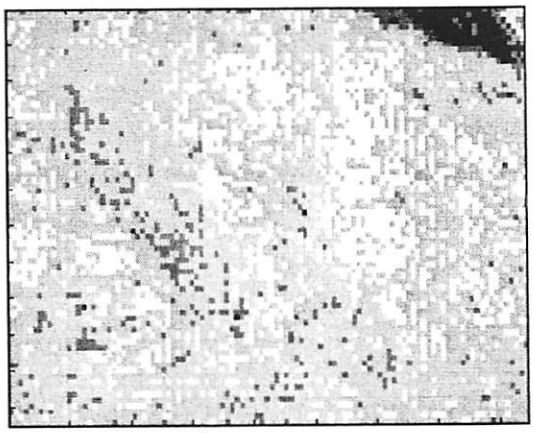
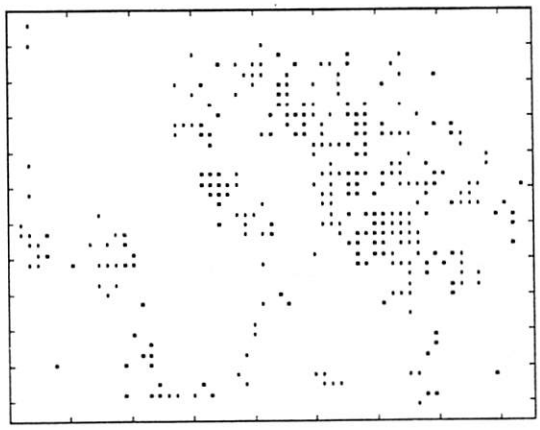
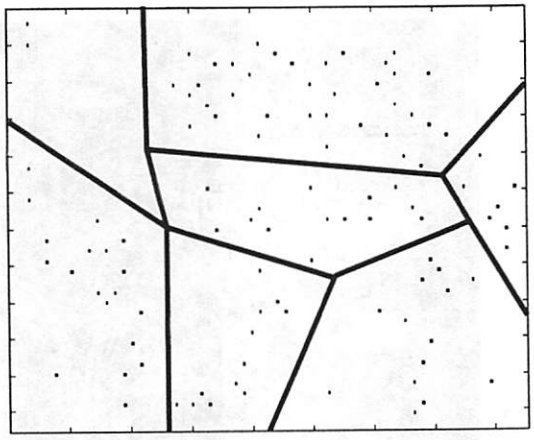


Figure 3







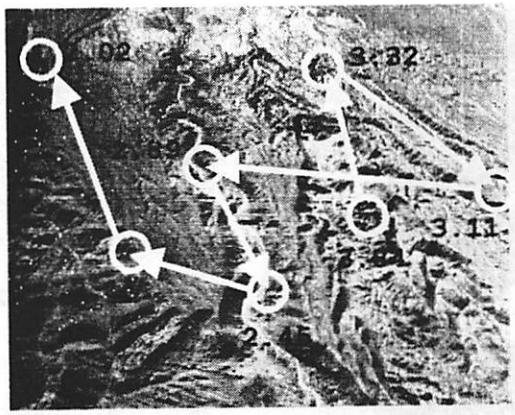


Figure 4

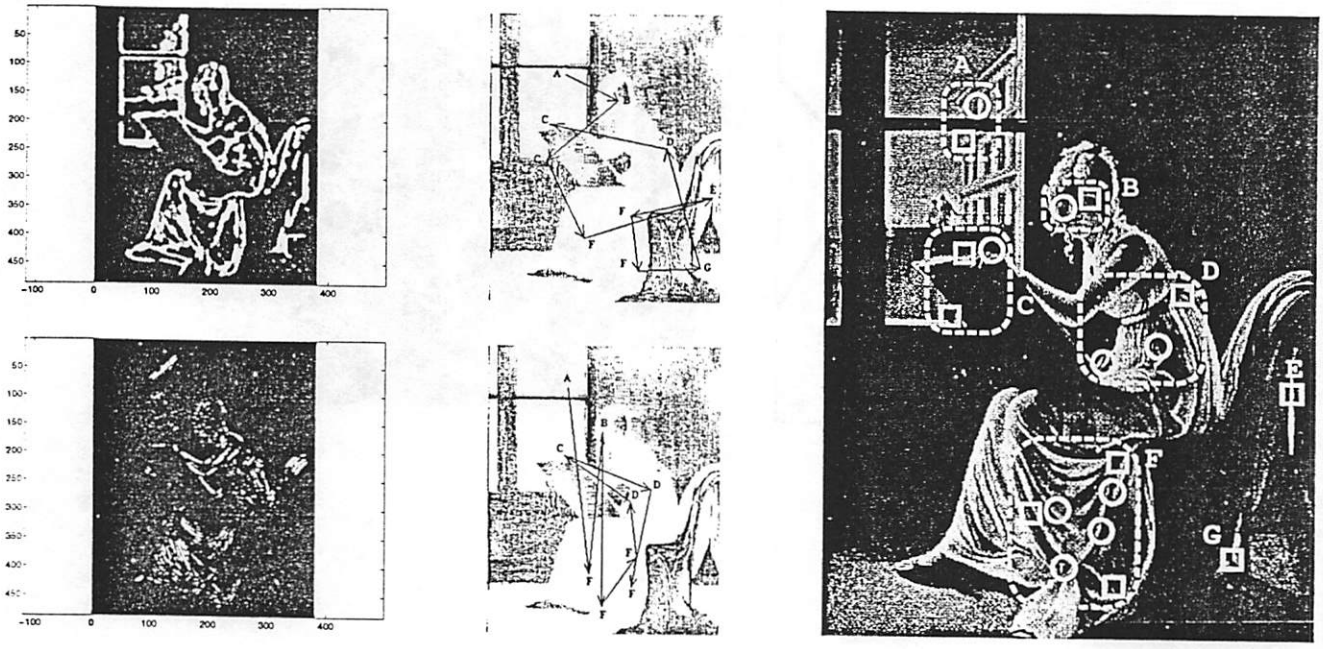
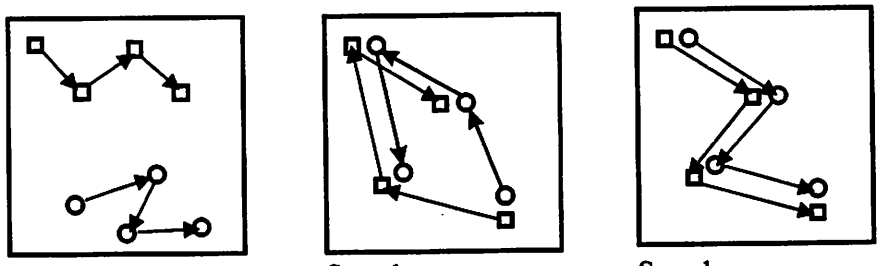


Figure 5



Sp = 0
Ss = 0

Sp = 1
Ss = 0

Sp = 1
Ss = 1

Figure 6

Y-MATRICES (partial: 2 subjects)

Sp	Subject1		Subject2	
	Pict1	Pict2	Pict1	Pict2
S1p1	0.65_R	0.38_I	0.54_L	0.18_G
S1p2		0.60_R	0.31_G	0.47_L
S2p1			0.69_R	0.33_I
S2p2				0.58_R

Ss	Subject1		Subject2	
	Pict1	Pict2	Pict1	Pict2
S1p1	0.40_R	0.24_I	0.31_L	0.08_G
S1p2		0.39_R	0.13_G	0.19_L
S2p1			0.43_R	0.21_I
S2p2				0.24_R

PARSING DIAGRAMS (complete: 7 subjects)

	Same Subject	Different Subjects
	Repetitive 0.64	**Local** 0.54
	Idiosyncratic 0.34	**Global** 0.28
Sp		**Random** 0.21

	Same Subject	Different Subjects
Same Picture	**Repetitive** 0.42	**Local** 0.28
Different Pictures	**Idiosyncratic** 0.21	**Global** 0.16
Ss		**Random** 0.04

Figure 7

Sp	*W*db	*W*sy	*L*	*N*	*F*	*C*
*W*db	-	0.57	0.37	0.21	0.28	0.15
*W*sy		-	0.39	0.18	0.24	0.22
L			-	0.19	0.25	0.16
N				-	0.69	0.61
F					-	0.63
C						-

Ss	*W*db	*W*sy	*L*	*N*	*F*	*C*
*W*db	-	0.31	0.05	0.02	0.05	0.02
*W*sy		-	0.09	0.00	0.01	0.02
L			-	0.01	0.01	0.05
N				-	0.23	0.22
F					-	0.33
C						-

Figure 8

Sp

Same pict.	**R** **1**	**L** A = **0.33** (0.04 **18.7**) A* = **0.36** (0.01 **27.0**)
Diff picts	**I** A = **0.20** (0.03 **1.62**) A* = **0.17** (0.05 **0.03**)	**G** A = **0.26** (0.05 **6.16**) A* = **0.24** (0.05 **4.16**)

Ra = 0.21

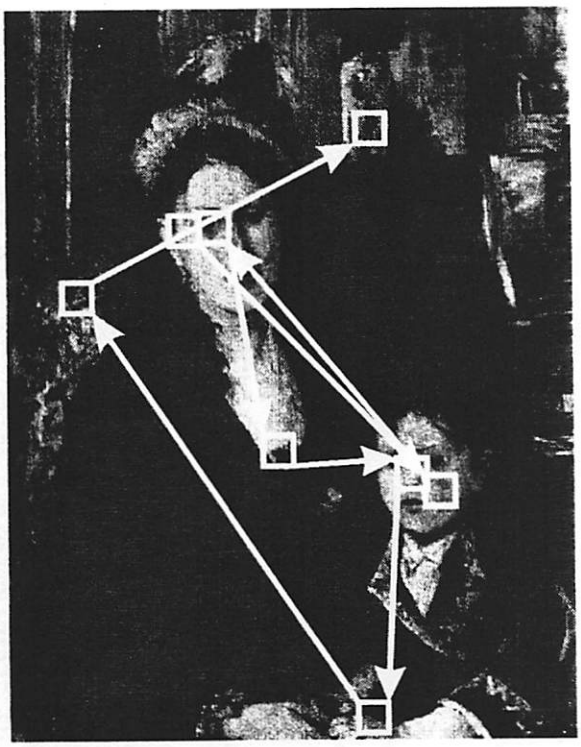
Algs vs. Algs Algs vs. Eye fixations

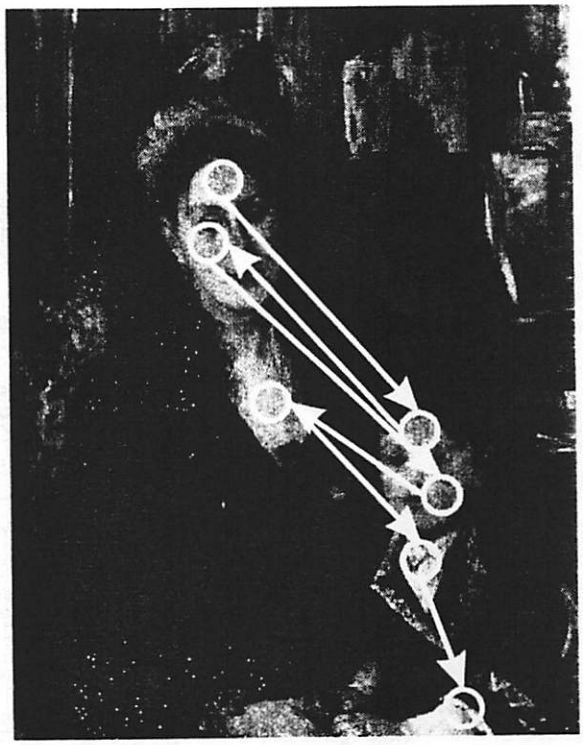
Ss

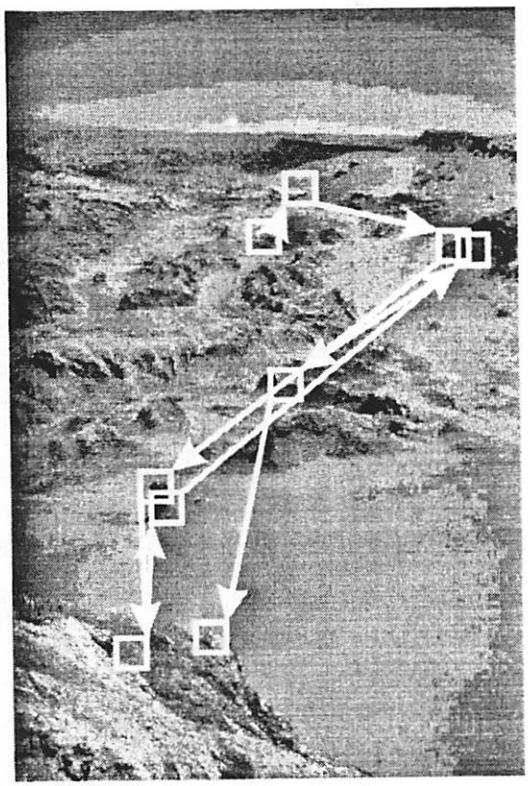
Same pict	**R** **1**	**L** A = **0.05** (0.01 **3.31**) A* = **0.05** (0.01 **2.21**)
Diff pict	**I** A = **0.03** (0.01 **0.12**) A* = **0.02** (0.01 **0.01**)	**G** A = **0.04** (0.01 **0.44**) A* = **0.04** (0.01 **0.32**)

Ra = 0.04

Figure 9







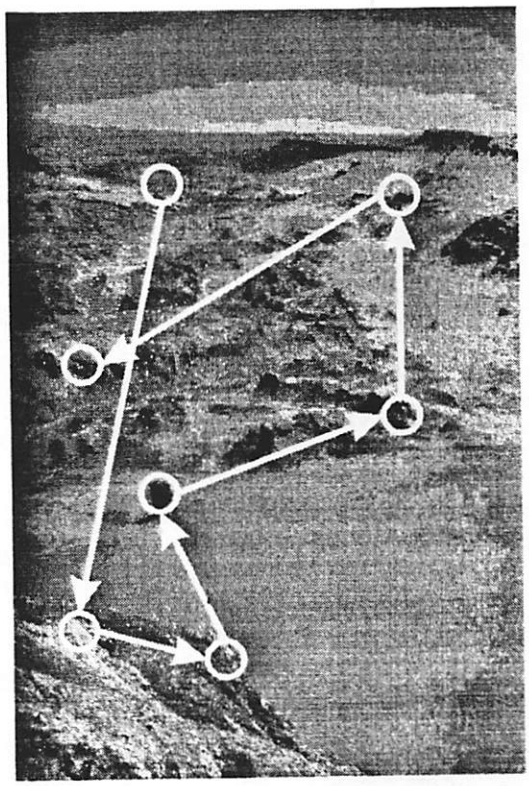


Figure 10

Sp	A	C	H	T	*X*	*S*	*W*	*F*
A	1	0.23	0.54	0.64	0.60	0.67	0.72	0.64
C		1	0.69	0.86	0.78	0.78	0.73	0.40
H			1	0.42	0.52	0.60	0.40	0.51
T				1	0.42	0.42	0.47	0.28
X					1	0.83	0.87	0.66
S						1	0.78	0.85
W							1	0.51
F								1

Figure 11

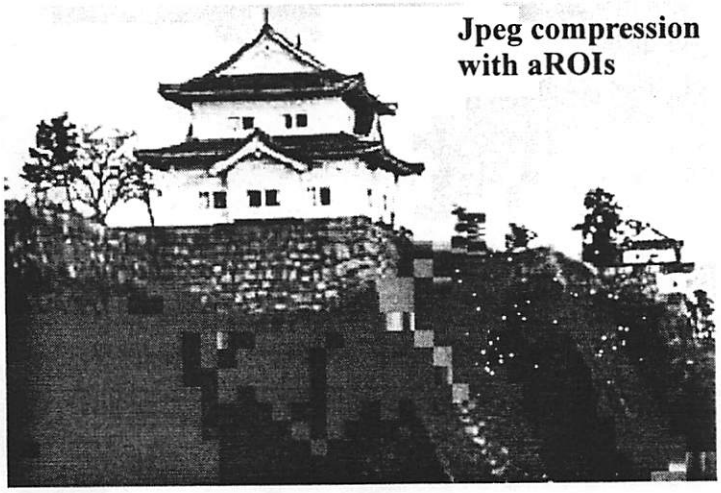


Figure 12