

# Using Human Observers' Eye Movements in Automatic Image Classifiers

Alejandro Jaimes<sup>1</sup>, Jeff Pelz<sup>2</sup>, Tim Grabowski<sup>2</sup>, Jason Babcock<sup>2</sup>, and Shih-Fu Chang<sup>1</sup>

<sup>1</sup>Dept. of Electrical Engineering, Columbia University, New York, NY 10027

<sup>2</sup>Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623

## ABSTRACT

We explore the way in which people look at images of different semantic categories (e.g., handshake, landscape), and directly relate those results to computational approaches for automatic image classification. Our hypothesis is that the eye movements of human observers differ for images of different semantic categories, and that this information can be effectively used in automatic content-based classifiers. First, we present eye tracking experiments that show the variations in eye movements (i.e., fixations and saccades) across different individuals for images of 5 different categories: *handshakes* (two people shaking hands), *crowd* (cluttered scenes with many people), *landscapes* (nature scenes without people), *main object in uncluttered background* (e.g., an airplane flying), and *miscellaneous* (people and still lives). The eye tracking results suggest that similar viewing patterns occur when different subjects view different images in the same semantic category. Using these results, we examine how empirical data obtained from eye tracking experiments across different semantic categories can be integrated with existing computational frameworks, or used to construct new ones. In particular, we examine the *Visual Apprentice*, a system in which image classifiers are learned (using machine learning) from user input as the user defines a multiple level object definition hierarchy based on an object and its *parts* (*scene*, *object*, *object-part*, *perceptual area*, *region*), and labels examples for specific classes (e.g., handshake). The resulting classifiers are applied to automatically classify new images (e.g., as handshake/non-handshake). Although many eye tracking experiments have been performed, to our knowledge, this is the first study that specifically compares eye movements across categories, and that links category-specific eye tracking results to automatic image classification techniques.

**Keywords:** eye tracking, automatic image classification, content based retrieval.

## 1. INTRODUCTION

Eye tracking studies have been performed for many years (e.g., one of the earliest reported in [2]), with many different purposes. One of the main goals of such studies has been understanding the human visual system and, in particular, the visual process itself. It is now well understood that humans move their eyes, in part, because visual acuity falls by an order of magnitude within degrees of central vision [1]. Therefore, for some tasks, eyes must be moved to shift the point of regard to regions requiring high spatial resolution. Humans, however, also move their eyes to objects or regions of interest even when foveal acuity is not required by the immediate task. Because these eye movements are made to visual and attentional targets, monitoring the eye movements of observers can provide an externally observable marker of subjects' visual strategies while performing tasks such as manual image indexing or passive image viewing. Analyzing eye movements for these tasks can be useful in understanding how humans look at images, not only in terms of the recognition strategy used (e.g., which areas are observed and in which order; how much time person spends looking at certain types of objects, etc.), but also in determining what is deemed as important during the process (i.e., areas looked at are probably more important than areas not looked at).

In the last few years, research in the field of Content Based Retrieval [3] has focused on facilitating access to multimedia information (e.g., images, video, etc.) in large digital databases. In particular, there has been a strong interest in being able to automatically classify multimedia data. Images and video, for example, can be placed into categories depending on their visual content, at several levels (e.g., *syntactic*: based on low-level features such as colors, etc.; and *semantic*: based on objects and scenes [11]) Automatically classifying images (e.g., photographs) and video is an important task since it facilitates indexing, which allows searching and browsing in large image collections (e.g., images on the internet). Various computational approaches that perform automatic classification (mostly in the field of Computer Vision) have drawn on theories of the functionality of the human visual system [14]. In order to limit the amount of information to be processed, for example, some techniques detect Regions of Interest (ROIs) so that only regions that may be relevant to the problem at hand

---

<sup>1</sup> E-mail: {ajaimes, sfchang}@ee.columbia.edu WWW: <http://www.ee.columbia.edu/~ajaimes, ~sfchang>

<sup>2</sup> E-mail: pelz@cis.rit.edu WWW: <http://www.cis.rit.edu/pelz>

are selected for analysis. Understanding the selection performed by humans, and the visual process, is deemed to be useful in the construction of algorithms to perform classification.

In spite of the similarities between human processing and automatic techniques, most computational approaches are based on general theories that do not directly link the specific problem with the information obtained from experiments involving human subjects. For example, when we observe an image of two people shaking hands, perhaps we always move our eyes in a specific path to fixate on areas that we deem important (e.g., two faces, handshake). The areas that we observe, and the order in which we make those observations depend highly on the content of the image (e.g., handshake vs. landscape), and the task (e.g., recognize a person; find an object in the image). It may be possible, however, to find patterns in the way different individuals look at images in the same category. Nonetheless, information on how humans perform these specific tasks is seldom included in computational approaches. Analyzing the way humans look at images, however, could lead to important improvements in the construction of such classifiers because, if class specific observation patterns exist, decisions regarding the computational recognition process could be made based on data collected from human observers.

In this paper, we present a study in which we analyze human observers' eye movements when observing color photographs of different semantic categories. Eye movement traces of ten subjects were recorded as they viewed a series of 250 randomly interleaved images from the following categories: *handshake* (two people shaking hands), *crowd* (e.g., many people), *landscape* (no people), *main object in uncluttered background* (e.g., an airplane flying), and *miscellaneous* (people and still lives). We analyze, in the viewing patterns: (1) *within image variations* (similar/dissimilar patterns for an image, across several subjects); (2) *across image variations* (subject's pattern depends strongly on the image); and (3) *within/across image category variations* (similar/dissimilar patterns for images in the same category, across several subjects). In addition, we explore different ways in which these results can be used directly in the construction of automatic classifiers in a framework like the *Visual Apprentice* [10]. In that system, image classifiers are learned from user input as a user defines a multiple level object definition hierarchy (scene, object, object-part, perceptual area, region), and labels examples for the specific class (e.g., handshake) he is interested in. Machine learning techniques are used, on the training data provided by the user, to construct classifiers that can be automatically applied to new images.

### 1.1. Related work

Eye tracking experiments have been performed for a many years. Several studies have examined eye movements of individuals as they perform natural tasks (e.g., [4][12][17]). Others have focused on the way humans observe pictures (e.g., photographs in [8], paintings in [24]). Many of these types of studies have been very useful in the development of theories of visual perception and recognition (e.g., [12][27]). For example, it has been suggested that humans move their eyes over the most informative parts of an image ([24]), and that eye movements (i.e., fixations and saccades, discussed later) are strongly influenced by the visual content of the image [7], and by the task being performed by the observer (e.g., describe image; search for an object [27]). No studies, to the best of the authors' knowledge, have tried to compare differences in eye movements across different semantic categories.

Computational techniques that use information from eye movements include [21] and [22]. Unlike the work presented in [21], and [22], we focus on studying the *differences* in the way humans look at images *across* different categories, and on the usefulness of those differences to construct automatic classifiers for the same categories. The relation between computational (i.e., by computer algorithms), and human selection of Regions of Interest (i.e., areas of an image deemed important by an observer) was studied in [19]. Our work differs from [19], since our goal is to use the results of category-specific eye tracking experiments to construct classifiers.

### 1.2. Outline

In section 2 we present a brief overview of eye tracking and automatic classification. In section 3 we describe the approach we used to track eye movements. Section 4 presents the human observer experiments we performed, and in section 5 we discuss ways in which those results could be used in automatic classifiers (e.g., the Visual Apprentice). We conclude in section 6.

## 2. EYE MOVEMENTS AND AUTOMATIC TECHNIQUES

### 2.1. Eye Tracking

The photoreceptor array in the image plane of the human eye (the retina) is highly anisotropic; the effective receptor density falls dramatically within one degree of the central fovea. The acuity demands of most visual tasks requires the high

resolution of the fovea, but observers move their eyes to objects or regions of interest even when foveal acuity is not required by the immediate task. People make well over 100,000 eye movements every day. When humans move their eyes they either hold their gaze at a stationary point (*fixations*) or move them quickly between those fixations (*saccades*). While observers could gather a great deal of information from images while holding fixation, subjects free to view images without instruction regarding movements of the eyes typically make several eye movements per second. It has been known for some time that eye movement patterns are image dependent and to some degree idiosyncratic [1][15]. In addition to image-dependence, the pattern of eye movements and spatial distribution of fixation points also varies with the instructed task. Yarbus (27) demonstrated that subjects adopted dramatically different eye movement patterns when viewing one image when the instructions were changed. For example, the pattern seen under free-viewing was different when a subject was asked to remember the location of objects in the image, or to estimate the ages of people in the image.

The eye movements necessitated by the limitations of the peripheral visual field are driven by the scene and task, but in general make approximately three to four saccadic eye movements per second. In between those eye movements that are made to shift the point of gaze from one point in the scene to another, the retinal image must be stabilized to ensure high acuity. When the observer and scene are static, the eyes are stationary in the orbit, resulting in a static image projected on the retina. These *fixations* allow high acuity vision. When the observer and/or the scene is in motion, other mechanisms are necessary to stabilize the retinal image. A number of oculomotor mechanisms provide this stabilization. Objects moving through the field can be tracked with *smooth pursuit* eye movements. Large-field motion also elicits smooth eye movements to stabilize the image on the retina. Image motion due to movement of the head and body are cancelled by the *vestibulo-ocular reflex*, which produces rotation of the eyes to compensate for head and body movements [9]. The *saccades* are rapid, ballistic movements that reach velocities of over 500 degrees/second. Saccades from less than one degree in extent to over 90 degrees are seen in subjects performing a number of tasks. The duration of the saccades varies, but are typically completed in less than 50 msec. Because of the speed with which the eyes move during a saccade, the retinal image is blurred during the eye movement. Subjects are not aware of the blurring caused during saccades because of a slight reduction in the systems sensitivity, but the effect is due primarily to a phenomenon termed *backwards masking*, in which the retinal image captured at the end of the saccade tends to mask the blur that would otherwise be evident.

## 2.2. Automatic Techniques

Several approaches have been proposed for automatically determining the visual content category of images. The techniques can be roughly separated into those that classify images according to scenes (e.g., indoor/outdoor [25][16], city/landscape [26], etc.), and according to the objects depicted in the images (e.g., sky, faces, etc.). Additionally, we can make a distinction between approaches that use global features (e.g., classify according to scene using global color and texture [25][26][16]), and those that use local features (e.g., regions [13][23][10]). Since our goal is to link human observers' eye movements to automatic classifiers, our concern is mainly with approaches that use local features to perform classification. In particular, we are interested in approaches that make use of local structure (see [11] for a discussion on different levels of indexing) to perform classification. Some of the approaches that use local structure are based on the automatic segmentation of images. Pixels within the image are automatically grouped according to color, texture, or other low-level features. The regions obtained from the segmentation are then used in automatic classifiers. In the work presented in [13][23], for example, the configuration of the regions is used to determine (in different ways) the image categories. In snowy mountain scenes, for example, it is common to find blue sky at the top of the image, directly above white regions (corresponding to snow). In other approaches, such as the *Visual Apprentice* [10], regions are grouped and structured (e.g., into hierarchies) for the detection of objects and scenes. Such structuring (e.g., a handshake can be modeled as an object that contains two faces and one handshake) is closely related to the way in which humans move their eyes when observing images (e.g., a viewer may look at the two faces and the handshake). In the following sections we discuss eye tracking and the experiments performed. In section 5 we discuss how those results could be incorporated into automatic techniques like the *Visual Apprentice*.

## 3. EYE MOVEMENT RECORDING

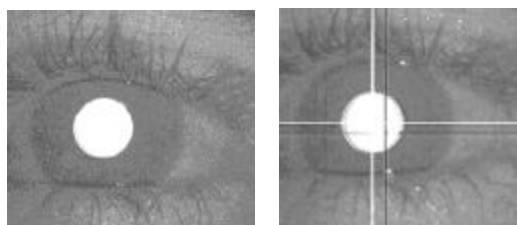
Several methods can be used to track a subject's gaze. Several systems are in use today, each offering advantages and disadvantages. One system uses coils of fine wire held in place on the eye with tight-fitting annular contact lenses [20]. Eye position is tracked by monitoring the signals induced in the coils by large transmitting coils in a frame surrounding the subject. *Scleral coil* eyetrackers offer high spatial and temporal resolution, but limit movement and require the cornea to be anesthetized to prevent pain due to the annular contact lens. Another system offering high spatial and temporal resolution is

the *dual-Purkinje* eyetracker [5]. These eyetrackers shine an infrared illuminator at the eye, and monitor the reflections from the first surface of the cornea and the rear surface of the eyelens (the second optical element in the eye). Monitoring both images allows eye movements to be detected independent of head translations, which otherwise cause artifacts. Limbus trackers track horizontal eye movements by measuring the differential reflectance at the left and right boundaries between the sclera (the ‘white of the eye’) and the pupil. Vertical eye movements are measured by tracking the position of the lower eyelid. While the limbus tracker provides high temporal resolution, the eye position signal suffers from inaccuracy, and there is significant cross-talk between horizontal and vertical eye movements. The class of eyetrackers used in this study illuminates the eye with infrared illumination, and images the eye with a video camera. Gaze position is then determined by analyzing the video fields collected at 60 Hz. Eye position data was collected with an Applied Science Laboratories Model ASL 504 Remote eyetracker. The system monitors eye position without any contact with the subject, an important factor to consider (Figure 1). The camera lens used to image the eye is surrounded by infrared emitting diodes (IREDs) providing illumination coaxial with the optical axis. The infrared, video-based eyetracker determines the point-of-gaze by using a video camera to extract the center of the subject’s pupil and a point of reflection on the cornea. Tracking both pupil and first-surface reflections (i.e., on the cornea) allows the image-processing algorithms to distinguish between eye-in-head movements and motion of the head with respect to the eyetracker. This infrared/video eyetracker is limited to 60 Hz sampling rate and provides accuracy of approximately one degree across the field. The system automatically tracks subjects’ head movements over a range of approximately 25 cm. Beyond that range, the tracker must be manually reset. The eyetracker signals such a track loss by setting the horizontal and vertical eye positions to zero.



**Figure 1.** The ‘remote’ eye camera (left) is placed just below the subject’s line of sight (right). The lens is surrounded by infrared emitting diodes to provide coaxial illumination.

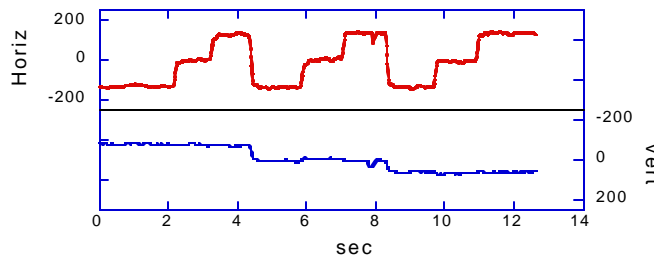
While the retina absorbs most light that enters the pupil, the retina is highly reflective in the far-red and infrared regions of the spectrum. This phenomenon, which leads to ‘red-eye’ in photographs taken with a flash near the camera lens, produces a ‘bright-pupil’ eye image. In this image, the iris and sclera are the darkest regions; the pupil is intermediate, and the first-surface reflection of the IR source off the cornea is the brightest. The eye image is processed in real-time to determine the pupil and corneal reflection centroids, which are in turn used to determine the line-of-sight of the eye with respect to the head. Figure 2 shows an eye image captured with the ASL bright-pupil system. The image on the left shows the raw IR illuminated image; the image on the right shows the image with the superimposed cursors indicating pupil and first-surface reflection centroids determined by thresholding the image and fitting a circle to the pupil and corneal reflection. As the observer moves his eyes, the shape of the pupil reflection changes, and so does its centroid. The difference between the centroid of the pupil and the centroid of the corneal reflection (two points indicated in Figure 2) is used to determine the actual eye movement.



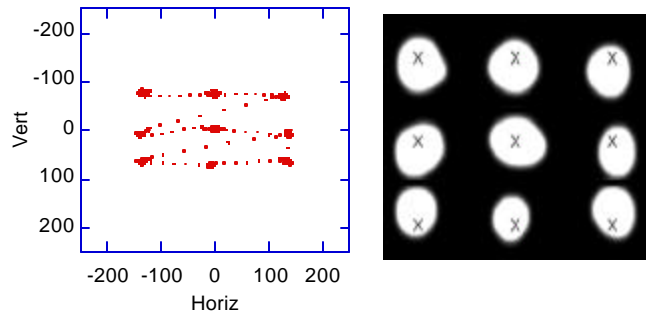
**Figure 2.** Image of the eye captured by the ASL eyetracking system (left); pupil centroid (white cross, right); and corneal reflection centroid (black cross, right)

Eye position is reported as a horizontal and vertical point of regard every 16.7 msec. The raw data is in arbitrary units based on display scaling, viewing distance, and subject calibration. The data is converted to image pixel units by scaling the output to the calibration points in pixel coordinates. The transformation corrected horizontal and vertical scaling, and offset the data to the center of the image display (i.e., [0,0] is the center of each image).

Figure 3 represents the horizontal (top of the figure) and vertical (bottom of the figure) eye position of a subject reviewing nine calibration points (see also Figure 4). Calibration is necessary to determine the observer's position in space, and must be performed once for each subject as long as the general setup does not change (e.g., observer's physical location with respect to the camera). The fixation sequence was left-to-right, top to bottom. The repeated 'step' pattern in the horizontal trace shows the sequence of three horizontal fixation points on each line of the calibration target. The vertical record indicates the three rows that are scanned in turn. The zero-slope portions of the graphs (Figure 3) represent fixations on the calibration points; the transitions between fixations represent the rapid saccadic eye movements between the calibration points.



**Figure 3.** Horizontal and vertical eye position during a 9-point calibration sequence in image pixel units.



**Figure 4.** Horizontal and vertical position during the 9-point calibration sequence, in image pixel coordinates (left), and, fixation density mask overlaid on calibration grid (rescaled) (right).

Figure 4 (left) shows the same data as in Figure 3 plotted in two dimensions to show the spatial distribution of the scanpath. Horizontal and vertical eye position during the 9-point calibration sequence are seen scaled to image display coordinates. The fixation pattern can also be visualized as an image in which the lightness of a given pixel is proportional to the fixation density in that region. Figure 4 (right) shows the data from Figure 3 displayed as an image mask. The fixated regions of the calibration target are visible through the mask; the dark regions are image locations that were not fixated by the subject during the data collection, and the x's correspond to the actual nine calibration points.

## 4. EXPERIMENTS

The goal of the eye tracking experiments was to determine whether individuals scan images of the same category in similar ways (and if there are differences across different categories).

### 4.1. Image Data set

For the experiments we selected 50 color images from each of five different categories (Figure 5): *handshake* (two people standing near each other, shaking hands); *main object in uncluttered background* (a prominent object around the center of the image, on an uncluttered background); *crowd* (cluttered scenes with many people); *landscape* (natural landscapes, without people); and *miscellaneous* (still lives, and people). The *handshake*, *crowd*, and *main object* images were

collected from an on-line news service. The *landscape* and *miscellaneous* images were obtained from the collection of a photographer (one of the authors). All of the images used in the experiments had a resolution of approximately 240 x 160 pixels. Note however, that the important parameter is the angle subtended by each image (discussed in the next section).



**Figure 5.** Example images from each of the five categories used in the experiments, from left to right: *handshake*, *main object*, *crowd*, *landscape*, and *miscellaneous*.

## 4.2. Subjects

Ten volunteers (four females and six males, all undergraduate students) participated in the experiment. All subjects were native English speakers, naïve as to the goals of the experiment, and had not participated in eyetracking experiments in the past. Before beginning the experiment, subjects read and signed an informed consent form describing the eyetracking apparatus. The observers were told to observe the images, but no explanations were given regarding the goal of the experiment, or the number of image categories. Since some of the images were obtained from a news source, it is possible that some of the subjects had familiarity with the persons and places in the photographs. However, no distinctions were made in this regard. The subjects viewed a total of 250 images. The images were interleaved in random order, and each was viewed for four seconds. The experiment was broken down into two sessions, each consisting of 125 images and lasting approximately 8.5 minutes. While subjects' heads were not restrained during the sessions, they were instructed to maintain their gaze on the TV display. Subjects took a self-timed break in between the sessions, typically lasting ~5 minutes before beginning the second set of 125 images.

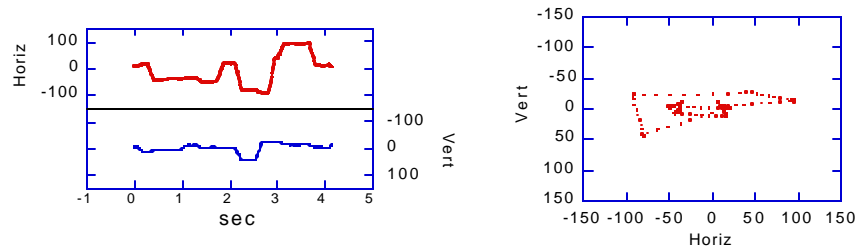
The goal was to determine the primary areas of interest in the image, so a viewing time was selected that was sufficient to allow several fixations (typically 8-12 fixations), yet not long enough to encourage the subject to scan the entire image. Pilot experiments indicated that a four-second exposure was appropriate. For visual examination of an image, the important viewing variable is the visual angle subtended by the image. The angular subtense of the image was selected to approximate that of an observer viewing a photograph in a magazine or newspaper (e.g., ~15 cm wide image field viewed at a distance of 33 cm).

Subjects were seated about 1 meter from an NTSC television monitor with screen dimensions ~53 cm x 40 cm. The images subtended a mean of 25 x 19 degrees of visual angle (the exact value varied with head position, but was within 10% of the stated value). The remote eyetracker system was adjusted to be just below the line-of-sight to get the best view of the eye without obscuring any portion of the monitor (Figure 1). Thresholds for pupil and corneal reflection discrimination were set for each subject to optimize the thresholding that is used to determine the pupil and corneal reflection centroids, as seen in Figure 2. The system was calibrated to each subject by instructing the subject to fixate each point in a rectangular calibration grid on the display. The raw pupil and corneal reflection centroids recorded at each calibration point, along with the location of those points in image coordinate space, are used to establish the relationship between measured pupil and corneal reflection position and point-of-gaze in the image plane.

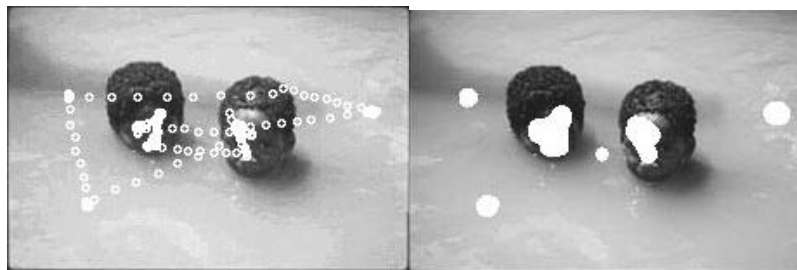
Figure 6 is a sample image from the 'miscellaneous' image class. Figure 7 shows the eye position as a function of time (left panel) and in two dimensions (right panel) for subject 6. While the pattern is less regular than the calibration set seen in Figure 3, it is clearly still made up of relatively long fixations separated by rapid saccadic eye movements. The two-dimensional plot also shows how the fixation patterns are tied to image content; while the eyes are moved across a broad area of the image, the majority of the trial is spent looking in a small number of image regions. The left panel of Figure 8 shows the fixation pattern superimposed over the target image, where each point represents gaze position at each video field (every 16.7 msec). The right panel shows the data mapped onto the image with one-degree circles indicating each fixation, defined as one-degree regions containing at least three data points (50 msec). Multiple fixations evident in the left panel result in larger indicators.



**Figure 6.** Example of image from ‘misc’ class (images were viewed in color)



**Figure 7.** Eye position trace for one subject for the image in Figure 6 in time (left) and image pixel units (right). Each point on the right represents a 16.7 msec sample



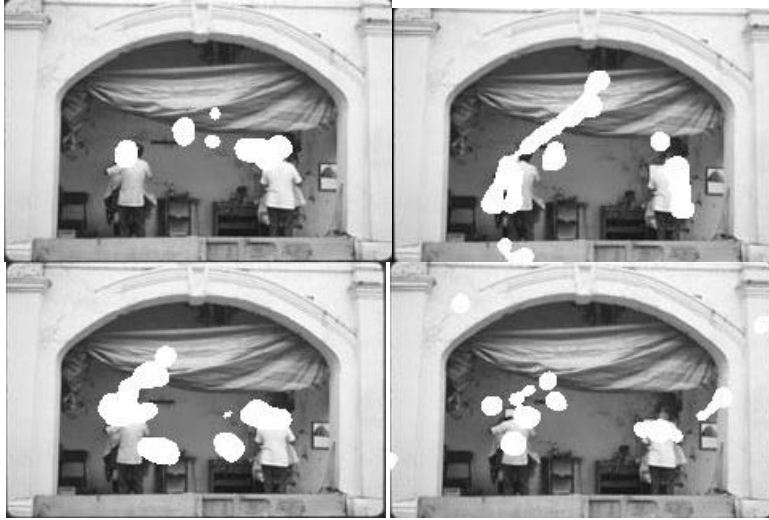
**Figure 8.** Scanpath overlay on display. Left panel indicates gaze position at 16.7 msec intervals. Individual fixations are indicated in the right panel with circles approximately one degree in diameter.

### 4.3. Eye Tracking Results

The experiments resulted in a fairly large amount of data, specifically eye tracking results for 10 subjects on 250 images in 5 categories. Since each subject viewed each image for approximately 4 seconds and the sampling rate of the eye tracker used is 60 Hz., we have an average of 250 data points per subject, per image, for a total of approximately 630,000 data points for the entire experiment.

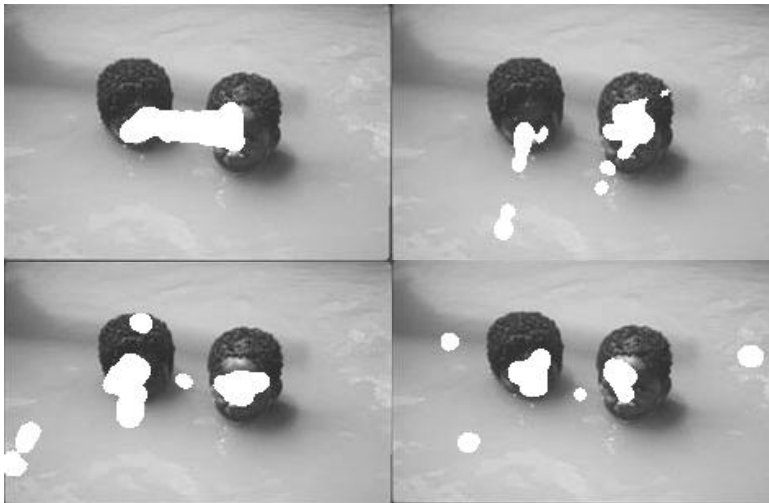
It is possible to mathematically process the acquired data (e.g., mathematically compare fixations of different subjects within an image category, etc.), as discussed in section 5. In this section, however, we focus on our own observations about the viewing patterns observed (i.e., scanpath, including fixations and saccades). In particular, we discuss, in the viewing patterns: (1) *within image variations* (similar/dissimilar patterns for an image, across several subjects); (2) *across image variations* (subject’s pattern depends strongly on the image); and (3) *within/across image category variations* (similar/dissimilar patterns for images in the same category, across several subjects). Observations based on the experiments are described next.

In general, viewing patterns were similar between subjects viewing the same image, though idiosyncratic behavior was evident. Figure 9 shows fixation plots for four subjects as they viewed the same image. All four subjects fixated on the two main figures in the image, but each made a number of other fixations not common with the other subjects. This example shows an image for which consistent patterns were found, across different individuals.



**Figure 9.** *Image 16* fixation mask overlay for subjects 2, 3, 5, and 6.

Figure 10 shows similar data for a different image. Again, the fixation density plots for the four subjects are similar.



**Figure 10.** *Image 20* fixation mask overlay for subjects 2, 3, 5, and 6.

As the two previous examples suggest, there are cases in which it is possible to find *consistent viewing patterns*, across *different individuals*, for a *given image*. The viewing patterns themselves, however, varied, for a single individual, depending on the specific image being observed. For example, Figure 11 shows the fixation patterns of a single subject viewing four different images, highlighting the *strong image-dependency of viewing patterns*.



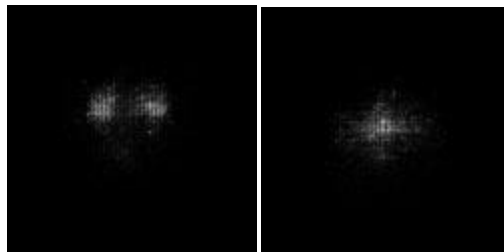


**Figure 11.** Fixation mask overlays for Subject 6 with *image 20, 16, 2\_2, & 35*.



**Figure 12.** Dissimilar viewing patterns for a given image. Fixations of four subjects on the same image.

In some cases, it was also evident that wide variations in viewing patterns occurred, for *different subjects* viewing the *same image*. Figure 12 illustrates one of those images for which there was a wide variation. In terms of categories, we found the most consistency in the *handshake* and *main object* classes (Figure 13), while there was very little consistency in viewing patterns in the remaining three categories (*landscape*, *crowd*, and *miscellaneous*). Note that in Figure 13, for illustration purposes, we plotted *all data points* for *all subjects*, for *all images* within each of the two categories. No distinction is made between saccades and fixations in these plots, but fixation concentrations are readily seen where there is a higher concentration of points. Note, for example, that for the *handshake* class there are two visible clusters, resulting from the two faces that appear in each image in that class, and that *patterns are different across categories*.



**Figure 13.** Data points for six subjects, for all images in the *handshake* category (left) and *main object* (right) category.

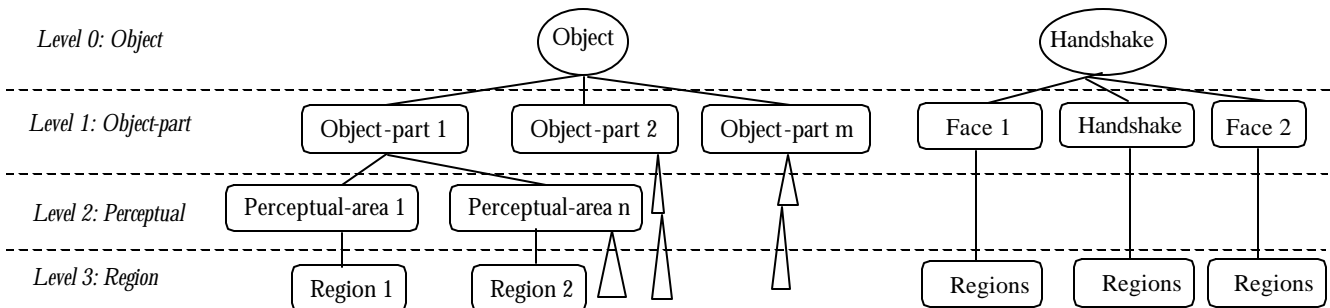
It is interesting to note that, in general, in the *handshake* class, subjects spent more time looking at the face on the left than at the face on the right. Additionally, it was somewhat surprising to find that in many cases the observers did not look (i.e., no data points occurred) at the handshake at all.

In summary, we found *images with consistent viewing patterns* (several subjects viewed the same image in a similar way), *images with inconsistent viewing patterns* (several subjects viewed the same image in different ways), and *strong image dependence* (the same subject used different patterns on different images). In addition, we found the most consistent viewing patterns in the *handshake* and *main object* image categories, and that there were *differences across categories*.

## 5. AUTOMATIC CLASSIFICATION

In the previous we discussed eye movement variations *for a subject, across images, and within/across categories*. One of the goals of our work is to determine whether results of eye tracking experiments like this one can be used in the construction of automatic classifiers. Therefore, data analysis across categories may be more useful because it could be used to construct classifiers for the classes studied (e.g., handshake). First, we will briefly describe the *Visual Apprentice* framework (VA) [10], and then explore different ways in which this data could be used to build classifiers with the framework.

One of the main principles of the VA is to allow users to construct their own classifiers through a simple training stage. First, the user collects example images/videos for the class he/she is interested in. Then, the images are automatically segmented (based on color and edge information) and provided to the user, who constructs an *object definition hierarchy* (using a simple graphical user interface) and labels the regions, in each example image/video, according to the hierarchy (see Figure 14). The hierarchy subjectively defined by the user models an object or scene using different levels: an *object* is composed of *object-parts*, which are composed of *perceptual areas*, which in turn contain *regions* (obtained from segmentation). The result of the training stage, then, is a set of labeled regions for each node of the hierarchy defined by the user. For example, if the user is constructing a handshake classifier, all regions corresponding to the handshake (e.g., two faces, handshake) would be labeled in every sample image. The system then computes feature vectors (e.g., color, texture, location, etc.) for each of the example regions, and uses different learning algorithms (e.g., decision trees, nearest-neighbor, etc.) to construct classifiers for each node of the hierarchy defined by the user. When new images are going to be classified, first they are automatically segmented, and the classifiers constructed from the training stage are applied according to the hierarchy. For example, face region classifiers are used to find candidate face regions, which are then grouped by the face object-part classifier, and so on (details in [10]).



**Figure 14.** A generic *object-definition hierarchy* in the *Visual Apprentice* (left), and an hierarchy for a handshake classifier (right).

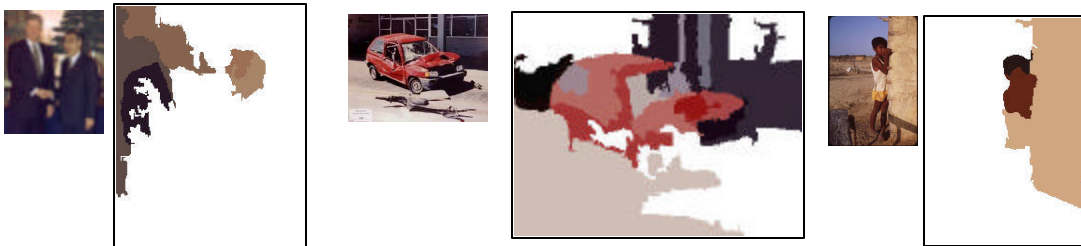
During training, the user manually clicks on regions that correspond to the *object definition hierarchy* he defined. Therefore, one possible use of the eye tracking data for each class (instead of manual labeling), consists of using the fixation points (e.g., for all subjects for each image) to select the relevant training regions. It would be expected (as shown by our experiments), in the handshake class, for example, for the observers to fixate their gaze on the faces of the people shaking hands, and possibly on the handshake itself. A preliminary analysis of the data, however, showed that such selection is not trivial and can present several complications. First, visual acuity and the accuracy of the eye tracker (in pixels) must be considered when performing the selection, so two fixation points within a certain distance (10 pixels in our setup) are indistinguishable. In the current VA setup, the user manually clicks anywhere inside the regions that he wants to label, so a single pixel location (in x,y image coordinates) is sufficient to accurately select and label regions. Since a fixation point in the eye tracking experiments corresponds to several data points from the eye tracker (e.g., a fixation point might be defined as lasting 167 msecs, or 10 data points), those points must be clustered in some way and a fixation center (or fixation area) must be computed. Figure 11 shows an example of fixation areas that were obtained from the eye tracking results, and Figure 15 shows a set of image regions (obtained from automatic segmentation) automatically selected by the fixations of 6 of the subjects. In other words, regions obtained from the automatic segmentation, that overlap (on the image) with fixations, are selected and shown in Figure 15. As it can be clearly seen in the figure, some of the relevant regions are not selected (i.e., those that would be selected by a human training the system, like the handshake regions missing from the image in Figure 15), while some irrelevant regions are selected. Errors in the segmentation algorithm are easily corrected by the human when he selects the correct regions for the objects/parts that he is interested in. Using the eye tracking results directly, however, poses a challenge in terms of the selection of the regions - the way in which the eye tracking data is used plays an important role. In [18], the authors compared automatically selected regions of interest with regions selected by human observers' eye

movements, and proposed a technique to cluster fixations and compare them across different viewers. Although in our experiments it is possible to apply the same approach to cluster points, decisions regarding the use of the data (e.g., clustering algorithm, criteria to group fixations of different observers, etc.) are not trivial and can have a strong impact on the construction of automatic classifiers. In the particular framework we are examining, variations in those criteria could result in the selection of different regions. One of the factors is that humans may fixate on certain areas of objects (e.g., eyes in a face), but those areas may not yield the best results in terms of selecting the relevant regions.

**Figure 15.** An example of the regions selected by the fixations of the human observers.

In addition to selecting regions automatically, based on fixations, it is possible to modify the algorithm of the VA and use the additional information provided by the eye tracking experiments. In that case, fixations could be used to give certain regions more weight than others (this could be easily included in the VA framework), and more importantly to also include regions selected by the saccades. Furthermore, scanpath order could be included in the classification strategy (i.e., classifiers would be applied according to scanpath order).

Alternatively, entirely new regions could be extracted from the training data (not using automatically segmented regions, but instead masks produced by the eye tracking experiments) and used to construct the classifiers. The actual classification approach, in that case, would also have to be modified so at the classification stage the regions would be extracted according



to the training data (instead of extracting regions without any class-specific knowledge and then trying to classify them).

## 6. CONCLUSIONS AND FUTURE WORK

We presented eye tracking experiments that show the variations in eye movements (i.e., fixations and saccades) across ten different individuals for color photographs of 5 different categories: *handshakes* (two people shaking hands), *crowds* (cluttered scenes with many people), *landscapes* (nature scenes without people), *main object in uncluttered background* (e.g., an airplane flying), and *miscellaneous* (people and still lives). In particular, we found, in the viewing patterns: (1) *within image variations* (similar/dissimilar patterns for an image, across several subjects); (2) *across image variations* (subject's pattern depends strongly on the image); and (3) *within/across image category variations* (similar/dissimilar patterns for images in the same category, across several subjects). Specifically, we found more consistent patterns within the *handshake*, and *main object* categories. More importantly, we found the patterns were different between those categories (*handshake/main object*).

Using results from the experiments, we suggested ways in which this type of data can be used in the construction of automatic image classifiers. In particular, we examined the *Visual Apprentice*, a system in which image classifiers are learned (using machine learning) from user input as the user defines a multiple level object definition hierarchy based on an object and its parts (*scene, object, object-part, perceptual area, region*), and labels examples for specific classes (e.g., handshake). The resulting classifiers are applied to automatically classify new images.

Although our analysis is preliminary, the results of the experiments are encouraging and suggest that it may be possible to use eye tracking data in the construction of automatic classifiers. Analysis of the data (e.g., selection and clustering of fixation points, use of scan order, etc.), however, plays a very important role since the criteria used can have a strong effect on the classifiers being built. Our future work includes more analysis of the data, and construction of automatic classifiers using these eye tracking results.

### 6.1. Acknowledgments

The authors wish to thank Diane Kucharczyk, Amy Silver, and the persons that participated in the experiments.

## 7. REFERENCES

- [1] Becker, W. "Metrics," In *The Neurobiology of Saccadic Eye Movements*. Goldberg, M.E. & Wurtz, R.H., Eds. Elsevier Science Publishers, 1989.
- [2] Buswell, G.T., *How People Look at Pictures*, University of Chicago Press, Chicago, 1920.
- [3] Chang, S.-F., Smith, J.R., Beigi, M. and Benitez, A. "Visual Information Retrieval from Large Distributed On-line Repositories," *Communications of the ACM*, 40(12):63-71, December, 1997.
- [4] Collewijn, H., Steinman, R.M., Erkelens, C.J., Pizlo, Z., & van der Steen, J., "Effect of freeing the head on eye movement characteristics during three-dimensional shifts of gaze and tracking," Chapter 64 in *The Head-Neck Sensory Motor System*, Berthoz, A., Graf, W., & Vidal, P.P., Eds. Oxford University Press, 1992.
- [5] Cornsweet; Tom N. ,Crane; Hewitt D., *US Patent US3724932*, 1973.
- [6] Epelboim J., Steinman R.M., Kowler E., Pizlo Z., Erkelens C.J., Collewijn H., "Gaze-shift dynamics in two kinds of sequential looking tasks", *Vision Res.* 37: 2597-2607, 1997.
- [7] Gaarder, K. R., *Eye Movements, Vision, and Behavior*, John Wiley & Sons, New York, 1975.
- [8] Gould, J. D., "Looking at Pictures", in *Eye Movements and Psychological Processes*, edited by R. A. Monty and J. W. Senders, John Wiley & Sons, New York, 1976.
- [9] Guedry FE, Benson AJ., "Tracking performance during sinusoidal stimulation of the vertical and horizontal semicircular canals," In: *Recent Advances in Aerospace Medicine*, Busby, D E (Ed.). D. Reidel Publ. Co., Dordrecht, Netherlands, 1970.
- [10] Jaimes, A. and Chang, S.-F. "Automatic Selection of Visual Features and Classifiers," in *proceedings of SPIE Storage and Retrieval for Media Databases 2000*, vol. 3972:346-358, San Jose, CA, January 2000.
- [11] Jaimes A. and Chang, S.-F., "A Conceptual Framework for Indexing Visual Information at Multiple Levels," in *proceedings of SPIE Internet Imaging 2000*, vol. 3964:2-15. San Jose, CA, January 2000.
- [12] Land M.F., Furneaux, S. *The knowledge base of the oculomotor system*. Phil Trans R Soc Lond B 352: 1231-1239, 1997.
- [13] Lipson, P., "Context and Configuration Based Scene Classification," *Ph.D. thesis*, MIT Electrical and Computer Science Department, September 1996.
- [14] Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, San Francisco, 1982.
- [15] Noton, D. and Stark, L.W. , "Scanpaths in Saccadic Eye Movements while Viewing and Recognizing Patterns," *Vision Research* 11 (9), 929-42, 1971
- [16] Paek, S. and Chang, S.-F. "InLumine: An Image Classification System Based on Probabilistic Reasoning," invited paper, *International Conference on Image Processing (ICIP 2000), special session on Image Content Extraction and Description for Multimedia*, Vancouver, Canada, September 10-13, 2000.
- [17] Pelz, J.B., Canosa, R., Babcock, J., Kucharczyk, D., Silver, A., and Konno, D., "Portable Eyetracking: A Study of Natural Eye Movements," in *proceedings of SPIE, Human Vision and Electronic Imaging* , San Jose, CA, 2000.
- [18] Privitera C.M., and Stark, L.W., "Evaluating Image Processing Algorithms that Predict Regions of Interest", in *Pattern Recognition Letters* 19, pp. 1037-1043, 1998.
- [19] Privitera, C.M., and Stark, L.W., "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(9):970-981, Sept. 2000.
- [20] Robinson, D.A., "A method for measuring eye movements using a scleral search coil in a magnetic field," *IEEE Trans Bio-Med Electron*, BME-10, 137-145, 1963.
- [21] Rybak, I.A., Gusakova, V.I, Golovan, A.V., Podladchikova, L.N., and Shevtsova, N.A., "A model of attention-guided visual perception and recognition," *Vision Research*. 38:2387-2400, 1998.
- [22] Schill K., Umkehrer E., Beinlich S., Zetzsche C, Deubel H, Pöppel E., "A hybrid system for scene analysis with saccadic eye movements: learning of feature relations", *European Conf. on Visual Perception*, Oxford, England, 1998.
- [23] Smith, J.R. and Chang, S.-F., "Multi-stage Classification of Images from Features and Related Text," in *proceedings Fourth DELOS workshop*, Pisa, Italy, August, 1997.
- [24] Solso, R.L., *Cognition and the visual arts*, MIT Press, Cambridge, Mass., 1994.
- [25] Szummer M. and Picard, R.W., "Indoor-Outdoor Image Classification," in *proceedings of IEEE International Workshop on Content-based Access of Image and Video Databases*, pages 42-51, Bombay, India, 1998.
- [26] Vailaya, A., Figueiredo, M., Jain, and Zhang, H.J., "Content-Based Hierarchical Classification of Vacation Images," in *proceedings of IEEE Multimedia Computing and Systems*, vol. 1:518-523, Florence, Italy, June 1999.
- [27] Yarbus, A.F. *Eye Movements and Vision*, New York, Plenum Press., 1967.