Human-Computer Interaction for Complex Pattern Recognition Problems

Jie Zou and George Nagy

Rensselaer Polytechnic Institute `zouj@alum.rpi.edu; nagy@ecse.rpi.edu`

Abstract

We review some applications of human-computer interaction that alleviate the complexity of visual recognition by partitioning it into human and machine tasks to exploit the differences between human and machine capabilities. Human involvement offers advantages, both in the design of automated pattern classification systems, and at the operational level of some image retrieval and classification tasks. Recent development of interactive systems has benefited from the convergence of computer vision and psychophysics in formulating visual tasks as computational processes. Computer-aided classifier design and exploratory data analysis are already well established in pattern recognition and machine learning, but interfaces and functionality are improving. On the operational side, earlier recognition systems made use of human talent only in preprocessing and in coping with rejects. Most current content-based image retrieval systems make use of relevance feedback without direct image interaction. In contrast, some visual object classification systems can exploit such interaction. They require, however, a domain-specific visible model that makes sense to both human and computer.

1 Introduction

The goal of visual pattern recognition during the past fifty years has been the development of automated systems that rival or even surpass human accuracy, at higher speed and lower cost. However, many practical pattern recognition applications involve: random noise and systematic variations in the patterns, inaccurate and incomplete prior information, limited and unrepresentative training samples, the mostly invincible challenge of segmentation, non-discriminating and unreliable features, many classes, as well as complex decision boundaries. Therefore, automatic recognition systems often require

years of research and development in order to achieve fast and accurate classification. Some applications, e.g., optical character recognition, fingerprint identification, and target recognition, have met with modest success after decades of research and development, but many theoretical and practical problems remain. Face recognition has been intensively studied since 60's, but is still considered unsolved [43]. Automated recognition in many other domains, such as petroglyphs, shards, arrowheads, flowers, birds, skin diseases, and so on, requires too much development for a limited market, or is too complex to be accommodated by the current methodologies.

A divide-and-conquer strategy for visual recognition should partition such domains into components that are relatively easier for both human and machine. There are pronounced differences between human and machine cognitive abilities. Humans excel in gestalt tasks, like object-background separation. We apply to recognition a rich set of contextual constraints and superior noise-filtering abilities. Computer vision systems, on the other hand, still have difficulty in recognizing "obvious" differences and generalizing from limited training sets [25]. We can also easily read degraded text on which the best optical character recognition systems produce only gibberish [1][15].

Computers, however, can perform many tasks faster and more accurately. Computers can store thousands of images and the associations between them, and never forget a name or a label. They can compute geometrical properties like higher-order moments whereas a human is challenged to determine even the centroid of a complex figure. Spatial frequency and other kernel transforms can be easily computed to differentiate similar textures. Computers can count thousands of connected components and sort them according to various criteria (size, aspect ratio, convexity). They can quickly measure lengths and areas. They can flawlessly evaluate multivariate conditional probabilities, decision functions, logic rules, and grammars. On the other hand, the study of psychophysics revealed that humans have limited memory and poor absolute judgment [30].

There is a growing consensus among experts to advocate interactive approaches to difficult pattern recognition problems. As early as 1992, a workshop organized by US National Science Foundation in Redwood, California, stated that "computer vision researchers should identify features required for *interactive image understanding*, rather than their discipline's current emphasis on automatic techniques" [27]. A more recent panel discussion at the 27th AIPR Workshop also emphasized "... the needs for computer-assisted imagery recognition technology" [29].

We concur with the suggestions of combining human and computer cognitive abilities to cope with the complexity of practical pattern recognition problems. To lay down some guidelines for integrating human-computer interaction with pattern recognition, we first briefly review human and machine visual perception and selected findings in psychophysics. We then discuss three human-computer interaction methodologies used in pattern recognition and image retrieval: Exploratory Data Analysis (EDA), Relevance Feedback

for content-based image retrieval, and Computer Assisted Visual InterActive Recognition (CAVIAR) for visual pattern classification.

Our conjectures on what aspects of visual pattern recognition are easy and difficult for humans and computers are set forth in Table 1. The remainder of this chapter attempts to justify some of these conjectures and explores their implications for the design of pattern recognition systems.

Table 1. Comparison of relative strengths of human and machine in diverse aspects of visual pattern recognition

Human	Machine
dichotomies	multi category classification
figure-ground separation	
part-whole relationships	
salience	
	nonlinear, high-dimensional classification boundaries
extrapolation from	
limited training samples	
broad context	
	store and recall many labeled reference patterns
	accurate estimation of statistical parameters
	application of Markovian properties
	estimation of decision functions from training samples
	evaluation of complex sets of rules
	precise measurement of individual features
	enumeration
gauging *relative* size and intensity	
detection of significant differences	
between objects	
	computation of geometric moments
	orthogonal spatial transforms (e.g., wavelets)
	connected component analysis
	sorting and searching
	rank-ordering items according to a criterion
	additive white noise
	salt & pepper noise
colored noise, texture	
non-linear feature dependence	
	determination of local extrama in high-D spaces
global optima in low dimensions	

2 Human and Machine Visual Perception

Visual perception is defined in [33] as: *"the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect."*

Visual perception has been studied separately by psychologists, performing experiments on sighted organisms; computer scientists, writing programs that extract and transform optical information; and neuroscientists, studying the structure and function of the visual nervous system. Recently, these three approaches converged to form a central idea of visual perception: *visual perception is a kind of computation*. In living organisms, eyes and brains perform visual perception through complex neural information processing, and in principle, visual perception can also be achieved by video cameras and programmed digital computers. This idea enables psychologists, computer scientists, and neuroscientists to relate their findings to each other in the common language of computation, and generates a new branch of cognitive science: *vision science* [33].

2.1 Machine Visual Perception

After Alan Turing defined the fundamental model of computation [42], Turing himself and many others realized that it may be possible for Turing machines to simulate human intelligence. This idea gave rise to the field of artificial intelligence.

The goal of the subfield of artificial intelligence called computer vision is to develop programmed computers, which can interpret the environment visually. The mathematical approach to creating working computer vision programs was most clearly and effectively articulated by David Marr and his colleagues at MIT [28]. Marr's work dominated computer vision research for the last two decades, and a great deal of progress has been made. Nevertheless, machine perception still lags far behind human visual perception with respect to breadth of visual stimuli, perspective invariance, partial occlusion, tracking, learning, and uneven illumination (highlights and shadows).

2.2 Human Visual Perception

Classical psychological theories about human visual perception include structuralism, gestaltism, ecological optics, and constructivism [33]. In the field of visual pattern recognition, there are two theories, recognition-by-components (RBC) [8] and view-based recognition [40]. Unfortunately, they do not agree with each other. The debate centers on the form of the representation mediating three-dimensional object recognition.

Recognition-by-components assumes that perceptual processes derive the constituent parts of an object and represent each of those parts with a simple geometric volume, or *geon*. An object representation, or geon structural

description, consists of geons corresponding to the two or three most salient parts of an object and the spatial configuration in which the geons are connected. This structural description is represented without regard for the specific viewpoint of the observer. Recognition is performed by recovering the 3D geon model from the input image.

In contrast, the key idea of the theory of view-based, or sometimes called image-based, recognition, is that object representations encode visual information as it appears to the observer from a specific vantage point.

After several years of debate between proponents of the two theories [9][40][10], most researchers now agree that these theories can be considered as different points in a single continuum. RBC, or viewpoint-invariant theory, does depend on viewpoint to some extent because single representations normally encode only some viewpoints of an object. A number of representations may be needed to cover all possible views of the object. Similarly, view-based theory doesn't propose that all view points are needed for recognition. In any case, how humans recognize objects is still not clearly understood.

2.3 Psychophysics

Image quality can be described in purely physical terms, but optimal image quality can only be described with reference to the performance of an imaging task. The relation between physical image quality and diagnostic performance is the borderland between physics and psychology known as psychophysics. Psychophysics is the quantitative branch of the study of perception, examining the relations between observed stimuli and responses and the reasons for those relations. Psychophysics is based on the assumption that the human perceptual system is a measuring instrument yielding results (experiences, judgments, responses) that may be systematically analyzed [6].

The psychophysical aspects of visual pattern recognition, including color, shape, perspective, and illumination, have been the objectives of sustained study for centuries. These studies revealed many facets of the amazing human capacity for visual perception, which are important guidelines for the design of systems that integrate human-computer interaction with pattern recognition.

Attneave pointed out the importance of redundancy in visual stimulation [4]. Visual perception is a kind of economical abstraction of the redundant visual stimuli. He proposed ten principles of abstraction in human visual perception, and mentioned that *"information is concentrated along contours."*

In a celebrated article, George A. Miller summarized many psychophysical experiments and claimed that human absolute judgment is poor, limited to distinguishing only about seven categories within any single dimension - tone, loudness, taste (saltiness), length, area, hue, brightness, curvature [30]. He also noted that we can accommodate only about seven objects in our span of attention, and that our short-term memory is limited to about seven items. Nevertheless, we can recognize hundreds or thousands objects because we can make relatively coarse absolute judgments of several features simultaneously.

We can also trick our short-term memory by recoding (so we can memorize a string of 30 zeros and ones by recoding it as 7 letters).

Ashby and Perrin argued that the perceptual effect of a stimulus is random, but that on any single trial it can be represented as a point in a multidimensional space [3]. The perceived similarity is determined by distributional overlap. The perceptual space itself is fundamental. The difference is in the nature of the response function of the subject. In a recognition task, the decision process divides the space into response regions, one associated with each response. On a particular trial, the subject's response is determined by the region into which the perceptual sample falls. This theory of human recognition is analogous to the theory of statistical pattern classification.

3 Exploratory Data Analysis

Human-computer interaction was first exploited for pattern recognition under the title of *Exploratory Data Analysis* (EDA). The increasing use of graphical user interfaces in the 70's attracted much research to visual data analysis for designing pattern classification systems.

The seminal works in EDA are those of Ball and Hall [5], Sammon [37], Tukey and Mosteller [41][31]. Chien summarized early work on interactive techniques in data acquisition, pattern analysis, and the design of pattern classification schemes in a monograph, *Interactive Pattern Recognition* [12]. Over the years, the techniques of EDA have been steadily enhanced [38], [46].

Most EDA techniques are graphical in nature, with only a few quantitative techniques. High-dimensional data is incomprehensible to humans, but we have superior ability to understand configurations of data in 1D, 2D, and 3D, and the evolution of changes over time. The primary goal of EDA is to maximize the analyst's insight into the underlying structure of a data set by projecting it into a 1D, 2D, or 3D subspace for ease of human visual assimilation. Exploratory Data Analysis facilitates understanding the distribution of samples in a fixed feature-space in order to design a classifier, but stops short of operational classification.

Recently, *Mirage*, an open source Java-based EDA software tool, was implemented at Bell Laboratories [23][24]. Besides supporting the basic EDA functions, i.e., projecting the data into one, two, or higher dimensional subspace, and displaying them in tables, histograms, scatter plots, parallel coordinate plots, graphs, and trees, Mirage facilitates the analysis and visualization of the correlation of multiple proximity structures computed from the same data. All functions are available through an elaborate Graphical User Interface, but a small interpretive command language is provided for repetitive, large-scale data analysis. In Mirage, the users can also configure several plots at the same time, and perform classification manually or automatically.

4 Relevance Feedback in Content-Based Image Retrieval

Content-based image retrieval (CBIR) has been the subject of widespread research interest. Many prototype systems have been implemented, such as QBIC [18], Virage [7], Photobook [34], MARS [26], PicToSeek [21], PicHunter [16], Blobworld [11], and so on. Several surveys have also been published [32][2][36][39] over the years. Content-based image retrieval attempts to retrieve images similar to the query image from an image database. It is motivated by the fast growth of image databases, which requires efficient search schemes.

Fully automatic content-based retrieval does not yet scale up to large heterogeneous databases. Human-computer interaction is an important component of all content-based image retrieval systems. *Relevance Feedback* is broadly adopted in content-based retrieval systems for human-computer interaction, and has been found effective [35][14].

A typical CBIR system with relevance feedback operates as follows: the user submits a query image, which is somewhat similar to the desired image (or a sketch of a desired image) and specifies which properties, e.g., overall color, overall texture, and so on, are important to the query. Upon seeing the query results, the user designates the retrieved images as acceptable or unacceptable matches in order to provide more information to the retrieval algorithm. This process is iterated until the user finds the desired image or gives up the task.

A major shortcoming of the above interface is that the user cannot share the computer's view of the image. Without knowing whether the query image was properly understood (processed) by the machine, the user can only wonder what went wrong when the retrieval result was unsatisfactory. The developers of "Blobworld" realized this drawback, and suggested that the CBIR systems should display its representation of the submitted and returned images and should allow the user to specify which aspects of that representation are relevant to the query. In the Blobworld image retrieval system, the user composes a query by submitting an image, then views its Blobworld representation, selects the blobs to match, and finally specifies the relative importance of the blob features.

5 Computer Assisted Visual InterActive Recognition

Reject correction may be the most common example of interacting with a classifier. Almost all classification algorithms admit some means of decreasing the error rate by avoiding classifying ambiguous samples. The samples that are not classified are called "rejects" and must, in actual applications, be classified by humans. Reject criteria are difficult to formulate accurately because they deal with the tails of the statistical feature distributions. Furthermore, most classifiers generate only confidence, distance, or similarity measures rather than

reliable posterior class probabilities. Regardless of the nature of the classifier, at least two samples must be rejected in order to avoid a single error, because any reject region must straddle the classification boundary, near which there must be a 50-50 mixture of two classes[1] [13].

The efficiency of handling rejects is important in operational character and speech recognition systems, but does not receive much attention in the research literature. Keeping the human in the loop was recently also demonstrated in the domains of face and sign recognition (the extraction and recognition of text in natural scenes). However, it was confined to preprocessing, i.e., establishing the pupil-to-pupil baseline [45] or a text bounding box [22][47]. In these approaches, human intervention occurs only at the beginning or at the end of the recognition process, i.e., segmenting objects or performing other kinds of preprocessing *before* machine operations, or handling rejects *after* machine operations. There is little communication between the human and the computer.

The motivation of our recently-proposed methodology for interactive visual pattern recognition, Computer Assisted Visual InterActive Recognition (CAVIAR), is simply that it may be more effective to establish a seamless human-computer communication channel to make parsimonious use of human visual talent *throughout* the process, rather than only at the beginning or end [48][49]. The vehicle for human-machine communication is a *visible model*.

Unlike content-based image retrieval, which is usually on a broad domain, each CAVIAR system addresses only a narrow domain. In the broad domain of content-based image retrieval, no effective way has been found so far to interact with arbitrary images. In pattern classification with CAVIAR, the domain-specific geometrical model, e.g., a set of contours and critical feature points, plays the central role in facilitating the communication (interaction) between the human and the computer. The key to effective interaction is the display of the automatically-fitted adjustable model that lets the human retain the initiative throughout the classification process.

CAVIAR is designed to allow the human to quickly identify an object with a glimpse at the candidate samples that were ranked near the top by the computer. Avoiding having to look at many low-ranked classes is clearly most effective in many-class classification problem. Because of the nature of the human-computer interaction, CAVIAR is more appropriate for low-throughput applications, where higher accuracy is required than is currently achievable by automated systems, but where there is enough time for a limited amount of human interaction.

---

[1] This is a lower bound under the assumption of uniform cost of errors, because some samples may occur near the intersection of more than two regions. Therefore error-reject curves have an initial slope of at least -0.5, which increases further as the fraction of rejects is increased to lower the error rate.

Traditionally, visual pattern recognition includes three subtasks: segmentation, feature extraction, and classification. As mentioned, psychophysical studies suggest that the information is concentrated along object contours [4], therefore the pattern contours are important for classification. Locating the precise object boundary (strong segmentation) is generally considered too difficult and unreliable [39]. On the other hand, it may not even be necessary for visual pattern recognition. Several content-based image retrieval systems circumvent strong segmentation by locating only the approximate object boundary (weak segmentation). CAVIAR also gives up strong segmentation for weak segmentation based on a family of rose curves specified by six parameters. If the automatically constructed rose curve does not fit well, the user can easily adjust the model parameters by dragging a few control points. In CAVIAR, this model describes not only the object contour, but also some components of the object (petals).

In Blobworld, the Blobworld representation, which is an approximate segmentation of the object, is displayed in order to avoid misunderstandings between the human and the computer. This is much better than leaving the users wonder what went wrong when a machine error occurs. However, apprehending the machine errors without being able to correct them is also frustrating. In CAVIAR, the user can not only *view* the machine's understanding (processing) of the image, but also *correct* the machine errors if necessary.

In CAVIAR, the first generic computer-vision task, segmentation, becomes model building. Therefore a CAVIAR process has three subtasks: *model building*, i.e., generating a model instance, which explains the image according to the domain model; *feature extraction*, i.e., measuring discriminative object properties according to the constraints provided by the model instance; and *classification*, i.e., assigning a category label to the object.

Model building in CAVIAR-flower consists of fitting a rose curve to the flower. First a circle is fitted to the foreground (the flower to be recognized) based on the expected difference in color between flowers and background (leaves, dirt). The boundary propagates to high-gradient locations penalized according to their distance from the initial circle [50]. Finally, a rose curve is fitted to the propagated boundary (Fig. 1). The area delineated by the rose curve constrains feature extraction to the discriminative parts of the picture.

The model instances constructed in this manner are not always correct (Fig. 2). After decades of extensive research on this topic, many researchers now agree that automatic image segmentation is not likely to correspond consistently to human expectations except in narrow domains. On the other hand, humans perform image segmentation smoothly, accurately, and with little ambiguity. So we believe that model building should be, at least for now, subject to human correction.

Most laypersons don't understand computer vision features like moment invariants or wavelets. Humans find it difficult to visualize computer vision feature vectors and the geometry and topology of high-dimensional feature spaces. Furthermore, lay users are seldom familiar with all the distinguishing
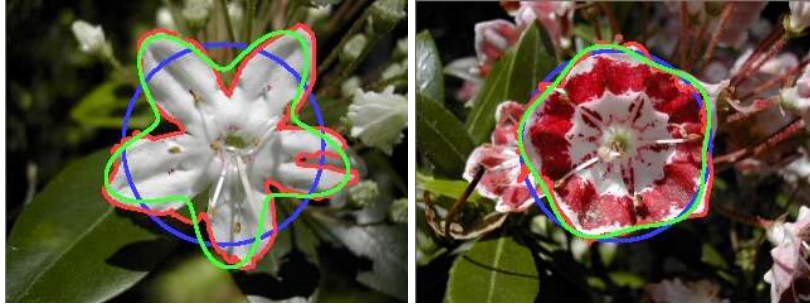
Fig. 1. Automated model construction in CAVIAR. Initial circle (blue), detailed segmentation (red), and parametric rose curve (green) segmentation of two flowers. The rose curve serves as a visible model of the computer's concept of the unknown flower. It guides the computer in the extraction of classification features.

properties of the various classes, and therefore cannot judge the adequacy of the machine-proposed decision boundary. As mentioned, psychophysical studies also point out that human absolute judgment is poor, effective only in an approximately seven-interval scale [30]. Machines, on the other hand, can compute complicated features very accurately and very fast. So, in CAVIAR, feature extraction should be performed primarily by machine, without human intervention. However, *indirect* human refinement of feature values, by adjusting the CAVIAR model instance throughout the process, does promote faster and more accurate classification.

The whole CAVIAR process can be modeled as a finite state machine (Fig. 3). The computer tries its best to estimate an initial model for the unknown sample and calculate its similarity to the training samples that belong to each class. Representative training pictures are displayed in the order of computer-calculated similarities. The current model is also displayed, so that the user can correct it if necessary. Any correction leads to an update of the CAVIAR state: the remaining unadjusted model parameters are re-estimated, and all the candidates are re-ordered. Fig. 2 shows a difficult example, where the picture is blurred.

In summary, CAVIAR operates on four entities: (1) the unknown image, (2) the parameters of a visible geometrical model instance, (3) the feature vector extracted from the image according to the model, and (4) the list of class labels ranked according to the similarity of the corresponding reference pictures to the query picture. Interaction takes place through the model. The process terminates when the user assigns a label to the unknown image.

The *image* is a 2D color or gray scale picture, as in the conventional visual pattern recognition systems.

The *geometrical model* consists of critical points and parametric curves, which are both abstract and visual descriptions of the contours of the pattern components and of the geometrical relations among them. The model estima-
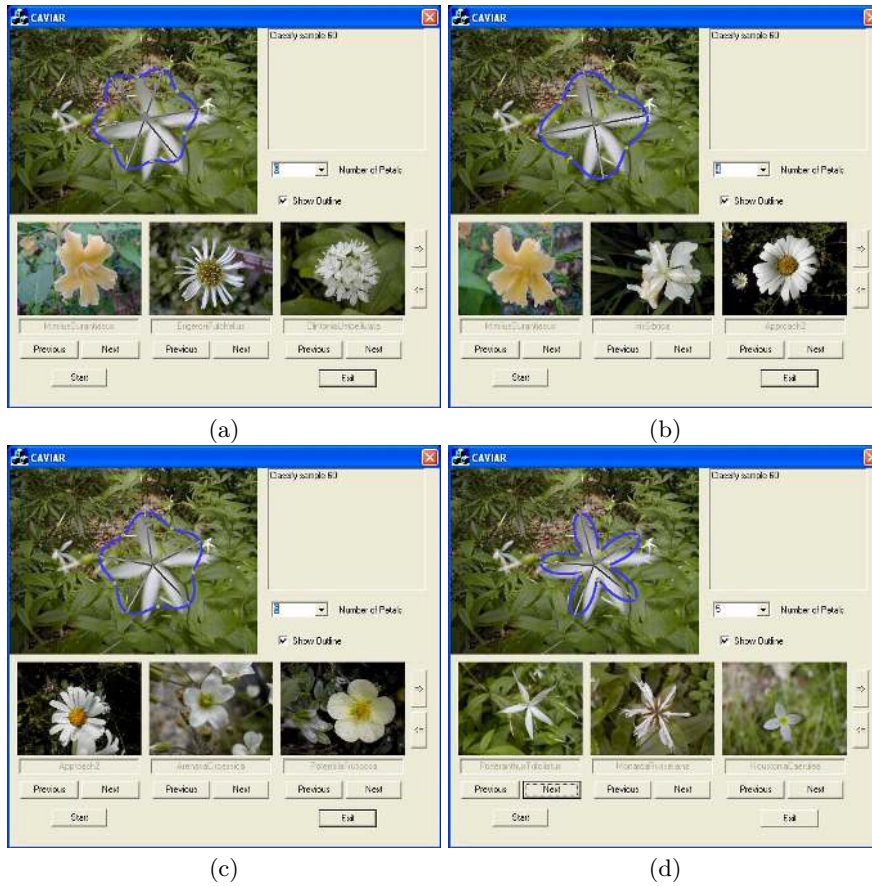
Fig. 2. An example of CAVIAR flower recognition. (a) The initial automatic rose curve estimation and indexing are bad because the picture is blurred: the correct candidate does not appear among the top three. (b) The user adjusts the center. (c) The user adjusts the petal number. (d) After the user adjusts the inner radius, the computer displays the correct candidate. (It is almost never necessary to make this many adjustments.)

tion algorithm can use any segmentation algorithm (edge based, region based, hybrid optimization) to locate these critical points and curves. The human's understanding of the image can be communicated to the machine by adjusting a few critical points. The machine can then re-estimate the remaining model parameters in a lower-dimensional space for improved classification.

The *feature vector* is a set of features for classifying patterns. It is extracted from a picture according to the model instance. The features, which may include shape (derived from the model parameters), color, texture, and other attributes, exist only in a high-dimensional space invisible to the user.
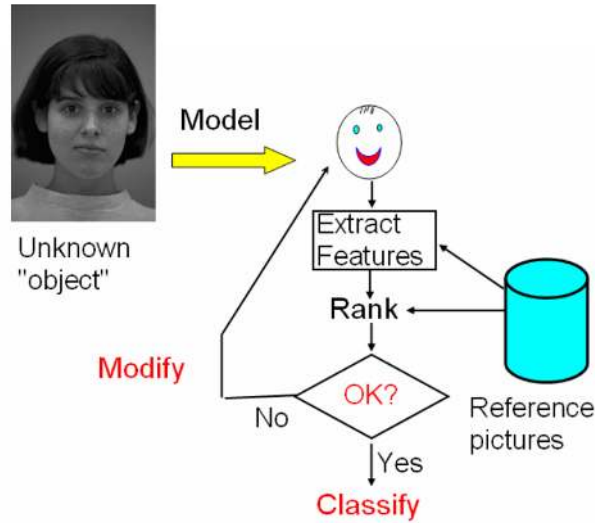
Fig. 3. CAVIAR Flowchart, showing transitions betweens automated modeling and human modification of the model, followed by browsing and classification.

The *class label list* is a machine-ordered list of candidates based on the feature vector. It governs the display of reference pictures. The user assigns a particular label to the unknown object by clicking on one of the displayed reference pictures.

The model parameters constitute a vector random variable. Human and machine observations of model parameters are also random variables, with human model estimates much better than machine estimates. The feature vector is related to the model parameters through a deterministic function. Human adjustments reduce the bias and variance of the feature vector by reducing the bias and variance of the model parameters. More accurate features generally improve classification.

The CAVIAR methodology has been applied to flower recognition on a database with 612 samples from 102 classes. Experiments with 36 naïve subjects show the following properties of CAVIAR systems [48][49].

- Human-computer communication through a geometrical model is effective. Combining human and machine can significantly reduce the recognition time compared to the unaided human, and significantly increase the accuracy compared to the unaided machine.
- The CAVIAR system can be initialized with a single training sample per class, but still achieve high accuracy (because there is a human in the loop).
- The CAVIAR system shows self-learning. The user classified samples, along with the user adjusted model instances, are added to the reference

set of labeled samples to effect unsupervised decision-directed approximation [17]. Although the samples may not be 100% correctly classified, automatic recognition still improves, which in turn helps users to identify the patterns faster. The performance based on just-classified samples is almost as good as with the same number of ground-truth training samples. Instead of initializing the CAVIAR system with many training samples, we can trust the system's self-learning ability (although, of course, the initial users would need more time).

- Users remember the examples to become "connoisseurs" of the specific family. With CAVIAR, lay persons need little practice to become faster than unaided "connoisseurs".

CAVIAR methodology can be applied to many other tasks. Interactive face recognition under head rotation, occlusion, and changes of illumination and facial expression is very challenging, but of great practical importance (Fig. 4). CAVIAR has also been ported to a stand-alone PDA, and to a pocket PC with a wireless link to a host laptop. Interaction with the visual model through a stylus is faster than with a mouse. We expect some applications, like the identification of skin diseases and other medical diagnoses based on visual observations, to be more appropriate for mobile versions of CAVIAR [51]. With mobile system, taking additional photos from a different perspective or distance, or under different illumination, could be extremely useful. Whether the resulting information should be combined at the pixel, feature, or classifier level is an unresolved research issue.

As do all classifiers, CAVIAR systems collect, in the course of operation, mostly-correctly-labeled samples. As more and more samples are accumulated, they can be used to improve the machine's performance either directly, by machine learning, or by studying the accumulated training samples and upgrading the classification and learning algorithms.

CAVIAR could offer suggestions to its users. For example, it could suggest which model parameters to adjust, or request the operator to inspect further candidates because the top candidates have low confidence values. We do not allow CAVIAR to make such suggestions, because its judgment so far is worse than the human's, therefore most of its suggestions would just annoy the user. Eventually machines may, of course, earn suggestion privileges.

6 Discussion

Fifty years of sustained research has increased our appreciation of the fundamental difficulty of some visual recognition tasks and our admiration for the complex, multi-level biological systems that accomplish these tasks with apparent ease. At the same time, technological developments have enabled human-computer interaction at a level that could be found earlier only in science fiction. Within the 0.5 second response time that experts consider acceptable, a laptop or a PDA can perform calculations that used to require
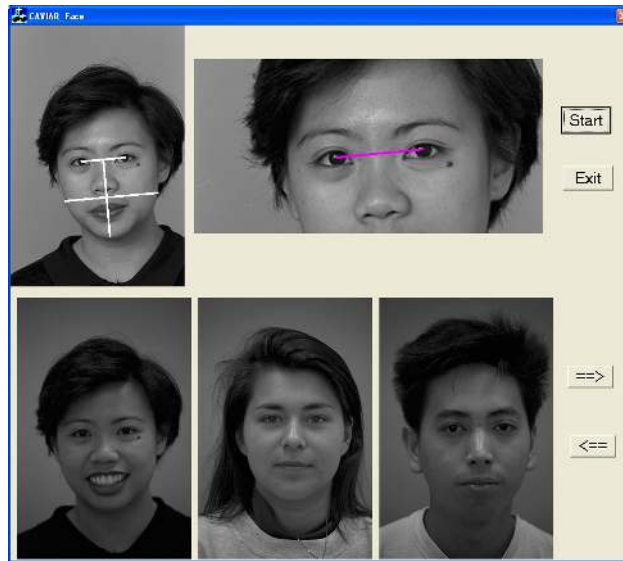
Fig. 4. CAVIAR-Face GUI and model. The eye region is enlarged to allow accurate location of the crucial characteristic points at the pupils.

hours or days on a mainframe, and display the results instantly at high resolution, color and, if need be, motion. It is therefore now highly appropriate to seek joint human-computer solutions, at least as a temporary expedient, to recognition problems that have so far eluded an entirely algorithmic approach.

An interactive solution is not appropriate for all classification tasks. Character and speech recognition require the rapid interpretation of long pattern sequences rather than isolated patterns, while "real time" in many military applications is much less than human reaction time. But there are also many applications, like face, fingerprint, or flower recognition and medical diagnosis, where isolated patterns are recognized only sporadically, and where image acquisition takes long enough to dominate any real need for quasi-instantaneous classification. The advent of PDAs and camera phones with internet access and plug-in cameras increases the scope for interactive personal recognition systems.

In many other fields of engineering, sophisticated and mature CAD software is widely used to mock-up proposed solutions, prepare and access test data, simulate experiments, check design constraints, perform routine calculations, and retrieve, modify and incorporate previously designed modules. As such systems evolve, more and more intricate tasks are relegated to the computer, but the design engineer always remains in charge. In pattern recognition and machine learning, specialized computer-aided design tools have been slow to emerge. Nevertheless, interactive design and analysis tools have

proved useful for improved understanding of the data even in domains where no intervention can be admitted at run time.

Interactive data analysis can lead to the selection of better features for classification, the identification of subsets of the data for which special provisions are necessary, the discovery of correlations or redundancies between features or patterns, and the detection of mislabeled items. Human-computer interaction is especially appropriate for discovering complex hidden information, and for accumulating training samples which, according to the no free lunch theorem [44] and the bias-variance dilemma [20][19], are the only two factors that can really improve classification performance.

Computer-assisted labeling has, of course, always been used to prepare training sets for classifier design, and often to classify rejects. It seems likely that with further advances in active and semi-supervised learning, these labeling operations will be more closely integrated with the algorithmic classification process itself. This may be most easily accomplished within the existing systems for exploratory data analysis.

At the operational, "real time" level, we have seen that there are two options. The more common one, almost universally used in content-based image retrieval, is to let the computer do the best it can, and tell it where it fails. The machine then can use the set of positive and negative samples that it has just acquired to improve its next try. The information provided by the user is limited to one bit per picture, because he or she has no knowledge of how the computer made its decision and where it went wrong. Some research attempts to organize postage-stamp displays of the retrieved images in a configuration that suggests their putative relationships.

The other paradigm is CAVIAR, where users interact with the picture directly through a parametric model. Such a model must be constructed for every new application domain. For applications that justify the investment of effort, it is an effective approach to interactive classification.

The differences between peripheral and in-the-loop human intervention exist in other fields as well. In chess and checkers, relevance feedback would only tell the machine whether it has won or lost the game (which of course it can deduce by itself), while a CAVIAR approach could offer comment on every move. Although using some computer help is quite popular in the current on-line format of postal chess competition, much AI research runs counter to our philosophy of letting the machine help the user, rather than vice-versa.

We summarize our main points. In some domains, the accuracy of automatic classification remains far below human performance. Human and computer abilities differ, and we are making progress in understanding the differences. A good interactive visual recognition system capitalizes on the strengths of both. It must establish effective two-way communication. In narrow domains simplified models of the real world can bridge the semantic gap between man and machine. The human must be able to exercise gestalt perception in his or her customary visual domain which, in addition to natural scenes, includes several well-established sets of symbols. The computer should

take full advantage of its almost unlimited memory and of its ability to solve huge but essentially repetitive problems. Further research is needed on how to translate complex multi-dimensional internal data to a form where any fallacies and failures of the current computer model can be readily apprehended and corrected.

References

1. von Ahn L, Blum M, Langford J (2004) Telling humans and computers apart automatically. Communications of ACM 47(2):57-60
2. Aigrain P, Zhang H, Petkovic D (1996) Content-based representation and retrieval of visual media: a state of the art review. Multimedia Tools and Applications 3:179-202
3. Ashby FG, Perrin NA (1988) Toward a unified theory of similarity and recognition. Psychological Review 95(1):124-150
4. Attneave F (1954) Some informational aspects of visual perception. Psychological Review 61:183-193
5. Ball GH, Hall DJ (1970) Some implications of interactive graphic computer systems for data analysis and statistics. Technometrics 12:17-31
6. Baird JC, Noma E (1978) Fundamentals of scaling and psychophysics. John Wiley & Sons
7. Bach J, Fuller C, Gupta A, Hampapur A, Horowitz B, Humphrey R, Jain R, Shu C, (1996) The Virage image search engine: an open framework for image management. Proc. SPIE Storage and Retrieval for Image and Video Databases IV, San Jose CA, 76-87
8. Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychological Review 94:115-147
9. Biederman I, Gerhardstein PC (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. Journal of Experimental Psychology: Human Perception and Performance 19:1162-1182
10. Biederman I, Gerhardstein PC (1995) Viewpoint-dependent mechanisms in visual object recognition: reply to Tarr and Bülthoff (1995). Journal of Experimental Psychology: Human Perception and Performance 21:1506-1514
11. Carson C, Belongie S, Greenspan H, Malik J (2002) Blobworld: image segmentation using expection-maximization and its application to image querying. IEEE Trans. Pattern Analysis and Machine Intelligence 24(8):1026-1038
12. Chien YT (1970) Interactive Pattern Recognition. Marcel Dekker Inc.
13. Chow CK (1970) On optimum recognition error and reject tradeoff. IEEE Trans. Information Theory 16:41-46
14. Ciocca G and Schettini R (1999) Using a relevance feedback mechanism to improve content-based image retrieval. Proc. Visual '99: Information and Information Systems 107-114
15. Coates AL, Baird HS, Fateman RJ (2001) Pessimal print: a reverse Turing test. Proc. Int. Conf. on Document Analysis and Recognition 1154-1159
16. Cox IJ, Miller ML, Minka TP, Papathomas TV, Yianilos PN (2000) The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. IEEE Trans. Image Processing 9(1):20-37

17. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley
18. Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P (1995) Query by image and video content: the QBIC system. IEEE Computer 28(9):23-32
19. Friedman JH (1997) On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discorvery 1(1):55-77
20. Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Networks 4(1):1-58
21. Gevers T, Smeulders AWM (2000) PicToSeek, combining color and shape invariant features for image retrieval. IEEE Trans. Image Processing 9(1):102-118
22. Haritaoglu I (2001) Scene text extraction and translation for handheld devices. Proc. IEEE conf. on Computer Vision and Pattern Recognition 2:408-413
23. Ho TK (2002) Exploratory analysis of point proximity in subspaces. Proc. 16th Int. Conf. on Pattern Recognition
24. Ho TK (2002) Mirage: a tool for interactive pattern recognition from multimedia data. Proc. of Astronomical Data Analysis Software and Systems XII
25. Hopgood A (2003) Artificial intelligence: hype or reality? IEEE Computer 36(5):24-28
26. Huang TS, Mehrotra S, Ramachandran K (1996) Multimedia analysis and retrieval system (MARS) project. Proc. 33rd Annual Clinic on Library Application of Data Processing Digital Image Access and Retrieval
27. Jain R (1992) US NSF workshop on visual information management systems
28. Marr D (1982) Vision: a computational investigation into the human presentation and processing of visual information. W.H. Freeman and Company, New York
29. Mericsko RJ (1998), Introduction of 27th AIPR workshop - advances in computer-assisted recognition. Proc. of SPIE 3584
30. Miller G (1956), The magical number seven plus or minus two; some limits on our capacity for processing information. Psychological Review 63:81-97
31. Mosteller F and Tukey J (1977) Data analysis and regression. Addison-Wesley
32. Nagy G (1986) Image database. Image and Vision Computing 3:111-117
33. Palmer SE (1999) Vision science, Photons to Phenomenology. MIT Press
34. Pentland A, Picard RW, Sclaroff S (1996) Photobook: content-based manipulation of image database. International Journal of Computer Vision 18(3):233-254
35. Rui Y, Huang TS, Ortega M, Mehrotra S (1998) Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Trans. Circuits and Systems for Video Technology 8(5):644-655
36. Rui Y, Huang TS, Chang SF (1999) Image retrieval: current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation 10:39-62
37. Sammon JW (1970) Interactive pattern analysis and classification. IEEE Trans. Computers 19:594-616
38. Siedlecki W, Siedlecka K, Sklansky J (1988) An overview of mapping techniques for exploratory pattern analysis. Pattern Recognition 21:411-429
39. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence 22(12):1349-1380
40. Tarr MJ, Bülthoff HH (1995), Is human object recognition better described by geon-structural-descriptions or by multiple-views? comment on Biederman and

Gerhardstein (1993). Journal of Experimental Psychology: Human Perception and Performance 21:1494-1505

41. Tukey J (1977) Exploratory data analysis, Addison-Wesley
42. Turing AM (1936) On computable numbers with an application to the entscheidungs problem. Proc. of the London Mathematical Society 42:230-265
43. Willing R (2003) Airport anti-terror systems flub tests face-recognition technology fails to flag 'suspects'. USA Today, 3A, Sept. 2, 2003. http://www.usatoday.com/usatonline/20030902/5460651s.htm
44. Wolpert DH (1995) The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In: Wolpert DH (eds) The Mathematics of Generalization. 117-214, Addison-Wesley, Reading, MA
45. Yang J, Chen X, Kunz W (2002) A PDA-based face recognition system. Proc. the 6th IEEE Workshop on Applications of Computer Vision 19-23
46. Vesanto J (1999) SOM-based data visualization methods. J. Intelligent Data Analysis 3:111-126
47. Zhang J, Chen X, Yang J, Waibel A (2002) A PDA-based sign translator. Proc. the 4th IEEE Int. Conf. on Multimodal Interfaces 217-222
48. Zou J (2004) Computer assisted visual interactive Recognition. Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, New York
49. Zou J, Nagy G (2004) Evaluation of model-based interactive flower recognition. Proc. of 17th Int. Conf. Pattern Recognition
50. Zou J (2005) A model-based interactive object segmentation procedure. Proc. of IEEE 2005 Workshop on Applications of Computer Vision
51. Zou J, Gattani A (2005) Computer-assisted visual interactive recognition and its prospects of implementation over the Internet. IS&T/SPIE 17th Annual Symposium, Internet Imaging VI