VISIBLE MODELS FOR INTERACTIVE PATTERN RECOGNITION

Jie Zou, National Library of Medicine, Bethesda, MD

George Nagy, DocLab, Rensselaer Polytechnic Institute, Troy, NY

ABSTRACT

The bottleneck in interactive visual classification is the exchange of information between human and machine. We introduce the concept of the *visible model*, which is an abstraction of an object superimposed on its picture. For a narrow domain, like flowers or faces, it may be as simple as an outline of the entire object to be classified or a set of characteristic points. For every new object to be classified, the machine proposes a model subject to constraints (learned from a training set) that help avoid implausible abstractions. The visible model is not by itself sufficient for classification, because it contains no intensity, color or texture information. However, using features extracted from the picture based on the model, the classes can be rank ordered according to the similarity of the unknown picture (in a hidden high-dimensional feature space) to the labeled reference pictures. If the rank ordering appears unsatisfactory, the operator may modify the model, resulting in the extraction of new features, and a new rank ordering. The interaction continues until the operator confirms a satisfactory match. The new object, with its model and label, is added to the reference database. Comprehensive experiments show that interactive recognition of flowers and faces is much more accurate than automated classification, much faster than unaided human classification, and that both machine and human performance improve with use.

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. But most operational systems still require human intervention at some stage, because more accurate classification is needed than is achievable by current algorithms. In this study, we focus on the role of interaction in narrow domains where higher accuracy is required than is currently achievable by automated systems, but where there is enough time for limited human interaction. The pronounced differences between human and machine capabilities suggest that a system that combines human and machine abilities can, in some situations, outperform both. The key to efficient interaction is a *visible model,* overlaid on the unknown picture, which provides two-way communication between man and machine.

As early as 1992, a workshop organized by The National Science Foundation in Redwood, California, recommended that “computer vision researchers should identify features required for *interactive image understanding*, rather than their discipline's current emphasis on automatic techniques” [Jain92]. A panel discussion at the 27th AIPR Workshop asserted “... the needs for Computer-Assisted Imagery Recognition Technology” [Mericsko98]. Kak’s ICPR’02 keynote emphasized the difficulties facing fully automated model-based vision [Kak02].

We did not find any previous work advocating image-based interaction through a domain-specific model, which is the principal contribution of this paper. To demonstrate the viability of this approach, we developed two *CAVIAR* (Computer Assisted Visual InterActive Recognition) systems where the operator may interact with the image

whenever the computer's response is unsatisfactory. All or part of the model may be constructed by the computer, under constraints derived from training data to avoid implausible models. The operator may correct the model at any time, and retains the initiative throughout the classification process. The interaction may extract directly some features for classification, and indirectly benefit the extraction of other features by improving the fit of the computer-proposed model.

The visibility of the model is critical, because most of us have no intuition for topological and geometric relations in high-dimensional feature space [Nagy04]. Even we did, we could not improve the decision boundaries without being familiar with all the classes. Although we cannot interact with the classifier itself, we can easily evaluate the fit of the visual model to the unknown object, and correct it if necessary. Moreover, our judgment of the adequacy of the machine-proposed class prototypes, compared visually to the unknown object, is far superior to any classifier-generated confidence measure. In contrast to classification, judging whether two pictures represent the same class does not require familiarity with every class.

The machine makes subsequent use of the refined model to improve not only its own statistical model-fitting process, but also its internal classifier. Classifier adaptation is based on decision-directed approximation [Duda01, Nagy 66, Baird94, Veera04]. The automated parts of the system gradually improve, decreasing the need for human intervention. As an important byproduct, the human's judgment of when interaction is beneficial also improves. Our experiments on a stress dataset of over 1000 flowers and on a standard face database [Feret] demonstrate both phenomena.

2. PRIOR WORK

In the broad domains of Content-Based Image Retrieval (CBIR), *relevance feedback* has been found effective [Rui98]. Interaction has been, however, necessarily limited to the initial query formulation and the selection of acceptable and unacceptable responses [Darman92, Cox00, Carson02], because no effective way has yet been found to interact with arbitrary images in a broad domain. A major shortcoming of the usual relevance feedback is the absence of information about the computer's view. Without knowing that the query image is not properly understood (and processed) by the machine, the user can only wonder what went wrong. The designer of "Blobworld" [Carson02] first realized this drawback, and suggested that the CBIR system should display its representation of the submitted and returned images. However, it is still frustrating to realize the errors in the machine presentation without being able to correct them. In CAVIAR, we not only let the user *view* the machine-processed image, but also provide means to *correct* errors. This two-way communication is achieved through the domain specific *visible model*.

Another relevant area of research is *Active Learning*, which makes use of human intervention to reduce the number of training samples that the classifier needs to achieve a target error rate [MacKay92, Cohn96, Lizotte03, Nguyen04]. In Active Learning, however, the operator does not interact with images of unknown objects, but only assigns a class label to patterns designated by the system. One approach is to select "informative" patterns close to the current decision boundary, because these patterns affect the boundary most. In CAVIAR, patterns near the boundary also require the operator's attention if their reference counterparts are not ranked among the top candidates.

However, the operator can opt to refine the model rather than assign a label to the query image.

In many operational systems, image acquisition, framing, and initial segmentation are the responsibility of the operator. For example, in document processing, every scanned page is checked by the operator before being sent to the modules for layout analysis, OCR, and document interpretation [Bradford91, Dickey91, Klein04]. In camera-based text recognition, the operator defines a bounding box [Haritaoglu01, Zhang02]. In face recognition, the operator sets the pupil-to-pupil baseline [Yang02]. The operator then intervenes again at the end, to deal with objects which cannot be confidently classified by the automated system, i.e., *rejects* [Sarkar02]. In contrast, we propose that the human and the machine take turns throughout the classification process, each doing what they do best.

The notion of characteristic points or landmarks is central in numerical taxonomy, where “the goal of data analysis is to identify the smallest subset of external body characteristics that uniquely diagnoses the new species as distinct from all of its known relatives” [Chen05]. Chen et al. give examples of the use of landmarks for classifying species of the fish genus Carpiodes. Homologous landmarks are also well established in aerial photography [Drewniok97], in radiography [Yue05], and in dactylography [Nilsson02]. The homologies or correspondences of interest range from translation, rotation, and scale invariance to affine and projective transformations.

The recognition of flowers has been investigated in [Das99, Saitoh04], while face recognition is a growth industry with entire conferences devoted to it [Zhao03]. Early attempts at face recognition in the 1960s and ‘70s were primarily based on the geometry

of local features [Goldstein71, Kanade77]. Modern face recognition methods can be generally divided into two categories: *holistic matching methods* and *local matching methods*.

The principle of holistic methods is to project and compare the faces in a low-dimensional subspace in order to avoid the curse of dimensionality. The subspace is constructed using Principal Component Analysis (PCA) [Kirby90, Turk91], Linear Discriminant Analysis (LDA) [Belhumeur97, Etemad97, Swets96], or Independent Component Analysis (ICA) [Bartlett98]. Local matching methods [Pentland94, Wiskott97, Ullman02, Martinez02] first locate several facial features (components), and then classify the faces by comparing the local statistics of the corresponding facial features.

We noticed that comparing corresponding local *regions* (instead of local facial features) uses shape information implicitly but efficiently (Figure 1) because it does not attempt to juxtapose facial features other than landmarks. Therefore, the automatic part of our CAIVAR-Face system compares many local regions. It outperforms the upper-bound of all four categories of FERET tests [Phillips00] and was reported in detail in [Zou06a, Zou06b].

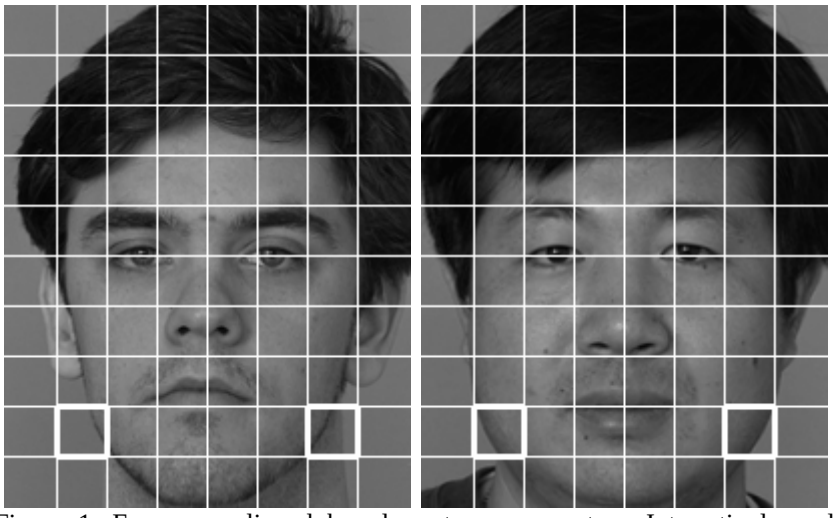

Figure 1. Faces are aligned based on two eye centers. Interestingly and somewhat counterintuitively, although the regions marked with thick boxes don't cover much of the face in the left image, they are actually very useful for distinguishing skinny faces from round faces.

3. THE CAVIAR MODEL

As mentioned, the important distinction between our approach and other research on image classification and content-based image retrieval is our reliance on a *visible model* to mediate human-computer communication. The visible model consists of a minimal set of perceptually salient landmark points (pixels) that establish a homology between two images. The desired homology is the structural correspondence between imaged objects of the same class. The computer vision and image processing communities have devoted considerable effort to automating the location of landmarks, but so far automated methods cannot generally match human accuracy. CAVIAR accepts the performance of the best available landmark-location algorithm, and lets the operator correct its output interactively when necessary.

The homology induced by the visible model allows mapping a region (a compact set of pixels) from one image into another. In CAVIAR, unlike in numerical taxonomy and some face classification methods [Wiskott97], the relative locations of the landmarks are not compared directly. They serve only to define the similarity transformation required for *registering* (juxtaposing) pairs of images. However, selected landmarks in a visible model may also serve as control points for parametric curves: in CAVIAR-Flower, they specify a *rose curve* that separates foreground from background. The landmarks may thus be considered as implicit specifications of structurally equivalent segmentations. Displaying line segments connecting or passing through landmarks makes these points more visible to the operator (as in CAVIAR-Face), which in turn promotes rapid adjustment.

As in the case of earlier manual approaches, the landmarks are used for computing *features* for determining the similarity of the query to each class of reference data. Each feature extracted from a selected region of the query image is compared to the (previously extracted) feature from the corresponding region of each reference image. What distinguishes CAVIAR from previous work is the interactive refinement of the visible model – the landmarks – *according to the results of the classification.*

Feature extraction in CAVIAR transforms a given set of pixels into a (scalar- or vector-valued) measurement. It performs dimensionality reduction with variable-length input and fixed-length output. As always, features are sought that preserve the distinguishing characteristics of the classes but are invariant to within-class variations. Due to the lack of universal methods for finding discriminative features, they are formulated by the system designer. The homology specified by the visible model ensures that the features extracted from images of the same class are *commensurable*. Ideally, the feature values extracted from same-class images are identical, but different from those of the other classes.

The model mediates only a restricted set of information. It does not tell the computer anything about the rich perceptions that lead the operator to correct or approve the model, and it does not tell the operator where the resulting feature vectors are in high-dimensional feature space. This design choice is the result of our awareness of the computer's lack of gestalt perception, and of human limitations in quantifying visual features and multivariate probabilities [Miller56]. The operator monitors gestalt tasks like object-background separation, and applies to recognition a rich set of contextual constraints and superior noise filtering abilities [Palmer99]. The machine computes

geometric and histogram moments, conditional probability distributions, and rank orders based on similarity or distance functions. It also stores all the reference images, labels, feature vectors and the associations between them. The interaction itself can be modeled by a simple finite-state machine.

The above formulation of the visible model leads to two evaluation criteria: (1) the Top-N error rate of automated classification with the features obtained using accurately instantiated visible models, and (2) the time required by the expected population of users to refine the automatically generated visible models as necessary. It is clear that both criteria depend not only on the characteristics of the visible model, but also on the selected feature-classifier combination. We can therefore evaluate a visible model only *in context*, as is inevitably the case for all the individual components of a classification system. Furthermore, the relative weight of the two criteria in the overall evaluation depends on the cost model of the underlying task. We report both the error rate of the classifier without human model refinement, and the time required for refinement.

Figure 2 shows examples of our flower and face models. We restrict the interaction to isolated points, because color and intensity are difficult to modify with a mouse or stylus. In effect, we allow the user to point and drag, but not to paint or shade. A line drawing is superimposed on the picture to let the operator judge whether a computer-suggested model fits the unknown object. These models are constructed automatically, and corrected interactively only when necessary. Recall, however, that the model is only an abstraction: for classification both the human and the computer must have access to the entire pixel array.

A model instance need not depict faithfully intensity, color, or texture edges. A poorly fitting model may suffice to classify an “easy” object. Conversely, even an

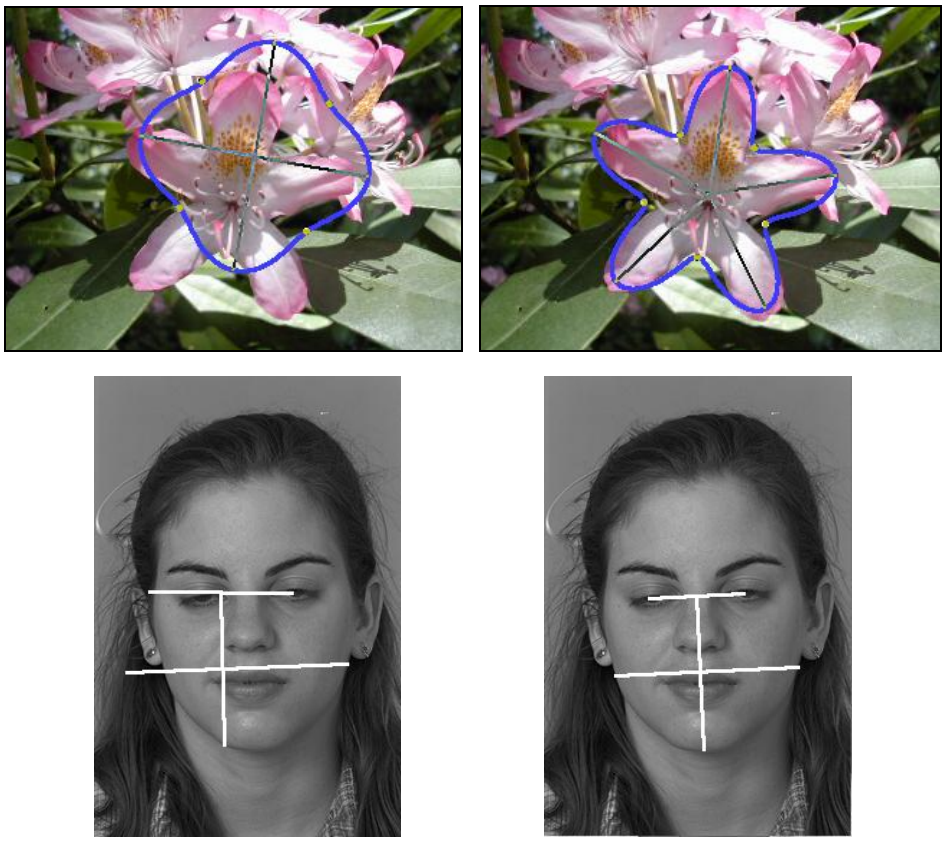

Figure 2. Examples of CAVIAR-flower (top) and CAVIAR-face (bottom) models, before and after human adjustment. Here automatic model construction failed because of overlapping flowers and partially closed eyes, respectively.

accurate model may result in ambiguous features. (One consequence of the role of the model in our system is that there can be no “ground truth” for it. Several models, or none, may lead to features that cause the correct candidate to be ranked on top.) The computer displays, in addition to the visible model, a set of reference pictures ranked according to the posterior class probabilities of the unknown object. The operator can then correct the model if a poorly specified similarity transformation caused obvious mismatches between the current model and the unknown object. Two CAVIAR graphic user interfaces (GUIs) are shown in Figure 3.

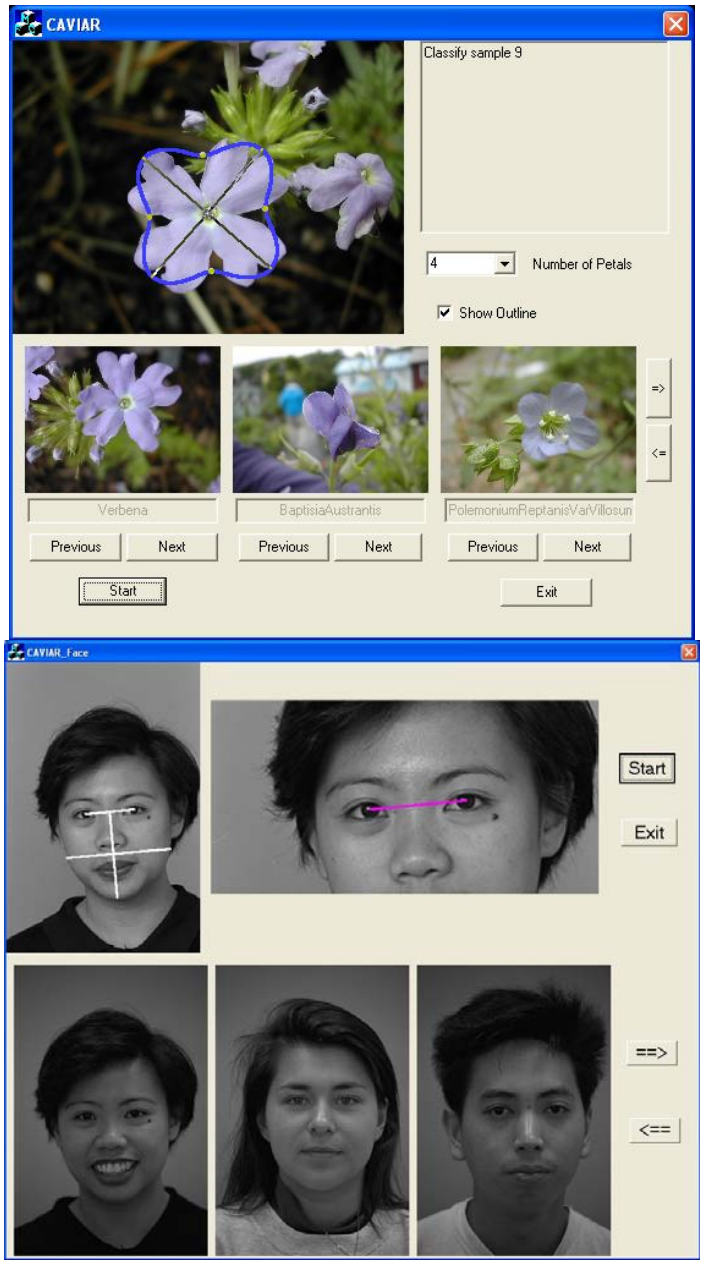

Figure 3. CAVIAR-flower (top) and CAVIAR-face (bottom) graphic user interface. In CAVIAR-Face, because accurate pupil location is important, an enlarged view is provided for adjustment.

4. CAVIAR-Flower and CAVIAR-Face systems

Model building

The visible model of CAVIAR-Flower (Figure 2) is a *rose curve* with 6 parameters, which are estimated with prior probabilities learned from a training set. First a circle is

fitted based on the expected difference between the foreground (flower) and background (leaves, dirt) colors. The boundary propagates to high-gradient locations penalized according to their distance from the initial circle [Zou05a]. Finally, a rose curve is fitted to the propagated boundary.

The visible model of CAVIAR-Face (Figure 2) consists of pixels at the centers of the eyes (*pupils*), at the bottom of the chin (*chin*), and below the ears (*jowls*, for lack of a better word). The configuration of these 5 characteristic points is restricted by ratios learned from a training set. The face region is found by a coarse template search (analogous to locating flowers with a circle template). Higher-resolution templates then locate (sometimes incorrectly) the five characteristic points of the visible face model in the vicinity of their expected locations.

In both CAVIAR-Flower and CAVIAR-Face, interactive model correction requires positioning the cursor to "acquire" some landmark, and then dragging it to a preferred location. The machine will then re-estimate the visible model based on the user's input.

Feature extraction

Features that characterize the unknown object are extracted from the image according to the model. Indeed, some model parameters may themselves serve as features. In CAVIAR-Flower, the 8 features are the two similarity-invariant parameters of the rose curve, and the first three moments of the hue and saturation histograms of the region enclosed by the curve [Zou04a].

In CAVIAR-face, the face is aligned based on the five landmarks, and then divided into a large number of local regions. These local regions serve as features.

Rank ordering (classification)

Strictly speaking, no automated classification takes place. The classes are ordered according to their features' similarity with the unknown. In CAVIAR-flower, they are ordered according to the Euclidian distance of the unknown features from the nearest feature vector of each class. In CAVIAR-face, the local regions from the unknown image are compared against corresponding local regions of every reference face. The local regions are sorted according to their match scores. The class with the best match is assigned Rank 1. The classes are then ordered by their total rank (the *Borda Count* [Ho94]) computed over all local regions.

In both CAVIAR-flower and CAVIAR-face, when the reference pictures of top 3 candidates are displayed, the operator decides whether to (1) accept one of the displayed classes by clicking on it, or (2) modify the model superimposed on the picture of the unknown object, or (3) inspect lower-ranked candidates ("browse") until a good match is found. The operator need not be able to classify the unknown object, but only to decide whether it matches one of the displayed reference pictures.

The model, the features, and even the classifier, are domain dependent. The *model* must be based on visible and readily discernible vertices and edges. The *features* must address properties of the objects that differentiate the classes. The choice of *classifier* is dictated by the number of classes, the number of features, the range and distribution of feature values, and the number of available reference samples per class. Only the interactive recognition system architecture that we propose (Figure 4) is general across different domains.

5. EVALUATION OF CAVIAR-FLOWER

We could not use pictures from any of the many excellent flower sites on the web because none have more than one or two samples per specie, and labeling conventions, background, and resolution differ too much from site to site. We therefore collected a database of 1078 flowers from 113 species, mostly from the New England Wildflower Garden [NEWFS]. Our system was developed on a subset of 216 flowers with 29 classes [Nagy02] and evaluated on a new subset of 612 flowers, consisting of 102 classes with 6 samples per class [Zou04b]. For classification, the photos, taken at the lowest resolution of a Canon Coolpix 775 camera, were further reduced to 320 x 240 pixels.

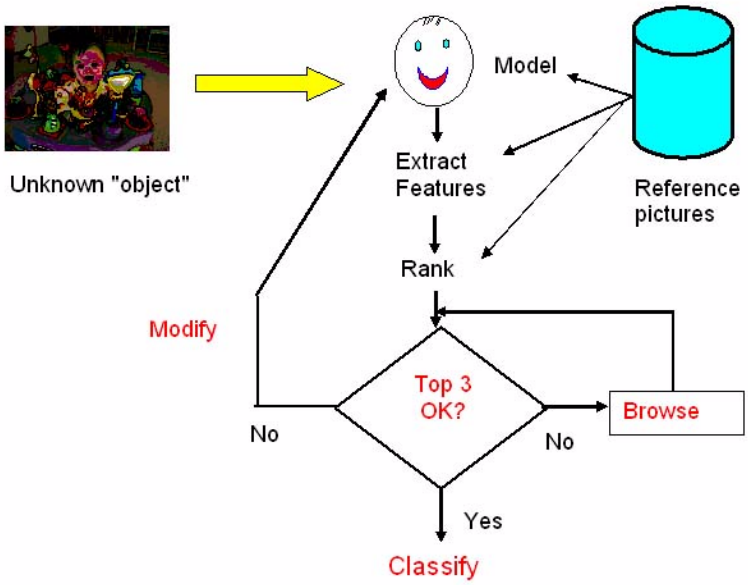

Figure 4. CAVIAR system architecture. Operator interactions shown in red.

The pictures were taken against complex backgrounds (dirt, weeds, and other flowers of the same or other species), under highly variable illumination (sharp shadows on foreground or background, specular reflections, saturation of some of the color channels), and poor imaging conditions (blur, incomplete framing), without necessarily a clear view of the camera viewfinder screen. The color distribution is not uniform, but most of our flowers are yellow, white, red, or blue. Several pictures contain multiple, tiny,

overlapping flowers. Our database, including labels and segmentation, is freely available on *http://www.ecse.rpi.edu/doclab/flowers*.

Experimental protocol

We asked 30 naive subjects (male and female adults without any connection to our department) to classify, as rapidly and as accurately as possible, one flower of each of 102 different categories. Each subject first viewed an instructional PowerPoint presentation, and then classified the flower images. The order of the 102 flower images was randomized for each subject to avoid class-specific effects of human learning. None of these test images was used in training the CAVIAR system used by that subject. Each session, consisting of instruction and classification, lasted about one hour.

The 30 sessions addressed five different tasks (T1-T5). Each task was replicated by 6 subjects, with images of different instances of the same 102 species. In baseline Task 1, neither the subjects nor the machine made use of a model. The subject just browsed the reference set (which was never reordered) to find an acceptable candidate class. In Task 2, all 5 of the available reference pictures were used to train the system. The remaining tasks (Task 3-5) explored semi-supervised learning, and differed only in the size and composition of the training sets used for the algorithmic components of CAVIAR. These training sets were generated by letting the users add the samples they classified to the training sample (we call samples with user-assigned labels and models *pseudo-training samples*). Table 1 shows the experimental design, based on 612 distinct flower images.

The times and locations of the mouse clicks of the subjects, and the responses of the system, were logged. This record was transferred to pre-formatted Excel worksheets after

completion of the experiments, and subsequently aggregated and tabulated as reported below.

Table 1. CAVIAR-Flower recognition experiments

Task	Purpose	Classification	Training set composition	
			# of labeled samples	# of pseudo-training samples
T1	Unaided classification	Browsing only	None	None
T2	Interactive classification	Rose curve adjustment + browsing	510	None
T3	Same as T2	Same as T2	102	None
T4	Same as T2	Same as T2	102	204
T5	Same as T2	Same as T2	102	408

Interactive accuracy and time compared to human alone and to machine alone

Table 2 shows the median performance of six subjects for human-alone (T1), machine-alone (T2 initial auto), and CAVIAR (T2). We observe that **(1)** there is no obvious difference between CAVIAR and human-alone in accuracy. However, with the machine's help, the median time spent on each test sample was reduced to less than half of the human-alone time; **(2)** the accuracy of the machine alone is low (39%). With a little human help (10 seconds per flower), the median accuracy increased to 93%; and **(3)** after the rose curves were adjusted to the extent the users deemed desirable, the initial automatic Top-3 accuracy increased from 55% to 79%.

Table 2. CAVIAR compared to human alone and to machine alone

	Time (s)	Top-1 accuracy (%)	Top-3 accuracy (%)	Rank Order
T1 (human alone)	26.4	94	N/A	51.0
T2 initial auto	0	39	55	6.6
T2 before labeling	8.5	52	79	4.0
T2 (CAVIAR)	10.7	93	N/A	N/A

From the above observations, we conclude that **(a)** 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; and **(b)** the visible rose-curve model mediates human-computer communication effectively.

Machine Learning

Table 3 shows the median values of the Top-1 accuracy, the Top-3 accuracy, the rank order after the initial automatic recognition, and the human time and the accuracy of the complete interactive recognition for T2, T3, T4, and T5. We observe that **(1)** the median Top-1 accuracy of the initial automatic recognition increases from T3 (27%) to T5 (37%), and approaches the median accuracy of T2 (39%), which is the best that we could hope for; **(2)** the median Top-3 accuracy of the initial automatic recognition increases from T3 (44%) to T5 (55%), which is the same as T2 (55%); **(3)** the median rank order after the initial automatic recognition decreases from T3 (12.7) to T5 (8.6), approaching the median rank order of T2 (6.6); **(4)** the median final accuracy of T3, where the classifier was trained on only one training sample, is still high (90%); **(5)** there is not much difference in accuracy among these four tasks: the median accuracies are all above 90%; and **(6)** the median time spent on each interactive recognition task decreases from T3 (16.4 seconds) to T5 (10.7 seconds), which is the same as the median time of T2.

Table 3. Machine learning.

	Initial Top-1 accuracy (%)	Initial Top-3 accuracy (%)	Initial rank order	Human time	Final Accuracy (%)
T3	27	44	12.7	16.4	90
T4	32	48	10.6	12.7	95
T5	37	55	8.6	10.7	92
T2	39	55	6.6	10.7	93

From these observations, we conclude that **(a)** the CAVIAR system can achieve high accuracy even when initialized with only a single training sample per class; **(b)** adding pseudo-labeled training samples improves automatic recognition, which in turn helps the subjects to identify the flowers faster; and **(c)** both automatic performance (initial rank order) and interactive performance (human time) for T5 are near the corresponding values of T2, which suggests that the adaptation mechanism is effective (although early users, working on a minimally trained system, will need more time).

Human recognition strategy and learning

Table 4 shows the average percentage of successive rose curve adjustments for T2-T5. We observe that **(1)** more than 90% of the samples are identified with 3 or fewer adjustments; and **(2)** there is little difference in the number of adjustments among T2, T3, T4, and T5. We therefore average them in Figure 5, and show that a geometric distribution with p=0.55 fits the curve well, i.e., the probability of success on each adjustment is just over one half.

Table 4. Average percentage of successive rose curve adjustments

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**>10**
T2	58.5	16.5	8.2	5.6	5.1	1.6	1.3	1.6	0.5	0.3	0.5	0.3
T3	45.9	19.3	10.5	9.5	4.4	3.8	1.5	1.8	1.1	0.7	0.2	1.5
T4	46.1	21.1	12.6	8.2	5.1	2.5	1.0	1.0	0.8	0.5	0.5	0.8
T5	57.4	16.8	8.5	7.2	4.6	1.6	1.1	0.3	0.5	0.5	0.3	1.1
Mean	52.0	18.4	9.9	7.6	4.8	2.4	1.2	1.2	0.7	0.5	0.4	0.9

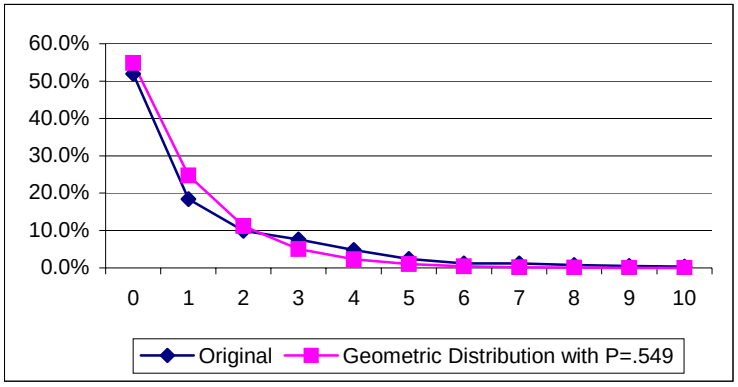

Figure 5. Successive rose curve adjustments.

Figure 6 shows the human time as a function of experience with the system, i.e., the number of samples that have been classified, for T1 and T2. We observe that **(3)** in T1, the human time decreased from 26 to 17 seconds as the subjects became more familiar with the database; and **(4)** in T2, the human time decreased from 9 to 5 seconds.

We conclude that **(a)** on average, 52% of the samples are immediately confirmed because the machine ranks the correct candidate very near the top; **(b)** subjects do adjust the rose curve when necessary and, on average, each sample requires 1.3 adjustments; **(c)** the interaction can be accurately modeled by a two-state Finite State Machine with the appropriate state transition probabilities; and **(d)** subjects remember the flowers to become "connoisseurs" of the flower database. With CAVIAR, lay persons need little practice to become faster than unaided "connoisseurs".

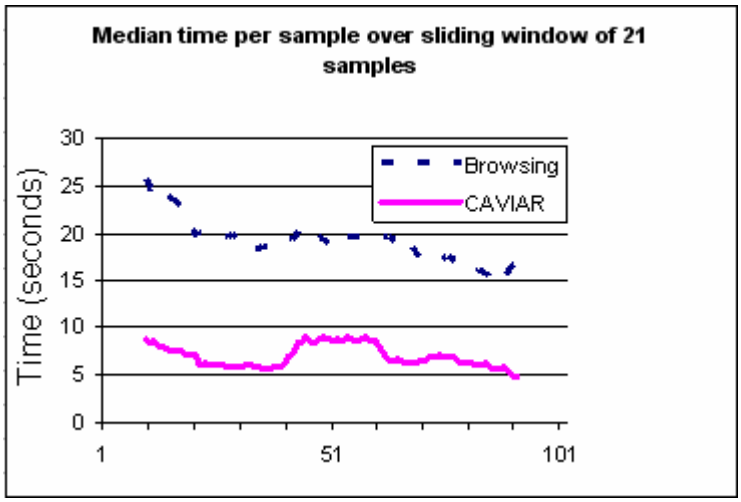

Figure 6. Recognition time as a function of experience with the system.

6. EVALUATION OF CAVIAR-FACE

We downloaded the FERET face database from the National Institute of Standards and Technology (NIST) [Feret]. Series BK was used for the “gallery” (reference) images. Part of the BA series was reserved for training, which requires pairs (BA and BK) of images of the same individual. Each of six subjects classified 50 randomly selected BA test pictures (different from the training set) against the same gallery of 200 BK pictures (taken on the same day as the test pictures but with a different camera and lighting). The faces vary in size by about 50%, and horizontal and vertical head rotations of up to 15° can be observed. Although the subjects had been asked to keep a neutral expression and look at the camera, some blinked, smiled, frowned, or moved their head. Because we had only two samples per face, here we could not test decision-directed machine learning.

The logging system was essentially the same as for flowers. Earlier experiments showed that human-alone (browsing only) required an average of 66 seconds per photo, and resulted with most subjects in perfectly accurate classification [Zou04a]. The results reported below are on the 50-picture interactive experiments that followed practice runs of 20 photos where we did not keep track of performance.

Figure 7 and Table 5 summarize the experimental results. We observe that over half (50.3%) the pictures were classified by clicking on one of the three candidates displayed as a result of the initial rank ordering. This took 2.3 seconds on average. If the correct candidate did not appear, most often the subjects adjusted the model: they resorted to time-consuming browsing only 4% of the time. The overall average human accuracy was 99.7%, and the average recognition time was 7.6 seconds per photo. Only 15% of the faces required more than

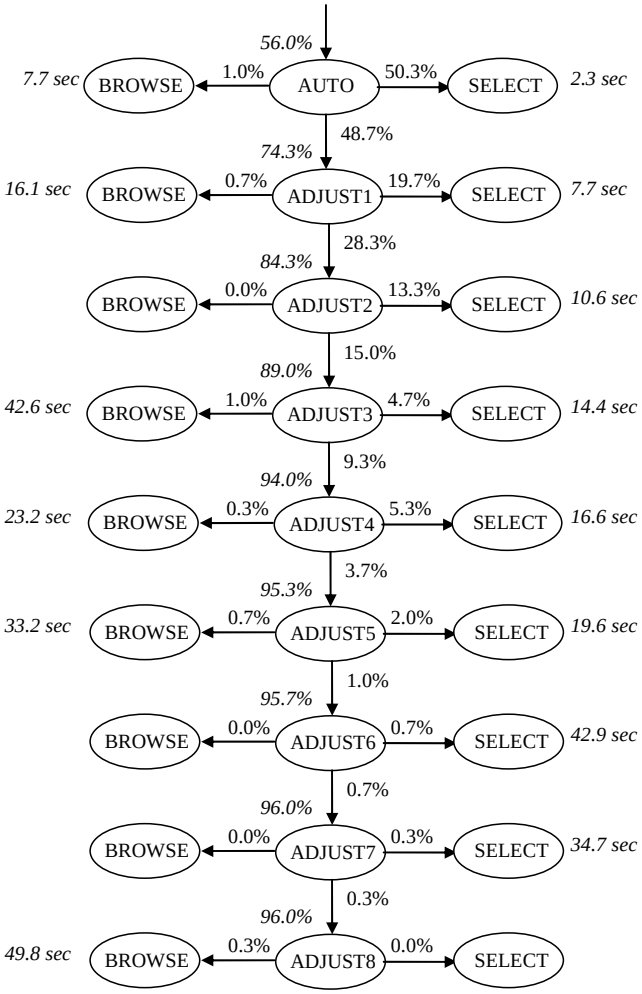

Figure 7. Interactions in CAVIAR-face (6 subjects). SELECT means choosing one of the displayed candidates. BROWSE means looking at more than 3 displayed faces before selecting a winner. Both SELECT and BROWSE terminate the interactive classification. ADJUST prompts automated rank ordering, which results in a new set of candidates for selection or browsing. Listed next to SELECT and BROWSE is the average (over subjects and pictures) human time required for final classification, including any adjustment or browsing. The percentage (in italic) indicated above AUTO or every ADJUST is the machine TOP-3 recognition accuracy.

two interactions. The top-3 accuracy of the initial automatic recognition was 56%, and it rose to 95% after five interactive model modifications. Table 5 clearly indicates that interactive recognition significantly reduces time without losing accuracy.

Table 5. CAVIAR-face compared to human alone and to machine alone.

	Accuracy	Time per face
Interactive	99.7%	7.6 sec
Machine Alone	48.0%	----
Human Alone	~100.0%	66.3 sec

7. MOBILE CAVIAR

CAVIAR-Flower was reprogrammed in Java on a camera-equipped Sharp Zaurus SL-5500 Personal Digital Assistant (PDA) at Pace University [Evans03]. Subsequently it was ported, in our laboratory, to a camera and Wi-Fi (IEEE 802.11b) equipped Toshiba e800 PDA dubbed M-CAVIAR. The work is shared between the PDA and the host computer (a laptop within ~100 meters). The PDA forwards, via the wireless network interface, each newly acquired image to the host. The laptop computes the initial visible model and rank order using its stored reference images, and returns the model parameters and index number of the top candidates to the PDA. The PDA then displays the top three candidates from its stored database of thumbnail reference pictures. If the user adjusts the model (using stylus or thumb), the adjusted model parameters are sent to the laptop and a new model and rank order is computed and communicated to the PDA [Gattani04, Zou05b]. The log file is kept on the PDA. The interface is illustrated in Figure 8.

We repeated on the PDA some of the earlier experiments with six new subjects. With this system, it was also possible to conduct field experiments to recognize flowers in situ. An additional 68 classes of flowers, with 10 samples of each, were collected with the new, lower-quality PDA camera. The principal findings were as follows. Recognition time per flower was over 20% faster than using the desktop, mainly because model adjustment was faster with either stylus or thumb than with a mouse. Recognition accuracy was slightly lower, because some reference flowers could not be easily distinguished on the small PDA display. The networked computation did not impose any significant delay: except for uploading each new flower picture to the laptop, only very short messages (model coordinates and rank orders) are exchanged.

Figure 8. M-CAVIAR graphic user interface.

8. SUMMARY

We presented a case for interaction *throughout* the recognition of visual objects, rather than only at the beginning or the end. CAVIAR-Flower and CAVIAR-Face show that the visible model mediates human-computer communication adequately, and that the CAVIAR system is much more accurate than the machine alone and significantly faster than the human alone. The CAVIAR system also improves with use.

It appears essential to let the human retain the initiative at all times and, for high accuracy, to be the final arbiter of correct classification (as opposed to merely proofreading already classified items). The principal cause of CAVIAR's success is the formulation of the visual model. Faces are sufficiently different from flowers to allow optimism with regard to finding such models for other domains.

Automated rank ordering can obviously be improved. With additional training samples, more elaborate estimation of prior probabilities will lead to improved initial models. Better features may be found for both flowers and faces. Simple nearest-neighbor classifiers may be replaced by more sophisticated classifiers. Logistic regression methods are superior to the Borda Count.

The CAVIAR architecture can scale up to a much larger number of classes. CAVIAR systems may prove valuable for constructing very large labeled training sets for automated algorithms. It is, however, obviously preferable to "grow" training sets by interactive classification under operational conditions, i.e., to rely on systems that improve with use (as does CAVIAR-flower).

The network protocol of M-CAVIAR will require some changes for camera-phone applications, which are gaining quickly on PDAs and stand-alone digital cameras. Careful interface design will be required to avoid disorienting the operator by zooming the tiny display, but stylus and thumb are better direct-action devices than a mouse. Portable, wireless CAVIAR systems offer the possibility of Internet-wide reference data collection and collaborative interactive recognition, including many educational applications.

Reference

[Baird94] H.S. Baird and G. Nagy, "A Self-Correcting 100-font Classifier," *Proc. SPIE Conference on Document Recognition, Volume SPIE-2181,* pp. 106-115, San Jose, CA, 1994.

[Bartlett98] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, "Face Recognition by Independent Component Analysis," *IEEE Trans. Neural Networks*, vol. 13, no.6, pp. 1450-1464, 2002.

[Belhumeur97] P.N. Belhumeur, J.P. Hespanha, and J.K. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[Bradford91] R. Bradford, "Technical Factors in the Creation of Large Full-Text Databases," *Proc. DOE Infotech Conference*, Washinton DC, May 1991.

[Carson02] C. Carson, S. Belongie, H. Greenspan and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026-1038, 2002.

[Chen05] Y. Chen, H.L. Bart, Jr., S. Huang and H. Chen, "A Computational Framework for Taxonomic Research: Diagnosing Body Shape within Fish Species Complexes," *Proc. of the Fifth IEEE International Conference on Data Mining (ICDM),* pp. 593-596, Houston, Texas, November 2005.

[Cohn96] D.A. Cohn, Z. Ghahramani, and M.I. Jordan, "Active Learning with Statistical Models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129-145, 1996.

[Cox00] I.J. Cox, M.L. Miller, T.P. Minka, T.V. Papathomas and P.N. Yianilos, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 20-37, 2000.

[Das99] M. Das, R. Manmatha, and E.M. Riseman, "Indexing Flower Patent Images Using Domain Knowledge," *IEEE Intelligent Systems*, vol. 14, no. 5, pp. 24-33, 1999.

[Dickey91] L.A. Dickey, "Operational Factors in the Creation of Large Full-Text Databases," *Proc. DOE Infotech Conference*, Washinton DC, May 1991.

[Duda01] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley, 2001.

[Darman92] D. Harman, "Relevance Feedback and Other Query Modification Techniques," *Information Retrieval: Data Structure and Algorithms*, W.B. Frakes and R. Baeza-Yates, eds., Prentice Hall, 1992.

[Drewniok97] C. Drewniok, K. Rohr, "Model-Based Detection and Localization of Circular Landmarks in Aerial Images," *International Journal of Computer Vision*, vol. 24, no. 3, pp. 187-217, 1997.

[Etemad97] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *Journal of the Optical Society of America*, vol. 14, pp. 1724-1733, 1997.

[Evans03] A. Evans, J. Sikorski, P. Thomas, J. Zou, G. Nagy, S.-H. Cha, C. Tappert, "Interactive Visual System," *CSIS Technical Report 196*, Pace University, 2003.

[Goldstein71] A.J. Goldstein, L.D. Harmon, and A.B. Lesk, "Identification of Human Faces," *Processings of IEEE*, vol. 59, pp. 748-760, 1971.

[Gattani04] A. Gattani, *Mobile Interactive Visual Pattern Recognition*, MS thesis, Rensselaer Polytechnic Institute, December 2004.

[Haritaoglu01] I. Haritaoglu, "Scene Text Extraction and Translation for Handheld Devices," *IEEE conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 408-413, December, 2001.

[Ho94] T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66-75, 1994.

[Jain92] R. Jain, *US NSF Workshop on Visual Information Management Systems*, 1992.

[Kak02] A.C. Kak and G.N. Desouza, "Robotic Vision: What Happened to the Visions of Yesterday," *Proc. Int. Conf. Pattern Recognition*, vol. 2, pp. 839-847, 2002.

[Kanade77] T. Kanade, *Computer Recognition of Human Faces*, Birkhauser, Basel, and Stuttgart, 1977.

[Klein04] B. Klein and A.R. Dengel, "Problem-Adaptable Document Analysis and Understanding for High-Volume Applications," *International Journal of Document Analysis and Recognition*, no. 6, pp. 167-180, 2004.

[Kirby90] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, 1990.

[Lizotte03] D. Lizotte, O. Madani, and R. Greiner, "Budgeted Learning of Naive Bayes Classifiers," *19th Annual Conference on Uncertainty in Artificial Intelligence*, 2003.

[MacKay92] D.J.C. MacKay, "Information-Based Objective Functions for Active Data Selection," *Neural Computation*, vol. 4 no. 4, pp. 590-604, 1992.

[Martinez02] A.M. Martinez, "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748-763, 2002.

[Miller56] G. Miller, "The Magical Number Seven Plus or Minus Two; Some Limits on Our Capacity for Processing Information," *Psychological Review*, vol. 63, pp. 81-97, 1956.

[Mericsko98] R. J. Mericsko, "Introduction of 27th AIPR Workshop - Advances in Computer-Assisted Recognition," *Proc. of SPIE*, vol. 3584, October 1998.

[Nagy66] G. Nagy and G.L. Shelton, "Self-Corrective Character Recognition System," *IEEE Transactions on Information Theory IT-12*, no. 2, pp. 215-222, April 1966.

[Nagy02] G. Nagy and J. Zou, "Interactive Visual Pattern Recognition," *Proc. International Conference on Pattern Recognition*, vol. 2, pp. 478-481, 2002.

[Nagy04] G. Nagy, "Visual Pattern Recognition in the Years Ahead," *Proc. International Conference on Pattern Recognition*, vol. 4, pp. 7-10, Cambridge, UK, August 2004.

[Nguyen04] H.T. Nguyen, A.W.M. Smeulders, "Active Learning Using Pre-Clustering," *Proc. of the Twenty-first International Conference on Machine Learning (ICML 2004)*, Banff, Alberta, Canada, July 4-8, 2004.

[Nilsson02] K. Nilsson, J. Bigun, "Prominent Symmetry Points as Landmarks in Finger Print Images for Alignment," *16th International Conference on Pattern Recognition*, 2002.

[Pentland94] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 84-91, 1994.

[Palmer99] S.E. Palmer, *Vision Science, Photons to Phenomenology*, MIT Press, 1999.

[Phillips00] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.

[Rui98] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644-655, 1998.

[Saitoh04] T. Saitoh, K. Aoki and T. Kaneko, "Automatic Recognition of Blooming Flowers," *International Conference on Pattern Recognition*, pp. 27-30, 2004.

[Sarkar02] P. Sarkar, H.S. Baird, J. Henderson, "Triage of OCR Output Using 'Confidence' Scores," *Proceedings of the SPIE/IS&T 2003 Document Recognition and Retrieval IX Conf (DR&R IX)*, San Jose, California, pp. 20-25, January 2002.

[Swets96] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 831-836, 1996.

[Turk91] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 13, no. 1, pp. 71-86, 1991.

[Ullman02] S. Ullman, M. Vidal-Naquet and E. Sali, "Visual Features of Intermediate Complexity and Their Use in Classification," *Nature Neuroscience*, vol. 5, no. 7, pp. 682-687, 2002.

[Veera04] S. Veeramachaneni and G. Nagy, "Adaptive Classifiers for Multi-Source OCR," *International Journal of Document Analysis and Recognition*, vol. 6, no. 3, pp. 154-166, 2004.

[Wiskott97] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 17, no. 7, pp. 775-779, 1997.

[Yang02] J. Yang, X. Chen, and W. Kunz, "A PDA-based Face Recognition System," *Proc. the 6th IEEE Workshop on Applications of Computer Vision*, pp. 19-23, December 2002.

[Yue05] W. Yue; D. Yin; C. Li; G. Wang; T. Xu, "Locating Large-Scale Craniofacial Feature Points on X-ray Images for Automated Cephalometric Analysis," *IEEE International Conference on Image Processing*, vol. 2, pp. 1246–1249, 2005.

[Zhang02] J. Zhang, X. Chen, J. Yang, and A. Waibel, "A PDA-based Sign Translator," *Proc. the 4th IEEE Int. Conf. on Multimodal Interfaces*, pp. 217-222, 2002.

[Zhao03] W. Zhao, R. Chellappa, A. Rosenfeld, P. Phillips, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, 2004.

[Zou04a] J. Zou, *Computer Assisted Visual InterActive Recognition*, Ph.D. thesis, ECSE Department, Rensselaer Polytechnic Institute, May, 2004.

[Zou04b] J. Zou, G. Nagy, "Evaluation of Model-Based Interactive Flower Recognition," *International Conference on Pattern Recognition*, 2004.

[Zou05a] J. Zou, "A Model-Based Interactive Image Segmentation Procedure," *IEEE Workshop on Applications of Computer Vision (WACV)* 2005.

[Zou05b] J. Zou, A. Gattani, "Computer Assisted Visual InterActive Recognition and Its Prospects of Implementation Over the Internet," *IS&T/SPIE 17th Annual Symposium Electronic Imaging, Internet Imaging VI*, 2005.

[Zou06a] J. Zou, Q. Ji, "Face Recognition through Selective Matching of Local Templates," *Submitted to International Conference Pattern Recognition*, 2006.

[Zou06b] J. Zou, Q. Ji, G. Nagy, "Face Recognition through Borda Count Combination of Local Regions," *Technical Report, Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute*, 2006.

[Feret] http://www.itl.nist.gov/iad/humanid/feret/feret_master.html; accessed on April 5, 2006.

[NEWFS] http://www.newfs.org/; accessed on April 5, 2006.