Computer Assisted Visual Interactive Recognition (*CAVIAR*) Technology

Arthur Evans, John Sikorski, Patricia Thomas, Sung-Hyuk Cha, Charles Tappert
evansart@prodigy.net, john.sikorski@regeneron.com, pt84575w@pace.edu, scha@pace.edu,
ctappert@pace.edu
Pace University, White Plains, NY
Jie Zou, Abhishek Gattani, George Nagy
Rensselaer Polytechnic Institute, Troy, NY
zouj@alum.rpi.edu, gattani@gmx.net, nagy@ecse.rpi.edu

The distinctive aspect of the CAVIAR technology is a visible, parameterized geometrical model that serves as the human-computer communication channel. Evaluation of CAVIAR flower and face recognition systems shows that their accuracy is much higher than that of the machine alone; their recognition time is much lower than that of the human alone; they can be initialized with a single reference sample per class; and they improve with use. CAVIAR-flower has been ported to stand-alone and to wireless laptop-client Personal Digital Assistants (PDAs).

1. Introduction

We describe a progression of computer-assisted visual interactive recognition (*CAVIAR*) systems. Our development was a Windows platform. We have ported our system to a stand-alone Personal Digital Assistant, and to a pocket PC configured as a wireless client to a laptop, both with plug-in cameras. The key difference between CAVIAR and most current classification systems is operator interaction based on a *visible model*. We report results on two very different applications, flower classification and face recognition, and comment on the merits of several recognition system architectures.

Automated visual recognition systems seldom achieve 100% correct classification on families of objects of interest. Most allow user interaction at the beginning, to locate or frame the object, and at the end, to classify "rejects" or "low confidence" items [Heritaoglu01, Zhang02]. Providing means of user interaction throughout the process, rather than only at the beginning or end, is more efficient in terms of human time. Leaving the operator fully in control of the classification process is more user-friendly than having to respond to machine-generated requests.

Mobile recognition systems offer obvious advantages for recognizing objects outside the office or home (like flowers or faces). However, perhaps their greatest potential advantage is that they provide an opportunity for taking additional pictures of a difficult object. Classification can then be based on several pictures: the simplest method is accepting the class that receives the most votes. More sophisticated methods will eventually be developed for merging relevant information at the image, feature, or classifier level. Although PDA cameras still lag stand-alone digital cameras in terms of optical and digital resolution (and convenience features), the greatest limitation of handheld systems compared to PC recognition platforms is their limited screen size, which prevents simultaneous display of several objects in adequate detail.

Although we experimented only with PDAs, it is clear that camera-phones will soon have enough storage and computing capacity, as well as appropriate operating systems, for interactive recognition. However, the display size limitation will be even more stringent. Some expect that it will be overcome by wireless access to large public displays [Raghunath04], but we can hardly expect a display to pop up whenever we wish to recognize a flower or a face.

2. Methodology

As in all classifiers, a set of labeled reference pictures, one or more per class, is stored in CAVIAR. Automated, but error-prone, algorithms segment each unknown picture, construct a visible model, and extract from the picture of the unknown object a set of preprogrammed discriminating features that can be compared with similar shape, color, or texture features extracted from the reference pictures. The candidates are then automatically ranked according to the similarity of their features to those of the unknown picture.

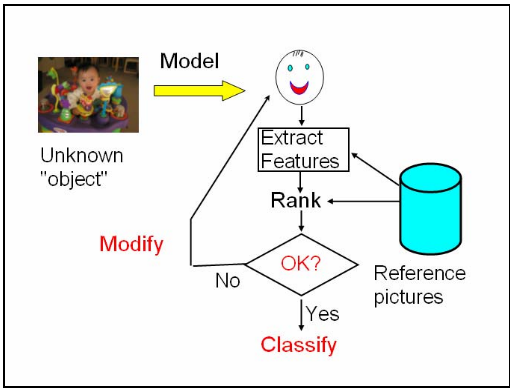


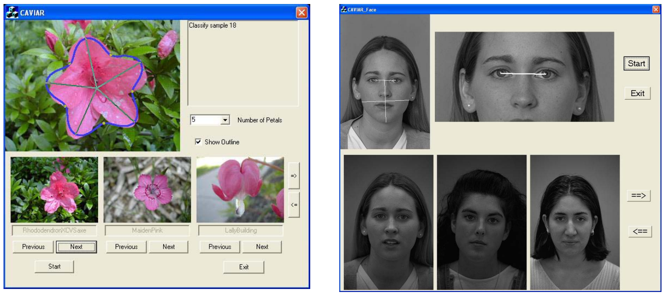
Fig 1 CAVIAR flowchart. Human actions shown in red.

If one of the displayed candidates matches the unknown picture, the operator simply clicks on it, thereby classifying the unknown. If not, the operator can adjust the visible model. The visible model guides the system in feature extraction. Therefore whenever the visible model is adjusted, new features are extracted, and all the candidates are automatically reordered. Occasionally, the correct candidate is not displayed even after adjustment of the visible model. In that case, operator can browse lower-ranked candidates by clicking on the NEXT button. If multiple reference pictures are available for each class, the operator can inspect them too.

The CAVIAR methodology is illustrated in Fig. 1. When a new picture is taken of an unknown object, the algorithmic part of the system ranks the candidates by comparing the features extracted from the new picture to all the stored reference pictures. It displays the top-n candidates (Fig. 2), as well as an automatically constructed visible model (Fig. 3) of the unknown object.

The visible models are simple line drawings where distinguished points can be acquired and moved by pointing and dragging. For flowers, our visible model is the *rose curve* of the 18th C Italian mathematician Guido Grandi. Our face model consists of five characteristic points. Because the pupil locations are critical for accurate registration, an enlarged view is provided in the GUI. The operator always classifies the object by clicking on the corresponding reference picture.

Pictures of just-classified objects can be added to the reference database along with their operator-assigned labels, and their visible models. They are subsequently used to improve model construction and rank ordering through decision-directed approximation [DHS00].


Classify sample 18

5 Number of Petals

☑ Show Outline

Fig 2 Desktop GUIs for CAVIAR-Flower (left) and CAVIAR-Face (right)

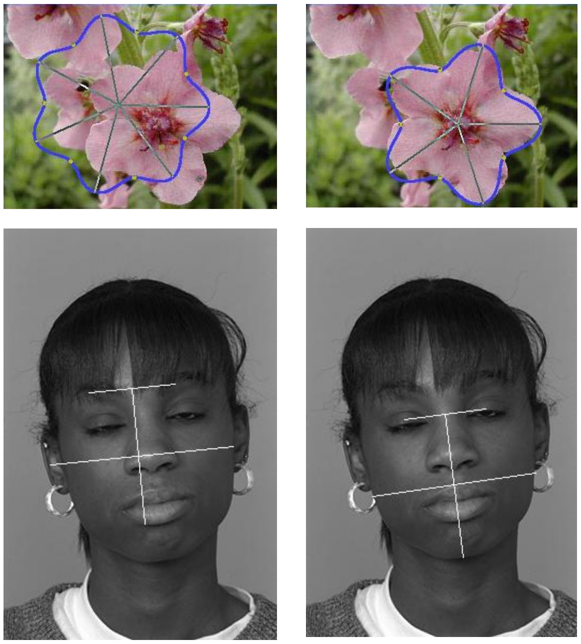


Fig 3 CAVIAR-Flower model (top) and CAVIAR-Face model (bottom): left, before adjustment; right, after adjustment by the operator. Here the automated model construction was confused by nearby flowers, and by closed eyes, respectively.

3. Mobile architectures

Our first portable CAVIAR, the Interactive Visual System (IVS), was a stand-alone implementation on the Sharp Zaurus SL-5500 with a 200MHz processor, 64 MB RAM, and Compact Flash and Serial Device ports [Evans03]. A Sharp CE-AG06 camera attachment, inserted into the Compact Flash port, allows direct capture of images (Fig. 4). Pictures from other cameras can also be uploaded through this port. We chose the Zaurus over the Ipaq 38xx series and the Sony Clie because it offered a high-end processor, a full-featured Linux OS with command line utilities, and Personal Java capabilities, in addition to the camera attachment. Personal Java has greater flexibility than MIDP (Mobile

Information Device Profile) that runs on low-end PDAs and cell phones. As an indication of the rapid advance of the technology, the Zaurus SL-5500 has already been superseded by the SL-6000 and the Ipaq 38xx series by the hx series, with 400MHz and 624MHz processors respectively.

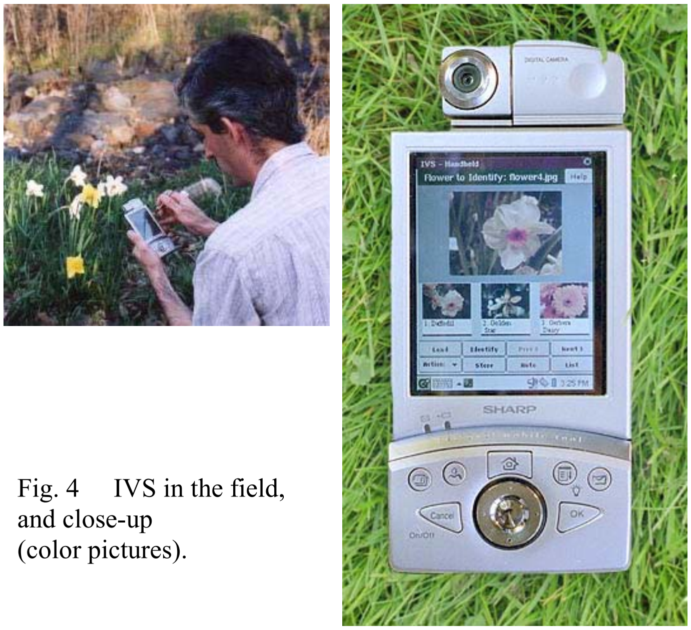


Fig. 4 IVS in the field, and close-up (color pictures).

The operating system is Embedix Linux with Personal Java support. All code was written to Personal Java specifications with code migration and extensibility in mind. The recognition engine is fully abstracted into generic and abstract classes, requiring only a few interfaces for data handling. Because generic classes handle user-image interactions, GUI abstraction requires the implementation of only a few methods.

IVS draws on human perceptual ability to group "similar" regions, perceive approximate symmetries, outline objects, and recognize "significant" differences. It exploits computer capability of storing image-label pairs, quantifying features, and computing distances in an abstract feature space of shapes and colors. The IVS architecture was developed specifically for isolated object recognition in the field, where the time available for classifying each image is comparable to that of image acquisition.

Next, CAVIAR was ported to a camera and Wi-Fi (IEEE 802.11b) equipped Toshiba e800 PDA dubbed M-CAVIAR (Fig. 5). The work is shared between the PDA and a nearby (i.e., within ~100meters) laptop host computer. The PDA forwards, via the wireless network interface, each newly-acquired picture to the host. The laptop computes the initial visible model, rank orders the candidates using its stored reference images, and returns the model parameters and index numbers of the top candidates to the PDA. The PDA then displays the top three candidates from its stored database of thumbnail reference pictures.

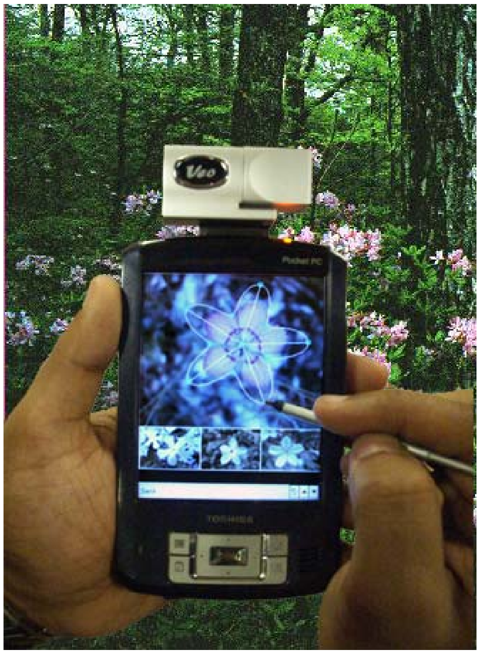


Fig 5 M-CAVIAR GUI.

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. The log file is kept on the PDA [Gattani04, Zou05].

With this system, it was possible to conduct field experiments on flowers in situ in addition to repeating some of the experiments on our flower database with six new subjects. 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.

The main drawback of current hand-helds is that they cannot be easily operated in broad daylight because the display is almost invisible in sunlight. Furthermore, under such conditions, the automatic camera exposure control does not always prevent overexposure. The lower-quality photos taken at dusk did prove adequate for interactive recognition in the field.

4. Experiments on flowers

Development and evaluation took place on a PC (Fig. 2). The software was written in C++ with INTEL Open Source Computer Vision Library routines. It includes several options for experimenting with different families of objects (flowers, fruit, cell micrographs, and Chinese ideographs), automatic and interactive segmentation methods [Zou05a], some additional features, and a choice of experimental protocols. Most importantly, it incorporates complete activity logging and statistical evaluation tools. Any testing session can be completely reconstructed and analyzed from the Excel worksheets created automatically from the log file.

For flower recognition, CAVIAR offers model-based automated and interactive segmentation, shape features based on moment invariants, hue and saturation histogram features, and an optimized Nearest Neighbor classifier [Nagy02]. Its browser can display not only alternative species, but also other instances of the same species. In anticipation of the port to a handheld, the CAVIAR window was restricted to 370 x 300 pixels. The simplicity of the interface draws on decades of HCI research by others [Myers99, Duric02].

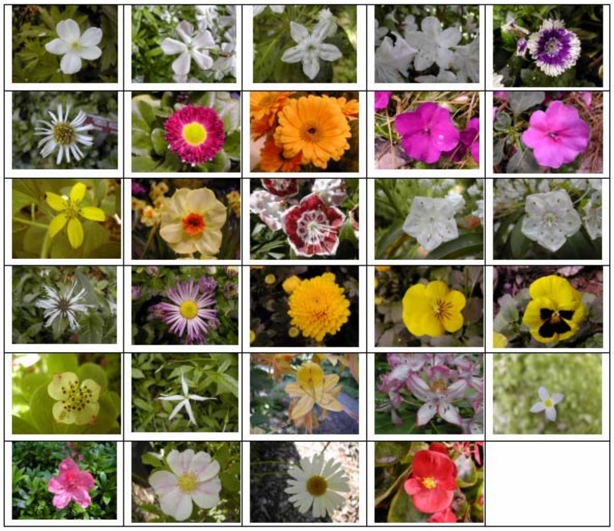


Fig 6 Development database of 29 species.

For experimental evaluation of classification accuracy, we were not able to use pictures from any of the many excellent flower sites on the web. None have more than one or two samples per specie, and labeling conventions, background, and resolution differs too much from site to site.

We tuned the system on several samples of each of 29 species photographed at nature gardens and flower shows (Figure 6). For classification, the photos, taken at the lowest resolution of our Canon Coolpix 775 camera, were further reduced to 320 by 240 pixels, with 3 bytes per pixel. Some of the species are quite similar (e.g. second row, 4 & 5; third row, 4 & 5), while different exemplars of the same species may exhibit marked differences in color and shape. This is by no means an easy recognition task for either laypersons or pattern recognition systems.

For an unbiased evaluation of human, machine, and interactive human-machine performance, we collected another dataset of 1078 flowers from 113 species, mostly from the New England Wildflower Garden, We tested CAVIAR on a subset of 612 flowers (from 102 classes which had at least 6 samples per class), which are freely available from author J. Zou.

Classification results based on 36 “naïve” subjects were reported in detail in [Zou04]. Interactive classification takes about 10 seconds per flower, and increases the accuracy from 39% for automated classification to 93% for interactive classification. Interactive classification is twice as fast as human classification without machine help. Unsupervised adaptation – adding classified flowers – boosted the top-3 accuracy of the automated rank ordering from 44% to 55%, further reducing human time.

5. Experiments on faces

For comparability with automatic methods, we downloaded the FERET mugshot database [Feret]: Each of six subjects classified 50 randomly selected test pictures against the same gallery of 200 pictures (one picture per individual, taken on the same day as the corresponding test pictures but with a different camera and lighting). Although the subjects were asked to keep a neutral expression and look at the camera, some blinked, smiled, frowned, or moved their head. The faces vary in size by about 50%, and horizontal and vertical head rotations of up to 15 degrees can be observed (Fig. 7).

In CAVIAR-face, the features are the match-scores of several templates extracted from the unknown image (11×11 pixels each) against each reference image. The score for each class-template pair is the value of the cosine between the 121-element template vector and the corresponding vector of the reference image. The maximum value is taken over a local 7×7 pixel registration window centered on points determined by the parameters of the similarity transformation obtained from the visible model. The reflectance values are subjected to local histogram equalization over a window four times as large as the templates (on the FERET database, this proved far superior to global intensity normalization). Discriminative locations for templates are obtained from the training set: many are near the eyebrows. The candidates are ranked according to the Borda Count of their match scores [Ho94]. As will be seen below, more match scores are computed for difficult faces.

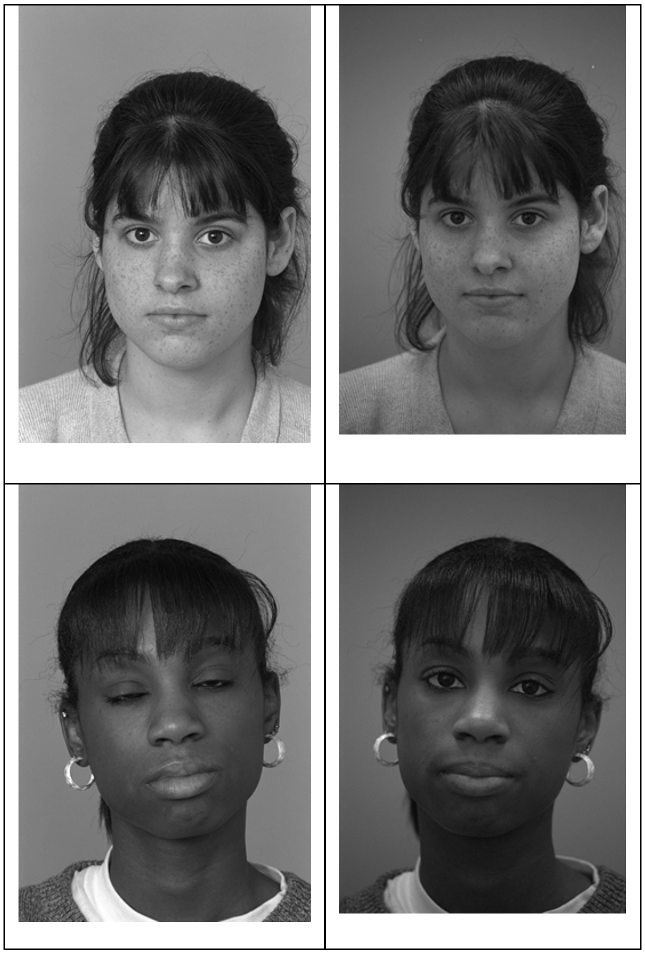


Fig 7 Easy and difficult Feret pairs.

A separate training set was used for (1) optimizing the initial model construction algorithms; (2) ordering, with a greedy feature-selection algorithm, the 255 potential locations of the templates used for classification;
(3) setting the size of the template registration window at the shift values where the accuracy leveled off; and
(4) determining the speed-up thresholds for deleting unlikely reference candidates and for halting the computation after testing enough templates. 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 [Zou04a] showed that human-alone (browsing only) required an average of 66.3 seconds per photo, and resulted with most subjects in perfectly accurate classification. 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.

The average interactive accuracy was 99.0%, and the average recognition time was 7.4 seconds per photo. 50% of the test pictures, over all subjects, required no interaction other than clicking on one of the displayed candidates (because the top-3 accuracy of the automated system was over 56%). The accuracy of the initial, automated classification (top 1) was 48%.

6. Conclusion

Our experiments demonstrate that interactive classification in many-class problems is much more accurate than current automated classifiers, and much faster than unaided human classification. Current camera, CPU and storage technology are adequate for supporting lightweight, highly portable computer vision systems that can operate either in a stand-alone mode or in wireless host-client mode. Within a year or two, we expect recognition systems to be embedded in camera phones. Widespread availability will trigger new applications in education, industrial inspection, and medical diagnosis of visible symptoms like skin lesions.

Acknowledgments

Hamei Jiang collected pictures of fruit, stamps, and coins for the early CAVIAR experiments. Greenie Cheng and Laura Derby photographed many flowers and helped build the database. Borjan Gagoski recruited subjects, conducted flower recognition experiments, and compiled the results. Rebecca Seth (City Naturalist, Lincoln, NE), Dr. Richard Mitchell (NY State Botanist) and Prof. Robert Ingalls (RPI CS Dept) gave us valuable advice about the classification of plants and flowers. Prof. Qiang Ji (RPI ECSE Dept) suggested that we apply CAVIAR to face recognition and freely shared with us his data and insights. We are greatly indebted to the many enthusiastic volunteers who helped us test the various incarnations of CAVIAR under agreement that they would remain anonymous.

References

[DHS00] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley, 2000.

[Duric02] Z. Duric, W.D. Gray, R. Heishman, F. Li, A. Rosenfeld, M.J. Schoelles, C. Chunn, and H. Wechsler, "Integrating Perceptual and Cognitive Modeling for Adaptive and Human-Computer Interaction," *Proceedings of the IEEE, Vol. 90,* no. 7, pp. 1272-1289, (Special issue on technology and tools for visual perception, C. Guerra and V. Cantoni, editors), July 2002.

[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.

[Feret04] http://www.itl.nist.gov/iad/humanid/feret/feret_master.html (*accessed January 30, 2005*).

[Gattani04] *Mobile Interactive Visual Pattern Recognition*, Rensselaer Polytechnic Institute MS Thesis, 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] Tin Kam Ho, Jonathan J. Hull, Sargur N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 16, no. 1, pp. 66-75, January 1994.

[Myers99] B.A. Myers, "A Brief History of Human Computer Interaction Technology," *ACM interactions,* vol. 5*,* no. 2, pp. 44-54, March, 1998.

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

[Raghunath04] Raghunath, Mandayam, C. Narayanaswami, and C. Pinhanez, "Fostering a Symbiotic Handheld Environment," *IEEE Computer,* pp. 56-65, Sept. 2004.

[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.

[Zou04] J. Zou, G. Nagy, "Evaluation of Model-Based Interactive Flower Recognition," *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2004.

[Zou04a] J. Zou, *Computer Assisted Visual Interactive Recognition: CAVIAR*, Rensselaer Polytechnic Institute PhD Thesis, May 2004.

[Zou05] 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.

[Zou5a] J. Zou, "A Model-Based Interactive Image Segmentation Procedure," *IEEE Workshop on Applications of Computer Vision (WACV)* 2005.