Industry and Object Recognition: Applications, Applied Research and Challenges

Yutaka Hirano[1], Christophe Garcia[2], Rahul Sukthankar[3,4], and Anthony Hoogs[5]

[1] Future Projects Division
Toyota Motor Corporation
Mishuku 1200, Susono-shi, Sizuoka-ken, 410-1193 Japan
yutaka@hirano.tec.toyota.co.jp
[2] France Telecom division R&D
4, rue du Clos Courtel
35512 Cesson Sevigne Cedex - France
christophe.garcia@francetelecom.com
http://perso.rd.francetelecom.fr/garcia/
[3] Intel Research Pittsburgh
4720 Forbes Ave. #410, Pittsburgh, PA 15213, USA
rahuls@cs.cmu.edu
[4] The Robotics Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[5] GE Global Research
One Research Circle
Niskayuna, NY 12309 USA
hoogs@research.ge.com

Abstract. Object recognition technology has matured to a point at which exciting applications are becoming possible. Indeed, industry has created a variety of computer vision products and services from the traditional area of machine inspection to more recent applications such as video surveillance, or face recognition. In this chapter, several representatives from industry present their views on the use of computer vision in industry. Current research conducted in industry is summarized and prospects for future applications and developments in industry are discussed.

1 Introduction

As visual recognition and computer vision in general have become more mature, industry has created an ever-increasing variety of computer vision products and services. From the traditional area of machine inspection, commercial vision applications have expanded into video surveillance, medical image analysis, face detection and recognition, and many others.

Significant challenges remain before generic, categorical object recognition can attain widespread commercial use. The major barriers include:

J. Ponce et al. (Eds.): Toward Category-Level Object Recognition, LNCS 4170, pp. 49–64, 2006.

- Robustness w.r.t. variation in viewpoint, illumination, scale and imaging conditions.
- Scaling up to thousands of object classes. While some applications may only require class libraries of dozens of objects, many require much larger class diversity requiring human-level performance.

In this chapter, the visual recognition needs, challenges and current research of four industrial labs are described. These corporations – Intel, Toyota, General Electric, and France Telecom – are some of the largest technology and engineering firms in the world. Additional corporations and organizations sponsored the workshop, including Northup Grumman, Xerox, Lockheed Martin, Microsoft and DARPA.

2 Visual Recognition at France Telecom Research and Development

In the last decade, we have entered the digital era, with the convergence of telecommunication, video and informatics. Our society (press agencies, television channels, customers) is producing daily extremely large and increasing amounts of digital images and videos, making it more and more difficult to track and access this content via database search engines that rely mostly on manually annotated information. Content-based indexing via automatic object detection and recognition techniques has become one of most important and challenging issues for the years to come, in order to face the limitation of traditional information systems. Some expected applications include [34]: information and entertainment video production and distribution; professional video archive management including legacy footages; teaching, training, enterprise or institutional communication; TV program monitoring; self-produced content management; internet search engines; video surveillance and video conference archiving and management; and advanced object-based image coding.

There has been much research over the last decade to develop image and video content-based indexing systems [43,35]. Most existing commercial products rely on searching images "looking like" others, using global descriptions, by extracting feature vectors that summarize the content of the images in terms of luminance, color or texture. These feature vectors are traditionally statistical summaries of color distribution in different color spaces, textures in the form of histograms of gradient directions or Gabor coefficient statistics. These descriptors offer the advantage of being invariant to global image transformations such as warping or object motion. On the other hand, structural information in the image is not captured and different images can have very similar global feature descriptors. In order to take into account the heterogeneous nature of an image and somehow its structure, more advanced systems are based on the detection of patches or salient points (local contrast, edges, corners and junctions,etc.) where local signatures are computed around each patch to characterize the more visually discriminant parts of the image. Image comparison is then performed by matching patches.

Both approaches, using no a priori knowledge on the image content, allow treating images regardless of their specific semantic content. But, one can wonder what global similarity between images means, which is extremely subjective and application-dependent. For a large range of applications, accessing the semantic content and identifying high-level indices is a pre-requisite, regardless of the global context of the image. This is the goal of object detection and recognition techniques that aim at locating faces, human bodies, cars, buildings, etc. They can be successfully applied to adult content filtering in the web, traffic surveillance, security access control, visual geo-localization, visio-conferences or intelligent man-machine communication. More generally, key object detection in collections of images or video sequences may provide easy, accurate and more natural ways of indexing and retrieving information ("find my photos in front of the Eiffel tower", "find the photos of Barbara", etc.). Given the growing volume of personal digital pictures, and the rapid development of Peer-to-Peer applications, one of the key applications is the management of self-produced content, where collections of personal digital pictures have to be stored, shared, sorted and retrieved according to the presence of specific persons, of specific objects or buildings. They may be tagged with meta-data when recorded, or indexed a posteriori when users formulate a specific request, like finding photos of friends, family, etc.

If object detection and recognition methods have long been limited to the "world of cubes", using low-level image analysis and heuristics, new supervised learning-based appearance methods have appeared recently and proved to be very efficient for several specific applications. For instance, human face detection can be considered as a mature tool, even though progress must be made for full-profile view detection and accurate facial feature detection, for allowing efficient face recognition. The method proposed by Viola and Jones [44], relying on a boosted cascade of simple classifiers based on Haar low level features seems very appealing given its speed and its good detection rate. More recently, Garcia and Delakis [7] proposed a near-real time neural-based face detection scheme, named "Convolutional Face Finder" (CFF) that has been designed to precisely locate multiple faces of 20x20 pixel minimum size and variable appearance, rotated up to ± 30 degrees in image plane and turned up to ± 60 degrees, in complex real world images. As a generic object detection method, the proposed system automatically synthesizes simple problem specific feature extractors and classifiers from a training set of faces, without making any assumptions or using any hand-made design concerning the features to extract or the areas of the face pattern to analyze. Moreover, global constraints encoding the face model are automatically learnt and used implicitly and directly in the detection process. After training, the face detection procedure acts like a pipeline of simple convolution and subsampling modules that treat the raw input face image as a whole in order to locate faces, without requiring any local contrast preprocessing in the input image. Experiments have shown high detection rates with a particularly low number of false positives, on difficult test sets, without requiring the use of multiple networks for handling difficult cases. For instance, a good detection

rate of 90.3% with 8 false positives have been reported on the CMU test set, which are the best results published so far on this test set (Figure 1).

Fig. 1. Some results of CFF on the CMU test set and a CFF-based face recognition system at France Telecom

One can notice that, for the time being, most approaches tackle detection of single objects with stable 2D appearances. There is still much to be done in the case of deformable 3D object detection. Moreover, most state-of-the-art successful methods rely on large training data sets, in order to infer the object class boundaries in discriminant feature space. Generative methods requiring fewer object examples must be investigated and combined with these discriminative methods, in order to ease the development of the pattern classifiers. More than single object detection, more general object category recognition techniques (like vehicles, buildings, etc.) have also to be considered in order to reduce the number

of specific object detection methods and allow more powerful and natural user queries.

Beyond the development of specific algorithms, performance evaluation of object recognition techniques for content-based image and video indexing is still a critical issue. A few specific frameworks have been organized, for specific recognition tasks such as face recognition [1] or for more global video indexing [2], where measures such as good detection/false alarm rates, or accuracy/recall rates are estimated using test data with ground truth. Some challenges tend to be organized like the "Pascal Visual Object Classes Challenge" (VOCC), organized by Mark Everingham, Luc Van Gool, Chris Williams and Andrew Zisserman, in March 2005 [3]. The goal of this challenge was to assess different object recognition approaches for different visual object classes (motorbikes, bicycles, people and cars) in images of realistic scenes. A training set of labelled images and various test sets were provided to assess the generalization capabilities of supervised algorithms trained with a reduced and unique set of examples. Among the competing methods, a modified version of the CFF system [7] has been applied to the detection and localization of cars and motorbikes, showing good generalization capabilities, given the small number of examples and the variability of the objects to detect.

But, in general, given the very large number of possible applications and the very specific research projects, most approaches are tested on "home-made" reduced data sets, where the proposed techniques perform reasonably and that are not easily shared among the research groups. Evaluating each approach and comparing it with others is therefore difficult. Moreover, developers of industrial applications obviously require successful techniques, but also clearer insight regarding the limits of the approaches, i.e. when and why they fail, in order to offer reliable solutions.

3 Visual Recognition at Intel

Intel Research engages in a variety of research projects that address real-world problems using techniques from object recognition. In addition to conducting work that relates directly to its product roadmaps, Intel is also active in exploratory research, particularly in the context of open collaborative projects pursued with faculty and students in academia. These projects typically generate implementations that are released as open source. This section presents an overview of three selected projects: efficient sub-image retrieval using local descriptors; object-based image retrieval; and computational nanovision.

3.1 Efficient Sub-image Retrieval Using Local Descriptors

The goal of sub-image retrieval is to find all of the images in a database that have features in common with a query image. Applications include content-based image retrieval (CBIR), identifying copyright violations on the web and detecting image forgeries. However, unlike traditional CBIR, the query image

cannot be matched against the database using global features. Our system [19] builds a parts-based representation of images using distinctive local descriptors which give high quality matches even under severe transformations. To cope with the large number of features extracted from the images, we employ locality-sensitive hashing [10] to index the local descriptors. This allows us to make approximate similarity queries that only examine a small fraction of the database. Although locality-sensitive hashing has excellent theoretical performance properties, a standard implementation would still be unacceptably slow for this application. By optimizing layout and access to the index data on disk, we can efficiently query indices containing millions of keypoints.

Figure 2 illustrates the system architecture. As images are added to the database, we perform feature extraction using the SIFT [25] detector and PCA-SIFT descriptor [20]. These keypoints are stored in a collection of LSH hashtables on disk. In typical experiments on a fine art collection, there are approximately 15 million keypoints for a set of 12,000 images. During retrieval, keypoints are extracted from the query image and the set of matching keypoints is efficiently retrieved from the database. The system employs a RANSAC-based geometric verification step (using an affine transform model) to eliminate false positives. Figure 3 shows a query generated by compositing patches from two images. The system correctly identifies both source images from a large collection of fine art images without finding any false positives. Our system achieves near-perfect accuracy (100% precision at 99.85% recall) on the tests presented in Meng *et al.* [28], and consistently strong results on our own, significantly more challenging experiments [19]. Query times are interactive even for collections of thousands of images.

3.2 Object-Based Image Retrieval

Object-based Image Retrieval is a collaborative effort between Intel Research Pittsburgh and Carnegie Mellon University, in the context of the Diamond project [13]. The goal is to create image retrieval systems based on the objects that appear in the images by learning the target concept *on-line* from a small set of examples provided by the user.

Unlike most existing systems that discriminate based on a histogram or clustering of color or texture features computed over the entire region, our system performs a windowed search over location and scale for each image in the database. This approach allows the retrieval of an image based on the presence of objects that may occupy only a small portion of the image. Also, we do not assume that a feature's value is independent of location within the window. This allows our system to retrieve images based on objects composed of colors and textures that are distinctive only when location within the window is considered, as is common with many man-made objects.

The system consists of two stages. An exhaustive windowed scan over scale and position generates a set of subimages. The first stage classifies and ranks

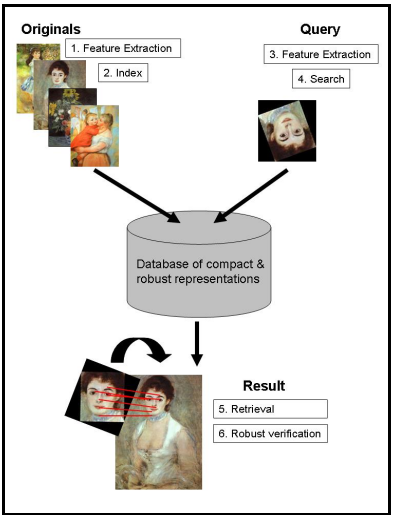

Fig. 2. System diagram: subimage retrieval relies on efficient near-neighbor searches of PCA-SIFT descriptors

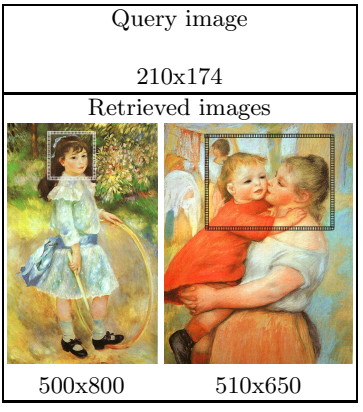

Fig. 3. Given a composite query, the system correctly retrieves the two source images, without false matches from a database containing over 6000 similar paintings

subimages using the posterior probability, computed from the estimated unconditional density and the object class conditional density. The second stage, trained using relevance feedback, reduces false positives by classifying subimages that are labeled as positive by the first stage. If a subimage passes both stages, the image is returned to the user.

3.3 Computational Nanovision

Silicon manufacturing technology is now able to shrink critical dimensions of structures down to scales well below 100 nm. These nanostructures are too small to see, even with the most sophisticated imaging equipment. This presents a challenge for Intel engineers who examine microprocessors to identify which nanostructures are defective, so repairs can be made. The Computational Nanovision research project addresses the challenge posed by the low resolution and the low signal-to-noise ratio of nano-imaging tools. Researchers apply computer vision techniques based on sophisticated mathematical models to measure and create visual representations of these structures, and to automate image and data analysis.

Computational nanovision exploits the availability of detailed models of microprocessor layouts and manufacturing processes. By integrating this information with knowledge of the physics underlying image formation, one can develop new model-based techniques for analyzing nanostructures in images. This has

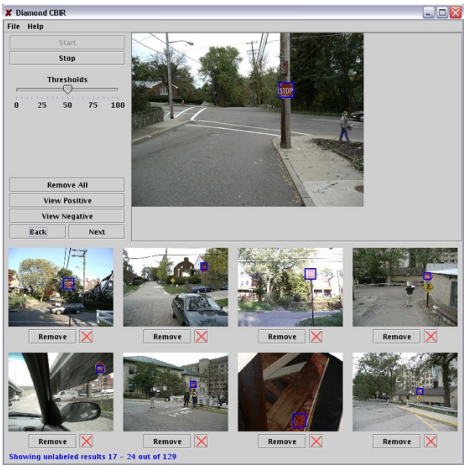

Fig. 4. Results of object-based image retrieval search for stop signs. The system was trained on a total of only 12 stop sign images after 1 round of feedback.

led to tools for image reconstruction, feature detection and classification in noisy images.

Image and Surface Reconstruction. The smaller the structure, the noisier the image. Researchers have developed probabilistic techniques for real-time reconstruction of noisy images of nano-structures. This enables a user to observe features that are otherwise not visible through the noise, as shown in Figure 5.

To do this, researchers first create statistical models of the noise distribution of specific tools and incorporate them into a Bayesian de-noising framework. This allows them separate the real image structure from the noise, and provides the user with a significantly enhanced image. In some cases, users wish to see the real 3D structure of an object instead of a scanning beam image. To enable this, the research group has recently developed a novel technique for rapidly generating three-dimensional structures from two-dimensional images of scanning electron microscopes, which was computationally intractable in the past. Using this new technique, 3-D reconstructions of nano-structures visible in an SEM image can be obtained within minutes.

Nanofeature Detection and Classification. Some applications, such as nano-machining, require real-time capability to allow for fast visual feedback from manufacturing tools. Even if a nanostructure image can be perfectly reconstructed, currently the tool operator must make a decision, such as determining when a structure of interest is visible in the noisy image. Researchers use probabilistic techniques to automatically detect and classify nano-features, to assist users and reduce the risk of human error. The eventual goal is to fully automate the process, removing humans from the loop.

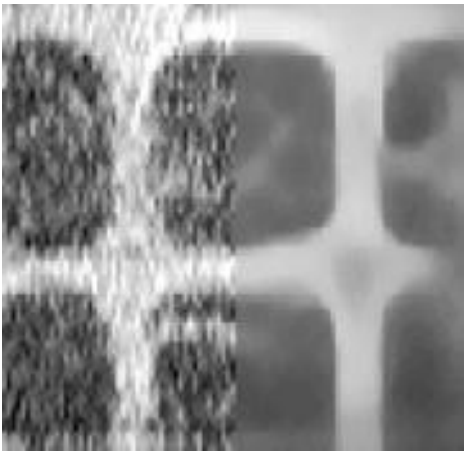

Fig. 5. Computational nanovision can perform image reconstruction on noisy input images (left), enable users to observe nanostructures through the noise (right)

4 Visual Recognition at General Electric Global Research

The Visualization and Computer Vision Lab within General Electric Global Research is currently engaged in the development of computer vision technology in video surveillance, medical image analysis, biological image analysis, industrial inspection, and broadcast media. In all of these applications, visual recognition is a critical enabling technology. For video surveillance, we are conducting research in a variety of areas, including: tracking people and vehicles [29,16,22]; segmenting moving crowds into individuals [32]; person re-identification [9]; detecting events and behaviors of interest [5,4]; camera calibration [21,41]; scene understanding [18,14]; and face analysis and recognition [24,40].

For broadcast video, we have developed methods for semantic object recognition using context established by the transcript [15,36,33]. For industrial inspection, we have focused on the problem of curved surfaces with complex reflectance [37]. In medical imaging, we have developed algorithms to automatically screen diagnostic images for early cancer detection [17,27]. We have also conducted research in recognizing partial or low-quality fingerprints [38,39], which has been used by the FBI.

In all of these areas, significant progress in visual recognition has been essential for developing prototypes and for transitioning algorithms into operations and products. As recognition technology improves, we envision significantly enhanced applications in many of the GE businesses.

Our recent efforts in scene content classification and person re-identification are perhaps the most relevant to this workshop, and are summarized briefly here.

4.1 Scene Content Classification

The goal of scene content classification is to label every pixel in an image with its category. We define this problem using a small number of broad categories, such that every pixel can be correctly classified into some category.

In our approach, we perform an initial, dense region segmentation on the image to form "superpixels", which are then attributed with a feature vector. We have explored and compared two formulations of this feature vector. First, following the work of [14,18,31], which we will call *perceptual features*, we include superpixel contrast, parallelism and continuity features derived from the region graph using adjacent regions. Second, following [42,**?**], which we will call *texton-based features*, a texture filter bank is computed at each pixel, quantized into textons, and histogrammed over the region.

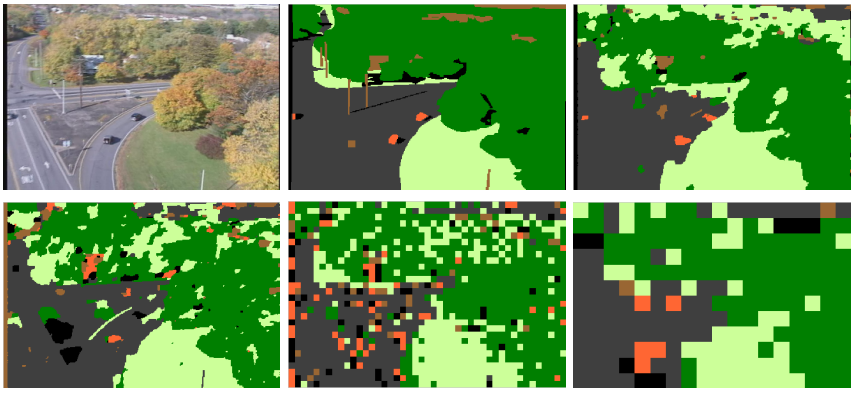

Fig. 6. Example results of our segmentation and classification methods, compared to image block classification. Top left: an image (not used in training); top middle: manual segmentation and labeling. The six classes are: dark gray=road, orange=vehicle, brown=building, dark green=tree, light green=grass, black=shadow. Four classification results from: perceptual features with AdaBoost.MV, with 75.2% pixel-wise correct classification (top right); texton features 66.0% (bottom left); 8x8 image blocks 62.6% (bottom middle); 20x20 image blocks 63.8% (bottom right).

Each attributed region is classified using a novel, generic extension of boosting for a multiclass problem, AdaBoost.MV. We treat the output of an ensemble of binary classifiers as a derived "vote" feature vector, performing MAP classification in this more discriminating space using a Gaussian distribution over classes.

Comparative results are shown in Figure 6. We compared our region-based methods to block-based methods, where each block is characterized by its texton histogram. On a set of 25 images, with 10 training and 15 test, AdaBoost.MV with perceptual features outperforms texton features 75.2% to 66% in pixel-wise classification accuracy. Textons on 20x20 image blocks scored 64%.

These scene classification methods have been applied to broadcast news content annotation [15] and tracking vehicles [16]. In the latter, knowledge of scene content is used to improve stabilization, moving object detection and track loss due to occlusion.

4.2 Person Re-identification Using Spatio-temporal Appearance

In many surveillance applications it is desirable to determine if a given individual has been previously observed over a network of cameras. This is the person

Fig. 7. Top ten person matches using the model-based algorithm. The query image is shown in the left column, and the remaining columns are the top matches ordered from left to right. The query and matching images are taken from different cameras. A box is used to highlight when a match corresponds to query. Third row shows an example where the correct match is not present in the top ten matches.

reidentification problem. Our approach focuses on reidentification algorithms that use the overall appearance of an individual as opposed to passive biometrics such as face and gait [9]. Person reidentification approaches have two aspects: (i) establish correspondence between parts, and (ii) generate signatures that are invariant to variations in illumination, pose, and the dynamic appearance of clothing. A novel spatiotemporal segmentation algorithm is employed to generate salient edgels that are robust to changes in appearance of clothing. The invariant signatures are generated by combining normalized color and salient edgel histograms. Two approaches are proposed to generate correspondences: (i) a model based approach that fits an articulated model to each individual to establish a correspondence map, and (ii) an interest point operator approach that nominates a large number of potential correspondences which are evaluated using a region growing scheme. These approaches were evaluated on a 44 person database across 3 disparate views.

5 Visual Recognition at Toyota

Recently the application of object recognition to real-world systems for cars and also to autonomous robots is rapidly growing. For cars, there already are some systems utilizing visual recognition as follows:

- A lane departure warning and lane-keeping assist system using white line detection.
- Detection of obstacles in front of the vehicle using stereo images.
- A pedestrian detection and warning system using infrared images.

Many more applications for future intelligent vehicles to prevent potential traffic accidents and also to assist driving are expected. For the realization of those future systems, recognition and prediction of the motion of pedestrians, other cars, other bikes etc. will be necessary. Also for autonomous and semi-autonomous driving, as well as for driving support, recognition of traffic signs, signals etc. and also segmentation and categorization of road area, sidewalks, guardrails, crosswalks, crossroads etc. will be necessary. Furthermore, to predict future possible dangers and prevent them, scene understanding considering the context of the scene will become important. However, there still are many difficulties for these tasks such as large occlusions, very large variations in weather, lighting conditions, shape of the objects, and so on.

One of the most challenging applications of visual recognition is pedestrian detection, because of 1) large appearance change with changes in posture and viewpoint; 2) large self occlusions and overlap between multiple people; 3) large variation of appearance due to clothes, age, gender, etc. By the recent development of many kinds of local feature descriptors and also combining those technologies with statistical learning technologies, some of these difficulties are gradually being addressed. [8] showed excellent performance of an Adaboost-based algorithm for this problem. [6] showed the potential of combining affine

invariant local features and statistical learning. Recently [23] gave a more robust solution for pedestrian detection in cluttered scenes. But these methods still fall short of what is required for commercial systems.

On the other hand, for autonomous robots such as future service robots, object recognition in outdoor and indoor scenes is very important. One necessary capability of these robots is recognition of objects to be handled and also of obstacles and the 3D environment for autonomous navigation in cluttered scenes. Also categorization of floor, wall, furniture, moving human and so on is necessary for simultaneous localization and mapping. If shape and specific local descriptors for a 2D[30] or 3D[11] image are extracted, it is possible to detect and recognize objects by matching those descriptors between the input camera image and a database. It is also possible to estimate position and orientation of the known objects by the same way as camera pose estimation using corresponding feature points between the input image and the database image. For non-textured objects, a descriptor using the contour information can be used. On the other hand, if there is no such database but most of the objects can be fitted to simply shaped primitives, only separating each object and estimating the position and orientation make sense for grasping those objects by robot hand. We at first developed a technology based on 3D reconstruction of the object shape and then separate each object using 3D shape information [12]. For separating objects, an algorithm using graph-cut [26] was developed. And to fit the shape of the each object to primitives, clustering of normalized vectors of each surface is used. We plan to recognize objects with more complex shapes in cluttered environments by combining two approaches mentioned above.

There still are problems to be solved for these methods when applying them to operational systems. One is how to improve the accuracy of matching descriptors especially for low resolution images. Speedup and improvement of 3D reconstruction is also a big problem for the actual implementation. Sensor fusion of range sensor and multi-view vision is one possibility for addressing this problem.

Recently, thanks to the rapid growth of computation power and also to the development of mathematical theories, statistical methods are becoming more useful in all of the related engineering areas, including computer vision. Application of object classification as pedestrian detection owes much success to this trend. However there still are many unresolved problems such as error by over learning, how to construct a proper learning dataset, etc. On the other hand, the robustness and speed of local feature detectors and descriptors are still hot topics. Also for the matching problem, there still needs to be improvement in outlier rejection. There also remains the problem of how to speed up the matching for huge object databases. Now that computer vision is becoming useful for various real applications, the expectation of industry for academia to solve remaining problems is very strong. To accelerate this movement, frank and deep discussion about the matching of technical needs with academic research is becoming increasingly important.

Acknowledgments

Section 4 of this report was prepared by GE GRC as an account of work sponsored by Lockheed Martin Corporation. Information contained in this report constitutes technical information which is the property of Lockheed Martin Corporation. Neither GE nor Lockheed Martin Corporation, nor any person acting on behalf of either; a. Makes any warranty or representation, expressed or implied, with respect to the use of any information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or b. Assume any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method or process disclosed in this report.

References

1. The NIST humanid evaluation framework. `www.frvt.org` (2003)
2. The TREC video retrieval evaluation. `www-nlpir.nist.gov/projects/trecvid` (2003)
3. The Pascal visual object classes challenge. `www.pascal-network.org/challenges/VOC` (2005)
4. M. Chan, A. Hoogs, J. Schmiederer, M. Petersen. Detecting rare events in video using semantic primitives with HMM. In: Proc. ICPR. Volume 4. (2004) 150–154
5. M. Chan, A. Hoogs, A. Perera, R. Bhotika, J. Schmiederer, G. Doretto. Joint recognition of complex events and track matching. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2006)
6. R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR2003).
7. C. Garcia, M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2004) 1408–1423
8. D. Gavrila. Pedestrian detection from a moving vehicle, in Sixth European Conference on Computer Vision (ECCV2000), Springer, pp. 37-49.
9. N. Gheissari, T.B. Sebatian, P.H. Tu, J. Rittscher, R. Hartley. A novel approach to person reidentification. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2006.
10. A. Gionis, P. Indyk, R. Motwani. Similarity search in high dimensions via hashing. In: Proc. Conference on Very Large Databases. (1999)
11. A. Johnson and M. Hebert: Using spin images for efficient object recognition in cluttered 3D scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 5, May, 1999, pp. 433 - 449.
12. Y. Hirano, K. Kitahama, and S. Yoshizawa. Image-based Object Recognition and Dexterous Hand/Arm Motion Planning Using RRTs for Grasping in Cluttered Scene, in IEEE/RSJ Conference on Intelligent Robots and Systems (IROS2005), Edmonton, Canada.
13. D. Hoiem, R. Sukthankar, H. Schneiderman, L. Huston. Object-based image retrieval using the statistics of images. In: Proc. Computer Vision and Pattern Recognition. (2004)

14. A. Hoogs, R. Collins, R. Kaucic, J. Mundy. A common set of perceptual observables for grouping, figure-ground discrimination and texture classification. T. PAMI **25** (2003) 458–475
15. A. Hoogs, J. Rittscher, G. Stein, J. Schmiederer. Video content annotation using visual analysis and large semantic knowledgebase. In: Proc. CVPR, IEEE (2003)
16. R. Kaucic, A.G.A. Perera, G. Brooksby, J. Kaufhold, A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In: Proc. CVPR. (2005) 990–997
17. R.A. Kaucic, C.C. McCulloch, P.R.S. Mendonça, D.J. Walter, R.S. Avila, J.L. Mundy. Model-based detection of lung nodules in CT exams. In Lemke, H.U., Vannier, M.W., Inamura, K., Farman, A.G., Doi, K., Reiber, J.H.C., eds.: Computer Assisted Radiology and Surgery. Volume 1256 of International Congress Series., London, UK, Elsevier 990–997, 2003.
18. J. Kaufhold, A. Hoogs. Learning to segment images using region-based perceptual features. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, IEEE (2004)
19. Y. Ke, R. Sukthankar, L. Huston. Efficient near-duplicate and sub-image retrieval. In: Proc. ACM Multimedia. (2004)
20. Y. Ke, R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc Computer Vision and Pattern Recognition. (2004)
21. N. Krahnstoever, P. Mendonca. Bayesian autocalibration for surveillance. In: Proc. ICCV, IEEE (2005)
22. N. Krahnstoever, T. Kelliher, J. Rittscher. Obtaining pareto optimal performance of visual surveillance algorithms. In: Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2005.
23. B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR2005), San Diego, CA.
24. X. Liu, T. Chen, J. Rittscher. Optimal pose for face recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2006.
25. D.G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004)
26. J. Shi and J. Malik: Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), pp. 888-905, August 2000.
27. C.C. McCulloch, R.A. Kaucic, P.R.S. Mendonça, D.J. Walter, R.S. Avila. Model-based detection of lung nodules in computed tomography exams. Academic Radiology **11** (2004) 258–266
28. Y. Meng, E. Chang, B. Li. Enhancing DPF for near-replica image recognition. In: Proc. Computer Vision and Pattern Recognition. (2003)
29. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, W. Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2006.
30. F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce: 3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2003), Madison, WI, June 2003, Vol. II, pp. 272-277.
31. X. Ren, J. Malik. Learning a classification model for segmentation. In: Proc. IEEE International Conference on Computer Vision, 2003.
32. J. Rittscher, P. Tu, N. Krahnstoever. Simultaneous estimation of segmentation and shape. In: Proc. CVPR, IEEE, 2005.

33. J. Rittscher, A. Blake, A. Hoogs, G. Stein. Mathematical modeling of animate and intentional motion. Philosophical Transactions of the Royal Society of London: Biological Sciences **358** 475–490, 2003.
34. H. Sanson. Video indexing: Myth and reality. In: Fourth International Workshop on Content-Based Multimedia Indexing, Riga, Latvia (2005)
35. C. Snoek, M. Worring. Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications **25** (2005) 5–35
36. G. Stein, J. Rittscher, A. Hoogs. Enabling video annotation using a semantic database extended with visual knowledge. In: Proceedings of the International Conference on Multimedia and Expo, IEEE, 2003.
37. P. Tu, P. Mendonca. Surface reconstruction via helmholtz reciprocity with a single image pair. In: Proc. CVPR. (2003)
38. P. Tu, J. Rittscher, T. Kelliher. In: Challenges to Fingerprints, 2005.
39. P. Tu, R. Hartley. Statistical significance as an aid to system performance evaluation. In: European Conference On Computer Vision. Volume II, 366–378, 2000.
40. P. Tu, R. Hartley, A. Allyassin, W. Lorensen, R. Gupta, L. Heier. Face reconstructions using flesh deformation modes. In: International Association for Craniofacial Identification, 2000.
41. P. Tu, J. Rittscher, T. Kelliher. Site calibration for large indoor scenes. In: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, IEEE (2003)
42. M. Varma, A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In: Proc. European Conference on Computer Vision. Volume 3, 255–271, 2002.
43. R.C. Veltkamp, M. Tanase. Content-based image retrieval systems: A survey. IEEE Image Processing **1** (2001) 100–148
44. P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features. In: Proceedings of IEEE Int. Conf. on Computer Vision and Patttern Recognition, Hawaii, US, 511–518, 2001.