Mobile museum guide based on fast SIFT recognition

Boris Ruf[1], Effrosyni Kokiopoulou[1], and Marcin Detyniecki[2]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL)
{boris.ruf,effrosyni.kokiopoulou}@epfl.ch
[2] Laboratoire d'Informatique de Paris 6 (LIP6)
marcin.detyniecki@lip6.fr

Abstract. This article explores the feasibility of a market-ready, mobile pattern recognition system based on the latest findings in the field of object recognition and currently available hardware and network technology. More precisely, an innovative, mobile museum guide system is presented, which enables camera phones to recognize paintings in art galleries.
After careful examination, the algorithms Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) were found most promising for this goal. Consequently, both have been integrated in a fully implemented prototype system and their performance has been thoroughly evaluated under realistic conditions.
In order to speed up the matching process for finding the corresponding sample in the feature database, an approximation to Nearest Neighbor Search was investigated. The k-means based clustering approach was found to significantly improve the computational time.

1 Introduction

1.1 Motivation

Worldwide, sales of camera phones are skyrocketing. Almost every new cellphone purchased today is equipped with a built-in camera, and camera phones are projected to outsell digital standalone cameras within a few years. The Gartner Group estimates that in 2006 nearly 460 million camera phones were shipped and it forecasts that number to hit one billion devices by 2010 [1].

Cellphones have clearly evolved beyond mere conversational communication devices to ubiquitous imaging devices that support various forms of multimedia. This prevalence, coinciding with rapidly advancing communication infrastructures, initiated a growing interest in the application of image recognition on mobile devices. Using them as interactive user interfaces and image sensors has the great potential to augment the user's reality.

Several applications have already been envisioned such as bar code scanners [2], image-based object search [3] and an urban navigation system [4].

The domain this project deals with is the appealing idea of an enhanced museum tour guide. Today, museums and art galleries usually provide visitors

either with paper booklets or with audio guides providing an contrived identification of system. The prototype presented here enables a camera phone to act as a museum guide: the user points with his camera phone to the painting of interest and takes a picture. Image processing technology recognizes the input picture and provides multi-modal, context-sensitive information regarding the identified painting. Details such as title, artist, historical context, critical review can be easily communicated to the visitor in the language of his choice. Such an augmented reality application could assist to appreciate art more deeply and also make it more accessible to everyone.

Using cell phones as a platform for personal museum guides would have several advantages over current audio guide systems: the interaction of taking a snapshot is found more intuitive than finding an object's number and typing it into the device. Moreover, the identification can be performed not only for the global painting, but also for details. For instance particular faces or sub-scenes of large painting or frescoes can, if the description is available, be identified.

Finally from an economical point of view, either museum operators profit by significantly reducing maintenance and specific infrastructure costs or tourist operators can develop their own products, since the visitor can use his own mobile device.

1.2 Problem statement

Object recognition is still an open problem in computer vision, and the reasons for this are numerous. Images may be subject to variations in point of view, illumination and sharpness; different camera characteristics can also be an issue. Moreover, the museum environment has some unique properties: indoor lighting in museums can be insufficient and museum rules may prohibit using a flash. Reflection of security glass which protects pieces of art is another challenge. Camera phones still tend to have cheap lenses that produce noisy photographs of poor quality. As cell phones are not primarily designed for taking pictures they are more difficult to hold steady which in turn increases the likelihood of camera shake. In a crowded museum paintings might be partly occluded by other visitors or even cropped if the piece of art is too vast to be captured at once. Also, more than one painting may appear on the image if the paintings have been arranged close together. Frames can vary from bold, rectangular ones to subtle, oval ones and cast significant, shadowed regions. Both the shape and shadows of the frame complicate a possible segmentation of the painting incredibly. More difficulties become obvious when considering the content of the painting: the uniqueness of features is reduced as paintings from the same epoch show recurring styles and similar color schemes. In fact, in the case of studies, whole patches of some paintings can be found repeated in other paintings.

The aim of this work was to overcome these problems in a mobile real-life image matching application.

Most systems presented in related work in mobile visual communication have actually been simulated on desktop PCs. This project firmly intended to deploy the client software on a real hand-held device and evaluate its handling under

the most realistic conditions possible. For the same reason, a large database and many test samples were chosen. These requirements bear additional challenges to the implementation.

1.3 Related work

Object recognition Two major families of methods have evolved in the field of object recognition. The holistic *global* feature approach handles the entire image as one entity, while the *local* feature approach selects distinctive regions in the image.

The most obvious *global* features are color histograms. A recognition system based on color histograms was presented by Swain and Ballard [5] in 1991. Face recognition is a well explored domain which often relies on global features [6], [7]. Some of the most popular algorithms in this field include Eigenfaces [8], which uses Principal Component Analysis (PCA), and Fisherfaces [9], which adopts Fisher Linear Discriminant Analysis (FLD). PCA methodologies select a dimensionality reducing linear projection which models the data by maximizing its scatter. FLD techniques attempt to improve reliability for classification problems by taking into account the classes and maximizing the ratio between the intra-class and the extra-class.

In 1988, Harris and Plessey introduced the Harris corner detector [10] to find local interest points. Mohr et al. later applied this concept to locate invariant features and matched them against a large database [11]. In 1996, van Gool introduced generalized color moments that represent the shape and the intensities of different color channels in a local region [12]. In 1999, Lowe presented the Scale-Invariant Feature Transform algorithm (SIFT) which achieved scale invariance using local extrema detected in Gauss-filtered difference images for object recognition [13]. In 2002, Siggelkow showed methods to use local feature histograms for content-based image retrieval [14]. In the same year Schaffalitzky and Zisserman investigated how a combination of image invariants, covariants, and multiple view relations can be used for efficient multiple view matching [15]. Mikolajczyk and Schmid used the differential descriptors to approximate a point neighborhood [16]. In 2004, Till Quack et al. introduced Cortina, a large-scale image retrieval system for images of the Web based on low-level MPEG-7 visual features and indexed keywords as additional high-level feature.[17] Combined with association rule mining this concept sucessfully improved the quality of the search results.

Experimental museum guide systems In 2002, Kusonoki et al. presented a location-aware sensing board for kids which gives visual and auditory feedback to attract users' interests. Interactive museum tour-guide robots have been proposed by Burgard [18] in 1998 and Thrun [19] in 2000.

In January 2005, Adriano Alberti et al. described an augmented reality system using video see-through technology that provides contextual information for details of one painting [20]. The system is trained with a large number of

Fig. 1. High-level architecture of the prototype system

synthetically generated images. The recognition process utilizes a set of multidimensional receptive field histograms that represent features such as hue, edginess and luminance.

An Interactive Museum Guide [21] that is capable of recognizing objects in the Swiss National Museum in Zurich was proposed by Herbert Bay et al. in September 2005. In order to reduce the search space, Bluetooth emitters were installed on site. Objects are recognized with an approximated SIFT algorithm.

In October 2005 Erich Bruns et al. from the Bauhaus University in Weimar presented the PhoneGuide [22]. Two-layer neural networks are used in combination with Bluetooth emitters and trained directly on the mobile phone. All computation for object recognition is carried out on the device.

The French-Singapore IPAL Joint Lab presented in July 2007 the Snap2Tell prototype [23] which recognizes tourist attractions and provides multi-modal descriptions. Scenes are recognized by distinguishing local discriminative patches described by color and edge information. As discriminative classifiers Support Vector Machines (SVMs) are used. The reference database contains a notable number of images per object and GPS was evaluated as additional feature.

2 System description

2.1 Architecture

A PDA with integrated camera and Internet connection was enabled to act as a universal museum guide for paintings in art galleries. In contrast to conventional audio museum guides or booklets, objects are selected by simply taking a picture of them.

The major advantage of the system presented here over other experimental systems that have been proposed in previous work on the same subject is that

it does not depend on additional infrastructure on site. Neither barcode labels nor extra hardware such as Bluetooth emitters need to be set up.

The architecture follows the classical server-client approach: the client only acts as periphery which acquires and sends sample data and eventually receives the results. No additional computation such as feature extraction is executed on the client. This decision has been taken for several reasons: the CPU of mobile clients is generally very slow and running the feature extraction on the mobile client might result in unbearably long waiting times for the user. Also, the recognition performance is good even at low resolution. Transmitting scaled down images of small data size is sufficient for successful operation of the system.

Mobile clients have been developed for Windows Mobile and the Android operating system by Google. Furthermore, a web browser-based interface was implemented to enable access to the painting recognition system through the Internet.

A very basic, high-level description of the architecture of the system is shown in Figure 1.

2.2 Hardware

All images were captured with a HP iPAQ hw6900 handheld device and have an original resolution of 1280×1024. The experiments were conducted on a virtual private server equipped with an Intel(R) Xeon(TM) CPU 2.80GHz, 384 MB RAM and Debian Linux 3.1.

3 Feature extraction

After evaluating several methods for object recognition, Scale-Invariant Feature Transform (SIFT), conceived by David G. Lowe et al. in 1999, and Speeded Up Robust Features (SURF), introduced by Herbert Bay et al. in 2006, were identified as most appropriate for museum-inherent challenges. Both are robust regarding scale, lighting and perspective distortion. But, again, their greatest benefit is the use of local features. When employing algorithms with global features, the objects of interest first need to be clipped away from any background. In this case, the reference samples in the database show only the painting with neither frame nor background. The test samples taken in the museum, however, include parts of the environment: often, paintings are surrounded by massive frames. The wall does not always contrast clearly with the piece of art. Visitors or objects besides the painting of interest may appear on the photos. If the image was taken from a distance, the size of the painting proportional to the total image size can vary significantly. Detecting the painting becomes particularly challenging when it is surrounded by shadowed regions or if the frame is of unusual shape like oval. Segmentation techniques for clipping away the background before classifying the foreground are expensive and prone to failure due to these factors. This step can be skipped when using local features.

3.1 Scale-Invariant Feature Transform (SIFT)

The Scale-Invariant Feature Transform (SIFT) [13] algorithm provides a robust method for extracting distinctive features from images that are invariant to rotation, scale and distortion. In order to identify invariant keypoints that can be repeatably found in multiple views of varying scale and rotation, local extrema are detected in Gauss-filtered difference images. Stability of the extrema is further ensured by rejecting keypoints with low contrast, and keypoints localized along edges. As keypoint descriptor, an orientation histogram is computed for the area around the keypoint location. Gradient magnitude and the weight of a Gaussian window originating at the keypoint add to the value of each sample point within the considered region.

3.2 Speeded Up Robust Features (SURF)

The Speeded Up Robust Features (SURF) [24] algorithm is a variation of the SIFT algorithm. Its major differences include a Hessian matrix-based measure as an interest point detector and approximated Gaussian second order derivatives using box type convolution filters. Here, the use of integral images [25] enables rapid implementation.

4 Matching process

4.1 Nearest Neighbor Search (NNS)

A straightforward approach to find the match of a sample keypoint within the reference keypoints is Nearest Neighbor Search (NNS). Here, the closest candidate measured by Euclidean distance is found by linearly iterating over all reference keypoints in no particular order. This method results in finding the exact nearest neighbor to the sample keypoint. Two keypoints are considered a match if the distance between them is closer than 0.6 times the distance of the second nearest neighbor [26] [27]. However, for large data sets and high-dimensional spaces this is an inefficient approach due to the time complexity of $O(N \cdot d)$ where N is the number reference keypoints and d is the dimensionality of a keypoint vector.

4.2 Best-Bin-First (BBF)

Jeffrey S. Beis and David G. Lowe proposed an approximation to NNS called Best-Bin-First (BBF) [28].

The index structure used to store the keypoints is a k-d tree. When creating the tree, the data set is recursively subdivided into even groups on iterating dimensions. At each split, the keypoint which contains the median becomes a new internal node. This step is repeated for the children at the next dimension on the elements of the subgroups. The resulting tree is balanced and binary with a depth $d = \lceil \log_2 N \rceil$ where N is the number reference keypoints.

In order to find the nearest neighbor of a sample keypoint, the tree is first traversed to locate the bin which contains the sample keypoint. The algorithm backtracks from this bin, considering each node along the way for comparison. If the distance to a node is greater than the shortest distance found so far, the subtree of this node can be ignored.

According to [28], with x=200, this approximation provides a 2 order of magnitude speed-up over exhaustive NNS, and still returns the correct nearest neighbor more than 95% of the time. In our case, however, preliminary tests on a subset of the data revealed unacceptable loss of performance.

4.3 K-means based tree

A different tree-based clustering approach adopted from the paper "Tree-Based Pursuit: Algorithm and Properties" by Jost et al. [29] was evaluated.

Here, clustering is achieved based on the Euclidean distance between the vectors. The k-means algorithm [30] is used to cluster the data set into k subgroups. The centroids found become internal nodes of the tree. Recursively, the clusters are subdivided in the same manner until they consist of less than k elements. Once this state is reached, the elements of the cluster become children of their centroid node and leaf nodes of the tree. The resulting tree is not balanced and its shape highly depends on the data set, the quality of the initial centers and the value of k. The matching process of a new element breaks down to tree traversal from the root node to the bottom of the tree always choosing the node of lowest Euclidean distance. The leaf node is then considered the nearest neighbor.

5 Experimental results

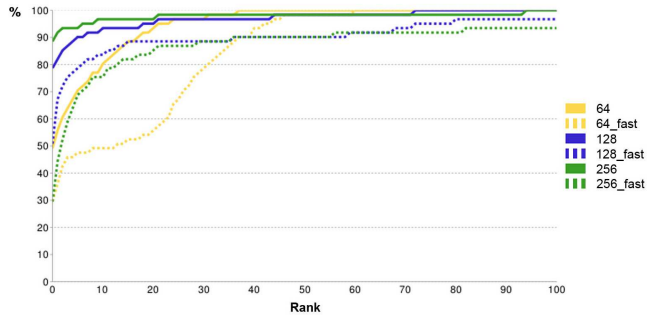
5.1 Setup

Training sample data has been extracted from the online archive Web Gallery of Art [31]. More precisely, all 1,002 works available from the Louvre Museum were considered in the experiment. Each reference painting is represented by one sample. The paintings from the online source have been digitalized without frame.

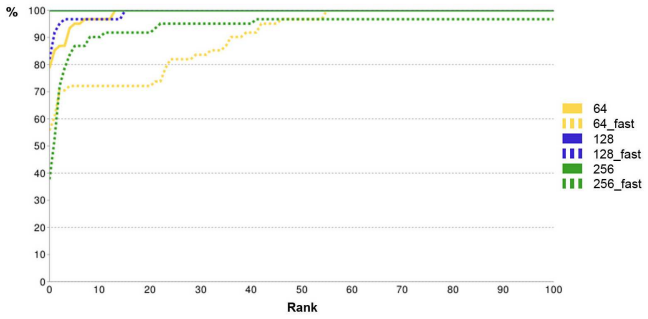
The test sample data consists of photo series of 48 paintings taken in the Louvre Museum (in total 200 images). Four different types of perspective have been considered to stress test the algorithms and also evaluate their robustness under extreme perspectives: frontal, left, right, distant.

In order to remove noise, the images have been converted to gray-level representation. To evaluate the correlation between resolution and performance of the algorithms, the images have been downsampled to 4 different resolutions: 512×410, 256×205, 128×103, 64×51.

Cumulative Match Characteristic (CMC) curves summarize the accuracy of a recognition system: for each test sample, its rank is determined by finding the position of the hypothesis for the desired, correct reference sample on a sorted



(a) All perspectives



(b) Frontal

Fig. 2. Performance comparison for approximated NNS

list of all hypotheses constructed for this sample. Ideally, the rank is 0. In this case, the hypothesis for the correct painting also received the most votes, and the sample could be identified successfully. The CMC chart integrates these results and depicts the probabilities of identification for all ranges of ranks.

5.2 SIFT vs. SURF

Figure 3 shows CMC curves for the results of linear matching using NNS, grouped by perspective. The charts on the left side result from employing the SIFT algorithm, corresponding curves for the SURF algorithm can be found on the opposite side.

It can be seen that higher resolution does not necessarily correspond to better recognition rate. For the frontal perspective, even very low resolution yields satisfying results.

5.3 Approximated SIFT

The k-means based clustering approach has been coined *SIFT_fast* and was implemented with $k = 15$.

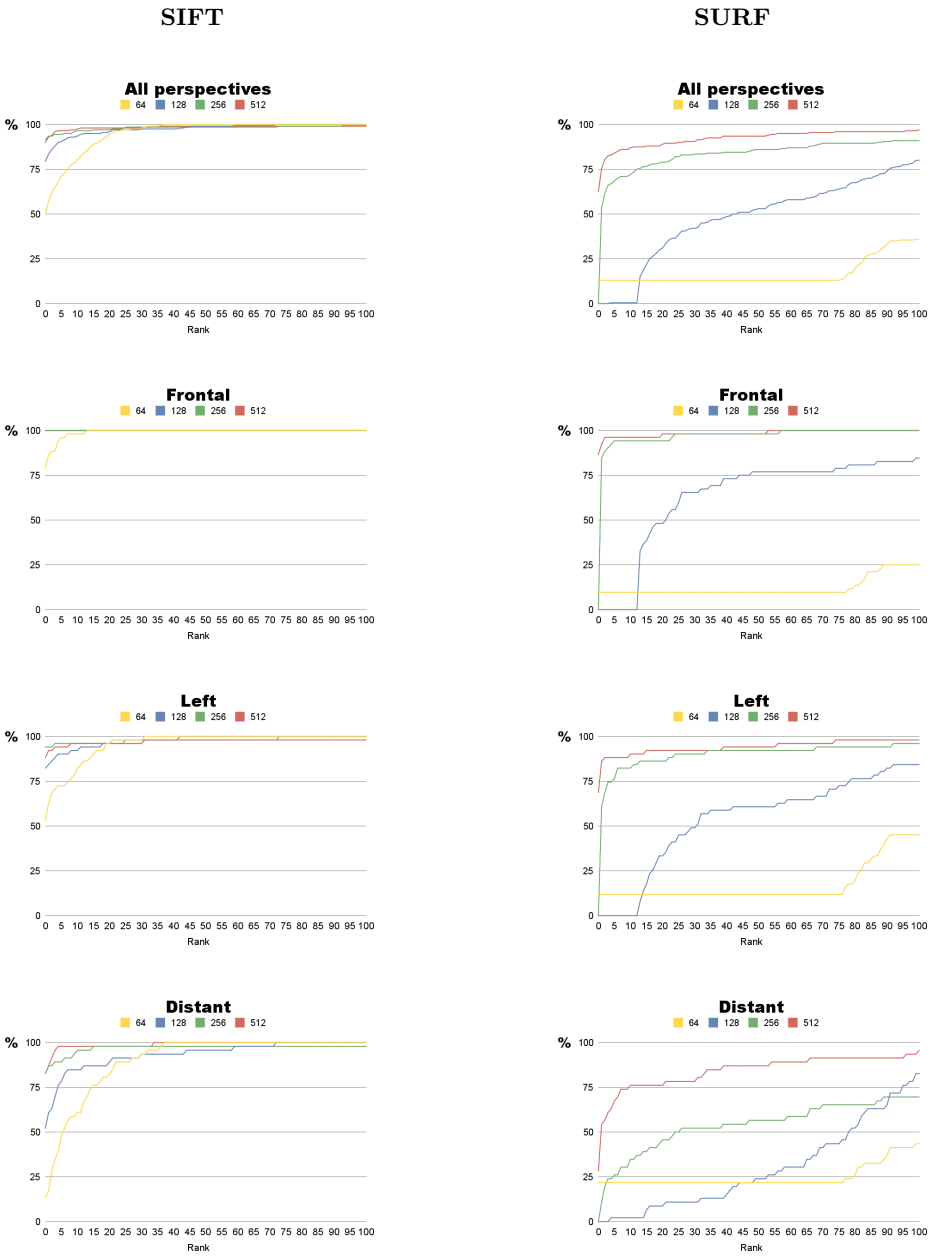


Fig. 3. CMC curves illustrate the probabilities of identification for SIFT and SURF

Figure 2 shows CMC curves for the experiment results using a linear matching approach (solid lines) and the approximated k-means tree approach introduced in Section 4.3 (dashed lines). Performance losses compared to the exhaustive approach are obvious, however, for a real-time application that deals mainly with frontal views (as the museum guide in this work does), the algorithm *SIFT_fast* for resolution 128 offers an acceptable trade-off between speed and performance.

5.4 Processing time

The table in Figure 4 lists the average times of the matching process depending on algorithm and resolution; Figure 5 clarifies the proportions graphically.

The runtime computational complexity of SURF is lower for all resolutions. This is due to the fact that SURF descriptor vectors are of dimension 64 in contrast to 128 components contained in the descriptor vectors of SIFT. However, the median of the number of keypoints is lower, too, which has direct influence on the recognition performance: SURF is inferior to SIFT in any experiment.

The large time increase of the conventional SIFT algorithm between resolution 128 and 256 can be explained by the huge variance of keypoints of this algorithm. The variance of SURF keypoints is much smaller in comparison. This is beneficial as it makes the runtime of the matching process more predictable.

The gain of time achieved when matching SIFT keypoints using a k-means tree compared to linear NNS is significant: with resolution 128, the approximated approach takes 45 seconds instead of about 306 seconds using linear NNS matching. The downside clearly is a loss of performance as shown in Figure 2.

	64	128	256
SIFT	144.25	305.73	1440.36
SURF	78.54	95.56	198.57
SIFT_fast	13.85	44.65	150.68

Fig. 4. Table of average processing times in seconds

6 Discussion

In general, the evaluation reveals that the SIFT algorithm outperforms the SURF algorithm for any resolution considered. However, the runtime computational complexity of SURF is lower due to the fact that SURF descriptor vectors are of lower dimension than descriptor vectors of SIFT. The variance of the number of keypoints found with SURF is much smaller compared to the distribution of SIFT keys. This is advantageous as it makes the runtime of the matching process more predictable. However, the median is lower, too, which has direct influence on the recognition performance. In fact, the strength of the SURF algorithm only becomes apparent at the highest resolution of 512 $\times$ 410 tested in the

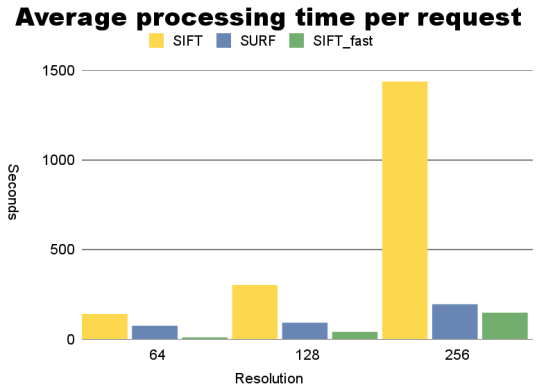


Fig. 5. Average processing times visualized for comparison

experiments. SIFT features, on the other hand, show sufficient distinctive power even for images of significantly lower resolution than used in the experiment section of the SIFT paper (600×315). Our experiments show that input images of 128×103 already deliver reasonable performance.

These findings, and the fact that programming on mobile platforms is rather cumbersome, implies an architecture on which the feature extraction part is done on the server.

Analysis of the experimental data also clearly showed that perspective distortion is still an issue. However, for an application as described in this project, it is acceptable to assume a frontal perspective and to choose rather low resolution parameters in order to strike a balance between efficiency and accuracy.

Moreover, clustering methods which approximate the conventional Nearest Neighbor Search are an important extension to a recognition system, in particular to a real-life application such as this one. In fact, they enormously speed up the response time. The tests show that the k-means based tree approach provides an acceptable trade-off between performance loss and gain of time.

7 Conclusion

The results presented in this article demonstrate the feasibility of a market-ready mobile pattern recognition system in the form of a universal museum guide. Several prototype clients were fully implemented and have been subject to thorough evaluation under realistic conditions.

Our tests showed the advantages of an architecture where the feature extraction part is done on the server. Such a setup requires uploading images and favors low resolutions, as this decreases the response time. Although the SURF algorithm is faster than the SIFT one, for low resolution images SURF's performance is unacceptable.

Our tests further showed that methods which approximate the conventional Nearest Neighbor Search can also reduce response times. The k-means based tree approach provided an acceptable trade-off between performance loss and gain of time.

Finally, based on this study, we conclude that a combination of client-server architecture, the use of a SIFT algorithm with a resolution of 128 × 103, combined with the k-means based tree approach is most appropriate for deployment.

The extension of the presented framework to standard representations such as MPEG-7 would require deeper examination and remains as interesting objective for the future.

Acknowledgments

Many thanks to Prof. Pascal Frossard for providing valuable feedback on this article. Special thanks to Dr. Emil Krén and Dr. Dániel Marx without their generous permission to use the entire image data from Web Gallery of Art [31] the experiments would not have been possible.

References

1. G. Group, "2006 Press Releases," November 2, 2006, http://www.gartner.com/it/page.jsp?id=498310.
2. M. Rohs and B. Gfeller, "Using camera-equipped mobile phones for interacting with real-world objects," in *Advances in Pervasive Computing*. Vienna, Austria: Austrian Computer Society (OCG), 2004, pp. 265–271.
3. T. Yeh, K. Grauman, K. Tollmar, and T. Darrell, "A picture is worth a thousand keywords: image-based object search on a mobile platform," in *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, 2005, pp. 2025–2028.
4. D. Robertsone and R. Cipolla, "An image-based system for urban navigation," in *The 15th British Machine Vision Conference (BMVC04)*, 2004, pp. 819–828.
5. M. J. Swain and D. H. Ballard, "Color indexing," vol. 7, no. 1. Hingham, MA, USA: Kluwer Academic Publishers, 1991, pp. 11–32.
6. R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey." in *Proceedings of the IEEE*, vol. 83, no. 5, May 1995, pp. 705–740.
7. A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," in *Pattern Recogn.*, vol. 25, no. 1. New York, NY, USA: Elsevier Science Inc., 1992, pp. 65–77.
8. M. Turk and A. Pentland, "Eigenfaces for recognition," in *CogNeuro*, vol. 3, no. 1, 1991, pp. 71–96.
9. P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul 1997.
10. C. Harris and M. Stephens, "A combined corner and edge detection," in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147–151. [Online]. Available: http://www.csse.uwa.edu.au/~pk/research/matlabfns/Spatial/Docs/Harris/A_Combined_Corner_and_Edge_Detector.pdf

11. C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–535, 1997.
12. L. J. V. Gool, T. Moons, and D. Ungureanu, "Affine/ photometric invariants for planar intensity patterns," in *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume I*. London, UK: Springer-Verlag, 1996, pp. 642–651.
13. D. Lowe, "Object recognition from local scale-invariant features," *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, pp. 1150–1157 vol.2, 1999.
14. S. Siggelkow, "Feature histograms for content-based image retrieval," Ph.D. dissertation, University of Freiburg, Institute for Computer Science, 2002.
15. F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or 'How do i organize my holiday snaps?'," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*. London, UK: Springer-Verlag, 2002, pp. 414–431.
16. K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of the 7th European Conference on Computer Vision-Part I*, 2002, pp. 128–142.
17. T. Quack, U. Monich, L. Thiele, and B. Manjunath, "Cortina: A system for large-scale, content-based web image retrieval," in *ACM Multimedia 2004*, Oct 2004.
18. W. Burgard, A. Cremers, D. Fox, D. Hhnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The interactive museum tour-guide robot," in *Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.
19. S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hhnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "Probabilistic algorithms and the interactive museum tour-guide robot Minerva," in *International Journal of Robotics Research, 19(11)*, 2000, pp. 972–999.
20. A. Albertini, R. Brunelli, O. Stock, and M. Zancanaro, "Communicating user's focus of attention by image processing as input for a mobile museum guide," in *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2005, pp. 299–301.
21. H. Bay, B. Fasel, and L. V. Gool, "Interactive museum guide: Fast and robust recognition of museum objects," in *Proceedings of the first international workshop on mobile vision*, 2006.
22. E. Bruns, B. Brombach, T. Zeidler, and O. Bimber, "Enabling mobile phones to support large-scale museum guidance," in *IEEE MultiMedia*, vol. 14, no. 2. Los Alamitos, CA, USA: IEEE Computer Society, 2007, pp. 16–25.
23. J.-H. Lim, Y. Li, Y. You, and J.-P. Chevallet, "Scene recognition with camera phones for tourist information access," *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 100–103, 2-5 July 2007.
24. H. Bay, T. Tuytelaars, and L. J. V. Gool, "SURF: Speeded Up Robust Features." in *ECCV (1)*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951. Springer, 2006, pp. 404–417.
25. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, vol. 1, 2001, pp. I–511 – I–518.
26. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
27. A. Baumberg, "Reliable feature matching across widely separated views," in *Computer Vision and Pattern Recognition, 2000. Proceedings*, vol. 1, 2000, pp. 774–781.

28. J. Beis and D. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 1000–1006, 17-19 Jun 1997.
29. P. Jost, P. Vandergheynst, and P. Frossard, "Tree-Based Pursuit: Algorithm and Properties," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4685–4697, 2006.
30. J. B. Macqueen, "Some methods of classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
31. E. Kren and D. Marx, "Web Gallery of Art," http://www.wga.hu.