

IDENTIFYING PAINTINGS IN MUSEUM GALLERIES USING CAMERA MOBILE PHONES

BORIS RUF[†]

*LIP6, University Pierre and Marie Curie (UPMC), 104, av. du Président Kennedy
Paris, 75015, France*

MARCIN DETYNIECKI

*LIP6, Centre national de la recherche scientifique (CNRS), 104, av. du Président
Kennedy
Paris, 75015, France*

This work focuses on the viability of using a cell-phone as mobile museum guidance. The integrated cell-phone camera is used to recognize the paintings in the gallery.

The chosen solution is based on a client-server architecture and the object recognition is based on local features. The study focuses on the comparison, in terms of time and performance, of the Scale-Invariant Feature Transform (SIFT), the Speeded Up Robust Features (SURF), the Nearest Neighbor Search (NNS) match and a k-means trees based search. It was found that SIFT outperforms SURF in terms of performance but is dominated in terms of time. Finally, the combination of SIFT and k-means based search provides a good compromise for the low-resolution images necessary in this setup.

The study was performed using a windows mobile operated cell-phone and the 200 test images were taken on site from 4 different perspectives. The reference data set consisted of 1002 different art works of the Louvre.

Today, museums and art galleries usually provide visitors either with paper booklets or with audio guides providing a 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, and 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.

[†] This work was done in collaboration with the Ecole Polytechnique Fédérale de Lausanne (EPFL) under the supervision of Prof. Effrosyni Kokiopoulou and Prof. Pascal Frossard.

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 benefit 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.

1.1. *Related experimental museum guide systems*

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.

An Interactive Museum Guide [1] 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 [2]. 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 [3] 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.

The major advantage of the system presented in this paper over other experimental systems that have been proposed 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.

2. Object recognition using a cell phone

2.1. *The Museum Environment*

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.

In order to successfully recognize an object with a cell phone four steps have to be performed. First an image containing the object has to be acquired. We used a PDA phone with an integrated camera. Since images cannot be compared as they are, in a second step, a signature has to be extracted. Third, a matching process is run against a database containing the signatures of all the paintings of the museum. Finally the object id is returned.

2.2. *Signature extraction*

After evaluating several methods for object recognition, Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) 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.

2.2.1. Scale-Invariant Feature Transform (SIFT)

The Scale-Invariant Feature Transform (SIFT) [4] 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.

2.2.2. Speeded Up Robust Features (SURF)

The Speeded Up Robust Features (SURF) [5] 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 enables rapid implementation.

2.3. Matching process

2.3.1. Nearest Neighbor Search (NNS)

A straightforward approach to find the match of a sample keypoint within the reference keypoints is Nearest Neighbor Search (NNS) [6]. Here, the closest candidate measured by Euclidean distance is found by linearly iterating over all reference keypoints. This method results in finding the exact nearest neighbor to the sample keypoint. However, for large data sets and high-dimensional spaces this is an inefficient approach due to the time complexity.

2.3.2. *Best-Bin-First (BBF)*

Jeffrey S. Beis and David G. Lowe proposed an approximation to NNS called Best-Bin-First (BBF) [7]. An index structure is used to store the keypoints: the k -d tree. According to [7], 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.

2.3.3. *K-means based tree*

A different tree-based clustering approach adopted from the paper "Tree-Based Pursuit: Algorithm and Properties" by Jost et al. [8] was evaluated. Here, clustering is achieved based on the Euclidean distance between the vectors. The k -means algorithm [9] 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.

2.4. *System Architecture*

A PDA with integrated camera and Internet connection was enabled to act as a universal museum guide for paintings in art galleries.

Some preliminary tests showed that the CPU of mobile clients is generally very slow and running the feature extraction on the mobile client results in unbearably long waiting times for the user. Thus the classical server-client architecture was chosen: the client only acts as periphery, which acquires and sends sample data and eventually receives the results.

In the one hand, since in this configuration images have to be sent over the phone connection, we are interested in transmitting scaled down images representing smaller data size. In the other hand, smaller resolution may imply a decrease in recognition performance. In the following we study the corresponding trade-off on set of experiments.

3. Experiments

The reference database was extracted from the online archive Web Gallery of Art [10]. More precisely, all 1,002 works available from the Louvre Museum were considered in the experiment. Each reference painting is represented by one *single* sample. The paintings from the online source are 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.

All images were captured with a HP iPAQ hw6900 handheld device and have an original resolution of 1280x1024. 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.

In order to remove noise, the images were converted to gray-level representation. To evaluate the correlation between resolution and performance of the algorithms, the images have been down-sampled to 4 different resolutions: 512x410, 256x205, 128x103, 64x51.

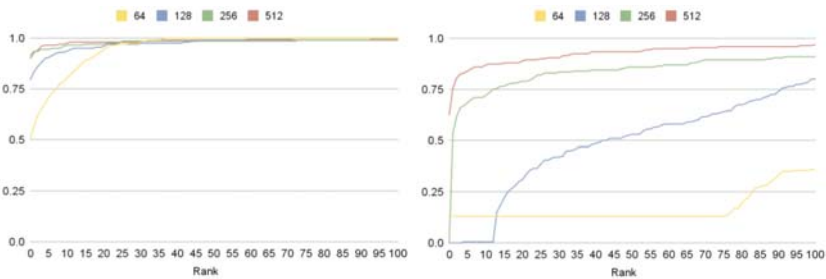


Figure 1. CMC curves comparing (on the left) SIFT based NNS matching and (on the right) SURF based NNS matching, for all perspectives taken together. Four different resolutions are considered.

3.1. *SIFT* vs. *SURF*

A set of test was run to compare the compare accuracy when using SIFT or SURF. To avoid errors due to the matching process the linear NNS was used.

The resulting Cumulative Match Characteristic (CMC) curves (Figure 1) show that SIFT out performs SURF for all of the considered resolutions and for all perspectives (frontal, left, right, distant). We also observed this tendency for any individual perspective. Moreover, SURF performs always *very* poorly for the

lowest resolution (64x51). Finally, for the frontal perspective the use of SIFT yields to satisfying results, even for very low resolutions.

3.2. *K-means based matching vs. NNS matching (on SIFT)*

The k-means based clustering approach has been coined *SIFT_fast* and was implemented with $k=15$. Figure 2 shows CMC curves for the experiment results using a linear matching approach (solid lines) and the approximated k-means tree approach. Performance losses compared to the exhaustive approach are obvious, however, for a real-time application dealing mainly with frontal views (as the museum guide in this work does), the algorithm *SIFT_fast* for resolution 128x103 offers an acceptable trade-off between speed and performance.

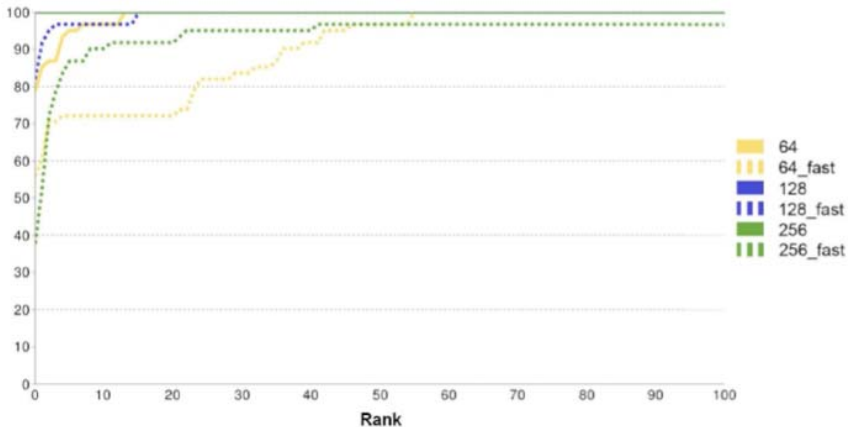


Figure 2. CMC curve comparing for the frontal perspective and on SIFT features, the accuracy for different resolutions and different matching algorithms: the linear NNS and the fast k-means tree.

3.3. *Processing time*

Table 1 compares the average times of the matching process for three resolution and for different approaches.

The runtime computational complexity of SURF is lower for all resolutions, but as we observed earlier the recognition performance is inferior. The large time increase of the conventional SIFT algorithm between resolution 128x103 and 256x205 can be explained by the huge variance of keypoints of this algorithm.

The gain of time achieved when matching SIFT keypoints using a k-means tree compared to linear NNS is significant: with resolution 128x103, the approximated approach takes 45 seconds instead of about 306 seconds using

linear NNS matching. The downside is a loss of performance as shown in Figure 2

Table 1. Average processing times per request (in second).

Resolution	SIFT	SURF	SIFT_Fast
64x51	144	79	14
128x103	306	96	45
256x205	1440	199	150

4. 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 tested resolution: 512x410. 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 (600x315). Our experiments show that input images of 128x103 already deliver reasonable performance.

These findings, and the fact that programming on mobile platforms is rather cumbersome, confirms the choice of the architecture, where 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.

5. 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 128x103, combined with the k-means based tree approach is most appropriate for deployment.

Based on this two mobile clients have been developed: one for Windows Mobile and one for 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.

Acknowledgments

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 [10] the experiments would not have been possible.

References

1. 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).
2. 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, pp. 16–25 (2007).
3. J.-H. Lim, Y. Li, Y. You, and J.-P. Chevallet, "Scene recognition with camera phones for tourist information access," in *proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 100–103 (2007).

4. D. Lowe, "Object recognition from local scale-invariant features," in the *proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 (1999).
5. H. Bay, T. Tuytelaars, and L. J. V. Gool, "SURF: Speeded Up Robust Features." in *ECCV, Lecture Notes in Computer Science*, vol. 3951. Springer, pp. 404–417 (2006).
6. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630 (2005).
7. J. Beis and D. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in the proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1000–1006 (1997).
8. 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)
9. S. P. Lloyd. "Least squares quantization in PCM," in *special issue on quantization, IEEE Transactions in Information Theory*, 28:129–137 (1982)
10. E. Kren and D. Marx, "Web Gallery of Art," <http://www.wga.hu>