INDOOR LOCATION RECOGNITION USING FUSION OF SVM-BASED VISUAL CLASSIFIERS

Mats Sjöberg, Markus Koskela, Ville Viitaniemi, Jorma Laaksonen

Adaptive Informatics Research Centre
Aalto University School of Science and Technology
P.O. Box 15400, FI-00076 Aalto, Finland
firstname.lastname@tkk.fi

ABSTRACT

We apply our general-purpose algorithm for visual category recognition using bag-of-visual-words and other visual features and fusion of SVM classifiers to the recognition of indoor locations. This is an important application in many emerging fields, such as mobile augmented reality and autonomous robots. We evaluate the proposed method with other location recognition systems in the ImageCLEF 2010 RobotVision contest. The results show that given a large enough training set, a purely appearance-based method can perform very well – ranked first for one of the contest's training sets.

1. INTRODUCTION

Visual category recognition has recently attracted a lot of attention in the computer vision research community. This is largely due to the emergence and success of *bag-of-features* approaches, in which objects and scenes are represented as unordered sets of feature descriptors. A popular technique is to extract a set of local image descriptors and represent these image-wise descriptors as histograms via clustering. These are known as *bag-of-visual-words* (BoV) methods, in analogy to the *bag-of-words* approach in textual information retrieval [1].

For building detectors for visual categories, Support Vector Machines (SVMs) can be considered as the current *de facto* standard. The basic SVM is a binary classifier, but there are several approaches to extend SVMs for multi-class classification. The most common approaches are *one-versus-the-rest* and *one-versus-one* or pairwise classification. In the former one trains a binary classifier for each class taking all the other classes as negative examples, while in the latter the system learns to separate each possible pair of classes. In any non-trivial case the one-versus-one approach requires many more classifiers than the one-versus-the-rest approach, however the binary problems are substantially smaller.

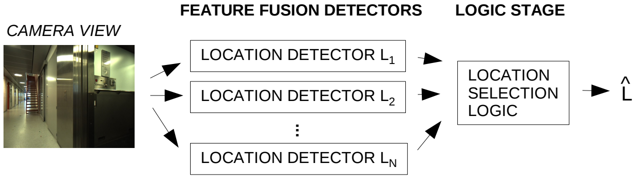


Fig. 1. General architecture for predicting location $\hat{L}$ based on a camera view.

Indoor localisation is a fundamental task for autonomous robots [2, 3]. A number of different approaches have been proposed, but arguably the prevailing method is to combine camera-based visual information to some additional input modalities [4, 5], such as laser range sensors, sonar, stereo vision, temporal continuity, odometry, and floorplan of the environment. In addition to location category recognition, an alternative vision-based method is to match the query image directly to the images in the training set, using e.g. pairwise matching of interest points [6].

In this paper, we apply our general-purpose algorithm for visual category recognition using low-level features and SVMs [7] to the classification of indoor locations. The only modality we consider in this work is the current view from one or more forward-pointing cameras. In addition to autonomous robots, this kind of setup arises, for example, in many applications of mobile augmented reality [8]. In fact, indoor localisation constitutes also one of the sub-tasks in the development of our research platform for accessing abstract information in real-world environments through augmented reality displays [9]. We evaluate our method by comparing with the state-of-the-art systems taking part in the ImageCLEF@ICPR 2010 RobotVision contest[1].

Fig. 1 illustrates our approach. Given training images

[1] http://www.imageclef.org/2010/ICPR/RobotVision/

with location labels, we first train a separate detector for each location L_i. Section 2 describes these single-location detectors that employ fusion of several SVM detectors, each based on a different visual feature. The probabilistic outcomes of the detectors are then used in Section 3 as inputs to multi-class classification step that determines the final location label $\hat{L}$ for a test image. $\hat{L}$ is one of the known locations L_i. Alternatively, the system can predict that the image is taken in a novel unknown location, or declare the location to be uncertain.

In Section 4 we describe our experiments in RobotVision 2010 and summarise the results. Finally, conclusions are drawn in Section 5.

2. SINGLE-LOCATION DETECTORS

For detecting a single location L_i, our system employs the architecture illustrated in Fig. 2. The training phase begins with the extraction of a large set of low-level visual features. The features and binary location labels of the training images (L_i or non-L_i) are then used to train a set of probabilistic two-class SVM classifiers. A separate SVM is trained for each visual feature. The training images are also used in a supervised fusion stage for combining the outcomes of the SVM detectors.

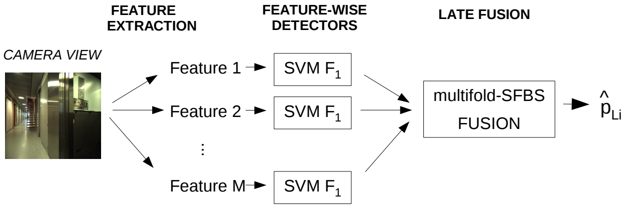


Fig. 2. Architecture for estimating the probability $\hat{p}_{Li}$ that the given camera view is from location L_i.

After training, the detector can estimate the probability of a novel test image depicting location L_i. This is achieved by first extracting the same set of visual features from the test image that were extracted from the training images. The trained feature-wise SVM detectors produce a set of probability estimates that are combined to a final probability estimate in a fusion stage. The location-wise estimates are then combined in a multi-class classification stage to determine the location of the test image.

2.1. Feature extraction

From each image, a set of low-level visual features are extracted. We use our own implementations of the following MPEG-7 features: *Color Layout*, *Dominant Color*, *Scalable Color*, and *Edge Histogram*. Furthermore, we employ some additional features, viz. *Average Color*, *Color Moments*, *Texture Neighbourhood*, *Edge Histogram*, *Edge Co-occurrence* and *Edge Fourier*. These features are described in more detail in [7].

Eight different bag-of-visual-words (BoV) features have also been extracted. In the BoV model images are represented by histograms of local image features. The features result from combining three independent design choices. First, we use either the *SIFT* [10] or the opponent color space version of the *Color SIFT* [11] descriptor. Second, we employ either the Harris-Laplace detector or use dense sampling of points as the interest point detector. For some of the features, we have additonally used the spatial pyramid extension of the BoV model. Third, we optionally use soft-histogram refinement of the BoV codebooks [11].

2.2. Feature-wise detectors

In our location recognition system, the association between an image's visual features and its location is learned using the SVM supervised learning algorithm. The SVM implementation we use in our system is an adaptation of the C-SVC classifier of the LIBSVM[2] software library. For all histogram-like visual features we employ the χ^2 kernel

$$g_{\chi^2}(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{x_i + x_i'}\right). \qquad (1)$$

The radial basis function (RBF) SVM kernel

$$g_{\mathrm{RBF}}(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right) \qquad (2)$$

is used for all the other features. The motivation for this is the well-known empirical observation that χ^2 distance is well-suited for comparing histograms.

The free parameters of the SVMs are selected with an approximate 10-fold cross-validation search procedure that consists of a heuristic line search to identify a promising parameter region, followed by a grid search in that region. To speed up the computation, the data set is radically downsampled for the parameter search phase. Further speed-up is gained by optimising the C-SVC cost function only very approximately during the search.

For the final detectors we also downsample the data set, but less radically than in the parameter search phase. Usually there are much fewer annotated example shots of a location (positive examples) than there are example shots not exhibiting that location (negative examples). Consequently, for most of the locations, the sampling is able to retain all the positive examples and just limit the number of negative examples. The exact amount of applied sampling varies according to the computation resources available and the required accuracy of the outputs. Generally we have observed the downsampling to degrade detection accuracy.

[2] http://www.csie.ntu.edu.tw/˜cjlin/libsvm

2.3. Fusion

The supervised fusion stage of our location recognition system is based on the geometric mean of feature-wise detector outcomes. However, instead of calculating the mean of all feature-wise detectors we select the set using sequential forward-backward search (SFBS). This supervised variable selection technique requires detector outcomes also for training images. These outcomes are obtained via 10-fold cross-validation.

Our search technique refines the basic SFBS approach by partitioning the training set into multiple folds. In our implementation we have used a fixed number of six folds. The SFBS algorithm is run several times, each time leaving one fold outside the training set. The final fusion outcome is the geometric mean of the fold-wise geometric means.

3. MULTI-CLASS CLASSIFICATION

The fusion of the feature-wise detector scores described in the previous section provides probability estimates for each location given a particular image. The final classification step is a traditional multi-class classification, where we combine several one-versus-the-rest SVM classifiers. The straightforward solution is to classify the image to the class with the highest probability estimate.

However, in the RobotVision scenario, we must also be able to detect *unknown* categories, i.e. images of new locations that have not been seen before. We have implemented this by tresholding; if all probability estimates are below a given threshold, the location is considered unknown. Furthermore it is possible to *decline* classification entirely for a particular image if we cannot determine the class with high confidence.

4. EXPERIMENTS

4.1. RobotVision competition

In the ImageCLEF RobotVision competition setting, there is a real mobile robot moving through an office environment and the recognition system should be able to answer the simple question “where are you?” when presented with a new sequence of camera images aquired by the robot. In the competition, the participants were asked to classify rooms and areas in a sequence captured at 5 fps by a mounted stereo camera rig.

Two training datasets (*easy* and *hard*) and a validation set, all from the COLD-Stockholm database [12], were released in connection with the competition. All these three sets contain a total of nine locations shown with example images in Fig. 3. The easy training set (4074 frame pairs) differs from the hard one (2267 frame pairs) by showing each location from multiple points and angles, thus giving more training data for each location. Furthermore the hard set was acquired by driving in the opposite direction from all the other sets.

Both easy and hard training sets were acquired during the day, with cloudy weather outside. The validation set was created under similar conditions as the easy set, but during the night. In addition to changing illumination, variations in the visual scene were also caused by people or various objects being variably present or absent. The test sequence has 2551 frame pairs of the same locations as the training sequence plus four previously unseen locations.

The system was expected to be trained and evaluated on both training sets separately. Each test frame pair is to be assigned to a known location or as “unknown”, or the system can refrain from making a classification. A score of $+1.0$ is awarded for each correctly classified frame. A misclassification is scored -0.5 and no score is given for unclassified frames. The final result of a run is the sum of the scores for the easy and hard sets.

In the obligatory part of the competition, the recognition system was allowed to use only the current pair of frames for the classification. In addition, it was optional to submit results using the whole sequence seen until the frame in question so that the temporal continuity of the sequence could be utilised. In this paper we, however, consider only the instantaneous case.

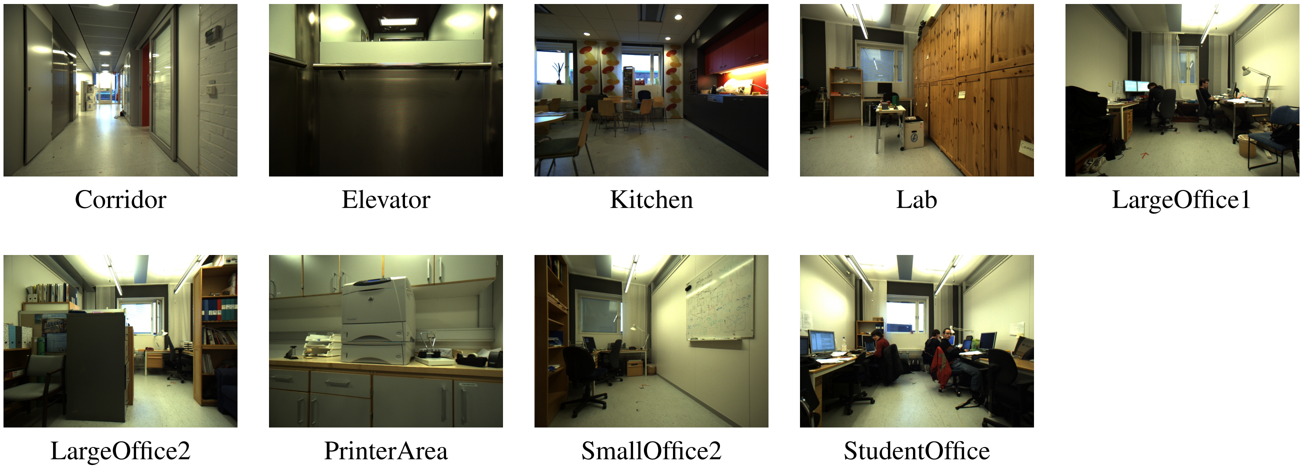
4.2. Recognition with stereo images

In the RobotVision setup, the presence of a stereo image pair demands some additional considerations. We may e.g. learn an independent model for each camera or use all images as common training data discarding the left/right distinction. The stereo image pair could also be utilised for stereo imaging, and depth information would undoubtedly be an useful feature for location recognition. In fact, the contest organisers provided the camera calibration data for the image sequences.

In our work, we are however focusing mainly on monocular camera recognition, so we ignore the stereo information. Instead, we experiment with using the two cameras separately, with averaging their detection results, and with using images from both cameras to train a single detector for each location.

4.3. Parameter selection

As discussed in Section 3, the final recognition result is obtained by selecting the location with the highest detector score from the single-location detectors. However, if all scores are below a given threshold, the frame is then assigned to the “unknown” location. This threshold was obtained by optimising the performance score in the validation



Corridor Elevator Kitchen Lab LargeOffice1

LargeOffice2 PrinterArea SmallOffice2 StudentOffice

Fig. 3. The nine known locations in the training set.

set. We also experimented with a second threshold to decline from classification when the two highest scores were too similar. However, this did not change the results significantly, and is not included in the results presented here.

Because the testing set included four unknown locations (out of 13), we simulated this situation in training by considering three random locations (roughly the same ratio of unseen to seen) as unknown and use this setup when determining the thresholds.

4.4. Results

We submitted a total of eight runs to the RobotVision contest with the group name "PicSOM TKK". Our best submitted result for the easy set received a score of 2176.0, which is 85% of the best possible score. This result was based on detectors trained on the left camera data only, and it obtained the overall highest score in the competition in the obligatory task (i.e. the instantaneous case). The same setup achieved our best result (1117.0) for the hard set as well. Fig. 4 visualises this run compared to the ground truth. For the hard set, our result was slightly above the median of the submitted results using the hard training data. The overall best submitted run to the hard set was 1777.0. The best submitted results for all participating groups are shown in Fig. 5 for the two training sets, and the overall score. The overall score in the competition is the sum of the score from the easy and the hard sets.

Some of our submitted runs and also some additional runs are summarised in Table 1, with "•" denoting that the run was submitted to the competition. The additional runs could be performed as the participants were given access to the class labels for the testing dataset after the competition. The first column in the table specifies how the training data was selected with regard to the cameras. The word "separate" indicates that separate models were trained for each camera and then averaged, while "both" uses all images to train a single model. The second column states whether fusion of single-feature classifiers (Section 2.3) or just the single best performing feature (ColorSIFT with dense sampling, soft clustering with a spatial pyramid) is used.

Somewhat surprisingly, using information from both cameras does not improve the results, in fact using a single camera works better than using a single model trained on all images. This difference is especially notable on the hard training set. Also, using the separate left and right models together gives no improvement over using just one of them. This result cannot be explained with poor parameter selection; after the competition we also tried to optimise the thresholds in the test set directly (not shown here) but with similar results. Furthermore, we observed that such optimised results are only slightly better than the submitted ones, indicating that the system performance is not very sensitive to the threshold parameters.

Finally, in Table 1 we can also see that the feature fusion is highly beneficial: with a single well-performing feature the results are significantly weaker.

Table 1. RobotVision recognition scores.

cameras	features	easy	hard	total
• left only	fusion	2176.0	1117.0	3293.0
right only	fusion	2210.5	1072.0	3282.5
separate	fusion	2207.5	1057.0	3264.5
• both	fusion	2065.0	665.5	2730.5
• both	single	964.0	554.5	1518.5

5. CONCLUSIONS

Our results indicate that a general-purpose algorithm for visual category recognition can perform well on indoor lo-

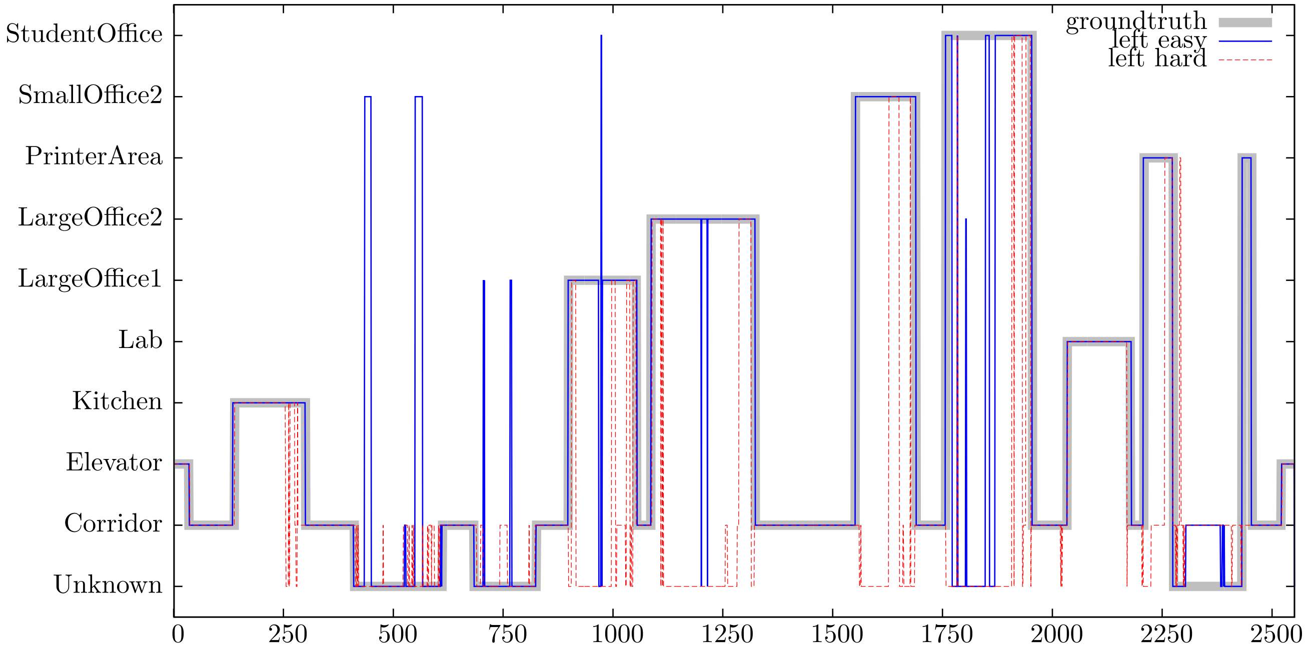


Fig. 4. Recognition results over time (frame indices) based on left camera images, trained on easy (blue line) and hard (red dashed) sets. The ground truth is shown as a thick grey line.

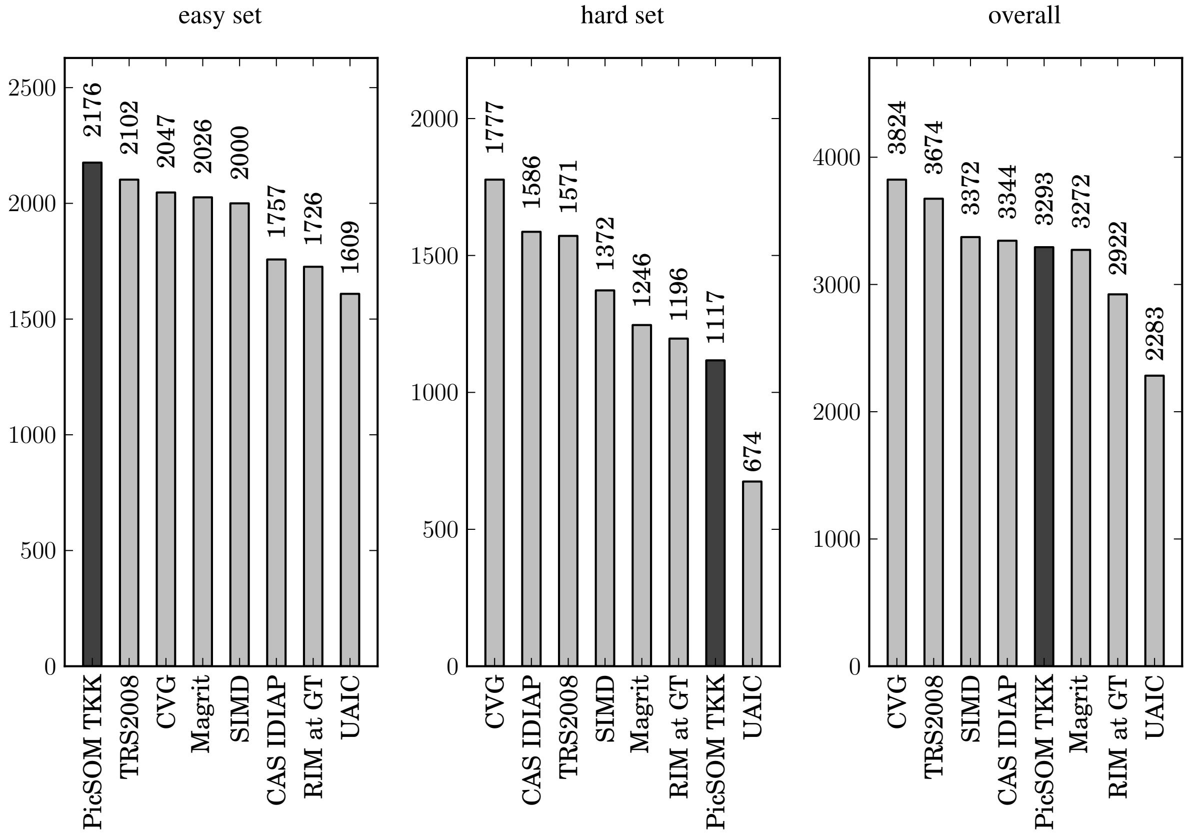


Fig. 5. Scores of participating groups for the easy and hard sets, and the overall scores.

cation recognition, given that enough training data is available. The generality of our approach is illustrated e.g. by its successful application to image and video retrieval [7]. However, with limited training data the performance of a purely appearance-based method is less competitive. There are several possible explanations for this. It might be that the generic scene appearance features utilise the limited training data uneconomically and other domain-specific modalities would be needed to make best use of the scarce training examples. For location recognition, these could include the depth information, the temporal continuity of the frame sequence, and using image-matching-based techniques.

Yet, it is also possible that better performance could be achieved on basis of the generic appearance features by better system design. In particular, there might be some overlearning issues. With the larger training set, just memorising all the camera views appearing in the training material might be a viable strategy, whereas the smaller training set calls for generalising between views. A naive use (such as here) of a rich and distinctive scene representation might actually lead to worse performance than a feature extraction scheme with more limited distinguishing power if the interview generalisation issue is not properly taken care of. Our experiments reported here are insufficient to confirm either one of these hypotheses.

Furthermore, the results confirm our earlier findings that fusion of a large set of visual features can consistently result in a much better category recognition accuracy than the use of any single feature.

6. ACKNOWLEDGEMENTS

This work has been funded by Aalto MIDE programme (project UI-ART) and in part by Finnish Funding Agency for Technology and Innovation (TEKES) under the project DIEM/MMR and by the PASCAL2 Network of Excellence, ICT 216886.

7. REFERENCES

[1] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid, “Local features and kernels for classification of texture and object categories: a comprehensive study,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[2] J. Borenstein, H. R. Everett, and L. Fen, *Navigating Mobile Robots: Sensors and Techniques*, A. K. Peters, Ltd., 1996.

[3] Dieter Fox, Wolfram Burgard, and Sebastian Thrun, “Markov localization for mobile robots in dynamic environments,” *Journal of Artificial Intelligence Research*, vol. 11, pp. 391–427, 1999.

[4] Stephen Se, David Lowe, and Jim Little, “Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks,” *The International Journal of Robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.

[5] A. Pronobis, O. Martínez Mozos, and B. Caputo, “SVM-based discriminative accumulation scheme for place recognition,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA08)*, Pasadena, CA, USA, May 2008.

[6] Yue Feng, Martin Halvey, and Joemon M. Jose, “University of Glasgow at ImageCLEF 2009 Robot Vision task,” in *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September-October 2009.

[7] Mats Sjöberg, Ville Viitaniemi, Markus Koskela, and Jorma Laaksonen, “PicSOM experiments in TRECVID 2009,” in *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[8] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster, “A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment,” *Personal and Ubiquitous Computing*, vol. 1, no. 4, pp. 208–217, December 1997.

[9] Antti Ajanki, M. Billinghurst, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Teemu Ruokolainen, and Timo Tossavainen, “Contextual information access with augmented reality,” in *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, Kittilä, Finland, August-September 2010.

[10] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. in press, 2010.

[12] A. Pronobis and B. Caputo, “COLD: COsy Localization Database,” *The International Journal of Robotics Research (IJRR)*, vol. 28, no. 5, May 2009.