# Real time automatic scene classification

Menno Israël[a] Egon L. van den Broek[b]
Peter van der Putten[c] Marten J. den Uyl[a]

[a] ParaBot Services bv / Vicar Vision bv,
Singel 160, 1054 TA Amsterdam
[b] NICI, Radboud University Nijmegen,
P.O. Box 9104, 6500 HE Nijmegen
[c] LIACS, P.O. Box 9512, 2300 RA Leiden

This work has been done as part of the EU VICAR (IST) project and the EU SCOFI project (IAP). The aim of the first project was to develop a real time video indexing classification annotation and retrieval system.

For our systems, we have adapted the approach of Picard and Minka [3], who categorized elements of a scene automatically with so-called 'stuff' categories (e.g., grass, sky, sand, stone). Campbell et al. [1] use similar concepts to describe certain parts of an image, which they named "labeled image regions". However, they did not use these elements to classify the topic of the scene. Subsequently, we developed a generic approach for the recognition of visual scenes, where an alphabet of basic visual elements (or "typed patches") is used to classify the topic of a scene.

We define a new image element: a patch, which is a group of adjacent pixels within an image, described by a specific local pixel distribution, brightness, and color. In contrast with pixels, a patch as a whole can incorporate semantics.

A patch is described by a HSI color histogram with 16 bins and by three texture features (i.e., the variance and two values based on the two eigen values of the covariance matrix of the Intensity values of a mask ran over the image. For more details on the features used we refer to Israel et al. [2].

We aimed at describing each image as a vector with a fixed size and with information about the position of patches that is not strict (strict position would limit generalization).

Therefore, a fixed grid is placed over the image and each grid cell is segmented into patches, which are then categorized by a patch classifier. For each grid cell a frequency vector of its classified patches is calculated. These vectors are concatenated. The resulting vector describes the complete image.

Several grids were applied and several patch sizes with the grid cells were tested. Grid size of 3x2 combined with patches of size 16x16 provided the best system performance.

For the two classification phases of our system, back-propagation networks were trained: (i) classification of the patches and (ii) classification of the image vector, as a whole.

The system was tested on the classification of eight categories of scenes from

the Corel database: interiors, city/street, forest, agriculture/countryside, desert, sea, portrait, and crowds. Each of these categories were relevant for the VICAR project. Based upon their relevance for these eight categories of scenes, we choose nine categories for the classification of the patches: building, crowd, grass, road, sand, skin, sky, tree, and water. This approach was found to be successful (for classification of the patches 87.5% correct, and classification of the scenes 73.8% correct).

An advantage of our method is its low computational complexity. Moreover, the classified patches themselves are intermediate image representations and can be used for image classification, image segmentation as well as for image matching. A disadvantage is that the patches with which the classifiers were trained had to be manually classified. To solve this drawback, we currently develop algorithms for automatic extraction of relevant patch types.

Within the IST project VICAR, a video indexing system was built for the Netherlands Institute for Sound and Vision[1], consisting of four independent modules: car recognition, face recognition, movement recognition (of people) and scene recognition. The latter module was based upon the afore mentioned approach.

Within the IAP project SCOFI, a real time Internet pornography filter was built, based upon this approach. The system is currently running on several schools in Europe. Within the SCOFI filtering system, our image classification system (with a performance of 92% correct) works together with a text classification system that includes a proxy server (FilterX, developed by Demokritos, Greece) to classify web-pages. Its total performance is 0% overblocking and 1% underblocking.

**Acknowledgments:** This work was partially supported by the EU projects VICAR (IST-24916) and SCOFI (IAP-2110; http://www.scofi.net/). Further, we thank the reviewers, for their comments on the manuscript, and Robert Maas for his work on the texture algorithm.

# References

[1] N. W. Campbell, W. P. J. Mackeown, B. T. Thomas, and T. Troscianko. Interpreting image databases by region classification. *Pattern Recognition*, 30(4):555–563, 1997.

[2] M. Israël, E. L. van den Broek, P. van der Putten, and M. J. den Uyl. Automating the construction of scene classifiers for content-based video retrieval. In L. Khan and V. A. Petrushin, editors, *Proceedings of the Fifth International Workshop on Multimedia Data Mining (MDM/KDD'04)*, pages 38–47, Seattle, WA, USA, 2004.

[3] R. W. Picard and T. P. Minka. Vision texture for annotation. *Multimedia Systems*, 3(1):3–14, 1995.

[1] http://www.beeldengeluid.nl