Technical Report 58

Classification of Semantic Concepts to Support the Analysis of the Inter-Cultural Visual Repertoires of TV News Reviews

Martin Stommel
Martina Dümcke
Otthein Herzog
TZI, Universität Bremen

TZI-Bericht Nr. 58
2011

Universität Bremen

TZI-Berichte

Herausgeber:
Technologie-Zentrum Informatik und Informationstechnik
Universität Bremen
Am Fallturm 1
28359 Bremen
Telefon: +49-421-218-7272
Fax: +49-421-218-7820
E-Mail: info@tzi.de
http://www.tzi.de

ISSN 1613-3773

Classification of Semantic Concepts to Support the Analysis of the Inter-Cultural Visual Repertoires of TV News Reviews

M. Stommel, M. Duemcke and O. Herzog

TZI Center for Computing and Communication Technologies,
University Bremen, Am Fallturm 1, 28359 Bremen, Germany
mstommel@tzi.de, mduemcke@googlemail.com, herzog@tzi.de

Abstract. TV news reviews are of strong interest in media and communication sciences, since they indicate national and international social trends. To identify such trends, scientists from these disciplines usually work with manually annotated video data. In this paper, we investigate if the time-consuming process of manual annotation can be automated by using the current pattern recognition techniques. To this end, a comparative study on different combinations of local and global features sets with two examples of the pyramid match kernel is conducted. The performance of the classification of TV new scenes is measured. The classes are taken from a coding scheme that is the result of an international discourse in media and communication sciences. For the classification of studio vs. non-studio, football vs. ice hockey, computer graphics vs. natural scenes and crowd vs. no crowd, recognition rates between 80 and 90 percent could be achieved.[1]

1 Analysis of Visual Repertoires in Media and Communication Sciences

The development of our society as documented in TV news reports is subject to research in media and communication sciences. While the contents of a news report itself is of high importance, media and communication scientists are aware of more subtle but also crucial sources of information: The structure of the scene setup may for example suggest a certain social role of the actors. The meaning of a scene also does not only depend on the video data but also on the cultural background of the viewer. And often it is more conclusive to identify issues that have been omitted compared to those actually addressed.

TV news are suited well to study such questions. The constant process of production, repetition and summarisation of TV news and news reviews results in video representations of the most relevant events of our society in very concise form [2]. The symbolic value as well as the high spread of these representations make them interesting for comparison across countries or years.

[1] A short version of this article has been published at the KI 2011 conference [1].

The analysis usually includes a lot of manual video annotation. Research efforts in different countries resulted in a coding sheet that states the most important items for annotation [3]. Additional items are included to handle specific research questions. To reduce the influence of personal background and understanding, the annotation is conducted by specialists that have been trained for a high inter-coder reliability, i.e. a high agreement in the annotations. The inter-coder reliability, measured as Krippendorff's alpha, reaches an agreement of more than 70 percent, under good conditions. The annotation is used to compare the depictions of people and events over different countries or years.

In this paper, we study if the process can be facilitated by using current pattern recognition techniques. To this end, we chose four items with low symbolic connotation from the annotation scheme. The items are studio/non-studio, football/ice hockey, computer graphics/natural scenes and crowd/no crowd. The pyramid match kernel is trained to classify these items based on a set of local and global detectors and descriptors. Using the optimal feature configurations, we achieve excellent recognition rates for all classes.

2 Computational Approaches

Computational approaches consist of preprocessing, feature extraction and classification steps [4]. For some industrial computer vision applications this may be a straight process chain. The classification of TV material with its contextual cross references and rich semantics requires a more complex procedure in multiple stages. The idea of a multi-stage or hierarchical procedure can already be found in earlier connectionist approaches [5]. The approaches are justified biologically [6], psychologically [7] or statistically [8]. The structure and understanding of the hierarchy is application dependent. For the case of TV material, Dorai and Venkatesh [9] distinguish between a high and a low level in their theoretical framework. The high level deals with the narrative form and the arrangement of scenes and effects by the filmmaker. Low level features on the other hand are characterised as rather formal properties that can be extracted from single frames or shots.

Practical efforts to reach the high level are connected to the notion of a semantic concept [10]. On an intermediate level, semantic concepts are named objects or scene types. The name distinguishes them from strictly syntactical low-level features. Finer, sometimes recursive subdivisions of objects into their parts have been proposed (e.g. [11, 6, 12, 7]). Hauptmann et al. [13] extrapolate from measurements on 300 TRECVid concepts and conclude that a few thousand concepts with moderate recognition accuracy might be sufficient to reliably retrieve news videos.

While low-level syntactical features do not allow for a reliable scene classification [13], they achieve a certain invariance against illumination and deformation. The influence of illumination and pose on the object appearance has been visualised by Murase and Nayar [14], allowing them to model the appearance directly by using principle component analysis. Garg et al. [15] provide theoretical and practical results that the dimensionality of scene appearances under natural conditions can be reduced to a number of 10 to 30 dimensions without visual loss using principle component analysis.

In most cases the scene appearance is not modelled directly. Instead, semantic concepts are usually represented by sets of local feature vectors [16–19] trained by machine learning algorithms [20]. A popular approach is to subdivide the feature space into bins that can be used to compute histograms over the feature space or to span simplified new feature spaces [21–23, 17]. The subdivision can be general purpose or optimised to a particular semantic concept [24]. To a certain degree, the trained feature sets resemble the alphabet of moderately complex features found by Tanaka [25] in the inferior temporal cortex.

Because geometrical dependencies often cause high computational costs, these approaches often follow the bag-of-features principle. However, experiments on different types of constellation models indicate advantages for the use of geometry [26] depending on the level of abstraction [12]. Some studies therefore aim at incorporating geometrical information [27–30]. Yang et al. [31] propose a scene classification based on motion features.

Recent results indicate that the time consuming clustering of local features can be simplified by creating a random alphabet of visual words given a sufficient size of the alphabet [32, 33] and a proper pooling function [11].

3 Experimental Setup

In our analysis we evaluate two versions of Grauman and Darrell's Pyramid Match Kernel [34, 35] in combination with four interest point detectors, four feature descriptors, and three global features.

The Pyramid Match Kernel compares histograms of the input data based on the simultaneous histogram intersection at multiple bin widths. The kernel function is then be used with a Support Vector Machine. While the original version uses bins that are aligned to regular grids, a later version [35] performs a hierarchical clustering to align the bins to the distribution of the data.

The Pyramid Match Kernel is used to classify local and global image features both separately as well as in combination. Feature combinations are represented by concatenating their descriptors. Local features are computed at interest points detected by Speeded Up Robust Features (SURF) [36], Maximally Stable Extremal Regions (MSER) [37], and Harris corner points obtained in the Harris-Affine or Hessian-Affine version [38].

These local detectors are combined with four feature descriptors. The descriptors are the one proposed in the Speeded Up Robust Features, then the location of a feature point (i.e. the image coordinate), Steerable Filters [39] and Shape Context [40].

As global features we use colour histograms in two versions: Global colour histograms are build by concatenating the intensity histograms of the three colour channels. Local colour histograms are the concatenation of all colour histograms computed in the cells of a regular grid with a spacing of 16 pixels placed over

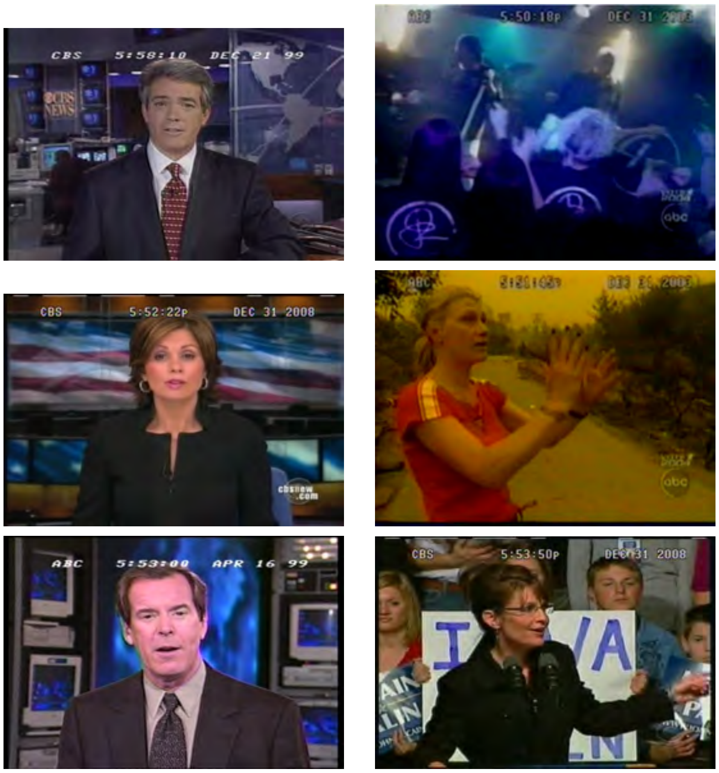

Fig. 1. Three samples from the studio (on the left) and non-studio class (on the right).

the image. The presence or absence of faces is used as a third global feature [41]. The aim of this setup is to benefit from complementary information, e.g. colour and texture.

The classification is conducted on single frames that are representatively chosen. Every frame stands for a shot in a TV news review and is annotated by the binary categories studio vs. non-studio, football vs. ice hockey, computer graphics vs. natural scenes and crowd vs. no crowd. The sample sizes are each 200 frames for studio and no studio. The images are taken from 400 shots of ABC and CBS TV news reviews from 1999, 2001, and 2003–2009. For the categories football and ice hockey, each 50 frames are chosen from ARD and ZDF news reviews from 2008–2010. The categories computer graphics and natural are represented by each 50 frames from ABC and CBS news reviews from 1999–2000, 2003–2006 and 2008. The recognition of crowds is tested with each 40 positive and negative samples of ABC and CBS news reviews from 1999, 2001, 2005 and 2008. The images are randomly split into equally sized training and test samples. Special care is taken that no frames of the same video are present in the training and test set at the same time. This is to exclude spurious matches between

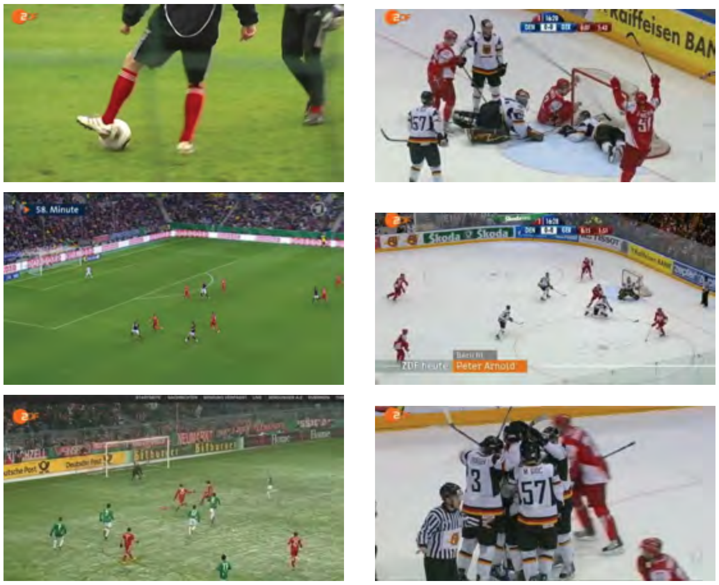

Fig. 2. Three samples from the football (on the left) and ice hockey class (on the right).

related shots of a longer scene. The figures 1, 2, 3 and 4 show three samples for each class.

4 Experimental Results

Figure 5 shows the accuray of the classification of studio scenes using the original Pyramid Match Kernel. Comparatively high results of up to 77 per cent are obtained for the SURF descriptor in combination with MSER or one of the corner detectors. Texture and edges therefore seem more important for the studio class than colour. Faces also appear as a good feature and it seems that the classifier recognises studio frames by the anchor person. However, most combinations yield only recognition rates slightly better than random.

The hierarchical clustering introduced later [35] leads to a significant improvement for almost all feature types. Figure 6 shows the accuracy. Experiments on the number and depth of the branches of the cluster hierarchy show that a proper alignment of the Match Kernel to the data distribution is indeed crucial. Our results therefore validate the observations by Grauman and Darrell [35]. With the better alignment, the best results are now obtained for feature configurations including the shape context. In the following, all results are obtained using the hierarchical clustering in the preprocessing.

As fig. 7 shows, the combination of multiple detectors increases the accuracy to more than 81 per cent. However, the increase in accuracy is balanced by the computational cost to handle a higher number of interest points. The figure also

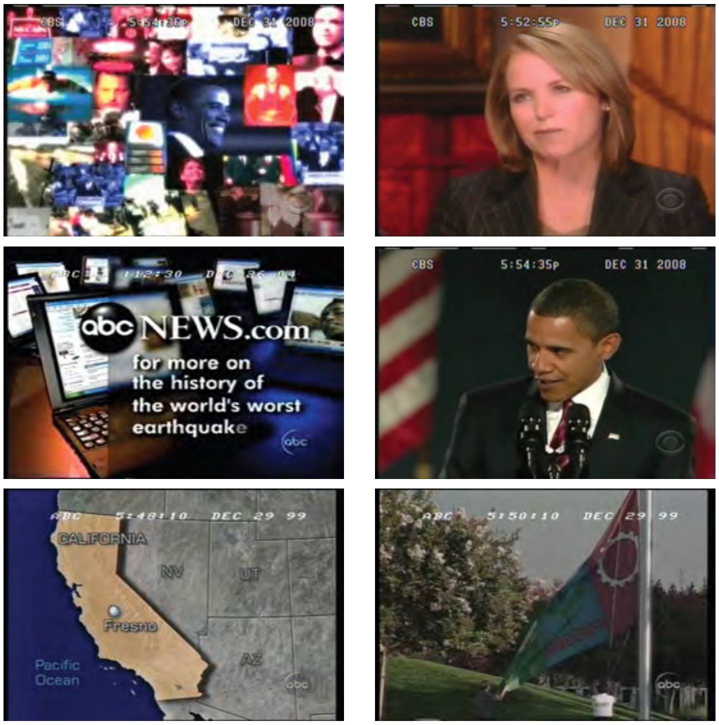

Fig. 3. Three samples from the computer graphics (on the left) and natural scene class (on the right).

shows that the combination of multiple descriptors instead of multiple detectors decreases the accuracy. The result shows that the trade-off between the fusion of complementary information and numerical stability is still a non-trivial problem. This is also in accordance to obervations by Haupmann et al. [13] on the combination of semantic concepts.

The classification of the sport type can be handled very well by the experimental setup. The best feature combination reaches an accuracy of 98 per cent (see fig. 8). The highly dynamic scenes are handled best by the SURF detector and descriptor, while the feature location proves inappropriate here. The predominance of either white or green background (see fig. 2) is reflected in the good results for the colour histograms. The frequent occurrence of the audience at the top margin of the images might explain the advantage of the local colour histograms.

The good contrast of the computer generated TV news shots seems to match the MSER detector combinations best with an accuracy of 72 per cent on the average (see fig. 9). The high performance of the location descriptor with the best accuracy of up to 77 per cent in combination with the Harris-Affine interest operator can be explained by the static nature of the video type. Computer

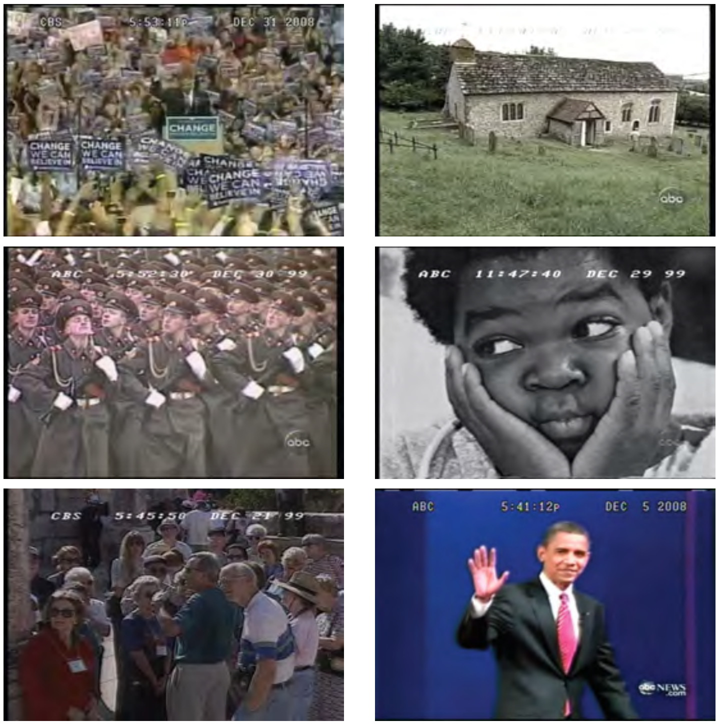

Fig. 4. Three samples from the crowd (on the left) and non-crowd class (on the right).

animations are also frequently repeated without significant change since they form a distinguishing feature of a TV news show.

The accuracy for the recognition of crowds is shown in fig. 10. The results are good for most local features including local colour histograms. A maximum of more than 89 per cent is reached for the SURF descriptor combined with either the SURF or Hessian-Affine interest point detector. The clear advantage over the results for the global colour histogram indicates that geometry is a crucial feature for this class. The face detector performs bad in the recognition of a crowd. Although many faces are present, faces are often occluded or too small to be detected. Also, the skin colour analysis might be disturbed by badly illuminated faces and faces that blur with the background.

5 Conclusion

In this paper, we study the classification of four semantic concepts used in current media research. Two versions of the Pyramid Match Kernel are combined with four feature detectors and four descriptors. Our experiments show, that the best classifier setups achieve a high accuracy of 77% to 98% depending on the class.

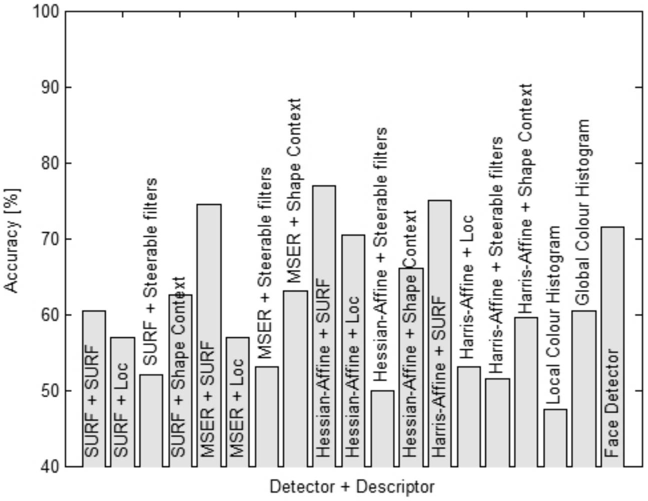

Fig. 5. Results for the scene type 'studio' using the pyramid match kernel and different detector-descriptor combinations.

The good classification results prove a growing importance of computer vision methods for media interpretation.

References

1. Stommel, M., Duemcke, M., Herzog, O.: Classification of Semantic Concepts to Support the Analysis of the Inter-Cultural Visual Repertoires of TV News Reviews. In: 34th German Conference on Artificial Intelligence (KI), Berlin, Germany, October 4–7, 2011. Lecture Notes in Artificial Intelligence, Springer (2011)
2. Ludes, P.: Visual Hegemonies: An Outline = Volume 1 of The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences. LIT, Muenster (2005) (Translations into Portuguese in 2007 and Chinese in 2008.).
3. Hanitzsch, T.: Codebook for Content Analysis Foreign TV News Project. Worlds of Journalisms Project (February 2010)
4. Rosenfeld, A.: Picture Processing by Computer. ACM Computing Surveys (CSUR) **1**(3) (1969) 147–176
5. Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review **65**(6) (1958) 386–408
6. Serre, T., Wolf, L., Poggio, T.: A new biologically motivated framework for robust object recognition. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2005)
7. Ommer, B., Sauter, M., Buhmann, J.M.: Learning Top-Down Grouping of Compositional Hierarchies for Recognition. Proc. of the Conference on Computer Vision and Pattern Recognition (2006) 194–202
8. Stommel, M., Kuhnert, K.D.: Part Aggregation in a Compositional Model based on the Evaluation of Feature Cooccurrence Statistics. Int'l Conf. on Image and Vision Computing New Zealand (IVCNZ), Christchurch, New Zealand, Nov. 26-29 (2008) 26–29

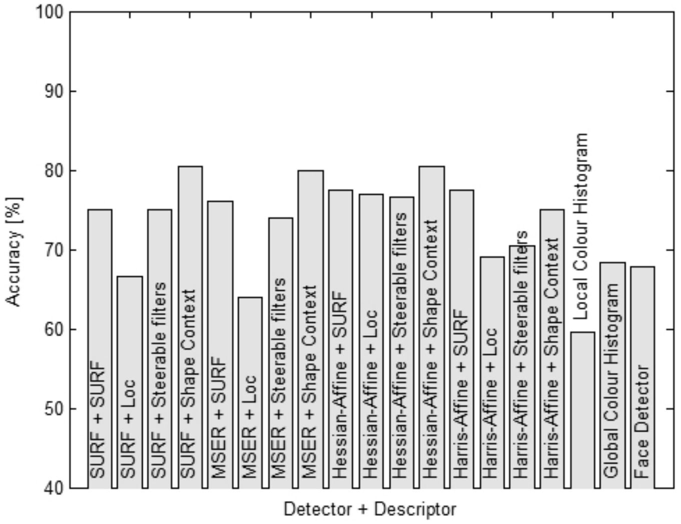

Fig. 6. Classification of studio scenes using the hierarchical clustering.

9. Dorai, C., Venkatesh, S.: Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics. Computational Semiotics (COSIGN) (2001) 94–99
10. Smeaton, A.F., Over, P., Kraaij, W.: High level feature detection from video in TRECVid: a 5-year retrospective of achievements. In Divakaran, A., ed.: Multimedia Content Analysis, Theory and Applications. Springer (2008)
11. Jarrett, K., Kavukcuoglu, K., Ranzato, M.A., LeCun, Y.: What is the Best Multi-Stage Architecture for Object Recognition? IEEE International Conference on Computer Vision (ICCV) (2009)
12. Stommel, M., Kuhnert, K.D.: Visual Alphabets on Different Levels of Abstraction for the Recognition of Deformable Objects. Joint IAPR International Workshop on Structural, Syntactic and Statistical Pattern Recognition (S+SSPR), Cesme, Izmir, Turkey, August 18-20 **LNCS 6218** (2010) 213–222
13. Hauptmann, A., Lin, W.H., Yan, R.: How Many High-level Concepts Will Fill the Semantic Gap in News Video Retrieval? In Proceedings of ACM International Conference on Image and Video Retrieval (2007) 627–634
14. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. Int. Journal of Computer Vision **14**(1) (January 1995) 5–24 Kluwer.
15. Garg, R., Du, H., Seitz, S.M., Snavely, N.: The Dimensionality of Scene Appearance. IEEE International Conference on Computer Vision (ICCV) (2009)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
17. Mikolajczyk, K., Leibe, B., , Schiele, B.: Local Features for Object Class Recognition. In: International Conference on Computer Vision (ICCV). (2005)
18. Stark, M., Schiele, B.: How good are local features for classes of geometric objects. IEEE 11th International Conference on Computer Vision ICCV (2007) 1–8
19. Ke, Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. Computer Vision and Pattern Recognition (CVPR) **2** (2004) 506–513

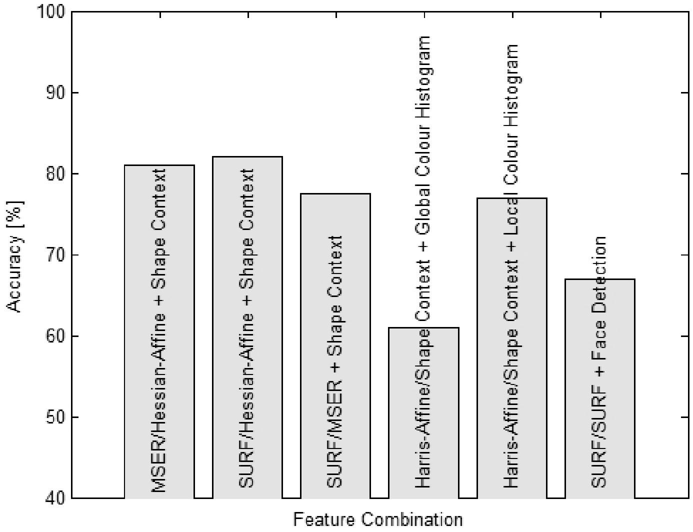

Fig. 7. Classification of the studio scenes using feature combinations.

20. Jain, A.K., Duin, R., Mao, J.: Statistical Pattern Recognition: A Review. In IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1) (2000) 4–37
21. Liu, J., Yang, Y., Shah, M.: Learning Semantic Visual Vocabularies Using Diffusion Distance. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2009)
22. Lin, Z., Hua, G., Davis, L.: Multiple Instance Feature for Robust Part-based Object Detection. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2009)
23. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple Object Class Detection with a Generative Model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06). (June 2006)
24. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008)
25. Tanaka, K.: Inferotemporal cortex and object vision. Annual Reviews of Neuroscience **19** (1996) 109–139
26. Crandall, D.J., Felzenszwalb, P.F., Huttenlocher, D.P.: Spatial Priors for Part-Based Recognition Using Statistical Models. Computer Vision and Pattern Recognition (CVPR) (2005) 10–17
27. Perdoch, M., Chum, O., Matas, J.: Efficient Representation of Local Geometry for Large Scale Object Retrieval. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2009)
28. Chum, O., Perdoch, M., Matas, J.: Geometric min-Hashing: Finding a (Thick) Needle in a Haystack. IEEE Conf. on Computer Vision and Pattern Recognition (2009)
29. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) **2** (2006) 2169–2178

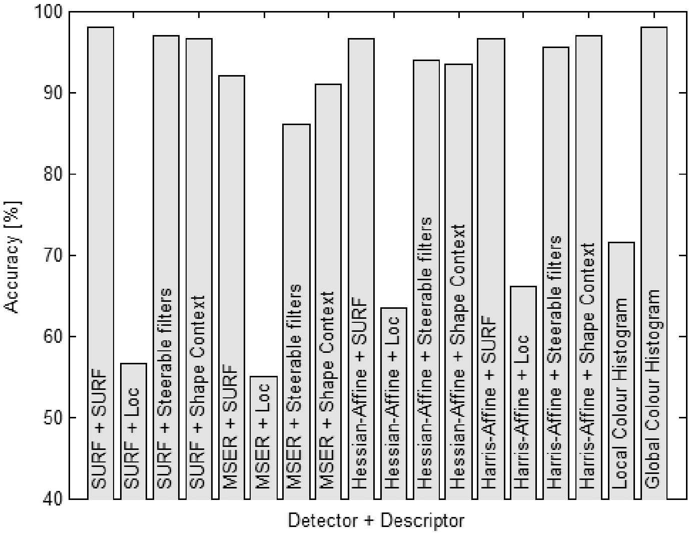

Fig. 8. Accuracy for the scene type 'sport'.

30. Zhang, E., Mayo, M.: Improving bag-of-words model with spatial information. Int'l Conf. on Image and Vision Computing New Zealand (IVCNZ) (2010)
31. Yang, Y., Liu, J., Shah, M.: Video Scene Understanding Using Multi-scale Analysis. IEEE International Conference on Computer Vision (ICCV) (2009)
32. Stommel, M., O.Herzog: Sift-based object recognition with fast alphabet creation and reduced curse of dimensionality. Int'l Conf. on Image and Vision Computing New Zealand (IVCNZ) (2009)
33. Ilies, I., Jacobs, A.: Automatic Image Annotation through Concept Propagation. In Ludes, P., Herzog, O., eds.: Algorithms of Power - Key Invisibles, The World Language of Key Visuals: Computer Sciences, Humanities, Social Sciences. Volume 3. (2011) 67–82
34. Grauman, K., Darrell, T.: The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. IEEE International Conference on Computer Vision (ICCV) **2** (2005) 1458–1465
35. Grauman, K., Darrell, T.: Approximate correspondences in high dimensions. In Advances in Neural Information Processing Systems (NIPS) (2006)
36. Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU) **110**(3) (2006) 346–359
37. Forssen, P.E.: Maximally stable colour regions for recognition and matching. Computer Vision and Pattern Recognition (CVPR) (2007)
38. Mikolajczyk, K., Schmid, C.: An Affine Invariant Interest Point Detector. European Conference on Computer Vision (ECCV) (2002) 128–142
39. Freeman, W.H., Adelson, E.H.: The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991) 891–906
40. Belongie, S., Mori, G., Malik, J.: Matching with shape contexts. IEEE Workshop on Content-based access of Image and Video-Libraries (CBAIVL) **13** (2000) 20–26
41. Vezhnevets, V., Sazonov, V., Andreeva, A.: A Survey on Pixel-Based Skin Color Detection Techniques. Proc. Graphicon-2003 **13** (2003) 85–92

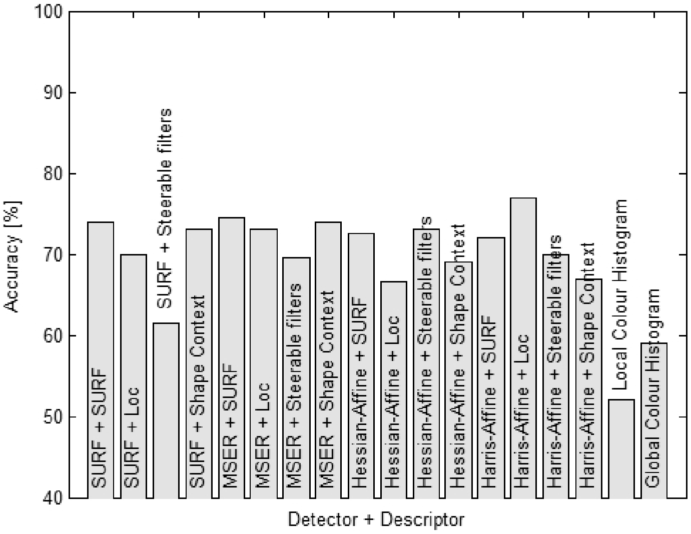

Fig. 9. Detection of computer graphics.

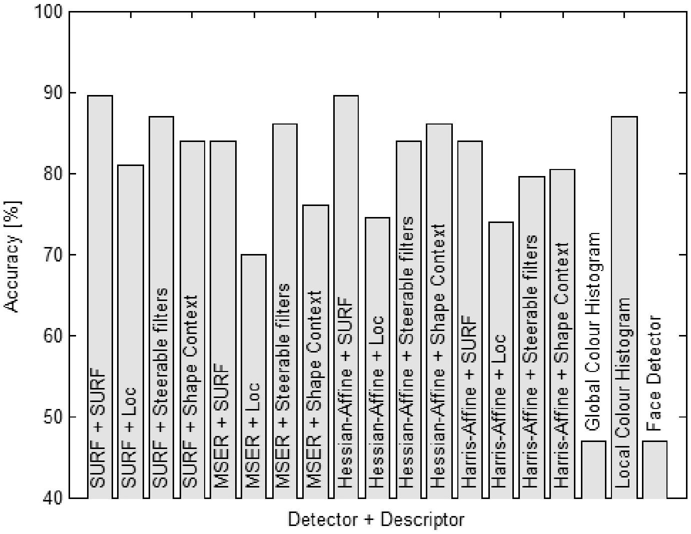

Fig. 10. Accuracy for the class 'crowds'.