# The Visual Concept Detection Task in ImageCLEF 2008

Thomas Deselaers[1] and Allan Hanbury[2,3]

[1] RWTH Aachen University, Computer Science Department, Aachen, Germany
deselaers@cs.rwth-aachen.de
[2] PRIP, Inst. of Computer-Aided Automation, Vienna Univ. of Technology, Austria
hanbury@prip.tuwien.ac.at
[3] CogVis GmbH, Vienna, Austria

**Abstract.** The Visual Concept Detection Task (VCDT) of ImageCLEF 2008 is described. A database of 2,827 images were manually annotated with 17 concepts. Of these, 1,827 were used for training and 1,000 for testing the automated assignment of categories. In total 11 groups participated and submitted 53 runs. The runs were evaluated using ROC curves, from which the Area Under the Curve (AUC) and Equal Error Rate (EER) were calculated. For each concept, the best runs obtained an AUC of 80% or above.

## 1 Introduction

Searching for images is, despite intensive research on alternative methods in the last 20 years, still a task that is mainly done based on textual information. For a long time, searching for images based on text was the most feasible method because on the one hand, the number of images to be searched was rather restricted, and on the other hand, only few people needed to access huge repositories of images. Both of these conditions have changed. The number of available images is growing more rapidly than ever due to the falling prices of high-end imaging equipment for professional use and of digital cameras for consumer use. Publicly available image databases such as Google picassa and Flickr have become major sites of interest on the Internet.

Nevertheless, accessing images is still a tedious task because sites such as Flickr do not allow images to be accessed based on their content but only based on the annotations that users create. These annotations are commonly disorganised, not very precise, and multilingual. Access problems can be addressed by improving the textual access methods, but none of these improvements can ever be perfect as long as the users do not annotate their images perfectly, which is very unlikely. Therefore, content-based methods have to be employed to improve access methods to digitally stored images.

A problem with content-based methods is that they are often computationally costly and cannot be applied in real-time. An intermediate step is to automatically create textual labels based on the images' content. To make these labels

as useful as possible, frequently occurring visual concepts should be annotated in a standard manner.

In the visual concept detection task (VCDT) of ImageCLEF 2008, the aim was to apply labels of frequent categories in the photo retrieval task to the images and evaluate how well automated visual concept annotation algorithms function. Additionally, participants of the VCDT could create annotations for all images used in the photo retrieval task, which were provided to the participants of this task. In the following, we describe the visual concept detection task of ImageCLEF 2008, the database used, the methods of the participating groups, and the results.

Other tracks in ImageCLEF 2008 were the photo retrieval task [1], the medical retrieval task [2], the Wikipedia multimedia retrieval task [3], and the medical image annotation task [4].

## 2 Database and Task Description

As database for the ImageCLEF 2008 visual concept detection task, a total of 2,827 images were used. These are taken from the same pool of images used to create the IAPR-TC12 database [5], but are not included in the IAPR-TC12 database used in the ImageCLEF photo retrieval task.

The visual concepts were chosen based on concepts used in previous work on visual concept annotation. In particular they are an extension of the hierarchy used in the attribute recognition task of the ImagEVAL 2006 campaign [6]. They are organised hierarchically, as shown in Figure 1. An image can be labelled by a group of concepts. The hierarchy demonstrates the interdependency between some concepts, e.g. if the *sunny*, *partly cloudy* or *overcast* concept applies to an image, then the *sky* concept must apply too.

As for the ImageCLEF object detection task in 2007 [7], a web interface was created for manual annotation of the images by the concepts. Annotation was mainly carried out by undergraduate students at the RWTH Aachen University and by the track coordinators. A general opinion expressed by the annotators was that the concept annotation required more time than the object annotation of 2007. The number of images that were voluntarily annotated this year was also significantly less than the 20,000 images annotated by object labels in 2007.

Of the 2,827 manually annotated images, 1,827 were distributed with annotations to the participants as training data. The remaining 1,000 images were provided without labels as test data. The participants' task was to apply labels to these 1,000 images. Table 1 gives an overview of the frequency of the 17 visual concepts in the training data and in the test data and Figure 2 gives an example image for each of the categories.

## 3 Results from the Evaluation

In total 11 groups participated and submitted 53 runs. For each run, results for each concept were evaluated by plotting ROC curves. The results for each

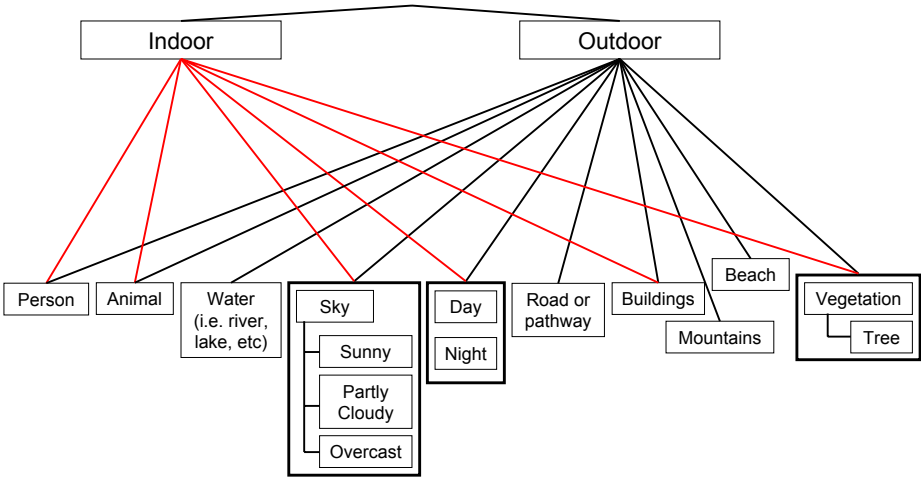


Fig. 1: Visual concept hierarchy used in the visual concept detection task of ImageCLEF 2008.

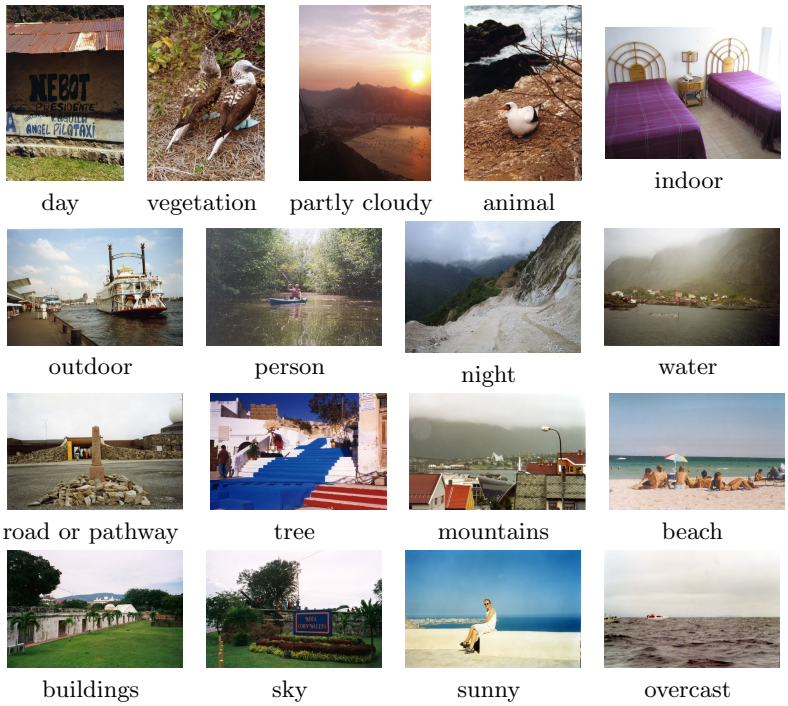


Fig. 2: Example images for each of the concepts.

Table 1: Statistics on the frequency of concepts in the training and test data.

| number | category | train [%] | test [%] |
|---|---|---|---|
| 00 | indoor | 9.9 | 10.2 |
| 01 | outdoor | 88.0 | 88.1 |
| 02 | person | 43.8 | 44.9 |
| 03 | day | 82.0 | 81.9 |
| 04 | night | 3.7 | 2.3 |
| 05 | water | 23.1 | 21.7 |
| 06 | road or pathway | 20.0 | 19.4 |
| 07 | vegetation | 52.5 | 51.7 |
| 08 | tree | 29.3 | 30.8 |
| 09 | mountains | 14.3 | 13.8 |
| 10 | beach | 4.4 | 3.7 |
| 11 | buildings | 45.5 | 43.6 |
| 12 | sky | 66.9 | 69.3 |
| 13 | sunny | 12.3 | 13.1 |
| 14 | partly cloudy | 22.7 | 22.2 |
| 15 | overcast | 19.6 | 21.4 |
| 16 | animal | 5.6 | 5.8 |

concept were summarised by two values: the area under the ROC curve (AUC) and the Equal Error Rate (EER). The latter is the error rate at which the false positive rate is equal to the false negative rate. Furthermore, for each run, the average AUC and average EER over all concepts were calculated.

Below, we briefly describe the methods employed by each group:

**CEA-LIST.** The Lab of Applied Research on Software-Intensive Technologies of the CEA, France submitted 3 runs using image features accounting for color and spatial layout with nearest neighbour and SVM classifiers.

**HJ FA.** The Microsoft Key Laboratory of Multimedia Computing and Communication of the University of Science and Technology, China submitted one run using color and SIFT descriptors which are combined and classified using a nearest neighbour classifier.

**IPAL I2R.** The IPAL French-Singaporean Joint Lab of the Institute for Infocomm Research in Singapore submitted 8 runs using a variety of different image descriptors.

**LSIS.** The Laboratory of Information Science and Systems, France submitted 7 runs using a structural feature combined with several other features using multi-layer perceptrons.

**MMIS.** The Multimedia and Information Systems Group of the Open University, UK submitted 4 runs using CIELAB and Tamura features and combinations of these.

**Makerere.** The Faculty of Computing and Information Technology, Makerere University, Uganda submitted one run using luminance, dominant colors, and texture and shape features classified using a nearest neighbour classifier.

Table 2: Summary of the results of the VCDT in ImageCLEF 2008

| | | best run | | | average | | |
|---|---|---|---|---|---|---|---|
| | runs | rank | EER | AUC | rank | EER | AUC |
| XRCE | 2 | 1 | 16.7 | 90.7 | 1.5 | 18.0 | 89.7 |
| RWTH | 1 | 3 | 20.5 | 86.2 | 3.0 | 20.5 | 86.2 |
| UPMC | 6 | 4 | 24.6 | 82.7 | 11.0 | 27.2 | 65.2 |
| LSIS | 7 | 5 | 25.9 | 80.5 | 20.3 | 32.8 | 71.8 |
| MMIS | 4 | 13 | 28.4 | 77.9 | 23.3 | 32.6 | 73.0 |
| CEA-LIST | 3 | 17 | 29.0 | 73.4 | 26.3 | 33.4 | 59.7 |
| IPAL-I2R | 8 | 19 | 29.7 | 76.4 | 32.1 | 36.0 | 68.3 |
| budapest | 13 | 20 | 31.1 | 74.9 | 31.8 | 35.2 | 68.6 |
| TIA | 7 | 24 | 32.1 | 55.6 | 39.6 | 39.9 | 36.3 |
| HJ-FA | 1 | 47 | 45.1 | 20.0 | 47.0 | 45.1 | 20.0 |
| Makere | 1 | 51 | 49.3 | 30.8 | 51.0 | 49.3 | 30.8 |

**RWTH.** The Human Language Technology and Pattern Recognition Group from RWTH Aachen University, Germany submitted one run using a patch-based bag-of-visual words approach using a log-linear classifier.

**TIA.** The Group for Machine Learning for Image Processing and Information Retrieval from the National Institute of Astrophysics, Optics and Electronics, Mexico submitted 7 runs using global and local features with SVMs and random forest classifiers.

**UPMC.** The University Pierre et Marie Curie in Paris, France submitted 5 runs using fuzzy decision forests.

**XRCE.** The Textual and Visual Pattern Analysis group from the Xerox Research Center Europe in France submitted two runs using multi-scale, regular grid, patch-based image features and a Fisher-Kernel Vector classifier.

**budapest.** The Datamining and Websearch Research Group, Hungarian Academy of Sciences, Hungary submitted 13 runs using a wide variety of features, classifiers, and combinations.

Table 2 gives an overview of the submissions and results for the task. The table is ranked by the performance of the best run submitted by the groups. It can be seen that the XRCE runs perform best.

Table 3 shows a breakdown of the results per concept. For each concept, the best and worst EER and AUC are shown, along with the average EER and AUC over all runs submitted. The best results were obtained for all concepts by XRCE, with budapest doing equally well on the *night* concept. The AUC per concept for all the best runs is 80.0% or above. Among the best results, the concepts having the highest scores are *indoor* and *night*. The concept with the worst score among the best results is *road or pathway*, most likely due to the high variability in the appearance of this concept. The concept with the highest average score, in other words, the concept that was detected best in most runs is *sky*. Again, the concept with the worst average score is *road or pathway*.

Table 3: Overview of the results per concept.

| | best | | | average | | worst | |
|---|---|---|---|---|---|---|---|
| # concept | EER | AUC | group | EER | AUC | EER | AUC |
| 00 indoor | 8.9 | 97.4 | XRCE | 28.0 | 67.6 | 46.8 | 2.0 |
| 01 outdoor | 9.2 | 96.6 | XRCE | 30.6 | 70.5 | 54.6 | 13.3 |
| 02 person | 17.8 | 89.7 | XRCE | 35.9 | 62.2 | 53.0 | 0.4 |
| 03 day | 21.0 | 85.7 | XRCE | 35.4 | 64.9 | 52.5 | 9.7 |
| 04 night | 8.7 | 97.4 | XRCE/budapest | 27.6 | 72.5 | 73.3 | 0.0 |
| 05 water | 23.8 | 84.6 | XRCE | 38.1 | 57.8 | 53.0 | 3.2 |
| 06 road/pathway | 28.8 | 80.0 | XRCE | 42.6 | 50.7 | 56.8 | 0.0 |
| 07 vegetation | 17.6 | 89.9 | XRCE | 33.9 | 67.4 | 49.7 | 30.7 |
| 08 tree | 18.9 | 88.3 | XRCE | 36.1 | 62.8 | 59.5 | 1.0 |
| 09 mountains | 15.3 | 93.8 | XRCE | 33.1 | 61.2 | 55.8 | 0.0 |
| 10 beach | 21.7 | 86.8 | XRCE | 35.8 | 57.6 | 51.4 | 0.0 |
| 11 buildings | 17.0 | 89.7 | XRCE | 37.4 | 60.8 | 64.0 | 0.5 |
| 12 sky | 10.4 | 95.7 | XRCE | 24.0 | 78.6 | 50.8 | 37.3 |
| 13 sunny | 9.2 | 96.4 | XRCE | 30.3 | 66.5 | 55.4 | 0.0 |
| 14 partly cloudy | 15.4 | 92.1 | XRCE | 37.5 | 58.9 | 55.5 | 0.0 |
| 15 overcast | 14.1 | 93.7 | XRCE | 32.1 | 67.6 | 61.5 | 0.0 |
| 16 animal | 20.7 | 85.7 | XRCE | 38.2 | 54.2 | 58.4 | 0.0 |

Two automatic runs provided by participants of the VCDT were made available to ImageCLEF participants. These provide annotations of the 20,000 ImageCLEF photo images with the VCDT concepts. Two groups participating in the photo retrieval task of ImageCLEF made use of these annotations, while one group used VCDT annotations provided by their own algorithm.

The group from Université Pierre et Marie Curie in Paris, France made use of their own VCDT algorithm to provide concepts for the photo retrieval task. They used the detected visual concepts to re-rank the first 50 results returned using text retrieval approaches. The concepts to use for the re-ranking were chosen using two approaches: (i) the concept word appears in the query text and (ii) the concept word appears in the list of synonyms (obtained using WordNet) of the words in the query text. The first approach improved the results of all the queries for which it was applicable, while the second resulted in worse results for some topics. Both approaches resulted in better overall performance than using text alone: the F-measure for the best text only run (using TF-IDF) is 0.273, while the F-measure for the run re-ranked using the first approach is 0.289.

The group from the National Institute of Informatics in Tokyo, Japan made use of both provided VCDT concept annotations. They also used the concepts to re-rank results returned by a text retrieval approach, where the best results were obtained by re-ranking based on a hierarchical clustering using distances between vectors encoding the VCDT concepts. This re-ranking decreased the P20 metric while increasing the CR20 metric, resulting in an increase of the F-measure from 0.224 for text only to 0.230 after the re-ranking.

Although the TIA-INAOE group in Puebla, Mexico also made use of one of the provided VCDT concept annotations, this was as part of a group of visual retrieval algorithms whose results were used in a late fusion process. It is therefore not possible to determine the effect of only the VCDT concepts on the results.

## 4 Conclusion

This paper summarises the ImageCLEF 2008 Visual Concept Detection Task. The aim was to automatically annotate images with concepts, with a list of 17 hierarchically organised concepts provided. The results demonstrate that this task can be solved reasonably well, with the best run having an average AUC over all concepts of 90.66%. Six further runs obtained AUCs between 80% and 90%. When evaluating the runs on a per concept basis, the best run also obtained an AUC of 80% or above for every concept. Concepts for which automatic detection was particularly successful are: *indoor/outdoor*, *night*, and *sky*. The worst results were obtained for the concept *road or pathway*.

## A Results for all submissions

The results for each submitted run are given in Table 4.

## References

1. Arni, T., Clough, P., Sanderson, M., Grubinger, M.: Overview of the ImageCLEFphoto 2008 photographic retrieval task. In: Proceedings of the CLEF Workshop 2008. Lecture Notes in Computer Science, Aarhus, Denmark (2008 (printed in 2009))
2. Müller, H., Kalpathy-Cramer, J., Kahn Jr., C.E., Hatt, W., Bedrick, S., Hersh, W.: Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Proceedings of the CLEF Workshop 2008. Lecture Notes in Computer Science, Aarhus, Denmark (2008 (printed in 2009))
3. Tsikrika, T., Kludas, J.: Overview of the wikipediaMM task at ImageCLEF 2008. In: Proceedings of the CLEF Workshop 2008. Lecture Notes in Computer Science, Aarhus, Denmark (2008 (printed in 2009))
4. Deselaers, T., Deserno, T.M.: Medical image annotation in ImageCLEF 2008. In: Proceedings of the CLEF Workshop 2008. Lecture Notes in Computer Science, Aarhus, Denmark (2008 (printed in 2009))
5. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR TC-12 benchmark - a new evaluation resource for visual information systems. In: Proceedings of the International Workshop OntoImage'2006. (2006) 13–23
6. Fluhr, C., Moëllic, P.A., Hède, P.: ImagEVAL: Usage-oriented multimedia information retrieval evaluation. In: Proceedings of the second MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation, Alicante, Spain (2006) 3–8
7. Deselaers, T., Hanbury, A., Viitaniemi, V., Benczúr, A., Brendel, M., Daróczy, B., Balderas, H.E., Gevers, T., Gracidas, C.H., Hoi, S.C.H., Laaksonen, J., Li, M., Castro, H.M., Ney, H., Rui, X., Sebe, N., Stöttinger, J., Wu, L.: Overview of the imageclef 2007 object retrieval task. In: Proceedings of the CLEF Workshop 2007. (2007 (printed in 2008))

Table 4: Average EER and Average AUC over all concepts for all runs of all participating groups.

| group | run | EER [%] | AUC [%] |
|---|---|---|---|
| CEA_LIST | CEA_LIST_2 | 29.71 | 71.44 |
| CEA_LIST | CEA_LIST_3 | 41.43 | 34.25 |
| CEA_LIST | CEA_LIST_4 | 29.04 | 73.40 |
| HJ_FA | HJ_Result | 45.07 | 19.96 |
| IPAL_I2R | I2R_IPAL_Cor_Run1 | 40.02 | 62.62 |
| IPAL_I2R | I2R_IPAL_Edge_Run2 | 45.71 | 55.79 |
| IPAL_I2R | I2R_IPAL_HIST_Run4 | 31.83 | 73.80 |
| IPAL_I2R | I2R_IPAL_Linear_Run5 | 36.09 | 68.65 |
| IPAL_I2R | I2R_IPAL_Texture_Run | 39.22 | 62.93 |
| IPAL_I2R | I2R_IPAL_model_Run6 | 33.93 | 72.01 |
| IPAL_I2R | IPAL_I2R_FuseMCE_R7 | 31.17 | 74.05 |
| IPAL_I2R | IPAL_I2R_FuseNMCE_R8 | 29.71 | 76.44 |
| LSIS | GLOT-methode23_LSIS_evaOK | 26.56 | 79.92 |
| LSIS | new_kda_results.txt | 25.88 | 80.51 |
| LSIS | FusionA_LSIS.txt | 49.29 | 50.84 |
| LSIS | FusionH_LSIS.txt | 49.38 | 50.20 |
| LSIS | MLP1_LSIS_GLOT | 25.95 | 80.67 |
| LSIS | MLP1_vcdt_LSIS | 25.95 | 80.67 |
| LSIS | method2_LSIS | 26.61 | 79.75 |
| MMIS | MMIS_Ruihu | 41.05 | 62.50 |
| MMIS | ainhoa | 28.44 | 77.94 |
| MMIS | alexei | 28.82 | 77.65 |
| MMIS | combinedREPLACEMENT | 31.90 | 73.69 |
| Makerere | MAK | 49.25 | 30.83 |
| RWTH | PHME | 20.45 | 86.19 |
| TIA | INAOE-kr_00_HJ_TIA | 42.93 | 28.90 |
| TIA | INAOE-kr_04_HJ_TIA | 47.12 | 17.58 |
| TIA | INAOE-lb_01_HJ_TIA | 39.12 | 42.15 |
| TIA | INAOE-psms_00_HJ_TIA | 32.09 | 55.64 |
| TIA | INAOE-psms_02_HJ_TIA | 35.90 | 47.07 |
| TIA | INAOE-rf_00_HJ_TIA | 39.29 | 36.11 |
| TIA | INAOE-rf_03_HJ_TIA | 42.64 | 26.37 |
| UPMC | LIP6-B50trees100C5N5 | 27.32 | 71.98 |
| UPMC | LIP6-B50trees100C5N5T25 | 28.93 | 53.78 |
| UPMC | LIP6-B50trees100COOC5T25 | 28.83 | 54.19 |
| UPMC | LIP6-B50trees100pc | 24.55 | 82.74 |
| UPMC | LIP6-B50trees100pc_COOC5 | 27.37 | 71.58 |
| UPMC | LIP6-B50trees100pc_T25 | 26.20 | 57.09 |
| XRCE | TVPA-XRCE_KNN | 16.65 | 90.66 |
| XRCE | TVPA-XRCE_LIN | 19.29 | 88.73 |
| budapest | acad-acad-logreg1 | 37.36 | 66.39 |
| budapest | acad-acad-logreg2 | 37.12 | 66.53 |
| budapest | acad-acad-lowppnn | 36.07 | 67.15 |
| budapest | acad-acad-lowppnpnn | 32.46 | 73.05 |
| budapest | acad-acad-medfi | 32.47 | 73.57 |
| budapest | acad-acad-mednofi | 32.10 | 74.18 |
| budapest | acad-acad-medppnn | 37.01 | 59.30 |
| budapest | acad-acad-medppnpnn | 32.47 | 73.61 |
| budapest | acad-acad-mixed | 38.34 | 63.80 |
| budapest | acad-budapest-acad-glob1 | 45.72 | 52.78 |
| budapest | acad-budapest-acad-glob2 | 31.14 | 74.90 |
| budapest | acad-budapest-acad-lowfi | 32.48 | 73.03 |
| budapest | acad-budapest-acad-lownfi | 32.44 | 73.32 |