

The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks

Stefanie Nowak, Karolin Nagel and Judith Liebetrau

Fraunhofer Institute for Digital Media Technology (IDMT)
Ehrenbergstr. 31, 98693 Ilmenau, Germany
research@stefanie-nowak.de, judith.liebetrau@idmt.fraunhofer.de

Abstract. The ImageCLEF 2011 Photo Annotation and Concept-based Retrieval Tasks pose the challenge of an automated annotation of Flickr images with 99 visual concepts and the retrieval of images based on query topics. The participants were provided with a training set of 8,000 images including annotations, EXIF data, and Flickr user tags. The annotation challenge was performed on 10,000 images, while the retrieval challenge considered 200,000 images. Both tasks differentiate among approaches that consider solely visual information, approaches that rely only on textual information in form of image metadata and user tags, and multi-modal approaches that combine both information sources. The relevance assessments were acquired with a crowdsourcing approach and the evaluation followed two evaluation paradigms: per concept and per example. In total, 18 research teams participated in the annotation challenge with 79 submissions. The concept-based retrieval task was tackled by 4 teams that submitted a total of 31 runs. Summarizing the results, the annotation task could be solved with a MiAP of 0.443 in the multi-modal configuration, with a MiAP of 0.388 in the visual configuration, and with a MiAP of 0.346 in the textual configuration. The concept-based retrieval task was solved best with a MAP of 0.164 using multi-modal information and a manual intervention in the query formulation. The best completely automated approach achieved 0.085 MAP and uses solely textual information. Results indicate that while the annotation task shows promising results, the concept-based retrieval task is much harder to solve, especially for specific information needs.

1 Introduction

With the increasing amount of digital information on the Web and on personal computers, the need for systems that are capable of automated indexing, searching, and organising multimedia documents incessantly grows. Automated systems have to retrieve information with high precision in order to be accepted by industry and end-users. Often, multimedia retrieval systems are evaluated on different test collections with different performance measures, which makes the comparison of retrieval performance impossible and limits the benefit of the approaches. Benchmarking campaigns counteract these tendencies and establish an

objective comparison among the performance of different approaches by posing challenging tasks and by distributing test collections, topics, and measures.

This paper presents an overview of the ImageCLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. The two tasks aim at the automated detection of visual concepts in consumer photos and the retrieval of photos based on a certain topic. Section 2 introduces the tasks and the evaluation methodology. Following, Section 3 discusses the visual concepts and query topics which simulate the user’s information need in image search. Then, Section 4 describes the test collection and the relevance assessment process. Section 5 summarizes the approaches of the participants. Following, the results for the annotation task and the concept-based retrieval task are presented and discussed in Section 6 and Section 7, respectively. Finally, Section 8 summarizes and concludes this paper.

2 Task Description

The Photo Annotation and Concept-based Retrieval Tasks pose an image analysis challenge which consists of two sub tasks. The annotation task aims at the automated annotation of consumer photos with multiple concepts. It is similar to the visual concept detection and annotation task (VCDT) as it was posed in the last years [1, 2]. This year, the participants are asked to annotate a test set of 10,000 Flickr images with 99 visual concepts. To solve this task, an annotated training set of 8,000 images is provided. The evaluation considers a fully assessed test collection to compare the approaches of the participants. The second challenge poses a concept-based retrieval task. The participants are asked to retrieve (up to) the 1,000 most relevant images in ranked order for a given topic out of a test collection of 200,000 images. In total 40 topics, each consisting of a logical connection of concepts from the annotation task, are provided. Concept detectors may be trained on the training set of the annotation task (8,000 images annotated with 99 visual concepts). The assessment incorporates a pooling strategy with crowdsourced relevance assessments. Both tasks can be solved by following three different approaches:

1. Automatic annotation with visual information only (“**V**”)
2. Automatic annotation based on Flickr user tags and image metadata (“**T**”)
3. Multi-modal approaches that consider visual information and/or Flickr user tags and/or EXIF information (“**M**”)

The participants can choose one task or participate in both. Both tasks make use of a subset of the MIR Flickr 1 Million image dataset [3]. The MIR Flickr collection supplies all original tag data provided by the Flickr users (further denoted as Flickr user tags). These Flickr user tags are made available for the textual and multi-modal approaches of both subtasks. For most of the photos, the EXIF data is included and may be used.

2.1 Evaluation Objectives

The main evaluation objectives of the two tasks in 2011 lie in the exploitation of different knowledge sources, the benefit of annotation approaches as part of the

retrieval process, and the automated prediction of subjective concepts such as sentiments. Moreover, participants need to deal with an unbalanced amount of data per concept, a varying number of labels per image, the diversity of image content per concept, and the different qualities of image metadata.

2.2 Ontology

The novel sentiment concepts are included in the Photo Tagging ontology [4] of the last years. The hierarchy allows making assumptions about the assignment of concepts to documents. Additionally, other relationships between concepts determine possible label assignments. The ontology restricts, for instance, the simultaneous assignment of some concepts (disjoint items) or defines that one concept postulates the presence of other concepts. The ontology allows the participants to incorporate semantic knowledge in their annotation algorithms, and to make assumptions about probable concept combinations.

2.3 Evaluation Measures

In the annotation task, the evaluation sticks to the concept-based and example-based evaluation paradigm. For the concept-based evaluation, the Mean interpolated Average Precision (MiAP) is utilized, while the example-based evaluation applies the example-based F-Measure (F-Ex). Additionally, we introduce a novel performance measure called Semantic R-Precision (SR-Precision) which is based upon the example-based R-Precision, but incorporates the Flickr Tag Similarity (FTS) [5] to determine the semantic relatedness of visual concepts in the case of misclassifications. R-Precision calculates Precision at perfect Recall in an example-based evaluation scenario. The SR-Precision variant assigns misclassification costs based on the semantic relatedness among misclassified concepts. The semantic relatedness is derived from the FTS measure. In contrast to the Ontology Score with Flickr Context Similarity (OS-FCS) which was used in 2010 [2], the SR-Precision is able to incorporate ranked predictions instead of forcing the systems to provide binary decisions. However, this measure requires a normalization of classifier scores over different classifier outputs to deliver meaningful results. This requirement was not explicitly posed to the participants and therefore algorithms might not be optimally parameterized for this measure.

The concept-based retrieval task evaluates performance on a test collection with incomplete relevance judgments. All submissions of the participants are pooled by using a pool depth of 100 documents per topic and run. Finally, the runs are evaluated with the Mean uninterpolated Average Precision (MAP), Precision@10 (P@10), Precision@20 (P@20), Precision@100 (P@100), and concept-based R-Precision (R-Prec) with the trec-eval 8.1 program.

3 Incorporation of user needs in the evaluation

Topics and visual concepts are strongly related to user needs and define the use cases of a system. While concepts are modality-independent (i.e., the event

“birthday” might be detectable in the visual modality (birthday cake, people celebrating) as well as in the auditory modality (people singing a birthday song)), visual concepts are solely described by the visual content of a photo and are therefore language independent. This section introduces the visual concepts that are applied in the annotation task and the derivation of query topics for the retrieval task based on these concepts, query logs, and related work. Please note that the process of collecting images and defining visual concepts is different from related work. While usually the concept lexicon exists before images are collected, in the case of the ImageCLEF VCDT test collection, this process is decoupled and the images have been collected first. This approach is much closer to reality and poses new challenges, as objects are not necessarily centred in the image and the distribution of images per concept varies considerably.

3.1 Definition of visual concepts

The test collection for the annotation task contains manual annotations for 99 visual concepts. These concepts describe the scene (*indoor, outdoor, landscape...*), depicted objects (*car, animal, person...*), the representation of image content (*portrait, graffiti, art...*), events (*travel, work...*), or quality issues (*overexposed, underexposed, blurry...*). This year, a special focus is laid on the detection of sentiment concepts. All in all, 49 concepts of the 53 concepts used in 2009 [1] were utilized again. The concept *Canvas* as well as the concepts *No_Visual_Season*, *No_Visual_Place*, and *No_Visual_Time* were discarded in this year’s challenge. The 41 concepts which were added in 2010 are all reused. In 2011, nine novel sentiment concepts were added to the test collection. For the definition of sentiments, we follow the approach of Russell [6], who defines an emotional space with two dimensions (arousal and valence) on which emotional adjectives/sentiments can be placed. Valence spans from the negative pole “misery” to the positive pole “pleasure” on the x-dimension, while arousal spans from “passive” to “active” in the y-dimension. In this model, adjectives are grouped into eight affect concepts in circular order. The model was slightly adapted and an additional concept *funny* was included. The eight sentiments are structured according to their degrees in the circle as proposed by Russel. Partly, the wording is changed as to better fit the sentiments to describe images (e.g., an image cannot be excited or astonished, but it may look exciting to a human being). Starting with *happy* at 0°, the circle is further composed of *funny* (about 30°), *euphoric* (70°), *active* (90°), *scary* (150°), *unpleasant* (180°), *melancholic* (210°), *inactive*, (270°), and *calm/comforting* (330°).

3.2 Definition of topics for the concept-based retrieval task

Based on the visual concepts, 40 topics for the concept-based retrieval task were constructed. We conceived that each topic contains a different number of relevant images, and that the topics comprise a range of difficulty levels. For the definition of relevant topics, we followed two approaches: First, we adapted topics from the ImageCLEF Wikipedia retrieval task [7], [8], as these topics

Table 1: Topics of the concept-based retrieval task. Topics, labelled with WR and an abbreviated year, are taken or adapted from the ImageCLEF Wikipedia retrieval tasks.

Topic No.	Topic title	Topic Source
1	graffiti on buildings or walls	WR 11
2	toy vehicle	
3	single person doing sports on the sea	WR 09/10
4	airplane in the sky	WR 10
5	rider on horse	WR 09/10
6	cyclist	WR 09/10
7	mountains with sky during night	WR 10
8	fish in water	WR 10
9	desert scenery	WR 10
10	single person playing a musical instrument	WR 10
11	animal in snow	
12	snowy winter landscape with trees	WR 10
13	female person(s) doing sports	
14	cities at night with cars	WR 09/10
15	sea sunset or sunrise	WR 10
16	outside view of a church	
17	waters in autumn	
18	female old person	
19	close-up of trees	WR 10
20	trains indoor	WR 10
21	scary dog(s)	
22	portrait that is out of focus	WR 10/11
23	bridges not over water	WR 10
24	funny baby	WR 09
25	melancholic photos in rain	
26	houses in mountains	WR 11
27	family holidays at the beach in summer	
28	fireworks	
29	close-up of flower(s) with rain drops	
30	cute toys arranged to a still-life	
31	ship or boat on a river	
32	underexposed photos of animals	
33	cars and motion blur	
34	unpleasant insects	
35	close-up of bird	
36	scary shadows of people	
37	painting of person(s)	
38	birthday or wedding cake	
39	house surrounded by a garden	
40	close-up of bodypart with depth of focus	

were designed based on web-query logs and because they range from simple to semantic (hence highly difficult) topics as described in [9]. A total of 17 topics were directly applicable to our test data. Second, we examined interesting queries for the test collection. Based on the output for each query, it was decided if the chosen topic comprises an adequate occurrence in the test collection. The 40 resulting topics and their source are shown in Table 1. Sample images of the dataset were taken for clarification and provided as examples for the topics.

4 Ground Truth Acquisition

The relevance assessments for the annotation task and the concept-based retrieval task were acquired with a crowdsourcing approach using Amazon Mechanical Turk¹ (MTurk). MTurk is an online marketplace which distributes mini-jobs to an undefined crowd of people. At MTurk these mini-jobs are called HITs (Human Intelligence Tasks). The workers at MTurk, called turkers, can choose the HITs they would like to perform and submit the results to MTurk. The requester of the work collects the results from MTurk and approves or rejects the work of the turkers. Experiences with MTurk from ImageCLEF 2010 show the applicability of crowdsourcing for ground truth acquisition of image labels. This year, additional quality assurance mechanisms were incorporated to reduce the impact of spammers on the annotations.

4.1 Design of the annotation HIT template

The assessment of the sentiment concepts was performed by asking the turkers what sentiments an image conveys. The HIT template includes a definition of sentiments, synonymous sentiments, and example images (see Figure 1). The definitions are derived from WordNet 3.0² and the Free Dictionary³. Each survey comprises ten images. The image is depicted on the left, while on the right the adapted circumplex model of Russel (see Section 3.1) is visualised, as illustrated on the example of one image in Figure 2. The option *no sentiment* should be chosen if no sentiment fits to the image. After selecting this checkbox, the turkers were asked to give a mandatory reason why no sentiment fits. We included this question in the survey to prevent turkers from clicking at this checkbox without thinking about the task. For all other sentiments, several choices could be selected at the same time. Additionally, the turkers were asked which reason let them decide for a sentiment: the motif or the overall impression of the image. They could choose on a five-point scale with the scales “motif” – “mostly motif” – “both equally” – “mostly overall impression” – “overall impression”.

The HIT template includes an automated verification procedure. For all ten images that belong to one HIT, it is verified that the survey is completely filled out before the submission of the task works. In the case of missing answers, the

¹ www.mturk.com

² <http://wordnetweb.princeton.edu/perl/webwn>, last accessed 20.07.2011

³ <http://www.thefreedictionary.com/>, last accessed 20.07.2011

Examples		
Happy	<p>Enjoying something or marked by joy or pleasure.</p> <p>Words used as synonyms are cheerful, beautiful, idyllic, sensual, heavenly.</p>	
Funny	<p>Causing laughter or amusement.</p> <p>Words used as synonyms are humorous, comical, merry, amusing.</p>	
Euphoric	<p>Feeling very happy, in high spirits.</p> <p>Words used as synonyms are expressive, inspiring, brilliant, joyful, exciting, elated, jocular.</p>	
Active	<p>Full of motion, energy and life. Describes dynamic situations or scenes.</p> <p>Words used as synonyms are adventurous, jaunty, vivid, gaily, swift, frolic, dynamic.</p>	
Scary	<p>Provoking fear or terror.</p> <p>Words used as synonyms are tragic, incriminating, painful, frightening, threatening, horrible, suffering.</p>	
Unpleasant	<p>Disagreeable to the senses, to the mind, or feelings.</p> <p>Words used as synonyms are sorrow, calamitous, ugly, nasty.</p>	
Melancholic	<p>A feeling of thoughtful sadness, a constitutional tendency to be gloomy and depressed, characterized by/feeling/repressing sadness.</p> <p>Words used as synonyms are depressing, gloomy, lousy.</p>	
Inactive	<p>Lacking in motion or dynamic. Full of passiveness.</p> <p>Words used as synonyms are static, passive, dull, quiet, monotonous, bleak.</p>	
Calm	<p>State of not being agitated, without losing self-possession. Providing freedom from worry.</p> <p>Words used as synonyms are relaxing, soothing, comfortable, cozy, easy.</p>	

Definitions are taken from [WordNet 3.0](#) and [The Free Dictionary](#).

Fig. 1: Definition of sentiment concepts in the HIT template.

turkers see the corresponding questions marked in red. This procedure ensures that it is not too easy to answer randomly and submit spam, and it helps reducing our work to filter out incomplete answers that need to be republished. While it does not assure that all random annotators are excluded (as turkers still can randomly answer each question), this at least assures that it also costs some amount of work to cheat compared to the time that is needed to answer honestly.

4.2 Assessment statistics of the annotation task

The ground truth was acquired in different annotation batches. The pretest included 400 images of the training set arranged in surveys of ten images per HIT. Each HIT was annotated three times by a total of 22 turkers in an average an-

What sentiments does this image convey?

Please choose at least one sentiment.

active euphoric
 scary MISERY PLEASURE funny
 unpleasant nothing happy
 melancholic PASSIVE calm/comforting
 inactive

What triggered that sentiment the most?

mostly overall impression

motif overall impression

Fig. 2: Example HIT for a complete annotation.

notation time of 3 minutes and 12 seconds and paid with 0.05\$. The purpose of the pretest was to understand if the template design and the task were understandable and if the turkers were able to solve the task. Results show that in about 50% of the images the turkers are agreeing on the sentiment (or choosing neighbouring sentiments) while in the other 50%, they chose opposite sentiments (like *happy* and *melancholic*). The rest of the training set of the Photo Annotation Task was annotated in altogether 4,225 HITs. Each HIT contained nine photos of the training set and one photo of the pretest as gold standard. The gold standard was built by a majority vote of the pretest images excluding the *no sentiment* concept and randomly placed into each HIT survey. Each HIT was annotated five times and rewarded with 0.07\$. On average, they were completed in 2 minutes and 36 seconds by a total of 258 distinct turkers. The test set was assessed in 5,560 HITs which each included nine images and one gold standard image of the training set. Each image of the test set was annotated five times. The HITs of the test set were divided into two batches (in order not to pose too many HITs at the same time) and annotated by 156 distinct turkers. Each HIT was rewarded with 0.07\$, again. For the first batch of 2,745 HITs, each HIT was annotated on average in 2 minutes and 8 seconds, while the 2,815 HITs of the second batch were annotated in average in 1 minute and 44 seconds.

The verification of the work of the turkers is difficult, as the task of sentiment annotation is very subjective. Therefore, we followed several strategies on how to compare the annotations. The verification of the HITs of the training and test set with the gold standard images lead to a direct acceptance of 3,204 and

What is the HIT about?

In the following HIT you will be given 24 images that were assigned to a certain topic. Please select any image that you think does not fit the topic and tell us the reason why you want to eliminate it.

Guidelines:

- Choose any image that does not fit the given topic.
- For any selected image, use the textbox to tell us why you chose it.

NOTE: You can only submit your answers if all of these steps are fulfilled. Incomplete answers are highlighted by a red border during submission.

Please be advised that occasionally there might be a small number of adult or disturbing images despite our effort to filter them out.

How are you paid?

You will be paid the amount that is defined in the HIT. Your submission will typically be approved/rejected within 7 days.

Topic: Airplane in the Sky

We'd like to find photos of real airplanes flying in the sky. Airplanes on the ground are not relevant, neither are airplanes pictured from the inside. Small models of airplanes are also not relevant.

Examples:




Fig. 3: Instructions in the HIT template for topic 4.

4,358 HITs, respectively, allowing a deviation of at most 90° on the affect circle. For HITs that did not pass the gold standard test, we compared the results of the HIT to the four answers of other turkers for the same HIT. For all images of the HIT, the deviation to the annotations of the other HITs was computed and the HITs were accepted when the deviation was equal or less to 90° on the affect circle per image. A total of seven out of the 10 images had to fit. With this procedure all remaining HITs could be accepted.

The final construction of the ground truth considers the majority vote for each image. In the case that no clear answer was given, we decided to discard any sentiment information for that image. In total, about 15% of the training set images and 14% of the test set images have no sentiment information. Interestingly, the *no sentiment* option was rarely chosen by the turkers. For none of the images of the training set and only for one image of the test set a majority of people decided for this concept.

4.3 Design of the topic HIT templates

In the relevance assessment of the concept-based retrieval task, the turkers were asked to mark all relevant images on the HIT template for a given topic. Each HIT template includes a definition of the topic and example images (see Figure 3). A HIT contains 22 images plus two gold standards images, which were used as a means of reliability control for the assessments. For each topic, we selected one image that fits the definition and one image that is not relevant for the given topic. Special attention was taken in the design of the irrelevant



Fig. 4: Sample images that are not in the scope of the topic *fish in the water* (top) and images that are covered by the topic (bottom).

images per topic. Instead of using images that are clearly out of the scope of the topic, images that match the meaning of the topic quite close, but not exactly, were chosen; see Figure 4 for sample images of the topic *fish in the water*. The gold images were placed randomly in the HIT templates.

4.4 Assessment statistics of the retrieval task

The number of HITs per topic is dependent on the total number of distinct images that were retrieved by the runs of the participants. Each HIT contained 24 images and was assessed by three turkers; so in total 7,868 HITs were processed. Accepted HITs were paid with 0.03\$. Each topic was processed by at least five (topic 29) and at most 41 (topic 12) distinct turkers. The average grading time varies per topic between 31 seconds (topic 25) and 1 minute and 19 seconds (topic 15). To increase the reliability of the relevance judgments, the results were subjected to a post screening procedure. The assessments of a turker were rejected if the gold standard images were not marked correctly. These HITs were published again until for all HITs three reliable results were available. In the next step, the assessments of the three turkers per HIT were compared with each other. As the task of selecting relevant images for a topic is a subjective task, its verification is difficult. In an additional step, we visualized the images that were assessed as relevant and estimated the number of false assignments and missing assignments. Depending on these results, the number of votes that were necessary to define an image as relevant were chosen for each topic. For most topics, a majority vote from at least two of the three assessors was necessary. A minority vote was used for only five topics, while all assessors had to agree on relevance for three topics.

5 Participation

Altogether, 48 groups registered for the challenge. 42 groups signed the license agreement and were provided with the test collections. For the annotation task, 18 groups submitted results in altogether 79 runs. The number of runs was restricted to a maximum of 5 runs per group. In total, there were 46 submissions using only visual information, 8 submissions using only textual information, and 25 submissions utilising multi-modal approaches. For the retrieval task, 4 groups submitted results in a total of 31 runs. The maximum number of runs per group was set to 10. The submissions include 14 visual runs, 7 textual runs, and 10 multi-modal runs. The runs can be subdivided into 16 runs that retrieved all images in a completely automated fashion and 15 runs that included a manual intervention in the query generation step or relevance feedback. All participants that submitted to the retrieval task also took the challenge in the annotation task. The teams and their approaches are briefly introduced in the following:

BPACAD [10]: The team of the Computer and Automation Research Institute of the Hungarian Academy of Science submitted one textual, two visual and three multi-modal runs to the annotation task. Their approach is based on a kernel weighting procedure using visual Fisher kernels and a Flickr-tag based Jensen-Shannon divergence based kernel. Classification uses a linear SVM trained for each concept separately.

BUFFALO: The team of the University at Buffalo, New York, USA submitted five visual runs for the annotation task. They follow two approaches: the first considers a local linear coordinate method to learn concepts with a regression method. The second uses a combination of GIST and colour features and classifies the images by a neural network.

CAEN [11]: The group of University of Caen, France participated with four visual runs in the annotation task. The proposed approach uses visual image features, such as SIFT, HOG, Texton, LAB, SSIM, and Canny, and aggregated them by a Bag-of-Words (BoW) model into a global histogram. Fisher Vectors and contextual information were used as enhancement of the BoW-models. The classification considers SVM models trained for each concept separately.

CEALIST [12]: The team from the Laboratory of Vision and Content Engineering, France submitted one textual, one visual, and three multi-modal runs to the annotation task. The textual descriptor is based on semantic similarity between tags and visual concepts. Two distances were used: one based on the Wordnet ontology and one based on social networks. The visual component considers various local and global features, such as Fisher vectors as well as colour and edge features. Late fusion was used to combine visual and textual modalities.

DBIS: The team of the Technical University of Cottbus, Germany submitted five runs in the visual configuration to the annotation task. They use various features and investigate the influence of several parameters in clustering on the annotation performance.

HHI [13]: The team of Fraunhofer HHI, Berlin, Germany submitted five visual runs to the annotation task. Their approach is based on the BoW model. A feature fusion of the opponent SIFT descriptor and the GIST descriptor was done

in order to improve the classification performance of scene-based concepts. HHI investigates a sampling of informative images in the training procedure, which resulted in qualitative as well as runtime performance gains. A post-classification processing step is incorporated, which refines classification results based on rules of inference and exclusion between concepts.

IDMT [14]: The group of Fraunhofer IDMT, Ilmenau, Germany submitted one textual and four multi-modal runs to the annotation task. Their approach focuses on the fusion of multi-modal information and the exploitation of Flickr user tags. As visual features, they employ RGB-SIFT features in a codebook approach and classify the images with a one-against-all strategy using a SVM with RBF kernel.

ISIS [15]: The Intelligent Systems Lab of the University of Amsterdam, The Netherlands participated with five runs in the annotation task (3V, 2M) and ten runs (10V) in the retrieval task. All runs of the annotation task use several colour SIFT features with Harris-Laplace and dense sampling, and apply the SVM classifier. The multi-modal runs further include binary vectors for the most frequent Flickr user tags. In the retrieval task, three runs are computed completely automated and seven include a manual intervention by following two approaches. In the fully automated runs, a combination of the provided positive example images and random irrelevant images were used to train the concept detector. For the human topic mapping, a human reads the topic and then selects the relevant concept(s). The probability scores of these concepts are then combined using either summation or multiplication. In the human topic inspection approach, relevance feedback was used to improve results.

LAPI [16]: The group of Laboratorul de Analiza si Prelucrarea Imaginilor, Universitatea Politehnica Bucuresti, Bucharest, Romania submitted two runs using a visual-only approach. They combine colour and structural features and adopt a Linear Discriminant Analysis for classification. Post-processing considers joint probabilities of concept occurrences in the training set for label elimination.

LIRIS [17]: The group of Université de Lyon, CNRS, France participated in the annotation task with two textual, one visual, and two multi-modal runs. They consider two textual descriptors: one is based on a semantic distance between the text and an emotional dictionary, the other one contains the valence and arousal meanings by making use of the Affective Norms for English Words dataset. In the visual approaches, different visual features including colour, texture, shape, and high level aesthetic features are applied. Performance is compared using different fusion strategies as well as Adaboost and SVM classifiers.

MEIJI [18]: The group of Meiji University, Kanagawa, Japan submitted five runs (2V, 1T, 2M) to the annotation task and ten completely automated runs (2V, 2T, 6M) to the retrieval task. Their approach is based on visual word co-occurrence using the BoW model and global colour features as well as textual features derived by tf-idf weights of Flickr user tags. Classification is performed by an adaptation of the so-called confabulation model.

MLKD [19]: The Machine Learning and Knowledge Discovery group of the Aristotle University of Thessaloniki, Greece participated in the annotation

task with five runs (1V, 1T, 3M) and in the retrieval task with two automated and eight semi-automated runs (2V, 4T, 4M). They approach the photo annotation challenge with multi-label learning algorithms based on Random Forests as base classifier. The visual features consider seven local descriptors with two sampling strategies. The textual models are based on a Boolean BoW representation including word stemming, stop words removal, and feature selection. The multi-modal approach considers a hierarchical late-fusion of the modalities. For the concept-based retrieval task two approaches were used: one based on the concept relevance scores in a manual configuration and one automated approach which is based solely on the sample images using textual information.

MRIM [20]: The team of Grenoble University, France submitted four runs (3V, 1M) to the annotation task. Classification considers multiple SVM classifiers with RBF kernel. In the visual runs, several global and local colour and texture descriptors are applied and dimension reduction techniques are investigated. The multi-modal run additionally considers Flickr user tags as simple textual features in a late fusion of SVM classifier scores.

MUFIN [21]: The Faculty of Informatics, Masaryk University, Brno, Czech Republic participated with four multi-modal runs in the annotation task. Their approach is based on a free-text annotation system that assigns arbitrary words to web images by visual and textual neighbour searching. For the textual search, the EXIF data and image descriptions were used, while the visual search considers different MPEG-7 descriptors. The search considers the Profimedia dataset to find the nearest neighbours and transfers its annotations to the ImageCLEF test collection including a removal of stopwords and names. The resulting words were transformed into the fixed set of 99 visual concepts of the annotation task with the help of WordNet and the provided ontology.

NII [22]: The team of the National Institute of Informatics, Tokyo, Japan participated with five visual runs in the annotation task. Their models are using global and local features. As for global features, colour moments, colour histogram, edge orientation histogram, and local binary patterns are applied. As for local features, keypoint detectors such as Harris Laplace, Hessian Laplace, Harris Affine, and Dense Sampling are used to extract SIFT-descriptors. Classification is performed with a SVM classifier.

REGIMVID [23]: The research group on Intelligent Machines, University of Sfax, Tunisia submitted one textual run to the annotation task and one textual, automated run to the retrieval task. Their approach focuses on the exploration of Flickr tags to extract contextual relationships of tag relations. Therefore, two types of contextual graphs are modeled: an inter-concepts graph and a concept-tags graph.

TUBFI [24]: The joint submission of the Machine Learning Group, Berlin Institute of Technology and Fraunhofer FIRST Berlin, Germany consists of four visual and one multi-modal run to the annotation task. Classification considers non-sparse multiple kernel learning and multi-task learning. Different extensions of the BoW models with respect to sampling strategies and BoW mappings were

Table 2: Summary of the results for the evaluation per concept. The table shows the MiAP for the best run per group and the averaged MiAP for all runs for each group and indicates the configuration of the run. The teams are sorted by the rank of their best run.

		BEST RUN			AVERAGE RUNS		
Team	Runs	MiAP	Rank	Config.	MiAP	Rank	Config.
TUBFI	5	1	0.443	M	11.00	0.394	V+M
LIRIS	5	2	0.437	M	26.00	0.372	V+T+M
BPACAD	5	3	0.436	M	13.40	0.401	V+T+M
ISIS	5	5	0.433	M	14.00	0.391	V+M
MLKD	5	9	0.402	M	31.00	0.349	V+T+M
CEALIST	5	11	0.384	M	32.80	0.339	V+T+M
CAEN	4	14	0.382	V	23.50	0.363	V
MRIM	4	16	0.377	M	43.75	0.289	V+M
IDMT	5	20	0.371	M	28.00	0.354	T+M
NII	5	34	0.337	V	42.80	0.321	V
HHI	5	36	0.335	V	40.00	0.328	V
MELJI	5	50	0.304	T	63.20	0.254	V+T+M
MUFIN	4	52	0.299	M	53.75	0.296	M
BUFFALO	5	60	0.249	V	62.60	0.236	V
DBIS	5	63	0.230	V	66.40	0.218	V
UNIKLU	4	70	0.207	V	71.50	0.206	V
REGIMVID	1	74	0.204	T	74.00	0.204	T
LAPI	2	77	0.177	V	77.50	0.177	V

proposed. The multi-modal run further considers frequent Flickr user tags based on a soft mapping for textual BoWs and Markov random walks over tags.

UNIKLU: The team of the Institute of Information Technology, Alpen-Adria University, Klagenfurt, Austria participated with four visual runs in the annotation task. They made use of the LIRE framework and applied several features such as SIFT, SURF, MSER, CEDD, FCTH, and colour histograms and classified the images with a linear SVM. Two of the runs incorporate an automated post-processing of the classification results.

6 Annotation Task: Results

This section illustrates the results for the annotation subtask. First, the overall results of all teams are presented, independent of the configuration. In the following subsections the results per configuration are highlighted. The results for all runs can be found at the Photo Annotation Task website⁴.

The task was solved best with a MiAP of 0.443 (TUBFI), followed by a MiAP of 0.437 (LIRIS) as illustrated in Table 2. Both runs make use of multi-modal

⁴ <http://www.imageclef.org/2011/photo>

Table 3: Summary of the results for the evaluation per example. The table shows the F-Ex, SR-Precision, and run configuration for the best run per group.

Team Rank F-Ex Config.				Team Rank SR-Prec. Config.			
ISIS	1	0.622	M	ISIS	1	0.742	M
CAEN	5	0.600	V	BPACAD	5	0.729	V
BPACAD	6	0.593	M	CAEN	6	0.727	V
HHI	8	0.588	V	LIRIS	7	0.725	V
LIRIS	14	0.576	M	HHI	11	0.718	V
TUBFI	17	0.566	M	IDMT	13	0.713	M
MLKD	19	0.560	V	CEALIST	19	0.711	M
MRIM	21	0.552	M	MRIM	23	0.706	M
IDMT	22	0.552	M	NII	26	0.702	V
BUFFALO	34	0.527	V	MLKD	31	0.698	M
DBIS	37	0.518	V	BUFFALO	39	0.683	V
CEALIST	42	0.508	M	UNIKLU	45	0.672	V
MEIJI	48	0.495	M	DBIS	47	0.671	V
UNIKLU	57	0.469	V	TUBFI	62	0.630	V
MUIN	62	0.462	M	MUFIN	63	0.628	M
LAPI	70	0.390	V	LAPI	71	0.551	V
NII	73	0.298	V	MEIJI	73	0.491	T
REGIMVID	78	0.141	T	REGIMVID	78	0.396	T

information. Table 3 depicts the overall rankings for the results of the evaluation per example. The best results in terms of F-Ex are achieved in a multi-modal configuration with 0.622 (ISIS), followed by 0.600 F-Ex (CAEN) which makes use of a visual configuration. In terms of SR-Precision, the best run scores with 0.742 SR-Precision (ISIS) in a multi-modal run, followed by 0.729 SR-Precision (BPACAD) in a visual run.

6.1 Results for the visual configuration

Table 4 shows the results of the best run of each group that participated in the visual configuration evaluated with all three evaluation measures. The best results in the visual configuration are achieved by the TUBFI team in terms of MiAP, closely followed by the team CAEN (0.388 vs 0.382 MiAP). In the example-based evaluation, ISIS scored best for both measures followed by CAEN (F-Ex) and BPACAD (SR-Precision).

6.2 Results for the textual configuration

The results for the textual runs are presented in Table 5. The best run scores with 0.346 MiAP (BPACAD), followed by 0.326 MiAP (IDMT, MLKD). In the example-based evaluation the best run scores with 0.525 F-Ex (IDMT) and 0.677 SR-Precision (IDMT) followed by 0.506 F-Ex (MLKD) and 0.676 SR-Precision (CEALIST, LIRIS).

Table 4: Summary of the results for the evaluation per concept in the visual configuration. The table shows the best run per group.

Team Rank MiAP			Team Rank F-Ex			Team Rank SR-Prec.		
TUBFI	1	0.388	ISIS	1	0.612	ISIS	1	0.734
CAEN	4	0.382	CAEN	4	0.600	BPACAD	4	0.729
ISIS	6	0.375	HHI	6	0.588	CAEN	5	0.727
BPACAD	9	0.367	BPACAD	11	0.568	LIRIS	6	0.725
LIRIS	11	0.355	MLKD	13	0.560	HHI	8	0.718
NII	14	0.337	TUBFI	14	0.552	MRIM	13	0.703
MRIM	15	0.336	MRIM	18	0.544	NII	14	0.702
HHI	16	0.335	LIRIS	19	0.539	CEALIST	16	0.700
MLKD	25	0.311	BUFFALO	21	0.527	MLKD	17	0.698
CEALIST	26	0.301	DBIS	22	0.518	BUFFALO	23	0.683
BUFFALO	28	0.249	CEALIST	25	0.503	UNIKLU	25	0.672
DBIS	31	0.230	MELJI	31	0.472	DBIS	27	0.671
UNIKLU	38	0.207	UNIKLU	33	0.469	TUBFI	38	0.630
MELJI	42	0.204	LAPI	39	0.390	LAPI	42	0.551
LAPI	44	0.177	NII	41	0.298	MELJI	44	0.452

6.3 Results for the multi-modal configuration

Table 6 depicts the results for the best multi-modal configuration of each group. As already stated, the run of TUBFI achieves the best overall results in terms of MiAP. In the example-based evaluation, ISIS scores best overall with 0.622 F-Ex and 0.742 SR-Precision.

6.4 Comparison of achievements with different information sources

Last year only two runs considered the textual configuration. In contrast, this year eight textual runs were submitted by seven teams. This allows for a more reliable analysis of the performance of textual runs in image annotation. The

Table 5: Summary of the results for the evaluation per concept in the textual configuration. The table shows the best run per group.

Team Rank MiAP			Team Rank F-Ex			Team Rank SR-Prec.		
BPACAD	1	0.346	IDMT	1	0.525	IDMT	1	0.677
IDMT	2	0.326	MLKD	2	0.506	CEALIST	2	0.676
MLKD	3	0.326	BPACAD	3	0.502	LIRIS	3	0.676
LIRIS	4	0.321	CEALIST	4	0.479	MLKD	5	0.653
MELJI	6	0.304	MELJI	5	0.459	BPACAD	6	0.635
CEALIST	7	0.292	LIRIS	6	0.432	MELJI	7	0.491
REGIMVID	8	0.204	REGIMVID	8	0.141	REGIMVID	8	0.396

Table 6: Summary of the results for the evaluation per concept in the multi-modal configuration. The table shows the best run per group.

Team Rank MiAP			Team Rank F-Ex			Team Rank SR-Prec.		
TUBFI	1	0.443	ISIS	1	0.622	ISIS	1	0.742
LIRIS	2	0.437	BPACAD	2	0.586	BPACAD	2	0.719
BPACAD	3	0.436	LIRIS	4	0.576	LIRIS	3	0.718
ISIS	5	0.433	TUBFI	6	0.566	IDMT	5	0.713
MLKD	9	0.402	MLKD	7	0.559	CEALIST	7	0.711
CEALIST	10	0.384	MRIM	8	0.552	MRIM	11	0.706
MRIM	11	0.377	IDMT	9	0.552	MLKD	15	0.698
IDMT	13	0.371	CEALIST	17	0.508	MUFIN	19	0.628
MUFIN	20	0.299	MEIJI	19	0.495	TUBFI	22	0.559
MEIJI	24	0.288	MUFIN	21	0.462	MEIJI	24	0.480

performance of textual runs is close to the results that can be achieved in the visual configuration. The best visual run achieves a MiAP of 0.388 in contrast to the best textual run, which scores with a MiAP of 0.346. The difference of 4.2% is rather small, especially when considering that not for all images EXIF data and Flickr user tags exist. In the example-based evaluation, the difference between visual and textual runs is more significant. Visual runs score better by about 9% in terms of F-Ex and about 6% in terms of SR-Precision. Results in the multi-modal configuration outperform classification with single modality information in the visual configuration by 5.5% and the textual configuration by about 10% in terms of MiAP. For the example-based measures F-Ex and SR-Precision, differences are very small with 1% for the visual configuration. Comparing the multi-modal to the textual configuration, differences are significant and lie by about 10% and 6.5% for F-Ex and SR-Precision, respectively.

6.5 Annotation performance per concept

In Table 7, the results for each concept are summarized independent of the configuration. On average, the concepts could be detected with a MiAP of 0.48 considering the best run per concept out of all configurations. In general, 79 concepts were best detected with a multi-modal approach, 17 concepts were detected best with a visual approach, and 3 concepts were detected best by a textual approach. High performance is achieved for the concepts *Neutral-Illumination*, *No-Persons*, *No-Blur*, and *Outdoor*. Following, the concepts *Sky*, *Day*, *Clouds*, and *Plants* were annotated with high scores. The worst annotation quality was achieved for the concept *abstract* followed by the concepts *work*, *graffiti*, *technical*, *old-person*, and *boring*.

In the evaluation in 2010, a great difference in prediction quality among the concepts from 2009 and the ones newly introduced in 2010 of 0.57 MiAP and 0.37 MiAP could be seen. This difference is still present in this evaluation cycle. The concepts from 2009 (number 1-49) could be detected with a MiAP

Table 7: This table presents the best annotation performance per concept, achieved by any team in any configuration, in terms of iAP. It lists the concept name, the iAP score, the team, and the configuration of the run.

Concept	iAP	Team	Config.	Concept	iAP	Team	Config.
Partylife	0.437	ISIS	M	Street	0.390	TUBFI	V
Family_Friends	0.564	TUBFI	M	Church	0.312	TUBFI	M
Beach_Holidays	0.581	TUBFI	M	Bridge	0.239	ISIS	M
Building_Sights	0.619	TUBFI	M	Park_Garden	0.491	TUBFI	M
Snow	0.539	ISIS	M	Rain	0.198	ISIS	M
Citylife	0.598	BPACAD	M	Toy	0.396	BPACAD	M
Landscape_Nature	0.812	BPACAD	M	MusicalInstrument	0.250	LIRIS	T
Sports	0.245	LIRIS	M	Shadow	0.195	TUBFI	V
Desert	0.356	TUBFI	M	bodypart	0.321	BPACAD	M
Spring	0.259	ISIS	M	Travel	0.213	BPACAD	M
Summer	0.351	TUBFI	M	Work	0.135	MLKD	M
Autumn	0.455	LIRIS	M	Birthday	0.172	ISIS	M
Winter	0.541	BPACAD	M	Visual_Arts	0.390	TUBFI	M
Indoor	0.643	BPACAD	M	Graffiti	0.139	MUFIN	M
Outdoor	0.911	BPACAD	M	Painting	0.305	TUBFI	M
Plants	0.814	TUBFI	M	artificial	0.217	TUBFI	M
Flowers	0.642	TUBFI	M	natural	0.740	LIRIS	M
Trees	0.687	ISIS	V	technical	0.149	MLKD	V
Sky	0.892	TUBFI	M	abstract	0.111	MRIM	V
Clouds	0.847	TUBFI	M	boring	0.158	ISIS	M
Water	0.736	ISIS	M	cute	0.640	CAEN	V
Lake	0.366	BPACAD	M	dog	0.712	LIRIS	M
River	0.349	ISIS	M	cat	0.403	IDMT	M
Sea	0.581	TUBFI	M	bird	0.631	LIRIS	M
Mountains	0.592	ISIS	M	horse	0.577	MLKD	M
Day	0.884	BPACAD	M	fish	0.519	IDMT	T
Night	0.658	BPACAD	M	insect	0.581	LIRIS	M
Sunny	0.518	TUBFI	V	car	0.475	TUBFI	M
Sunset_Sunrise	0.802	TUBFI	V	bicycle	0.535	TUBFI	M
Still_Life	0.455	ISIS	M	ship	0.431	LIRIS	M
Macro	0.539	BPACAD	M	train	0.347	BPACAD	M
Portrait	0.687	TUBFI	M	airplane	0.694	MLKD	T
Overexposed	0.241	TUBFI	V	skateboard	0.558	LIRIS	M
Underexposed	0.345	CAEN	V	female	0.531	TUBFI	M
Neutral_Illumination	0.984	CEALIST	V	male	0.295	ISIS	M
Motion_Blur	0.308	BPACAD	M	Baby	0.453	MLKD	M
Out_of_focus	0.315	CEALIST	M	Child	0.358	TUBFI	M
Partly_Blurred	0.778	BPACAD	M	Teenager	0.283	CAEN	V
No_Blur	0.916	BPACAD	M	Adult	0.601	TUBFI	M
Single_Person	0.613	TUBFI	M	old_person	0.151	BPACAD	M
Small_Group	0.397	TUBFI	M	happy	0.453	BPACAD	M
Big_Group	0.460	ISIS	V	funny	0.439	TUBFI	M
No_Persons	0.928	BPACAD	M	euphoric	0.188	MLKD	M
Animals	0.738	BPACAD	M	active	0.385	TUBFI	M
Food	0.646	LIRIS	M	scary	0.230	BPACAD	M
Vehicle	0.593	LIRIS	M	unpleasant	0.295	TUBFI	M
Aesthetic_Impression	0.340	TUBFI	V	melancholic	0.357	TUBFI	V
Overall_Quality	0.290	LIRIS	M	inactive	0.558	TUBFI	V
Fancy	0.250	TUBFI	V	calm	0.589	TUBFI	M
Architecture	0.366	LIRIS	M				

of 0.57 considering the best prediction for each concept out of all runs, while the 2010 concepts (numbers 50-90) improved minimally to a MiAP of 0.38. The new sentiment concepts (numbers 91-99) can be detected with a MiAP of 0.39. Although these are arguably very subjective concepts, the detection algorithms are capable of identifying a strong trend of sentiments correctly. Especially, the sentiments *calm* and *inactive* could be detected very well, while the sentiments *scary* and *euphoric* were annotated worst. However, one has to note that the 2010 concepts occur on average in 7.9% of the training set images, while the 2009 and 2011 concepts are visible in 18% and 14% of the training set, respectively. Therefore, the algorithms had more example images to learn the sentiments in comparison to the more object-based concepts introduced in 2010.

7 Concept-based Retrieval Task: Results

In the following, the results of the concept-based retrieval task are presented and discussed. The participation of four teams was lower with respect to the annotation task. Despite, 31 runs have been submitted in all configurations, consisting of 10 multi-modal, 7 textual, and 14 visual runs. Approximately half of the systems used a completely automated processing.

Table 8 depicts all runs and indicates the configuration and the degree of automation of each run. Results are sorted in terms of MAP. The MAP value ranges from 0.1640 to 0.0013. Overall, the task was solved best with a MAP of 0.164 by the MLKD group, who used a multi-modal configuration with manual query formulations. It can clearly be seen that the approaches using manual processing achieve better results than the automated versions. The best performing automated run achieves a MAP of 0.0849 (MLKD). The multi-modal and textual configurations of MLKD work significantly better than their visual ones. MLKD provides also the best MAP value for a textual configuration, which is only 0.0094 points lower than the overall best run. The best working visual configuration, also with manual query formulation, was submitted by ISIS, achieving a MAP of 0.0997. MELJI provided solely automated runs, for which the multi-model approaches outperform the textual and visual runs. REGIMVID provided one automated, text-based configuration, which achieves a MAP value of 0.0042.

Table 9 presents the best run per team in the three configurations. If a team submitted an automated and a manual run in the same configuration, the results for both runs are illustrated. The direct comparison of automated and manual runs per team shows that the manual runs work best, independent from the configuration (textual, visual, multi-modal). This is mostly apparent in the big difference – nearly factor two – of the textual configuration of MLKD. The results of the two approaches submitted by ISIS support this interpretation. The manually processed run outperforms the automated approach with a MAP of 0.0997 vs. 0.0430. The performance difference between the two multi-modal configurations submitted by MLKD and MELJI can partially be explained by the different degree of automation. MELJI uses a fully automated system resulting

Table 8: The table shows the results for all runs sorted by MAP. Automated runs are abbreviated with “A”, while manual query formulations are abbreviated with “M” in the column Automation.

	Run	MAP	P@10	P@20	P@100	R-Prec	Automation	Configuration
	MLKD_1307692157334_Sub4_Multi1_1000.txt	0.1640	0.3900	0.3700	0.3180	0.2467	M	M
	MLKD_1307692102536_Sub3_Tags_1000.txt	0.1546	0.4100	0.3838	0.3102	0.2366	M	T
	MLKD_1307692201394_Sub5_Multi2_1000.txt	0.1533	0.4175	0.3725	0.2980	0.2332	M	M
	MLKD_1308226941036_Sub9_Multi1_250.txt	0.1346	0.3900	0.3700	0.3180	0.2397	M	M
	MLKD_130822686454_Sub8_Tags_250.txt	0.1328	0.4100	0.3838	0.3102	0.2298	M	T
	MLKD_1308227066531_Sub10_Multi2_250.txt	0.1312	0.4175	0.3725	0.2980	0.2260	M	M
	ISIS_1308685876873_runcbr-UvA-enhancedplus100fast.txt	0.0997	0.3125	0.3050	0.2428	0.1712	M	V
	ISIS_1308685680597_runcbr-UvA-enhancedplus100.txt	0.0940	0.2950	0.2800	0.2335	0.1664	M	V
	ISIS_1308685828076_runcbr-UvA-enhancedfast.txt	0.0916	0.2700	0.2875	0.2277	0.1635	M	V
	ISIS_1308685497624_runcbr-UvA-multiconceptmul.txt	0.0888	0.2600	0.2738	0.2213	0.1597	M	V
	ISIS_1308685631678_runcbr-UvA-enhanced.txt	0.0867	0.2550	0.2725	0.2227	0.1597	M	V
	MLKD_1307692048001_Sub2_Tags_QueryByExample_1000.txt	0.0849	0.3000	0.2800	0.2188	0.1530	A	T
	ISIS_1308685542216_runcbr-UvA-multiconceptsum.txt	0.0803	0.2425	0.2600	0.2015	0.1500	M	V
	MLKD_1308226793805_Sub7_Tags_QueryByExample_250.txt	0.0708	0.3000	0.2800	0.2188	0.1486	A	T
	ISIS_1308685443537_runcbr-UvA-conceptmap.txt	0.0534	0.1775	0.1938	0.1675	0.1159	M	V
	MEJL_1308444633885_mejlVCTh.txt	0.0444	0.1625	0.1650	0.1465	0.1053	A	M
	MLKD_1308226793805_Sub7_Tags_QueryByExample_250.txt	0.0430	0.1675	0.1550	0.1270	0.0974	A	V
	MEJL_1308496071960_mejlVTTh.txt	0.0420	0.1725	0.1437	0.1417	0.1061	A	M
	MEJL_1308495839773_mejlVTh.txt	0.0408	0.1750	0.1513	0.1432	0.1053	A	M
	ISIS_1308685310238_runcbr-UvA-auto33.txt	0.0366	0.1400	0.1350	0.1020	0.0790	A	V
	MLKD_1307691968544_Sub1_Visual_1000.txt	0.0361	0.1525	0.1375	0.1080	0.0883	M	V
	MEJL_1308444822371_mejlVCTh.txt	0.0333	0.1300	0.1200	0.1130	0.0824	A	M
	MEJL_1308496150191_mejlVTTh.txt	0.0327	0.1275	0.1138	0.1135	0.0847	A	M
	MEJL_1308495988189_mejlVTTh.txt	0.0325	0.1425	0.1250	0.1140	0.0867	A	M
	MLKD_1308226593230_Sub6_Visual_250.txt	0.0295	0.1525	0.1375	0.1080	0.0863	M	V
	MEJL_1308444441206_mejlTr.txt	0.0227	0.0900	0.0962	0.0865	0.0628	A	T
	MEJL_1308443537669_mejlTh.txt	0.0213	0.0675	0.0862	0.0865	0.0648	A	T
	ISIS_1308685265788_runcbr-UvA-auto10.txt	0.0181	0.0875	0.0688	0.0530	0.0448	A	A
	REGINVID_1308548603241_REGINVIDTJFRT.txt	0.0042	0.0650	0.0550	0.0352	0.0200	A	T
	MEJL_1307986549452_mejlVh.txt	0.0017	0.0150	0.0150	0.0197	0.0151	A	V
	MEJL_1308443999150_mejlVr.txt	0.0013	0.0125	0.0125	0.0185	0.0122	A	V

Table 9: This table presents the best results per team in the single configurations differentiated by the degree of automation. In the case a team submitted an automated and a manual run in a particular configuration, both runs are listed. Results are sorted in terms of MAP.

Team	MAP	P@10	P@20	P@100	R-Prec	Automation
Visual						
ISIS	0.0997	0.3125	0.3050	0.2428	0.1712	M
ISIS	0.0430	0.1675	0.1550	0.1270	0.0974	A
MLKD	0.0361	0.1525	0.1375	0.1080	0.0883	M
MELJI	0.0017	0.0150	0.0150	0.0197	0.0151	A
Textual						
MLKD	0.1546	0.4100	0.3838	0.3102	0.2366	M
MLKD	0.0849	0.3000	0.2800	0.2188	0.1530	A
MELJI	0.0227	0.0900	0.0962	0.0865	0.0628	A
REGIMVID	0.0042	0.0650	0.0550	0.0352	0.0200	A
Multi-modal						
MLKD	0.1640	0.3900	0.3700	0.3180	0.2467	M
MELJI	0.0444	0.1625	0.1650	0.1465	0.1053	A

in a MAP value of 0.0444 and MLKD uses a manually query formulation which achieves a MAP value of 0.1640 (best run overall).

The boxplots of the MAP scores of all approaches per topic in Figure 5 allow for a more detailed examination. Most topics show a wide variation among the obtained MAP values, which indicates the differences in performance per topic. This can be seen, e.g., for the topics 33 (*cars and motion blur*) and 28 (*fireworks*). For topic 33, the lowest MAP scores are 0. The highest value (0.5218) for this topic is achieved by ISIS with an automated, visual configuration. MAP scores for topic 28 are in the range of 0 to 0.423. A closer look reveals that only for two topics (33 and 5) MAP values over 0.5 are reached. This is most surprising in the case of topic 5 (*riders on horse*), due to the consistent low MAP values of the other approaches. The six runs performing significantly better than the rest are provided by MLKD and use textual and multi-modal information. The figure also shows that the MAP scores for topic 18 (on the far right) are homogeneously low. It can be concluded that all approaches had great difficulties identifying relevant images for *female old person*. The topics 13, 21 and 26 were also hard to identify (MAP close to 0) for most of the approaches, but better performing outliers clearly exist. For topic 13 (*female person(s) doing sports*) some approaches of MLKD reach values above 0.14. The same observation, with values close to 0.1, applies to topic 21 (*scary dog(s)*). The most striking effect can be observed for *houses in mountains* (26): Only two ISIS approaches were able to achieve significantly higher MAP values than 0. The best sentiment topic

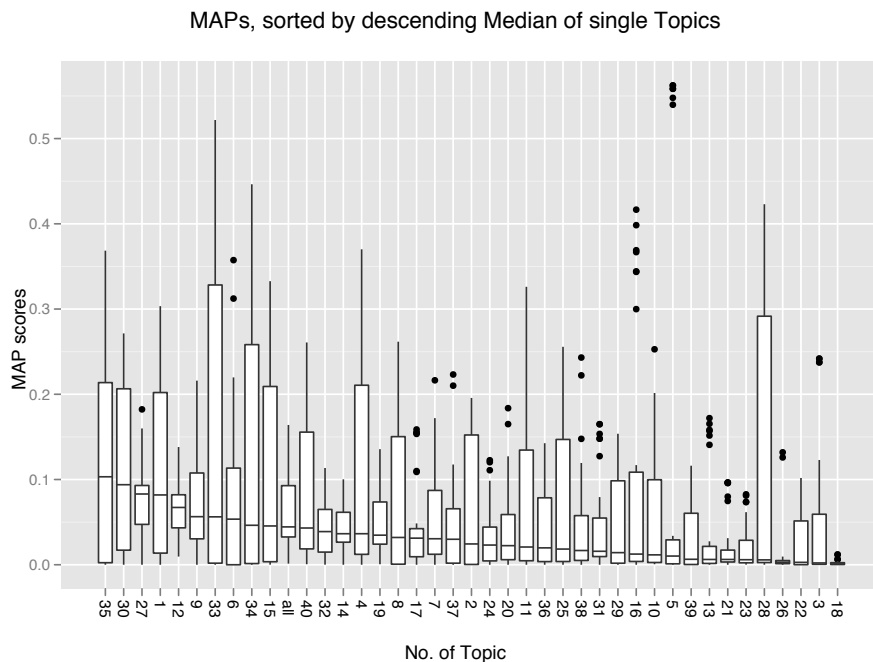


Fig. 5: Comparison of MAP scores per topic

in terms of MAP is topic 34 (*unpleasant insects*). Relatively high scores above 0.4 are achieved by several approaches.

8 Conclusions

The ImageCLEF 2011 Photo Annotation and Concept-based Retrieval Tasks posed two image analysis challenges that could be solved with three general configurations: textual-based analysis, visual-based analysis, and multi-modal analysis. The aim of the annotation task was to automatically annotate images with 99 concepts in a multi-label scenario. The task attracted a considerable number of international teams with a final participation of 18 teams that submitted a total of 79 runs. The results show that the annotation task could be solved reasonably well, with the best multi-modal run achieving a MiAP of 0.443 in the multi-modal configuration, a MiAP of 0.388 in the visual configuration, and a MiAP of 0.346 in the textual configuration. For the evaluation per example, the best multi-modal run achieves 0.62 F-Ex, the best visual run scores with 0.61 F-Ex, and the best textual run with 0.53 F-Ex. All in all, the multi-modal approaches got the best scores for 79 out of 99 concepts, followed by 17 concepts that could be detected best with the visual approach and 3 that won with a

textual approach. In general, the multi-modal approaches outperformed visual and textual configurations for nearly all performance measures of the teams that submitted results for more than one configuration.

The concept-based retrieval task asked participants to retrieve the most relevant images given certain topics. The topics were constructed based on user needs and query logs, and consist of a Boolean connection of several visual concepts. In total, 4 teams participated in this novel challenge and submitted 31 runs. 10 runs belong to the multi-modal configuration, 14 runs were submitted in the visual configuration, and 7 runs are based on textual information. The best multi-model configuration obtained a MAP value of 0.164, the textual configuration scored best with 0.1546 MAP, and the best visual run achieves a score of 0.0997 MAP. The task was solved by 16 completely automated approaches and 14 runs which include manual intervention. It was observed that most manually processed runs work best, independent from the configuration (textual, visual, or multi-modal). They achieve MAP values in the range of 0.164 and 0.0295, whereas the automated solutions range between scores of 0.0849 and 0.0013 MAP. A closer examination showed that all approaches had great difficulties to identify relevant images for the topic *female old person*. Also the topics 5, 13, 21, and 26 were hard to identify as well, but here some approaches were able to reach higher MAP values. Especially, the topic *riders on horse* shows very strong outliers with high MAP values above 0.5. The obtained MAP scores of the remaining topics vary widely which points to a large variation in the difficulty level of topics. Some configurations are able to achieve MAP values higher than 0.5 for individual topics. Considering the topic *female old person*, all runs show nearly the same low performance. This seems to be an extremely critical topic for concept-based image retrieval.

Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. This work was partly supported by grant 01MQ07017 of the German research program THESEUS funded by the Ministry of Economics.

References

1. Nowak, S., Dunker, P.: Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In Peters, C., Tsikrika, T., Müller, H., Kalpathy-Cramer, J., Jones, J., Gonzalo, J., Caputo, B., eds.: Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), Revised Selected Papers. Lecture Notes in Computer Science, Corfu, Greece (2010)
2. Nowak, S., Huiskes, M.: New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. Working notes of CLEF **2010** (2010)
3. Mark J. Huiskes, B.T., Lew, M.S.: New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In: MIR '10: Proceedings of the 2010

ACM International Conference on Multimedia Information Retrieval, New York, NY, USA, ACM (2010) 527–536

4. Nowak, S., Dunker, P.: A Consumer Photo Tagging Ontology: Concepts and Annotations. In: THESEUS/ImageCLEF Pre-Workshop 2009, Co-located with the Cross-Language Evaluation Forum (CLEF) Workshop and 13th European Conference on Digital Libraries ECDL, Corfu, Greece, 2009. (2009)
5. Jiang, Y., Ngo, C., Chang, S.: Semantic context transfer across heterogeneous sources for domain adaptive video search. In: Proceedings of the seventeen ACM international conference on Multimedia, ACM (2009) 155–164
6. Russell, J.: A circumplex model of affect. *Journal of personality and social psychology* **39**(6) (1980) 1161
7. Tsikrika, T., Kludas, J.: Overview of the wikipediamm task at imageclef 2009. In: Working notes of CLEF. (2009)
8. Tsikrika, T., Popescu, A., Kludas, J.: Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
9. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In: Working notes of CLEF. (2010)
10. Daróczy, B., Pethes, R., Benczúr, A.A.: SZTAKI @ ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
11. Su, Y., Jurie, F.: Semantic Contexts and Fisher Vectors for the ImageCLEF 2011 Photo Annotation Task. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
12. Znaidia, A., Le Borgne, H.: CEA LISTs participation to Visual Concept Detection Task of ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
13. Mbanya, E., Gerke, S., Hentschel, C., Ndjiki-Nya, P.: Sample Selection, Category Specific Features and Reasoning. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
14. Nagel, K., Nowak, S., Kühhirt, U., Wolter, K.: The Fraunhofer IDMT at ImageCLEF 2011 Photo Annotation Task. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
15. Van De Sande, K.E., Snoek, C.G.: The University of Amsterdam’s Concept Detection System at ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
16. Rasche, C., Vertan, C.: Testing a Method for Statistical Image Classification in Image Retrieval. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
17. Liu, N., Zhang, Y., Dellandréa, E., Bres, S., Chen, L.: LIRIS-Imagine at ImageCLEF 2011 Photo Annotation task. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
18. Izawa, R., Motohashi, N., Takagi, T.: Annotation and Retrieval System Using Confabulation Model for ImageCLEF2011 Photo Annotation. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
19. Spyromitros-Xioufis, E., Sechidis, K., Tsoumakas, G., Vlahavas, I.: MLKD’s Participation at the CLEF 2011 Photo Annotation and Concept-Based Retrieval Tasks. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
20. Albatal, R., Safadi, B., Quénot, G., Mulhem, P.: LIG-MRIM at Image Photo Annotation task in ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)

21. Budikova, P., Batko, M., Zezula, P.: MUFIN at ImageCLEF 2011: Success or Failure? In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
22. Le, D.D., Satoh, S.: NII, Japan at ImageCLEF 2011 Photo Annotation Task. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
23. Amel, K., Benammar, A., Amar, C.B.: REGIMvid at ImageCLEF2011: Integrating Contextual Information to Enhance Photo Annotation and Concept-based Retrieval. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)
24. Binder, A., Samek, W., Kawanabe, M.: The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands. (2011)