New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010

Stefanie Nowak[1] and Mark Huiskes[2]

[1] Fraunhofer IDMT, Ilmenau, Germany
[2] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
stefanie.nowak@idmt.fraunhofer.de, mark.huiskes@liacs.nl

Abstract. The ImageCLEF 2010 Photo Annotation Task poses the challenge of automated annotation of 93 visual concepts in Flickr photos. The participants were provided with a training set of 8,000 Flickr images including annotations, EXIF data and Flickr user tags. Testing was performed on 10,000 Flickr images, differentiated between approaches considering solely visual information, approaches relying on textual information and multi-modal approaches. Half of the ground truth was acquired with a crowdsourcing approach. The evaluation followed two evaluation paradigms: per concept and per example. In total, 17 research teams participated in the multi-label classification challenge with 63 submissions. Summarizing the results, the task could be solved with a MAP of 0.455 in the multi-modal configuration, with a MAP of 0.407 in the visual-only configuration and with a MAP of 0.234 in the textual configuration. For the evaluation per example, 0.66 F-ex and 0.66 OS-FCS could be achieved for the multi-modal configuration, 0.68 F-ex and 0.65 OS-FCS for the visual configuration and 0.26 F-ex and 0.37 OS-FCS for the textual configuration.

1 Introduction

The steadily increasing amount of multimedia data poses challenging questions on how to index, visualize, organize, navigate or structure multimedia information. Many different approaches are proposed in the research community, but often their benefit is not clear as they were evaluated on different datasets with different evaluation measures. Evaluation campaigns aim to establish an objective comparison between the performance of different approaches by posing well-defined tasks including datasets, topics and measures. This paper presents an overview of the ImageCLEF 2010 Photo Annotation Task. The task aims at the automated detection of visual concepts in consumer photos. Section 2 introduces the task and describes the database, the annotation process, the ontology and the evaluation measures applied. Section 3 summarizes the approaches of the participants to solve the task. Next, the results for all configurations are presented and discussed in Section 4 and Section 5, respectively. Finally, Section 6 summarizes and concludes the paper.

2 Task Description

The ImageCLEF Visual Concept Detection and Annotation Task poses a multi-label classification challenge. It aims at the automatic annotation of a large number of consumer photos with multiple annotations. The task can be solved by following three different approaches:

1. Automatic annotation with content-based visual information of the images.
2. Automatic annotation with Flickr user tags and EXIF metadata in a purely text-based scenario.
3. Multi-modal approaches that consider both visual and textual information like Flickr user tags or EXIF information.

In all cases the participants of the task were asked to annotate the photos of the test set with a predefined set of keywords (the concepts), allowing for an automated evaluation and comparison of the different approaches. Concepts are for example abstract categories such as *Family&Friends* or *Partylife*, the Time of Day (*Day*, *Night*, *sunny*,), Persons (*no person*, *single person*, *small group* or *big group*), Quality (*blurred*, *underexposed*) and Aesthetics; 52 from the 53 concepts that were used in the ImageCLEF 2009 benchmark are used again [1]. In total the number of concepts was extended to 93 concepts. In contrast to the annotations from 2009, the new annotations were obtained with a crowdsourcing approach that utilizes Amazon Mechanical Turk. The task uses a subset of the MIR Flickr 25,000 image dataset [2] for the annotation challenge. The MIR Flickr collection supplies all original tag data provided by the Flickr users (noted as Flickr user tags). In the collection there are 1386 tags which occur in at least 20 images, with an average total number of 8.94 tags per image. These Flickr user tags are made available for the textual and multi-modal approaches. For most of the photos the EXIF data is included and may be used.

2.1 Evaluation Objectives

This year the focus of the task lies on the comparison of the strengths and limitations of the different approaches:

- Do multi-modal approaches outperform text only or visual only approaches?
- Which approaches are best for which kind of concepts?
- Can image classifiers scale to the large number of concepts and data?

Furthermore, the task challenges the participants to deal with an unbalanced number of annotations per photo, an unbalanced number of photos per concept, the subjectivity of concepts like *boring*, *cute* or *fancy* and the diversity of photos belonging to the same concept. Further, the textual runs have to cope with a small number of images without EXIF data and/or Flickr user tags.

2.2 Annotation Process

The complete dataset consists of 18,000 images annotated with 93 visual concepts. The manual annotations for 52 concepts were acquired by Fraunhofer IDMT in 2009. (The concept *Canvas* from 2009 was discarded.) Details on the manual annotation process and concepts, including statistics on concept frequencies can be found in [3, 1]. In 2010, 41 new concepts were annotated with a crowdsourcing approach using the Amazon Mechanical Turk. In the following, we just focus on the annotation process of these new concepts.

Amazon Mechanical Turk (MTurk, `www.mturk.com`) is an online marketplace in which mini-jobs can be distributed to a crowd of people. At MTurk these mini-jobs are called HITs (Human Intelligence Tasks). They represent a small piece of work with an allocated price and completion time. 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 all results from MTurk after they are completed. The workflow of a requester can be described as follows: 1) design a HIT template, 2) distribute the work and fetch results and 3) approve or reject work from turkers. For the design of the HITs, MTurk offers support by providing a web interface, command line tools and developer APIs. The requester can define how many assignments per HIT are needed, how much time is allotted to each HIT and how much to pay per HIT. MTurk offers several ways of assuring quality. Optionally, the turkers can be asked to pass a qualification test before working on HITs, multiple workers can be assigned the same HIT and requesters can reject work in case the HITs were not finished correctly. The HIT approval rate each turker achieves by completing HITs can be used as a threshold for authorization to work. Before the annotations of the ImageCLEF 2010 tasks were acquired, we performed a pre-study to investigate if annotations from non-experts are reliable enough to be used in an evaluation benchmark. The results were very promising and encouraged us to adapt this service for the 2010 task. Details of the pre-study can be found in [4].

Design of HIT templates: In total, we generated four different HIT templates at MTurk. For all concepts, the annotations per photo were obtained three times. Later the final annotations are built from the majority vote of these three opinions. For the annotation of the 41 new concepts we made use of the pre-knowledge that we have from the old annotations. Therefore the 41 concepts were structured into four groups:

1. Vehicles
   The ImageCLEF 2009 dataset contains a number of photos annotated with the concept *Vehicle*. These photos were further annotated with the concepts *car*, *bicycle*, *ship*, *train*, *airplane* and *skateboard*. A textbox offered the possibility to input further categories. The turkers could select a checkbox saying that no vehicle is depicted in the photo to cope with the case of false annotations. The corresponding survey with guidelines can be found in Figure 1. Each HIT was rewarded with 0.01$.

Annotate this image

Guidelines:

- Please choose all applicable concepts for an image, but at least one.
- In case no definition is matching, use the text area to classify the vehicle.
- In case there are several vehicles shown, please answer for all of them.
- Please note that vehicles in paintings are also considered to be vehicles and should be classified.

Image:

Vehicles

1. Which vehicle(s) can you see at the photo (select all matching answers)

☐ car ☐ bicycle ☐ ship / boat ☐ train ☐ airplane ☐ skateboard

Another vehicle? ☐ There is no vehicle shown.

Fig. 1: MTurk HIT template for the annotation of specific vehicle concepts.

2. Animals
   The ImageCLEF 2009 photo collection already contains several photos that were annotated with the concept *animals*. The turkers at Amazon were asked to further classify these photos in the categories *dog*, *cat*, *bird*, *horse*, *fish* and *insect*. Again, a textbox offered additional input possibilities. For each HIT a reward of 0.01$ was paid.
3. Persons
   The dataset contains photos that were annotated with a person concept (*single person*, *small group* or *big group of persons*). These photos were further classified with human attributes like *female*, *male*, *Baby*, *Child*, *Teenager*, *Adult* and *old_person*. Each HIT was rewarded with 0.01$.

4. General annotations
   For the following 22 concepts, no prior information could be used. Therefore the concepts were annotated in all 18,000 photos. The HIT was designed as a survey with 6 questions aiming to annotate the categories "content elements" (*Architecture*, *Street*, *Church*, *Bridge*, *Park_Garden*, *Rain*, *Toy*, *Musical_Instrument* and *Shadow*), "persons" (*bodypart*), "events" (*Travel*, *Work*, *Birthday*), "image representation" (*Visual_Arts*, *Graffiti*, *Painting*), "impression" (*artificial*, *natural*, *technical*, *abstract*) and "feelings" (*boring*, *cute*). Each HIT was rewarded with 0.03$.

2.3 Ontology

The concepts were organised in an ontology. For this purpose the Consumer Photo Tagging Ontology [3] of 2009 was extended with the new concepts. The hierarchy allows making assumptions about the assignment of concepts to documents. For instance, if a photo is classified to contain *trees*, it also contains *plants*. Then, next to the is-a relationship of the hierarchical organization of concepts, also other relationships between concepts can determine label assignments. The ontology requires for example that for a certain sub-node only one concept can be assigned at a time (disjoint items) or that a special concept (e.g. *portrait*) postulates other concepts like *persons* or *animals*. The ontology allows the participants to incorporate knowledge in their classification algorithms, and to make assumptions about which concepts are probable in combination with certain labels. Further, it is used in the evaluation of the submissions.

2.4 Evaluation Measures

The evaluation follows the concept-based and example-based evaluation paradigms. For the concept-based evaluation the Average Precision (AP) is utilized. This measure showed better characteristics than the Equal Error Rate (EER) and Area under Curve (AUC) in a recent study [5]. For the example-based evaluation we apply the example-based F-Measure (F-ex). The Ontology Score of last year was extended with a different cost map that is based on Flickr metadata [6] and serves as additional evaluation measure. It is called Ontology Score with Flickr Context Similarity (OS-FCS) in the following.

2.5 Submission

The participants submitted their results for all photos in a single text file that contains the photo ID as first entry per row followed by 93 floating point values between 0 and 1 (one value per concept). The floating point values are regarded as confidence while computing the AP. After the confidence values for all photos, the text file contains binary values for each photo (so again each line contains the photo ID followed by 93 binary values). The measures F-ex and OS-FCS need a binary decision about the presence or absence of the concepts. Instead of

applying a strict threshold at 0.5 of the confidence values, the participants have the possibility to threshold each concept for each image individually. All groups had to submit a short description of their runs and state which configuration they chose (annotation with visual information only, annotation with textual information only or annotation with multi-modal information). In the following the visual configuration is abbreviated with "**V**", the textual with "**T**" and the multi-modal one with "**M**".

3 Participation

In total 54 groups registered for the visual concept detection and annotation task, 41 groups signed the license agreement and were provided with the training and test sets, 17 of them submitted results in altogether 63 runs. The number of runs was restricted to a maximum of 5 runs per group. There were 45 runs submitted in the visual only configuration, 2 in the textual only configuration and 16 in the multi-modal configuration.

BPACAD|SZTAKI [7]: The team of the Computer and Automation Research Institute of the Hungarian Academy of Science submitted one run in the visual configuration. Their approach is based on Histogram of Oriented Gradients descriptors which were clustered with a 128 dimensional Gaussian Mixture Model. Classification was performed with a linear logistic regression model with a χ^2 kernel per category.

CEA-LIST: The team from CEA-LIST, France submitted one run in the visual configuration. They extract various global (colour, texture) and local (SURF) features. The visual concepts are learned with a fast shared boosting approach and normalized with a logistic function.

CNRS|Telecom ParisTech [8]: The CNRS group of Telecom ParisTech, Paris, France participated with five multi-modal runs. Their approach is based on SIFT features represented by multi-level spatial pyramid bag-of-words. For classification a one-vs-all trained SVM is utilized.

DCU [9]: The team of Dublin City University, Ireland submitted one run in the textual configuration. They followed a document expansion approach based on the Okapi feedback method to expand the image metadata and concepts and applied DBpedia as external information source in this step. To deal with images without any metadata, the relationships between concepts in the training set is investigated. The date and time information of the EXIF metadata was extracted to predict concepts like *Day*.

HHI [10]: The team of Fraunhofer HHI, Berlin, Germany submitted five runs in the visual-only configuration. Their approach is based on the bag of words approach and introduces category specific features and classifiers including quality related features. They use opponent SIFT features with dense sampling and a sharpness feature and base their classification on a multi-kernel SVM classifier with χ^2 distance. Second, they incorporate a post-processing approach that considers relations and exclusions between concepts. Both extensions resulted in an increase in performance compared to the standard bag-of-words approach.

IJS [11]: The team of Jožef Stefan Institute, Slovenia and Department of Computer Science, Macedonia submitted four runs in the visual configuration. They use various global and local image features (GIST, colour histograms, SIFT) and learn predictive clustering trees classifiers. For each descriptor a separate classifier is learned and the probabilities output of all classifiers is combined for the final prediction. Further, they investigate ensembles of predictive clustering tree classifiers. The combination of global and local features leads to better results than using local features alone.

INSUNHIT [12]: The group of the Harbin Institute of Technology, China participated with five runs in the visual configuration. They use dense SIFT features as image descriptors and classify with a naïve-bayes nearest neighbour approach. The classifier is extended with a random sampling image to class distance to cope with imbalanced classes.

ISIS [13]: The Intelligent Systems Lab of the University of Amsterdam, The Netherlands submitted five runs in the visual configuration. They use a dense sampling strategy that combines a spatial pyramid approach and saliency points detection, extract different SIFT features, perform a codebook transformation and classify with a SVM approach. The focus lies on the improvement of the scores in the evaluation per image. They use the distance to the decision plane in the SVM as probability and determine the threshold for binary annotation from this distance.

LEAR and XRCE [14]: The team of LEAR and XEROX, France made a joint contribution with a total of ten runs, five submitted in the visual and five in the multi-modal configuration. They use SIFT and colour features on several spatial scales and represent them as improved Fisher vectors in a codebook of 256 words. The textual information is represented as a binary presence/absence vector of the most common 698 Flickr user tags. For classification a linear SVM is compared to a k-NN classifier with learned neighbourhood weights. Both classification models are computed with the same visual and textual features and late and early fusion approaches are investigated. All runs considering multi-modal information outperformed the runs in the visual configuration.

LIG [15]: The team of Grenoble University, France submitted one run in the visual configuration of the Photo Annotation task. They extract colour SIFT features and cluster them with a k-means clustering procedure in 4000 clusters. For classification a SVM with RBF kernel is learned in an one-against-all approach and based on the 4000 dimensional histogram of word occurrences.

LSIS [16]: The Laboratory of Information Science and Systems, France submitted two runs in the visual configuration. They propose features based on extended local binary patterns extracted with spatial pyramids. For classification they use a linear max-margin SVM classifier.

MEIJI [17]: The group of Meiji University, Kanagawa, Japan submitted in total five runs. They followed a conceptual fuzzy set approach applied to visual words, a visual words baseline with SIFT descriptors and a combination with a Flickr User Tag system using TF-IDF. Classification is based on a matching of visual word combinations between the training casebase and the test image.

For the visual word approach the cosine distance is applied for similarity determination. In total, two runs were submitted in the visual configuration and three in the multi-modal one. Their multi-modal runs outperform the visual configurations.

MLKD: The team of the Aristotle University of Thessaloniki, Greece participated with three runs; one in each configuration. For the visual and the textual runs ensemble classifier chains are used as classifiers. The visual configuration applies C-SIFT features with a Harris-Laplace salient point detector and clusters them in a 4000 word codebook. As textual features, the 250 most frequent Flickr user tags of the collection are represented in a binary feature vector per image. The multi-modal configuration chooses the confidence score of the model (textual or visual) per concept for which a better AP was determined in the evaluation phase. As a result, the multi-modal approach outperforms the visual and the textual models.

Romania [18]: The team of the University Bucharest, Romania participated with five runs in the visual configuration. Their approach considers the extraction of colour histograms and combine them with a method of structural description. The classification is performed using a Linear Discriminant Analysis (LDA) and a weighted average retrieval rank (ARR) method. The annotations resulting from the LDA classifier were refined considering the joint probabilities of concepts. As a result the ARR classification outperforms the LDA classification.

UPMC/LIP6 [19]: The team of University Pierre et Marie Curie, Paris, France participated in the visual and the multi-modal configuration. They submitted a total of five runs (3V, 2M). Their approach investigates the fusion of results from different classifiers with supervised and semi-supervised classification methods. The first model is based on fusing outputs from several RankingSVM classifiers that classified the images based on visual features (SIFT, HSV, Mixed+PCA). The second model further incorporates unlabeled data from the test set for which the initial classifiers are confident to assign a certain label and retrains the classifiers based on the augmented set. Both models were tested with the additional inclusion of Flickr user tags using the Porter stemming algorithm. For both models the inclusion of user tags improved the results.

WROCLAW [20]: The group of Wroclaw University, Poland submitted five runs in the visual configuration. They focus on global colour and texture features and adapt an approach which annotates photos through the search for similar images and the propagation of their tags. In their configurations several similarity measures (Minkowski distance, Cosine distance, Manhattan distance, Correlation distance and Jensen-Shannon divergence) are investigated. Further, an approach based on a Penalized Discriminant Analysis classifier was applied.

4 Results

This section presents the results of the Photo Annotation Task 2010. First, the overall results of all teams independent of the configuration are presented. In the following subsections the results per configuration are highlighted.

Table 1: Summary of the results for the evaluation per concept. The table shows the MAP for the best run per group and the averaged MAP for all runs of one group and indicates the configuration of the run.

		BEST RUN			AVERAGE RUNS		
TEAM	**RUNS**	**RANK**	**MAP**	**Conf.**	**RANK**	**MAP**	**Conf.**
XRCE	5	1	0.455	M	7.2	0.408	M+V
LEAR	5	3	0.437	M	7.8	0.392	M+V
ISIS	5	5	0.407	V	7.0	0.401	V
HHI	5	16	0.350	V	18.4	0.350	V
IJS	4	20	0.334	V	22.5	0.326	V
MEIJI	5	23	0.326	M	36.0	0.269	M+V
CNRS	5	28	0.296	M	30.0	0.293	M
BPACAD	1	33	0.283	V	33.0	0.283	V
Romania	5	34	0.259	V	43.8	0.221	V
INSUNHIT	5	36	0.237	V	41.0	0.230	V
MLKD	3	37	0.235	M	45.0	0.215	all
LSIS	2	38	0.234	V	38.5	0.234	V
DCU	1	44	0.228	T	44.0	0.228	T
LIG	1	46	0.225	V	46.0	0.225	V
WROCLAW	5	50	0.189	V	53.4	0.183	V
UPMC	5	54	0.182	M	59.0	0.160	M+V
CEA-LIST	1	61	0.147	V	61.0	0.147	V

Table 2: Summary of the results for the evaluation per example. The table shows the F-ex and the OS-FCS and the configuration used for the best run per group sorted by F-ex.

TEAM	**RANK**	**F-ex**	**Conf.**	**RANK**	**OS-FCS**	**Conf.**
ISIS	1	0.680	V	10	0.601	V
XRCE	5	0.655	M	1	0.657	M
HHI	8	0.634	V	3	0.640	V
LEAR	15	0.602	M	32	0.411	M
IJS	18	0.596	V	12	0.595	V
MEIJI	23	0.572	M	30	0.428	M
Romania	29	0.531	V	17	0.562	V
LSIS	30	0.530	M	21	0.536	V
WROCLAW	34	0.482	V	41	0.379	V
LIG	35	0.477	V	22	0.530	V
CEALIST	37	0.451	V	28	0.458	V
BPACAD	38	0.428	V	29	0.439	V
CNRS	43	0.351	M	31	0.421	M
MLKD	49	0.260	T	42	0.379	M
INSUNHIT	53	0.209	V	43	0.372	V
UPMC	55	0.186	M	55	0.351	M
DCU	60	0.178	T	60	0.304	T

In Table 1 the results for the evaluation per concept independent of the applied configuration are illustrated for the best run of each group. The results for all runs can be found at the Photo Annotation Task website[1]. The task could be solved best with a MAP of 0.455 (XRCE) followed by a MAP of 0.437 (LEAR). Both runs make use of multi-modal information. Table 2 illustrates the overall ranking for the results of the evaluation per example. The table is sorted descending for the F-ex measure. The best results were achieved in a visual configuration with 0.68 F-ex (ISIS) and in a multi-modal configuration with 0.66 OS-FCS (XRCE).

4.1 Results for the visual configuration

Table 3 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 were achieved by the ISIS team in terms of MAP and F-ex and the XRCE team in terms of OS-FCS. Both teams get close results in the concept-based evaluation (1.7% difference) while there is a bigger gap in the example-based evaluation (4.1% and 4.4%).

4.2 Results for the textual configuration

The results for the two textual runs are presented in Table 4. Both groups achieve close results in the concept-based evaluation. However, the example-based evaluation measures show a significant difference between the results of both teams.

4.3 Results for the multi-modal configuration

Table 5 depicts the results for the best multi-modal configuration of each group. As already stated the run of XRCE achieves the best overall results in terms of MAP and OS-FCS. In terms of OS-FCS, the results of XRCE in the multi-modal configuration are around 23% better than the second best configuration of the MEIJI team.

5 Discussion

The following section discusses some of the results in more detail. The best results for each concept are summarized in Table 6. On average the concepts could be detected with a MAP of 0.48 considering the best results per concept from all configurations and submissions. From 93 concepts, 61 could be annotated best with a multi-modal approach, 30 with a visual approach and two with a textual one. Most of the concepts were classified best by one configuration of the XRCE, ISIS or LEAR group.

[1] `http://www.imageclef.org/2010/PhotoAnnotation`

Table 3: Summary of the results for the evaluation per concept in the visual configuration. The table shows the MAP, F-ex and OS-FCS for the best run per group sorted by MAP.

TEAM	RANK	MAP	RANK F-ex		RANK OS-FCS	
ISIS	1	0.407	1	0.680	8	0.601
XRCE	6	0.390	6	0.639	1	0.645
LEAR	9	0.364	15	0.582	28	0.387
HHI	11	0.350	7	0.634	2	0.640
IJS	15	0.334	14	0.596	10	0.595
BPACAD	20	0.283	30	0.428	27	0.439
Romania	21	0.259	22	0.531	15	0.562
INSUNHIT	23	0.237	38	0.209	31	0.372
LSIS	24	0.234	23	0.530	19	0.536
LIG	30	0.225	27	0.477	20	0.53
MEIJI	31	0.222	18	0.559	34	0.363
WROCLAW	34	0.189	26	0.482	30	0.379
MLKD	40	0.177	37	0.224	37	0.359
UPMC	42	0.148	43	0.174	40	0.348
CEALIST	43	0.147	29	0.451	26	0.458

Table 4: Summary of the results for the evaluation per concept in the textual configuration. The table shows the MAP, F-ex and OS-FCS for the best run per group sorted by MAP.

TEAM	RANK	MAP	RANK F-ex		RANK OS-FCS	
MLKD	1	0.234	1	0.260	1	0.368
DCU	2	0.228	2	0.178	2	0.304

Table 5: Summary of the results for the evaluation per concept in the multi-modal configuration. The table shows the MAP, F-ex and OS-FCS for the best run per group sorted by MAP.

TEAM	RANK	MAP	RANK	F-ex	RANK OS-FCS	
XRCE	1	0.455	1	0.655	1	0.657
LEAR	3	0.437	3	0.602	5	0.411
MEIJI	6	0.326	6	0.573	3	0.428
CNRS	9	0.296	9	0,.351	4	0.421
MLKD	14	0.235	13	0.257	12	0.379
UPMC	15	0.182	15	0.186	15	0.351

Table 6: This table presents the best annotation performance per concept, achieved by any team in any configuration, in terms of AP. It lists the concept name, the AP score, the team that achieved the score and the configuration of the run.

Concept	AP	Team	Conf.	Concept	AP	Team	Conf.
Partylife	0.408	LEAR	M	Food	0.635	XRCE	M
Family_Friends	0.555	ISIS	V	Vehicle	0.546	XRCE	M
Beach_Holidays	0.531	LEAR	M	Aesthetic_Impression	0.339	ISIS	V
Building_Sights	0.609	ISIS	V	Overall_Quality	0.289	ISIS	V
Snow	0.530	XRCE	M	Fancy	0.245	LEAR	M
Citylife	0.566	XRCE	M	Architecture	0.361	ISIS	V
Landscape_Nature	0.816	ISIS	V	Street	0.398	ISIS	V
Sports	0.186	ISIS	V	Church	0.288	LEAR	M
Desert	0.210	MEIJI	M	Bridge	0.224	XRCE	M
Spring	0.229	XRCE	M	Park_Garden	0.476	XRCE	M
Summer	0.332	ISIS	V	Rain	0.167	LEAR	M
Autumn	0.438	XRCE	M	Toy	0.370	XRCE	M
Winter	0.522	XRCE	M	MusicalInstrument	0.179	CNRS	M
No_Visual_Season	0.965	ISIS	V	Shadow	0.194	ISIS	V
Indoor	0.639	ISIS	V	bodypart	0.320	XRCE	M
Outdoor	0.909	XRCE	M	Travel	0.199	ISIS	V
No_Visual_Place	0.634	ISIS	V	Work	0.131	XRCE	V
Plants	0.805	ISIS	V	Birthday	0.169	LEAR	M
Flowers	0.618	XRCE	M	Visual_Arts	0.389	ISIS	V
Trees	0.702	ISIS	V	Graffiti	0.145	XRCE	M
Sky	0.895	XRCE	M	Painting	0.281	LEAR	M
Clouds	0.859	XRCE	M	artificial	0.219	LEAR	M
Water	0.725	XRCE	M	natural	0.734	LEAR	M
Lake	0.353	XRCE	M	technical	0.142	ISIS	V
River	0.351	LEAR	M	abstract	0.046	DCU	T
Sea	0.568	XRCE	M	boring	0.162	ISIS	V
Mountains	0.561	ISIS	V	cute	0.632	XRCE	M
Day	0.881	XRCE	M	dog	0.702	XRCE	M
Night	0.646	XRCE	M	cat	0.374	LEAR	M
No_Visual_Time	0.811	XRCE	M	bird	0.589	XRCE	M
Sunny	0.496	ISIS	V	horse	0.521	MEIJI	M
Sunset_Sunrise	0.791	XRCE	M	fish	0.480	MEIJI	M
Still_Life	0.445	LEAR	M	insect	0.499	XRCE	M
Macro	0.529	ISIS	V	car	0.455	XRCE	M
Portrait	0.684	XRCE	M	bicycle	0.449	XRCE	M
Overexposed	0.225	XRCE	M	ship	0.237	MEIJI	M
Underexposed	0.328	XRCE	M	train	0.347	XRCE	M
Neutral_Illumination	0.982	XRCE	M	airplane	0.640	MEIJI	M
Motion_Blur	0.284	ISIS	V	skateboard	0.455	DCU	T
Out_of_focus	0.223	ISIS	V	female	0.616	ISIS	V
Partly_Blurred	0.769	ISIS	V	male	0.782	XRCE	M
No_Blur	0.915	ISIS	V	Baby	0.407	XRCE	M
Single_Person	0.582	XRCE	M	Child	0.312	XRCE	M
Small_Group	0.359	XRCE	M	Teenager	0.266	LEAR	M
Big_Group	0.466	ISIS	V	Adult	0.582	ISIS	V
No_Persons	0.919	XRCE	M	old_person	0.116	LEAR	M
Animals	0.708	XRCE	M				

The best classified concepts are the ones from the mutually exclusive categories: *Neutral-Illumination* (98.2% AP, 94% F), *No-Visual-Season* (96,5% AP, 88% F), *No-Persons* (91.9% AP, 68% F), *No-Blur* (91.5% AP, 68% F). Following, the concepts *Outdoor* (90.9% AP, 50% F), *Sky*, (89.5% AP, 27% F) *Day* (88.1% AP, 51% F) and *Clouds* (85.9% AP, 14% F) were annotated with a high AP. The concepts with the worst annotation quality were *abstract* (4.6% AP, 1% F), *old-person* (11.6% AP, 2% F), *work* (13.1% AP, 3% F), *technical* (14.2% AP, 4% F), *Graffiti* (14.5% AP, 1% F), and *boring* (16.2% AP, 6% F). The percentages in parentheses denote the detection performance in AP and the frequency (F) of the concept occurrence in the images of the test set. Although there is a trend that concepts that occur more frequently in the image collection can be detected better, this does not hold for all concepts. Figure 2 shows the frequency of concepts in the test collection plotted against the best AP achieved by any submission.

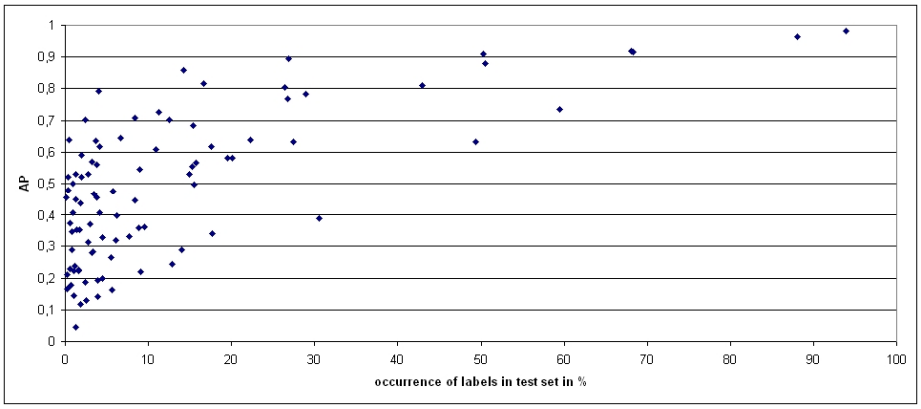

Fig. 2: Frequency of labels in test set plotted against best AP of submissions.

Although the performance of the textual runs is much lower in average than in the visual and textual runs, there are two concepts that can be annotated best in a textual configuration: *skateboard* and *abstract*. The concept skateboard was just annotated in six images of the test set and twelve of the training set. In the user tags of three images the word "skateboard" was present, while two images have no user tags and the sixth image does not contain words like "skateboard" or "skateboarding". It seems as if there is not enough visual information available to learn this concept while the textual and multi-modal approaches can make use of the tags and extract the correct concept from the tags for at least half of the images. The concept *abstract* was annotated more often (1,2% in the test set and 4,7% in the training set).

Further, one can see a great difference in annotation quality between the old concepts from 2009 that were carefully annotated by experts (number 1-52)

and the new concepts (number 53-93) annotated with the service of Mechanical Turk. The average annotation quality in terms of MAP for the old concepts is 0.57 while it is 0.37 for the new concepts. The reason for this is unclear. One reason may lie in the quality of the annotations of the non-experts. However, recent studies found that the quality of crowdsourced annotations is similar to the annotation quality of experts [21, 4, 22]. Another reason could be the choice and difficulty of the new concepts, as some of them are not as obvious and objective as the old ones. Further, some of the new concepts are special and their occurrence in the dataset is lower (7% in average) than the occurrence of the old concepts (17% in average).

One possibility to determine the reliability of a test collection is to calculate Cronbach's alpha value [23]. It defines a holistic measure of reliability and analyses the variance of individual test items and total test scores. The measure returns a value ranging between zero and one, for which bigger scores indicate a higher reliability. The Cronbach's alpha values show a high reliability for the whole test collection with 0.991, 0.991 for the queries assessed by experts and 0.956 for the queries assessed by MTurk. Therefore the scores point to a reliable test collection for both the manual expert annotations and the crowdsourced annotations and cannot explain the differences in MAP by the annotating systems.

6 Conclusions

The ImageCLEF 2010 Photo Annotation Task posed a multi-label annotation challenge for visual concept detection in three general configurations (textual, visual and multi-modal). The task attracted a considerable number of international teams with a final participation of 17 teams that submitted a total of 63 runs. In summary, the challenge could be solved with a MAP of 0.455 in the multi-modal configuration, with a MAP of 0.407 in the visual only configuration and with a MAP of 0.234 in the text configuration. For the evaluation per example 0.66 F-ex and 0.66 OS-FCS could be achieved for the multi-modal configuration, 0.68 F-ex and 0.65 OS-FCS for the visual configuration and 0.26 F-ex and 0.37 OS-FCS for the textual configuration. All in all, the multi-modal approaches got the best scores for 61 out of 93 concepts, followed by 30 concepts that could be detected best with the visual approach and two that won with a textual approach. As just two runs were submitted in the textual configuration, it is not possible to determine the abilities of purely textual classifiers reliably. In general, the multi-modal approaches outperformed visual and textual configurations for all teams that submitted results for more than one configuration.

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. Huiskes, M.J., Lew, M.S.: The MIR Flickr Retrieval Evaluation. In: Proc. of the ACM International Conference on Multimedia Information Retrieval. (2008)
3. 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)
4. Nowak, S., Rüger, S.: How reliable are Annotations via Crowdsourcing: a Study about Inter-annotator Agreement for Multi-label Image Annotation. In: MIR '10: Proceedings of the International Conference on Multimedia Information Retrieval, New York, NY, USA, ACM (2010) 557–566
5. Nowak, S., Lukashevich, H., Dunker, P., Rüger, S.: Performance Measures for Multilabel Evaluation: a Case Study in the Area of Image Classification. In: MIR '10: Proceedings of the International Conference on Multimedia Information Retrieval, New York, NY, USA, ACM (2010) 35–44
6. Nowak, S., Llorente, A., Motta, E., Rüger, S.: The Effect of Semantic Relatedness Measures on Multi-label Classification Evaluation. In: ACM International Conference on Image and Video Retrieval, CIVR. (July 2010)
7. Daróczy, B., Petrás, I., Benczúr, A.A., Nemeskey, D., Pethes, R.: SZTAKI @ ImageCLEF 2010. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
8. Sahbi, H., Li, X.: TELECOM ParisTech at ImageCLEF 2010 Photo Annotation Task: Combining Tags and Visual Features for Learning-Based Image Annotation. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
9. Li, W., Min, J., Jones, G.J.F.: A Text-Based Approach to the ImageCLEF 2010 Photo Annotation Task. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
10. Mbanya, E., Hentschel, C., Gerke, S., Liu, M., Nürnberger, A., Ndjiki-Nya, P.: Augmenting Bag-of-Words - Category Specific Features and Concept Reasoning. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
11. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Detection of Visual Concepts and Annotation of Images using Predictive Clustering Trees. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
12. Zhang, D., Liu, B., Sun, C., Wang, X.: Random Sampling Image to Class Distance for Photo Annotation. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
13. van de Sande, K.E.A., Gevers, T.: The University of Amsterdam's Concept Detection System at ImageCLEF 2010. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
14. Mensink, T., Csurka, G., Perronnin, F., Sánchez, J., Verbeek, J.: LEAR and XRCE's participation to Visual Concept Detection Task - ImageCLEF 2010. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
15. Batal, R.A., Mulhem, P.: MRIM-LIG at ImageCLEF 2010 Visual Concept Detection and Annotation task. In: Working Notes of CLEF 2010, Padova, Italy. (2010)

16. Paris, S., Glotin, H.: Linear SVM for LSIS Pyramidal Multi-Level Visual only Concept Detection in CLEF 2010 Challenge. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
17. Motohashi, N., Izawa, R., Takagi, T.: Meiji University at the ImageCLEF2010 Visual Concept Detection and Annotation Task: Working notes. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
18. Rasche, C., Vertan, C.: A Novel Structural-Description Approach for Image Retrieval. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
19. Fakeri-Tabrizi, A., Tollari, S., Usunier, N., Amini, M.R., Gallinari, P.: UPMC/LIP6 at ImageCLEFannotation 2010. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
20. Stanek, M., Maier, O.: The Wroclaw University of Technology Participation at ImageCLEF 2010 Photo Annotation Track. In: Working Notes of CLEF 2010, Padova, Italy. (2010)
21. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: SIGIR 2009 Workshop on the Future of IR Evaluation. (2009)
22. Hsueh, P., Melville, P., Sindhwani, V.: Data Quality from Crowdsourcing: a Study of Annotation Selection Criteria. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. (2009)
23. Bodoff, D.: Test theory for evaluating reliability of IR test collections. Information Processing & Management **44**(3) (2008) 1117–1145