

Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task

Bart Thomee¹ and Adrian Popescu²

¹ Yahoo! Research, Barcelona, Spain
bthomee@yahoo-inc.com

² CEA LIST, Vision & Content Engineering Laboratory, Fontenay-aux-Roses, France
adrian.popescu@cea.fr

Abstract. ImageCLEF’s Flickr Photo Annotation and Retrieval task aims to advance the state of the art in multimedia research by providing a challenging benchmark for visual concept detection, annotation and retrieval in the context of a diverse collection of Flickr photos. The benchmark consisted of two separate but closely connected subtasks, where the objective of the first subtask was to accurately detect a wide range of semantic concepts for the purpose of automatic image annotation, while the objective of second subtask was to correctly retrieve relevant images for concept-oriented queries inspired by what people actually search for on the internet. This paper presents an overview of the benchmark, summarizes the annotation and retrieval techniques proposed by the participating teams and evaluates their performance.

1 Introduction

Automatic recognition of photographic content is useful in a wide range of domains, ranging from specialized application, such as medical imagery, to large public applications, such as web content structuring and retrieval. Although considerable research efforts have been devoted to concept detection in public images, this task remains difficult because the number of possible concepts that can be depicted is boundless, where the visual aspect of each concept additionally can vary along numerous dimensions.

The Flickr Photo Annotation and Retrieval task we present in this paper is a multi-label classification challenge that offers a benchmark for testing novel visual concept detection, annotation and retrieval algorithms on a public collection containing photos gathered from the social sharing website Flickr¹. The aim is to analyze the images in terms of their visual and/or textual features in order to detect the presence of one or more semantic concepts. The detected concepts are then to be used for the purpose of automatically annotating the images or for retrieving the best matching images to a given concept-oriented query. The concepts are very diverse and range across categories such as people (e.g. teenager, female), scenery (e.g. lake, desert), weather (e.g. rainbow, fog) and even impressions (e.g. unpleasant, euphoric).

¹ <http://www.flickr.com>

This task has a longstanding tradition at ImageCLEF. Since 2009 the task has been based upon various subsets of the MIRFLICKR collection [1,2], where every year the list of concepts to detect was updated in order to cover a wider selection of concept types and to make the task more challenging. Last year a concept-based retrieval subtask was added to exploit the concept annotations in the context of image retrieval. In contrast with the closely related Scalable Web Image Annotation task [3], also held this year at ImageCLEF, our task involves a smaller collection of images but has been fully and manually annotated within the chosen concept space, a characteristic that favors experiment reproducibility. To this end, the annotations associated with the collection will be released after the campaign. The related PASCAL Visual Object Classes challenge [4] has as aim to accurately detect the bounding boxes and labels of objects in a set of images, whereas our focus is on both visual and textual information instead of visual information only and furthermore we offer a larger range of concepts to detect. Yet another related benchmark is the ImageNet Large Scale Visual Recognition Challenge², which is run over a larger dataset and with a larger number of concepts, but without focus on multi-label classification.

The remainder of this paper is organized as follows. First, in Section 2 we describe the dataset upon which the task is based, in Section 3 we present the concepts and the concept-oriented queries, and in Section 4 we discuss how we collected the ground truth. We then introduce the participating teams in Section 5 and evaluate the performance of their techniques on the annotation and retrieval subtasks in Sections 6 and 7 respectively. Finally, in Section 8 we offer outlooks for the future of visual concept detection, annotation and retrieval.

2 Dataset

The annotation and retrieval subtasks are based on the MIRFLICKR collection [1,2]. The entire collection contains 1 million images from the social photo sharing website Flickr and was created by downloading up to a thousand photos per day in the period 2008-2010 that at that moment were deemed to be the most interesting according to Flickr. All photos in this collection were released by their photographers under a Creative Commons license, allowing them to be freely used for research purposes. The annotation subtask was based on the first 25 thousand images of the MIRFLICKR collection, whereas the retrieval subtask involved a subset of 200 thousand images.

2.1 Textual features

Each of the images used in both subtasks was accompanied by descriptive metadata, within which we can distinguish the following textual features:

User tags: These are the tags that the users assigned to the photos they uploaded to Flickr.

² <http://www.image-net.org/challenges/LSVRC/2012/index>

User, photo and license information: These features contain details about the Flickr user that took the photo, the photo itself and the Creative Commons license associated with the photo.

EXIF metadata: If available, the EXIF metadata contains information about the camera that took the photo and the parameters used.

In Figure 1 we show an example photo and its associated textual features.



Fig. 1: An example photo from the MIRFLICKR collection and its associated user tags, user information, photo information, license information and EXIF metadata. Due to space considerations we only show part of the metadata.

2.2 Visual features

We noticed that often similar types of visual features were used by the participants in previous editions of the photo annotation task, in particular descriptors based on interest points and bag-of-words were popular. To allow the participants to direct their attention on the actual concept detection instead of having to compute common features, we extracted a number of descriptors for the participants beforehand and released them together with the dataset. We additionally gave the participants some pointers to toolkits that would allow them to extract the descriptors with a different set of parameters or to extract other related descriptors. Each of the images used in the annotation task was accompanied by the following visual features:

SURF [5]: The SURF technique uses a Hessian matrix-based measure for the detection of interest points and a distribution of Haar wavelet responses within the interest point neighborhood as descriptor. An image is analyzed at several scales, so interest points can be extracted from both global ('coarse') and local ('fine') image details. Additionally, the dominant orientation of each of the interest points is determined to support rotation-invariant matching. We used the OpenSURF toolkit³ to extract this descriptor.

³ <http://www.chrisevansdev.com/computer-vision-opensurf.html>

TOP-SURF [6]: A bag-of-words technique [7] was used to cluster the SURF interest points extracted from a representative collection of photographic images into a number of visual words. The interest points present in each Flickr image are then converted to their most closely matching visual words, after which a histogram is formed consisting of only the top few most dominant visual words. We used the TOP-SURF toolkit⁴ to extract this descriptor.

SIFT [8], C-SIFT [9], RGB-SIFT [10], OPPONENT-SIFT [10]: The SIFT descriptor describes the local shape of an image region using edge orientation histograms. The other three descriptors are variations that represent the image in different color spaces before computing the SIFT descriptor. We used the ISIS Color Descriptors toolkit⁵ to extract all these descriptors.

GIST [11]: The GIST descriptor is based on a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. To capture this image structure, oriented edge responses are aggregated at multiple scales into very coarse bins. We used the LabelMe toolkit⁶ to extract this descriptor.

3 Concepts and queries

Defining a compact yet representative list of concepts to annotate or queries to propose in a search engine are not trivial tasks, because the spaces to choose from and the user needs are virtually infinite. While in the past concepts and queries were chosen to represent different ontological fields and to be of variable difficulty, this year an usage-oriented constraint was added with the exploitation of image search query logs in order to define and/or refine the concepts and queries. We additionally took their ‘textualness’ and ‘visualness’ into account in order to offer a set of concepts and queries with varying difficulty and to accommodate for both the textual and visual techniques the participants may propose.

3.1 Concept definition

In this edition of the photo annotation and retrieval task we continued along the same lines as previous years in terms of concepts, where in total we defined a set of 94 concepts referring to nature, people, image quality and so on. In comparison with last year’s benchmark we removed a few of the concepts that were not sufficiently present in the dataset or ambiguously defined, based on feedback given by former participants. We furthermore introduced several new concepts that were inspired by popular queries issued to the Yahoo! image search engine⁷ in order to provide a more realistic context for the task. We

⁴ <http://press.liacs.nl/researchdownloads/topsurf>

⁵ <http://www.colordescriptors.com>

⁶ <http://labelme.csail.mit.edu/>

⁷ <http://images.yahoo.com>

show an overview of the concepts in Table 1, where we grouped them into related categories. The exact descriptions we used to represent each concept are listed in Appendix A.

Table 1: Overview of the concepts used in the photo annotation subtask, hierarchically grouped into categories.

Natural elements	
<i>time of day</i>	day, night, sunrise/sunset
<i>celestial bodies</i>	sun, moon, stars
<i>weather</i>	clear sky, overcast sky, cloudy sky, rainbow, lightning, fog/mist, snow/ice
<i>combustion</i>	flames, smoke, fireworks
<i>lighting effects</i>	shadow, reflection, silhouette, lens effects
Environment	
<i>scenery</i>	mountain/hill, desert, forest/park, coast, landscape, cityscape, graffiti
<i>water</i>	underwater, sea/ocean, lake, river/stream, other
<i>flora</i>	tree, plant, flower, grass
<i>fauna</i>	cat, dog, horse, fish, bird, insect, spider, amphibian/reptile, rodent
People	
<i>quantity</i>	none, zero, one, two, three, small group, large group
<i>age</i>	baby, child, teenager, adult, elderly
<i>gender</i>	male, female
<i>relationship</i>	family/friends, co-workers, strangers
Image elements	
<i>quality</i>	in focus, selective focus, out of focus, motion blur, noisy/blocky
<i>style</i>	picture-in-picture, circular warp, gray-color, overlay
<i>view</i>	portrait, close-up/macro, indoor, outdoor
<i>type</i>	city life, party life, home life, sports/recreation, food/drink
<i>impression</i>	happy, calm, inactive, melancholic, unpleasant, scary, active, euphoric, funny
Human elements	
<i>transportation</i>	bicycle/motorcycle, car/van/pick-up, truck/bus, rail vehicle, water vehicle, air vehicle

3.2 Query definition

Similar to how we defined the concepts for the annotation subtask, we removed a few of the queries of last year’s retrieval subtask and introduced several new ones that were also inspired by the most popular queries we found in the image search logs. In total we defined 42 concept-oriented queries, where many of them can be considered as linear combinations of the concepts used in this task, whereas others are formed by involving additional constraints. We show an overview of the queries in Table 2. The exact descriptions we used to represent each query are listed in Appendix B.

Table 2: Concept-oriented queries used in the photo retrieval subtask.

Query	Title	Query	Title
0	flying airplane	21	halloween costumes
1	horse riding	22	surf swim
2	mountain coast	23	flower field
3	single performer live music	24	foggy forest
4	snowy trees	25	grass field recreation
5	hot air balloon	26	woman short hairstyles
6	beach sunset sunrise	27	skyline fireworks
7	old men	28	full moon
8	train station	29	sleeping baby
9	sad dogs	30	graffiti artist
10	silence before the storm	31	fish tank
11	smooth water flow	32	people dancing at party
12	birds in a tree	33	beautiful sceneries
13	person silhouette	34	underwater sea life no divers
14	traffic light trails	35	euphoric people
15	city reflections by day	36	fire without smoke
16	close-up red roses	37	high speed cycling
17	double rainbow	38	water drops
18	fast car	39	dark clothing
19	autumn park leaves	40	above the clouds
20	close-up cupcakes	41	bride

4 Ground truth collection

We acquired the ground truth relevance annotations for the newly defined concepts and queries, as well as for the concepts and queries reused from last year, through crowd sourcing. We enlisted the help of many anonymous workers on Amazon’s Mechanical Turk⁸, which is an online marketplace for distributing small jobs to be performed by interested people for a small fee. Due to the presence of workers that do not have a genuine interest in performing the requested service, where such a worker may either be a real person or an automated service that pretends to be human, it is necessary to validate the quality of the performed work. To this end we used the intermediary service of CrowdFlower⁹ to obtain the relevance judgments, because this service automatically performs the filtering of the workers based on the quality of the work they perform by validating it against specific examples for which the correct answer is known. Such examples are commonly referred to as *gold* and need to be supplied in addition to the job. Ultimately, the CrowdFlower service guarantees that each *unit*, which in our case refers to a concept-image or query-image combination, ends up being assessed by at least three workers that exhibited good annotation behavior.

⁸ <http://www.mturk.com>

⁹ <http://www.crowdfunder.com>

4.1 Concept relevance assessment

The concept annotation jobs we created showed the workers a photo accompanied by a short list of related concepts and asked them to indicate all concepts that were clearly present in the image, see Figure 2 for an example. We created separate tasks for the different concept subcategories, because intuitively we felt it would be easier for a worker to perform relevance judgments faster and more accurately for related concepts (e.g. cat, dog, fish) than for arbitrary concepts (e.g. cat, reflection, city life).

Which of the following concepts are clearly present in the picture below? Tick all that apply

☐ Cat
☒ Dog
☐ Horse
☐ Fish
☐ Bird
☐ Insect
☐ Animal (other)
☐ None of the above



[click to view image in larger size](#)

Fig. 2: An example crowd sourcing job for the concept relevance assessment, showing an image and a list of concepts the worker can annotate the image with. Here, we have marked the correct answer in yellow.

Before starting a job, a worker was presented with a set of instructions, three example images for each concept included in the job description and three example images not containing any of the sought-for concepts. The gold we used for validating the performance of the workers was annotated by ourselves and either clearly exhibited a particular concept or it clearly did not exhibit it; to not mistakenly mark a worker as a bad annotator we did not include images as gold that ambiguously contained a concept. Due to the subjective nature of many of the concepts it was certainly possible for good annotators to disagree with each other about the presence of a concept in an image and therefore we applied the majority voting rule to the relevance judgments to make the final decisions. Using the earlier example, in Figure 3 we show which concepts were considered to be present by the workers in the image of the cake.

	concepts:	agreement:	vote:
	timeofday_day	1.00	X
	flora_flower	0.50	
	quantity_none	1.00	X
	quality_infocus	0.67	X
	quality_selectivefocus	0.33	
	style_circularwarp	0.20	
	style_overlay	0.80	X
	view_closeupmacro	1.00	X
	view_indoor	1.00	X
	setting_fooddrink	1.00	X
	sentiment_happy	1.00	X

Fig. 3: The concepts associated with the example image for which at least one worker indicated they were present. The figure further includes the relative agreement between the workers and the outcome of the majority vote.

After the ground truth annotations were collected, we divided the photo collection of 25 thousand images into a training set of 15 thousand images and a testing set of 10 thousand images. We ensured that the number of images assigned to each concept in both sets was roughly proportional to the quantity in which they were present in the entire collection, e.g. if concept A was present in a total of 250 images, then we aimed to assign 150 of these images to the training set and the remaining 100 to the testing set. We considered it to be of paramount importance to assure that concepts with few images were sufficiently present in both sets and in balance with each other, in effect mitigating the small sample size problem. This is also in response to the feedback received of some of last years’ participants, who indicated that previously some concepts were under-represented in the training set and overrepresented in the testing set, or vice versa. To this end, we used a greedy approach to perform the image assignments, where the algorithm iteratively distributed the images starting with the least represented concepts and ending with the most represented concepts. In Appendix A we have listed the number of times each concept is represented in both the training set and the testing set.

4.2 Query relevance assessment

For the query relevance assessment we used a similar setup as with the concept annotations, see Figure 4 for an example. We presented a worker with a set of instructions and with three example images for each of the 42 concept-related queries before they were able to start a job. The set of images that needed to be judged was formed by aggregating the top 100 images of all ranked retrieval results that were submitted by each participating team for each query. The gold we used for validating the performance of the workers was annotated by several trained editors of the National Institute of Standards and Technology (NIST), which we had access to through a collaboration with the TREC Crowdsourcing track. For the same reasons as for the concept relevance judgments, we applied the majority voting rule to the relevance judgments to determine whether or not an image was ultimately relevant to a particular query.

Is this photo relevant to the following query?

The user is looking for photos showing one or more airplanes flying in the sky. He is not looking for photos that show airplanes on the ground, taking off or landing, nor for pictures of airplanes from the inside. (query #0)

☒ Yes

☐ No



[click to view image in separate window](#)

Fig. 4: An example crowd sourcing job for the query relevance assessment, where the worker has to indicate whether the image is relevant to the query. Here, we have marked the correct answer in yellow.

5 Participation

This year, in total 100 teams registered for the benchmark and signed the license agreement to access the collections. Eventually, 18 teams submitted a total of 80 runs to the annotation subtask, where the maximum number of runs per team was limited to 5. For the retrieval subtask, 7 teams submitted a total of 47 runs to the retrieval subtask, where the maximum number of runs per team was limited to 10. In this section we will introduce the participating teams in alphabetical order and highlight the techniques they used to perform the image annotation and/or retrieval. We refer to their working notes with a superscript 'a' if the team participated in the annotation subtask and with a superscript 'r' if they participated in the retrieval subtask. In case of encountering unfamiliar acronyms in the technique summaries, please refer to the respective working notes, if available.

BUUA AUDR ^a[12]: This team proposed textual and multimodal runs. Bags of visual words based on SIFT, coupled with soft assignment, were used to represent visual content. Frequent tags were selected in order to form a textual vocabulary that was used to map visual concepts. SVM classifiers were then used to predict potentially relevant concepts for each image. Annotation refinement that accounts for concept correlation was introduced to improve results obtained with textual or visual schemes.

CEA LIST ^a[13]: This team concentrated on the combination of textual and visual image. Textual models were built by combining semantic and contextual information, respectively derived from WordNet and Flickr, that were consequently processed using pooling strategies. For visual information, a computationally efficient bag of multimedia words strategy was tested against classical bag of visual words approaches.

CERTH ^a[14]: This team tested two approaches for image annotation. The first was based on Laplacian Eigenmaps of an image similarity graph model and the second on a “same class” model. They tested different multimedia fusion schemes and reported that best results were obtained when both textual and visual information were combined using the first learning scheme.

DBRIS ^a[15]: This team focused on low-level image descriptions that combine different SIFT based features. Two image representations were compared: the first was based on spatial pyramids and the second on visual phrases. Results show that visual phrases clearly outperformed spatial pyramids and classical bags of visual words. Visual phrases alone also outperformed their combination with the other representation schemes.

DMS-SZTAKI ^a[16]: This team presented only multimodal runs. A fixed length visual descriptor with different similarity measures was devised and it allowed the early combination of textual and visual features. Gaussian Mixture Models were trained to define low-level features. Both global and local features were extracted from the images and a biclustering approach was adopted in order to represent Flickr tags. A three step fusion approach that included a transformation, a feature aggregation and a selection step was adopted.

IMU ^{a,r}[17]: This team focused on textual information modeling. Concept annotation was performed using maximum conditional probability to assess the probability of occurrence of a concept in an image based on already existing tags. Concept-based retrieval was performed using a classical language modeling technique.

IL ^a[18]: This team focused on the exploitation of textual features associated to images in the test dataset. They tackled tag noise and incompleteness by creating tag-concept co-occurrence models in a two phase procedure, which first removed noisy tags from the model and then enriched existing tags with related ones that were not filtered out during the first phase.

ISI ^a[19]: This team presented an approach that focused on scalability. Fisher Vectors and bag of visual words based on SIFTs were used to represent visual content, while classical bag of words with TF-IDF weighting was used to represent textual content. To achieve this, an online multi-label learning approach called Passive-Aggressive with Averaged Pairwise Loss was adapted from authors’ earlier work. Reported results showed that a combination of different visual features was beneficial for the overall performance of the system.

KIDS NUTN ^{a,r}[20]: This team proposed multimedia fusion techniques that exploited textual and visual features, whose processing was done using dimensionality reduction, random forest classifiers and semi-supervised learning strategies. They reported that simple visual feature combination did not improve results over the use of single visual descriptors and that semi-supervised learning did not outperform supervised learning. For the retrieval subtask, results were based on the annotation results and the best results were also obtained with a combination of textual and visual features.

LIRIS ^a[21]: This team modeled both textual and visual information and introduced a competitive fusion of the two modalities. A histogram of textual concepts that relied on semantic similarity between user tags and a concept dictionary was used to represent tags associated to Flickr images. Different global and local visual features were considered to model visual content. A Selective Weighted Late Fusion that iteratively selected and weighted the best features to use for each concept was introduced to combine textual and visual modalities, resulting in a significant improvement over monomodal runs.

MSATL ^{a,r}: This team used keyword and document representations of the concepts as textual features to match against the textual descriptions of the images for the purpose of annotation. In addition, they incorporated a random forest into which their visual features were embedded. For the retrieval subtask, they focused on textual features only and retrieved images based on the title of the query and a combination of descriptions and keywords associated with the query's concepts.

MLKD ^{a,r}[22]: This team proposed visual, textual and multimodal runs for the annotation subtask. Different visual representations, such as BOVW, VLAD and VLAT were tested on top of SURE, SIFT and color SIFT. Text was modeled using standard bag of words techniques with TF-IDF weighting. Multimodal combination was realized either by averaging or by selecting the best model for each concept. For retrieval, they introduced two methods, where the first involved the production of co-occurrence models from Flickr to score concepts, while the second was a standard vector space model for text retrieval.

NII ^a: This team used a combination of local visual features and global visual features to address the annotation subtask, where the local features included dense SIFT, color SIFT and PHOW.

NPDILIP6 ^a[23]: This team focused on visual processing and introduced Bossa Nova, a mid-level image representation, that enriched the classical bag-of-words image representations by adding a histogram of distances between the descriptors of the image and those in the codebook. This compact and efficient representation was a useful addition to Fisher Vector representations and the results were improved by combining these two techniques.

PRA ^a[24]: This team submitted only visual runs that combined different visual descriptors and furthermore proposed a dynamic fusion of visual classifiers. A combination of SVMs was used to obtain annotations, where the final decisions were obtained using the mean rule, through majority voting or according to a dynamic score selection approach.

RedCAD ^r[25]: This team used Latent Dirichlet Allocation to produce topic models and use the Jensen-Shannon Divergence measure for topic similarity to retrieve similar images.

REGIM ^r[26]: This team presented an approach that deals with query analysis and relevance-based ranking, two central problems in image retrieval. Their query analysis exploited both the textual and visual information provided with

the proposed queries. Ranking was performed by choosing an appropriate similarity measure and enhancing the results with a random walk with restart algorithm.

UAIC^a[27]: This team exploited both textual and visual features of the images to annotate. Tags were processed to extract most frequent elements and then processed using a linear kernel. Used visual features include both local ones, such as TOP-SURF, and global ones, such as Profile Entropy or Color Moments. SVM were used to fuse modalities and a post-processing step was added to check the consistency of predicted labels. In addition, face detection was used to increase the accuracy of person-related concepts.

UNED^{a,r}[28]: This team presented a system that exploited textual cues to pre-filter results before applying image processing techniques in a setting inspired from information retrieval algorithms. With minor adaptations the same system was used for both image annotation and retrieval. Logistic regression was used in order to predict the probability of occurrence of a concept in a given photo. Tag expansion techniques were used to improve image description prior to the annotation and retrieval processes.

URJCyUNED^a[29]: This team used multiple visual features to represent low-level image content and WordNet to derive similarities between user tags and the concepts to annotate. Fusion strategies that selected either textual or visual features were tested and the reported results showed that such strategies were superior to the use of both modalities. Nonetheless, their textual approach outperformed both fusion strategies.

6 Photo annotation evaluation

The runs submitted by the participants for the annotation subtask contained the relevance assessments for each concept-image combination, where a binary decision was made whether or not the concept was considered to be present in the image and additionally a real-valued score was supplied expressing the confidence of that decision. While the confidence scores are not comparable between runs of different teams or not even necessarily between runs of the same team, they can be seen as indicating a relative ordering of how the images are assigned to the concepts, allowing us to also apply evaluation measures to the confidence scores that are typically applied in the context of information retrieval. In this section we present only an evaluation for a selection of the runs, whereas detailed information on all runs can be found on the Photo Annotation subtask website¹⁰.

6.1 Evaluation measures

To assess the performance of the runs submitted by the teams, we used the following evaluation measures:

¹⁰ <http://imageclef.org/2012/photo-flickr/annotation/>

Mean Average Precision (MAP): This evaluation measure first ranks the images by their confidence scores, from high to low, for each concept separately. The images are inspected one by one and each time a relevant image is encountered the precision and recall values are computed. In case of ties we consider all the images with the same confidence score together at once and produce only a single precision and recall value for them using a tie-aware ranking approach [30]. We then interpolate the values so the recall measurements range from 0.0 to 1.0 with steps of 0.1; the precisions at these recall levels are obtained by taking the maximum precision obtained at any non-interpolated recall level equal or greater to the interpolated recall step level under consideration. To obtain the overall non-interpolated MAP (MnAP) value we average the non-interpolated precisions for each concept and then average these averages, whereas to obtain the overall interpolated MAP (MiAP) we instead average the average interpolated precisions over all concepts. In the analysis of the annotation runs we focus on the interpolated MAP, although for completeness we also report the non-interpolated MAP values in the detailed results available on the website.

Geometric Mean Average Precision (GMAP): This evaluation measure is an extension to MAP. When comparing runs with each other the GMAP specifically highlights improvements obtained on relatively difficult concepts, e.g. increasing the average precision of a concept from 0.05 to 0.10 has a larger impact in its contribution to the GMAP than increasing the average precision from 0.25 to 0.30. To compute the non-interpolated GMAP (GMnAP) and the interpolated GMAP (GMiAP), we follow the same procedure as with MnAP and MiAP, but we instead average the logs of the average precision for each concept, after which we exponentiate the resulting average back to obtain the GMAP. To avoid taking the log of an average precision of zero we add a very small epsilon value to each average precision before computing its log, which we remove again after exponentiating the averages of these logs; when the epsilon value is very small its effect on the final GMAP is negligible. In the analysis of the annotation runs we focus on the interpolated GMAP, although for completeness we also report the non-interpolated GMAP values in the detailed results.

F1: The F1-measure uses the provided binary scores to determine how well the annotations are. We have computed the instance-averaged, micro-averaged and macro-averaged F1 scores for the photos as well as for the concepts. The instance-F1 for the photos is computed by determining the number of true positives, false positives, true negatives and false negatives in terms of detected concepts and using this to compute the F1-score for each individual photo, after which the F1-scores are averaged over all photos. The micro-F1 for the photos is computed by averaging the precision and recall scores for each individual photo and then computing the F1-score from these averages. The macro-F1 for the photos is computed by aggregating the number of true positives, false positives, true negatives and false negatives over all photos and then computing the F1-score based on these numbers. The micro-F1 and macro-F1 for the concepts are computed in a similar fashion, swapping the roles of the photos and con-

cepts. In the analysis of the annotation runs we focus on the micro-F1 scores, although for completeness we also report all instance-F1 and macro-F1 values in the detailed results.

6.2 Results

The underlying techniques of the participants could use one of three possible configurations, namely textual features only, visual features only or a multi-modal combination of both. For this subtask, 18 teams submitted in total 80 runs, of which 17 runs exclusively used textual features, 28 runs exclusively used visual features and 35 runs used a multimodal approach. We present the overall evaluation results according to the MiAP, GMiAP and micro-F1 in Table 3 to get an understanding of the best results irrespective of the features used, where in the Feature column the letter T refers to the textual configuration, V to the visual configuration and M to the multimodal configuration. In the tables, the ranks indicate the position at which the best run appeared in the results. To compare only runs using the same configuration we present separate results for the textual features in Table 4, the visual features in Table 5 and the multimodal features in Table 6.

Table 3: Summary of the annotation results for the evaluation per concept and image for the best overall run per team per evaluation measure.

Team Rank MiAP Feature				Team Rank GMiAP Feature				Team Rank micro-F1 Feature			
LIRIS	1	0.4367	M	LIRIS	1	0.3877	M	LIRIS	1	0.5766	M
DMS-SZTAKI	3	0.4258	M	DMS-SZTAKI	3	0.3676	M	DMS-SZTAKI	3	0.5731	M
CEA LIST	6	0.4159	M	CEA LIST	5	0.3615	M	NII	6	0.5600	V
ISI	7	0.4136	M	ISI	7	0.3580	M	ISI	7	0.5597	M
NPDPIIP6	16	0.3437	V	NPDPIIP6	16	0.2815	V	MLKD	16	0.5534	V
NII	22	0.3318	V	NII	21	0.2703	V	CEA LIST	20	0.5404	M
CERTH	28	0.3210	M	MLKD	28	0.2567	V	CERTH	26	0.4950	M
MLKD	29	0.3185	V	CERTH	29	0.2547	M	IMU	30	0.4685	T
IMU	36	0.2441	T	IMU	35	0.1917	T	KIDS NUTN	34	0.4406	M
UAIC	38	0.2359	V	UAIC	39	0.1685	V	UAIC	35	0.4359	V
MSATL	41	0.2209	T	MSATL	42	0.1653	T	NPDPIIP6	37	0.4228	V
IL	46	0.1724	T	IL	45	0.1140	T	IL	49	0.3532	T
KIDS NUTN	47	0.1717	M	KIDS NUTN	49	0.0984	M	URJCyUNED	50	0.3527	T
BUAA AUDR	52	0.1423	V	BUAA AUDR	51	0.0818	V	PRA	54	0.3331	V
UNED	55	0.1020	V	UNED	55	0.0512	V	MSATL	57	0.2635	T
DBRIS	58	0.0976	V	DBRIS	57	0.0476	V	BUAA AUDR	58	0.2592	M
PRA	65	0.0900	V	PRA	66	0.0437	V	UNED	66	0.1360	V
URJCyUNED	77	0.0622	V	URJCyUNED	77	0.0254	V	DBRIS	69	0.1070	V

As we can see, the subtask was best solved with a MiAP of 0.4367 by LIRIS with the three runners up DMS-SZTAKI, CEA LIST and ISI all scoring above a MiAP of 0.4. The same ordering can be found when considering the GMiAP evaluations. As for the micro-F1 score, LIRIS performs best once again although its best run is closely followed by the runs of five other teams. If we look at the rank at which the best run of a team was placed, then the majority of the runs submitted by the teams that took the top 4 positions occupy the places 1-15, indicating that these teams submitted several variations of their runs that were all performing rather well.

Table 4: Summary of the annotation results for the evaluation per concept and image for the best textual run per team per evaluation measure.

Team Rank MiAP			Team Rank GMiAP			Team Rank micro-F1		
LIRIS	1	0.3338	LIRIS	1	0.2771	LIRIS	1	0.4691
CEA LIST	3	0.3314	CEA LIST	2	0.2759	IMU	2	0.4685
IMU	4	0.2441	IMU	4	0.1917	CEA LIST	5	0.4452
CERTH	6	0.2311	CERTH	7	0.1669	MLKD	7	0.3951
MSATL	8	0.2209	MSATL	9	0.1653	CERTH	8	0.3946
IL	11	0.1724	IL	11	0.1140	IL	10	0.3532
BUAA AUDR	13	0.1423	BUAA AUDR	13	0.0818	URJCyUNED	11	0.3527
UNED	14	0.0758	UNED	14	0.0383	MSATL	13	0.2635
MLKD	15	0.0744	MLKD	15	0.0327	BUAA AUDR	14	0.2167
URJCyUNED	17	0.0622	URJCyUNED	17	0.0254	UNED	16	0.0864

Table 5: Summary of the annotation results for the evaluation per concept and image for the best visual run per team per evaluation measure.

Team Rank MiAP			Team Rank GMiAP			Team Rank micro-F1		
LIRIS	1	0.3481	LIRIS	1	0.2858	NII	1	0.5600
NPDILIP6	2	0.3437	NPDILIP6	2	0.2815	MLKD	6	0.5534
NII	6	0.3318	NII	5	0.2703	ISI	7	0.5451
ISI	10	0.3243	ISI	10	0.2590	LIRIS	8	0.5437
MLKD	11	0.3185	MLKD	11	0.2567	CERTH	9	0.4838
CERTH	13	0.2628	CERTH	13	0.1904	UAIC	10	0.4359
UAIC	14	0.2359	UAIC	14	0.1685	NPDILIP6	11	0.4228
UNED	15	0.1020	UNED	15	0.0512	PRA	15	0.3331
DBRIS	16	0.0976	DBRIS	16	0.0476	URJCyUNED	18	0.1984
PRA	22	0.0873	PRA	23	0.0437	UNED	19	0.1360
MSATL	24	0.0868	MSATL	25	0.0414	DBRIS	22	0.1070
URJCyUNED	28	0.0622	URJCyUNED	28	0.0254	MSATL	23	0.1069

Table 6: Summary of the annotation results for the evaluation per concept and image for the best multimodal run per team per evaluation measure.

Team Rank MiAP			Team Rank GMiAP			Team Rank micro-F1		
LIRIS	1	0.4367	LIRIS	1	0.3877	LIRIS	1	0.5766
DMS-SZTAKI	3	0.4258	DMS-SZTAKI	3	0.3676	DMS-SZTAKI	3	0.5731
CEA LIST	6	0.4159	CEA LIST	5	0.3615	ISI	6	0.5597
ISI	7	0.4136	ISI	7	0.3580	CEA LIST	12	0.5404
CERTH	15	0.3210	CERTH	15	0.2547	MLKD	15	0.5285
MLKD	16	0.3118	MLKD	16	0.2516	CERTH	18	0.4950
UAIC	21	0.1863	UAIC	20	0.1245	KIDS NUTN	20	0.4406
KIDS NUTN	22	0.1717	KIDS NUTN	24	0.0984	UAIC	21	0.4352
BUAA AUDR	26	0.1307	BUAA AUDR	26	0.0558	BUAA AUDR	29	0.2592
MSATL	31	0.0867	MSATL	31	0.0408	URJCyUNED	30	0.2306
UNED	33	0.0756	UNED	33	0.0376	UNED	33	0.0849
URJCyUNED	34	0.0622	URJCyUNED	34	0.0254	MSATL	34	0.0319

Inspecting the results for the different feature configurations, we can see that the multimodal configuration was predominantly used in the best performing runs, where moreover the MiAP and GMiAP of the multimodal runs tended to be higher than those of the textual and visual runs. Nonetheless, the micro-F1 scores of the better performing visual runs come close to those of the multimodal runs, which means that they both were able to roughly equivalently well annotate the images with the correct concepts without including too many concepts that were incorrect. Overall, we can see across the different tables that that roughly the same ordering of the teams is maintained, suggesting that the better performing teams had a good underlying strategy that performed well irrespective of the features used. Note that the MLKD team discovered a bug in their textual runs that consequently also affected their multimodal runs. As this discovery happened past the submission deadline, we did not include their updated runs in the results out of fairness to the other teams. Nonetheless, please refer to their working notes [22] for the updated results, since the fixed runs yielded a substantial improvement over the original runs.

If we analyze the performance on the individual concepts instead of the performance over all concepts combined, as is shown in Figure 5, we can clearly observe that the concepts were of varying difficulty. In particular, the concepts `quantity_none` and `quality_infocus` appeared relatively easy to detect, with an average accuracy of around 0.8. For many concepts in the categories `natural elements` and `environment` the maximum attained accuracy exceeded 0.7, yet at the same time the minimum accuracy came close to or equaled zero, indicating that some of the runs were quite capable of detecting the concepts, whereas other runs were simply unable to detect the concepts at all. The concepts in the `impression` subcategory proved rather difficult, presumably due to their highly subjective nature, although the two concepts `impression_happy` and `impression_calm` were easiest to detect amongst them.

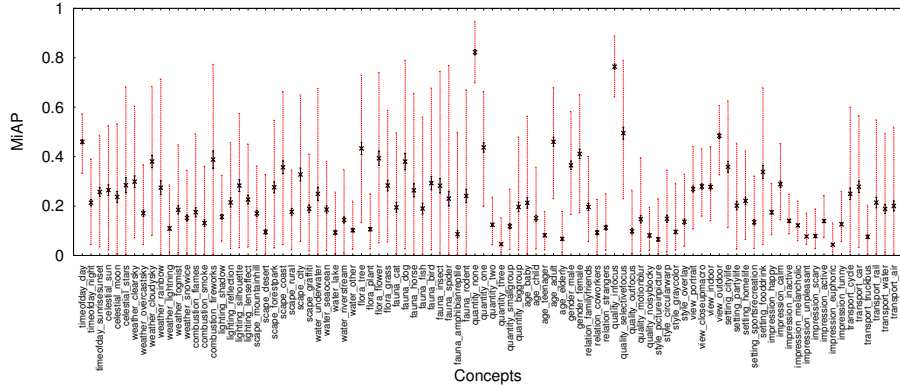


Fig. 5: Summary of the per concept annotation accuracy for all runs combined. The black bar lengths represent the standard deviations in accuracy, whereas the red bar lengths represent the minimum and maximum accuracy achieved. The concepts are list in order as specified in Appendix A.

7 Photo retrieval evaluation

The runs submitted by the participants for the retrieval subtask contained for each query a list of images they believed were the most relevant, where the list was ranked based on their included relevance scores. The number of images that could be returned for a query was restricted to be at most 1000, although eventually we only focused on the top 100 most relevant images returned by each team for each query. Due to the large collection of 200 thousand images within which the participants were to find all relevant images for a particular query, we did not obtain ground truth relevance assessments for each query-image combination beforehand. As already briefly mentioned in Section 4.2, we formed a *pool* for each query by aggregating all images that at least one team considered to be relevant for that query. For these pools we then obtained the relevance assessments using crowdsourcing, which we then used to evaluate the runs of each team. In this section we present only an evaluation for a selection of the runs, whereas detailed information on all runs can be found on the Photo Retrieval subtask website¹¹.

7.1 Evaluation measures

To assess the performance of the runs submitted by the teams, we used the following evaluation measures:

Mean Average Precision (MAP): This evaluation measure is the same as earlier defined in Section 6.1. In the analysis of the retrieval runs we focus on the non-interpolated MAP, although for completeness we also report the interpolated MAP values in the detailed results available on the website.

Geometric Mean Average Precision (GMAP): This evaluation measure is the same as earlier defined in Section 6.1. In the analysis of the retrieval runs we do not focus on GMAP at all, although we report both the interpolated and the non-interpolated GMAP values in the detailed results.

AP@X: This evaluation measure reports the average precision obtained once a certain number of images has been encountered. We have computed the scores for values ranging from 10 to 100 in steps of 10, although in the analysis of the retrieval runs we focus only on AP@10, AP@20 and AP@100.

7.2 Results

As with the annotation subtask, the techniques of the participants could use one of three possible configurations, namely textual features only, visual features only or a multimodal combination of both. For this subtask, 7 teams submitted in total 47 runs, of which 21 runs exclusively used textual features, 4 runs exclusively used visual features and 22 runs used a multimodal approach.

¹¹ <http://imageclef.org/2012/photo-flickr/retrieval/>

In addition, none of the runs this year included a manual intervention in the query generation step, such as explicitly specifying a boolean connection between concepts or using relevance feedback, and instead all retrieved the images in a completely automated fashion. We present the overall evaluation results according to the MnAP, AP@10, AP@20 and AP@100 in Table 7 to get an understanding of the best results indiscriminate of the features used, where as before in the Feature column the letter T refers to the textual configuration, V to the visual configuration and M to the multimodal configuration, while in the Type column the letter M refers to the manual query specification and the letter A to the automatic query specification. In the tables, the ranks indicate the position at which the best run appeared in the results based on MnAP. To compare the runs using the same configuration we present separate results for the textual features in Table 8, the visual features in Table 9 and the multimodal features in Table 10.

Table 7: Summary of the retrieval results for the evaluation per query for the best overall run per team.

	Team	Rank	MnAP	AP@10	AP@20	AP@100	Feature	Type
	IMU	1	0.0933	0.0187	0.0338	0.1715	T	A
	MLKD	10	0.0702	0.0214	0.0342	0.1495	M	A
	KIDS NUTN	19	0.0313	0.0051	0.0077	0.0729	M	A
	UNED	20	0.0295	0.0116	0.0223	0.0819	M	A
	MSATL	31	0.0138	0.0044	0.0077	0.0547	T	A
	ReDCAD	32	0.0129	0.0003	0.0042	0.0475	T	A
	REGIM	36	0.0031	0.0022	0.0039	0.0164	T	A

From the results we can immediately see that 9 runs of the IMU team were better than any run of the other teams, where all these runs addressed the subtask using textual features only. The runner-up MLKD was close in terms of performance to IMU and their 9 mainly multimodal teams were better than those of the remaining teams. Regarding performance, we recognize that the retrieval subtask was difficult, considering the highest MnAP was less than 0.1 and thus on average for every 10 images retrieved only at most one of them would be relevant to the query. However, if we look at the results from a user perspective, where search results are typically shown by the search engine in individual pages containing 10 or 20 images each, then based on the average precision at recall results we would have to conclude that in most instances the first two pages of results usually would not contain a single relevant image.

One of the most demanding aspects of the queries was that they were not composed of just a linear combination of individual concepts, but often had additional nuances associated with them that could be explained in multiple ways and constraints that required additional modeling. For example, the search query `close-up red roses` implicitly required the detection of the concepts `view_closeupmacro`, `quality_selectivefocus` and `flora_flower`, where additional constraints were placed on the shape (i.e. rose) and color (i.e. red) of the flower. Also, the query `flying airplane` had the remark that the airplanes

should not be taking off or landing, where photos of planes with their landing gear down but depicted in the middle of the sky were relevant or not was a point of contention amongst the annotators.

If we compare the runs per configuration we can see that even the best visual run did not come close to the performance of most of the textual or multimodal runs, indicating that textual features played an important role in addressing the query-based retrieval. Whereas the visual features could detect the individual concepts reasonably well, as judged by the results of the annotation subtask, they proved to be inadequately able to deal with additional nuances and constraints associated with the queries.

Table 8: Summary of the retrieval results for the evaluation per query for the best textual run per team.

	Team	Rank	MnAP	AP@10	AP@20	AP@100	Type
	IMU	1	0.0933	0.0187	0.0338	0.1715	A
	MLKD	10	0.0534	0.0111	0.0222	0.1335	A
	UNED	12	0.0250	0.0004	0.0019	0.0729	A
	MSATL	14	0.0138	0.0044	0.0077	0.0547	A
	ReDCAD	15	0.0129	0.0003	0.0042	0.0475	A
	REGIM	18	0.0025	0.0009	0.0024	0.0169	A

Table 9: Summary of the retrieval results for the evaluation per query for the best visual run per team.

	Team	Rank	MnAP	AP@10	AP@20	AP@100	Type
	MLKD	1	0.0244	0.0098	0.0176	0.0751	A
	IMU	2	0.0045	0.0030	0.0064	0.0316	A
	REGIM	3	0.0031	0.0022	0.0039	0.0164	A

Table 10: Summary of the retrieval results for the evaluation per query for the best multimodal run per team.

	Team	Rank	MnAP	AP@10	AP@20	AP@100	Type
	MLKD	1	0.0702	0.0214	0.0342	0.1495	A
	KIDS NUTN	8	0.0313	0.0051	0.0077	0.0729	A
	UNED	9	0.0295	0.0125	0.0206	0.0848	A
	REGIM	17	0.0020	0.0005	0.0019	0.0154	A

Analyzing the individual queries, as is shown in Figure 6, we can see that the average accuracy was highest for the query `skyline fireworks` with an average MnAP of 0.13, followed by `horse riding and full moon`. The highest accuracy was obtained for the query `hot air balloon` with a MnAP of 0.66. Yet, all queries had a minimum accuracy of zero, indicating that for each query at least one of the submitted runs was not able to retrieve any relevant images at all. Overall, whereas the average performance is quite low, the highest accuracies are much better although still far away from perfect.

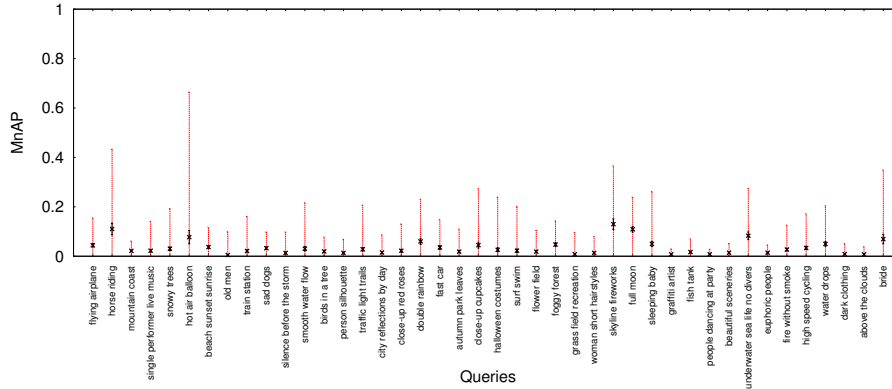


Fig. 6: Summary of the per query retrieval accuracy for all runs combined. The black bar lengths represent the standard deviations in accuracy, whereas the red bar lengths represent the minimum and maximum accuracy achieved. The queries are list in order as specified in Appendix B.

8 Conclusions

ImageCLEF’s Flickr Photo Annotation and Retrieval task is a multi-label classification challenge that offered a benchmark for testing novel visual concept detection, annotation and retrieval algorithms on a public collection containing photos gathered from the social sharing website Flickr. The task could be addressed by analyzing the textual features and/or visual features of the images in the dataset. The aim of the annotation subtask was to automatically annotate the images with one or more semantic concepts, whereas for the retrieval subtask the goal was to retrieve the most relevant images for concept-oriented search queries. The concepts and queries were to a certain extent similar to those used last year, although we removed several of them based on feedback from former participants and added new or refined existing ones based on an inspection of image search logs to provide a more realistic context for the task.

This year a total of 100 teams registered for the task, while eventually 18 teams together submitted 80 annotation runs and 7 teams together submitted 47 retrieval runs. The results indicate that the annotation subtask, like previous year, could be solved reasonably well, with the top runs achieving a MiAP of over 0.4 using multimodal features and a micro-F1 score of over 0.55 using visual or multimodal features. All in all, if we were to pick the best technique for each individual concept, i.e. we would only look at the maximum obtained accuracy of each concept, then for the majority of concepts an accuracy between 0.6 and 0.8 was achieved, which can certainly be called promising in light of the challenging set of concepts and the ambiguity involved in evaluating them. The retrieval subtask proved to be more difficult compared to last year, with the best run using textual features and achieving a MnAP of just under 0.1. None of the runs involved explicit manual intervention and thus the search queries were all automatically executed, which is somewhat surprising considering that last

year the manual runs performed generally better than the automatic runs. In comparison with the annotation subtask, the multiple concepts, nuances and constraints that were embedded into the queries made the retrieval subtask notably harder to solve.

Even though this year we had placed more emphasis on collecting more reliable ground truth annotations, the performance of crowdsource workers does not reach the same level as that of professionally trained editors, and as such the annotations may still have contained inaccuracies. In terms of evaluating the annotations, it is not necessarily an optimal strategy to apply the majority vote rule to the crowdsourced relevance assessments due to the subjective nature of many concepts, where the truth of whether a concept is present in an image or not is flexible and may depend on the viewer. Furthermore, it would be worthwhile looking into ‘gamifying’ the collection of the relevance assessment [31] to further boost their quality.

For next year’s task we plan to work more closely together with the participants, while at the same time reaching further out to what the world is really searching for, in order to redefine the aims of the annotation and retrieval subtasks and optimize the ground truth collection. This may mean breaking away from the set of concepts that have been the focus of this task during the previous years, and instead coming up with more realistic scenarios that can instantly be applied to the grand challenges our research community faces and have instant worldwide impact. Nonetheless, whereas several hurdles still will need to be crossed to get closer to a perfect precision, the contributions of the participants to both subtasks have raised the bar and advanced the state of the art in visual concept detection, annotation and retrieval.

Acknowledgements

We would like to express our deepest gratitude to the European Science Foundation for their financial support, which made the collection of the ground truth possible. Furthermore, we would like to thank the editors of the National Institute of Standards and Technology for assisting with the gold standard creation for the retrieval subtask. We were able to use their services through a successful collaboration with the TREC crowdsourcing track, kindly arranged by Gabriella Kazai from Microsoft Research. Finally, we are very grateful to the many crowdsource workers who performed the actual relevance assessments, as well as to the Flickr users whose photos we used in this task.

References

1. M.J. Huiskes and M.S. Lew. The MIR Flickr retrieval evaluation. In *Proceedings of the 10th ACM Conference on Multimedia Information Retrieval*, pages 39–43, Vancouver, BC, Canada, 2008.
2. M.J. Huiskes, B. Thomee, and M.S. Lew. New trends and ideas in visual concept detection. In *Proceedings of the 11th ACM Conference on Multimedia Information Retrieval*, pages 527–536, Philadelphia, PA, USA, 2010.

3. M. Villegas and R. Paredes. Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
4. M. Everingham, L. van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
5. H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.
6. B Thomee, E.M. Bakker, and M.S. Lew. TOP-SURF: A visual words toolkit. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1473–1476, Firenze, Italy, 2010.
7. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, USA, 2007.
8. D.G Lowe. Distinctive image features from scale-invariant keypoints. *Springer International Journal of Computer Vision*, 60(2):91–110, 2004.
9. G.J. Burghouts and J.M. Geusebroek. Performance evaluation of local color invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.
10. K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
11. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Springer International Journal of Computer Vision*, 42(3):145–175, 2001.
12. L. Huang and Y. Liu. BUAA AUDR at ImageCLEF 2012 Photo Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
13. A. Znaidia, A. Shabou, A. Popescu, and H. Le Borgne. CEA LIST’s participation to the Concept Annotation Task of ImageCLEF 2012. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
14. E. Mantziou, G. Petkos, S. Papadopoulos, C. Sagonas, and Y. Kompatsiaris. CERTH’s participation at the photo annotation task of ImageCLEF 2012. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
15. M. Rischka and S. Conrad. DBRIS at ImageCLEF 2012 Photo Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
16. B. Daróczy, D. Siklosi, and A.A. Benczúr. DMS-SZTAKI @ ImageCLEF 2012 Photo Annotation. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
17. X. Yan, W. Wu, G. Gao, and Q. Lu. IMU @ ImageCLEF 2012. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
18. M. Manzato. The participation of IntermediaLab at the ImageCLEF 2012 Photo Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.

19. Y. Ushiku, H. Muraoka, S. Inaba, T. Fujisawa, K. Yasumoto, N. Gunji, T. Higuchi, Y. Hara, T. Harada, and Y. Kuniyoshi. ISI at ImageCLEF 2012: Scalable System for Image Annotation. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
20. B.-C. Chien, G.-B. Chen, L.-J. Gaou, C.-W. Ku, R.-S. Huang, and S.-E. Wang. KIDS-NUTN at ImageCLEF 2012 Photo Annotation and Retrieval Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
21. N. Liu, E. Dellandrea, L. Chen, A. Trus, C. Zhu, Yu Zhang, C.-E. Bichot, S. Bres, and B. Tellez. LIRIS-Imagine at ImageCLEF 2012 Photo Annotation task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
22. E. Spyromitros-Xioufis, G. Tsoumakas, and I. Vlahavas. MLKD's Participation at Imageof the 2012 Conference and Labs of the Evaluation Forum Photo Annotation and Concept-based Retrieval Tasks. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
23. S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo. BossaNova at ImageCLEF 2012 Flickr Photo Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
24. L. Piras, R. Tronci, G. Murgia, and G. Giacinto. The PRA and AmILAB at ImageCLEF 2012 Photo Flickr Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
25. A. Hatem, M. Torjmen Khemakhem, and M. Ben Jemaa. Applying LDA in contextual image retrieval. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
26. F. Rim, F. Ghada, A. Ksibi A., Ben Ammar, and C. Ben Amar. REGIMvid at Imageof the 2012 Conference and Labs of the Evaluation Forum: Concept-based Query Refinement and Relevance-based Ranking. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
27. M. Pitu, D. Grijincu, and A. Iftene. UAIC participation at ImageCLEF 2012 Photo Annotation Task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
28. J. Benavent, A. Castellanos, X. Benavent, A. Garcia-Serrano, and E. de Ves Cuenca. Visual Concept Features and Textual Expansion in a Multimodal System for Concept Annotation and Retrieval with Flickr Photos at Imageof the 2012 Conference and Labs of the Evaluation Forum. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
29. J. Sánchez-Oro, S. Montalvo, A. Duarte, A.S. Montemayor, R. Cabido, J.J. Pantrigo, V. Fresno, and R. Martínez. URJCyUNED at ImageCLEF 2012 Photo Annotation task. In *Working Notes of the 2012 Conference and Labs of the Evaluation Forum*, Rome, Italy, 2012.
30. F. McSherry and M. Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *Proceedings of the 30th European Conference on Information Retrieval*, pages 414–421, Glasgow, Scotland, 2008.
31. C. Eickhoff, C.G. Harris, A.P. de Vries, and P. Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th ACM International Conference on Research and Development in Information Retrieval*, pages 871–880, Portland, OR, USA, 2012.

A Concept descriptions

In this appendix we list all concept definitions, which were part of the instructions we gave to the crowdsourcing workers. Between parenthesis we list the number of images in which a particular concept was considered to be present in the training and testing sets, respectively, based on applying the majority vote rule to the relevance assessments of the workers.

0 timeofday_day (4897,3325)

The picture shows that it was taken during the day.

1 timeofday_night (685,431)

The picture shows that it was taken during the night.

2 timeofday_sunrisesunset (508,348)

The picture shows that it was taken during the transition from night to day or from day to night, i.e. during sunrise, sunset, dusk, dawn or twilight.

3 celestial_sun (368,224)

The picture shows the sun. If only the effects of the sun are visible, e.g. clouds are lit up by the sunshine but the sun itself is not visible, then you should not mark the image as showing the sun.

4 celestial_moon (101,68)

The picture shows the moon.

5 celestial_stars (44,25)

The picture shows the stars in the sky.

6 weather_clearsky (1105,705)

The picture shows a sky that is completely clear of clouds, although a small amount of white puffs of water vapor in a clear sky is still acceptable.

7 weather_overcastsky (694,433)

The picture shows a sky that is completely covered in densely packed clouds, although a small amount of visible sky is acceptable.

8 weather_cloudysky (1196,812)

The picture shows a sky containing one or more large solid clouds. In principle it is the middle ground between a clear sky and an overcast sky.

9 weather_rainbow (33,18)

The picture shows a rainbow.

10 weather_lightning (167,125)

The picture shows a lightning strike.

11 weather_fogmist (168,100)

The picture shows a cloud of water, dust or sand particles suspended in the atmosphere at or near the earth's surface that obscures or restricts visibility. Note that this at times may look similar to smoke, so please do not confuse the two concepts.

12 weather_snowice (100,91)

The picture shows snow or ice. This also includes whiteness as a result of frost.

13 combustion_flames (68,35)

The picture shows flames emitted from a fire source.

14 *combustion_smoke* (71,47)

The picture shows smoke emitted from a fire source, such as smoke coming out of chimneys, cigarettes and airplanes.

15 *combustion_fireworks* (54,18)

The picture shows exploding fireworks.

16 *lighting_shadow* (861,576)

The picture shows a sharp and clearly visible shadow of something in the scene. This means that we are not looking for soft shadows, vague shadows and small shadows, and we are also not looking for shadows of something not visible in the scene.

17 *lighting_reflection* (448,273)

The picture shows a sharp and clearly visible reflection of something in the scene. This means that we are not looking for soft reflections, vague reflections and small reflections, and we are also not looking for reflections of something not visible in the scene.

18 *lighting_silhouette* (475,314)

The picture shows a silhouette of something in the scene. A silhouette refers to the dark shape and outline of someone or something visible against a lighter background. You should not be able to see any details of the shape due to it being so dark. For a silhouette to be present there must be some kind of background light visible. Pay attention with black and white images, because you may easily confuse actual dark shapes with silhouettes.

19 *lighting_lenseffect* (530,344)

The picture shows that the light sources visible in the image have been affected in some way. We are particularly looking for lens flares, where circular lighting effects are visible in the image, halos, where spiky lighting effects are visible around the light source, and bokeh, where the light sources are severely blurred.

20 *scape_mountainhill* (295,218)

The picture shows a mountain or hill.

21 *scape_desert* (73,36)

The picture shows a desert, containing sandy or rocky plains.

22 *scape_forestpark* (451,303)

The picture shows a forest or park, typically containing many trees and/or grass. We are not interested in people's gardens/backyards.

23 *scape_coast* (766,436)

The picture shows a coast, where the sea meets land. This includes photos showing cliffs, rocks sticking out of the sea close to land and the beach. In principle, it needs to be clear the image was taken at the coast or of the coast.

24 *scape_rural* (361,237)

The picture shows a landscape of the countryside, typically showing an open view of a rural or agricultural environment with at most a few man-made objects like cottages and small roads.

25 *scape_city* (906,572)

The picture shows a view of the city from the inside or the outside. We are interested in photos that let you get a clear impression of the city, so this means we are not looking for photos focusing on specific things in the city like shops, people and cars; rather, the scene will typically be taken with a zoomed out lens to capture as much of the scene as possible.

26 *scape_graffiti* (324,184)

The picture shows a large piece of graffiti that is the dominant feature of the scene.

27 *water_underwater* (53,44)

The picture shows an underwater scene.

28 *water_seaocean* (369,197)

The picture shows a sea or ocean.

29 *water_lake* (135,75)

The picture shows a lake.

30 *water_riverstream* (181,115)

The picture shows a river or stream.

31 *water_other* (399,255)

The picture shows liquid water not belonging to the other water categories. It can appear in all shapes and forms, such as a glass of water, water droplets, or a puddle of rain.

32 *flora_tree* (2129,1343)

The picture shows a tree or a close-up of a tree.

33 *flora_plant* (419,262)

The picture shows an indoor or outdoor plant, which typically has many leaves and small or no flowers. This includes cacti and close-ups of plants. As a rule of thumb, if you yourself would put what you see in a vase, then it is a flower. If you would put it in a pot, then it is a plant.

34 *flora_flower* (719,508)

The picture shows an indoor or outdoor plant, which typically has small or no leaves and a big flower. This includes close-ups of flowers. As a rule of thumb, if you yourself would put what you see in a vase, then it is a flower. If you would put it in a pot, then it is a plant.

35 *flora_grass* (858,548)

The picture shows a field of grass or a close-up of grass.

36 *fauna_cat* (106,72)

The picture shows a cat-like animal. This includes wild cats such as lions and cheetahs.

37 *fauna_dog* (361,267)

The picture shows a dog-like animal. This includes wild dogs such as wolves.

38 *fauna_horse* (64,40)

The picture shows a horse-like animal. This includes animals such as donkeys.

39 *fauna_fish* (49,39)

The picture shows a fish-like animal. This includes animals such as sharks. Photos showing underwater creatures that do not look like typical fish, such as jellyfish, seahorses, tortoises, etc. should not be marked with this concept.

40 *fauna_bird* (352,219)

The picture shows a bird-like animal. This includes animals such as pelicans, flamingos, ducks, geese and swans.

41 *fauna_insect* (137,114)

The picture shows an insect-like animal. This includes animals such as flies, wasps, bees, butterflies and moths.

42 *fauna_spider* (16,11)

The picture shows a spider-like animal.

43 *fauna_amphibianreptile* (40,27)

The picture shows an amphibian-like or reptile-like animal. This includes animals such as lizards, chameleons, frogs and crocodiles.

44 *fauna_rodent* (59,46)

The picture shows a rodent-like animal. This includes animals such as squirrels, hamsters, mice and rats.

45 *quantity_none* (10335,6989)

The picture shows no people.

46 *quantity_one* (3084,1990)

The picture shows one person.

47 *quantity_two* (682,432)

The picture shows two people.

48 *quantity_three* (203,127)

The picture shows three people.

49 *quantity_smallgroup* (313,239)

The picture shows a small group of people (4-9 persons).

50 *quantity_largegroup* (383,223)

The picture shows a large group of people (10+ persons).

51 *age_baby* (81,81)

The picture shows a baby (0-2 years of age).

52 *age_child* (400,256)

The picture shows a child (2-10 years of age).

53 *age_teenager* (313,220)

The picture shows a teenager (10-18 years of age).

54 *age_adult* (3536,2306)

The picture shows an adult (18-65 years of age).

55 *age_elderly* (225,127)

The picture shows an elderly person (65+ years of age)

56 *gender_male* (2484,1660)

The picture shows a male person.

57 *gender_female* (2619,1721)

The picture shows a female person.

58 *relation_familyfriends* (816,563)

The picture shows people that are likely friends or family of each other.

59 *relation_coworkers* (239,136)

The picture shows people that are likely co-workers of each other.

60 *relation_strangers* (335,212)

The picture shows people that likely do not know each other.

61 *quality_infocus* (9639,6421)

The picture shows a scene of which most, if not all, of the content is in focus. As a rule of thumb, the photographer that took the photo wanted to capture everything in the scene and did not focus on anything specifically.

62 *quality_selectivefocus* (3549,2293)

The picture shows a scene that partly is very much in focus and partly very much out of focus. You can clearly distinguish between an area in the image that is the 'foreground' and another area that is the 'background'. As a rule of thumb, the photographer that took the photo wanted to capture one particular part of the scene in particular, which is the part that is in focus, whereas the other part is out of focus.

63 *quality_outfocus* (100,83)

The picture shows a scene of which most, if not all, of the content is out of focus. As a rule of thumb, the photographer that took the photo made a mistake and did not properly set the lens focus, so the scene is a bit or very much blurred.

64 *quality_motionblur* (287,176)

The picture shows a scene of which part or all is blurred as a result of the camera moving or part of the scene moving while the photo was taken. This includes long exposure photos resulting in light trails. Typically when the camera was moved the entire scene looks streaky, whereas if part of the scene moved then only that part looks streaky.

65 *quality_noisyblocky* (318,199)

The picture is of low quality and very noisy or blocky. This is not because the scene is out of focus, but because the image was shot at a very low resolution or it was compressed afterwards. It may also be taken in low light conditions, introducing lots of tiny specks in the image.

66 *style_pictureinpicture* (113,64)

The picture is divided into multiple different photos.

67 *style_circularwarp* (167,141)

The picture shows a scene that is distorted/warped, so that straight lines in the scene look curved/round in the image and give a circular effect.

68 *style_graycolor* (306,219)

The picture shows a scene where most of the content is shown in black-and-white (grayscale), but only a small part is shown in its original color(s).

69 *style_overlay* (567,371)

The picture contains a piece of text or a logo that the photographer has added to the photo after it was taken, for example a copyright statement. Thus the text or logo was not present in the original scene.

70 *view_portrait* (1533,1069)

The picture shows a scene where one or more persons are the center of attention, typically facing the camera and aware that a photo is being taken of them. The photo normally captures at least their entire face, although a small part of it may be missing.

71 *view_closeupmacro* (2340,1589)

The picture shows a close-up of objects, where a lot of zoom has been used by the photographer, and includes macro shots, where things are shown much larger than they normally are. In contrast with a portrait a close-up can be of anything and not just people, although a photo showing only a face would be called a portrait.

72 *view_indoor* (2061,1399)

The picture shows an indoor scene.

73 *view_outdoor* (4856,3259)

The picture shows an outdoor scene.

74 *setting_citylife* (1676,1128)

The picture shows a typical scene from life in the city or town, showing for instance streets, shops, restaurants or bars. The picture may also show people doing their regular things such as shopping, talking and commuting. The image must clearly show it was taken in a city or town.

75 *setting_partylife* (368,256)

The picture shows scenes related to parties or celebrations, where people for instance are dancing, drinking or chatting, a music band is performing on stage or it may even show party-related equipment. Photos showing parties are typically taken during night time, although this does not always have to be the case.

76 *setting_homelife* (945,645)

The picture shows scenes of things taking place in or around one's home, such as having dinner, watching television, lying in bed, reading a book in the garden or even playing with the cat. The image must clearly show it was taken in or around someone's home.

77 *setting_sportsrecreation* (506,283)

The picture shows a scene where people are doing or watching sports or are enjoying themselves in a relaxed way, such as lying on the beach. Note that recreation is not the same as having or being at a party; a party is typically an organized event for dancing or celebrating, whereas recreation is more the opposite.

78 *setting_fooddrink* (626,430)

The picture shows a scene where food and/or drink play an important role. A photo showing only a little bit of food or drink, which does not have particular focus in the scene, should not be marked with this concept.

79 *impression_happy* (1146,840)

The picture shows a scene that looks happy and/or gives you a happy and warm feeling. Other words you can associate with this concept are joy and pleasure.

80 *impression_calm* (2119,1441)

The picture shows a scene that looks calm, quiet, peaceful and/or relaxed. Other words you can associate with this concept are soothing and comforting.

81 *impression_inactive* (1262,877)

The picture shows a scene where nothing exiting happens and it expresses a sense of timelessness. Other words you can associate with this concept are boring and passive.

82 *impression_melancholic* (880,594)

The picture shows a scene that looks gloomy and depressed, giving you a sense of sadness and darkness. An example would for instance be the feeling you get when you miss someone.

83 *impression_unpleasant* (623,447)

The picture shows a scene that feels uncomfortable, troublesome or nasty, giving you a general sense of unpleasantness. Other words you can associate with this concept are displeasing and dreadful.

84 *impression_scary* (377,278)

The picture shows a scene that is sinister or terrifying. Other words you can associate with this concept are spooky and horrifying.

85 *impression_active* (1087,735)

The picture shows a scene that is dynamic, colorful and gives you energy. Other words you can associate with this concept are uplifting, lively, sporty and busy.

86 *impression_euphoric* (189,140)

The picture shows a scene that bursts with energy and joy, giving you the impression of being on top of the world. Other words you can associate with this concept are excited, ecstatic and blissful.

87 *impression_funny* (765,557)

The picture shows a scene looks comical and makes you laugh.

88 *transport_cycle* (220,142)

The picture shows an unmotorized or motorized bike. This includes bikes such as push-bikes, scooters and motorbikes. A bike typically has two wheels, but can have three wheels or even have a sidecar attached.

89 *transport_car* (500,321)

The picture shows a car. This includes sedans, convertibles, vans, landrovers, suburbans, pick-up trucks, race cars and even taxis. In principle, vehicles that belong to this category have four wheels and have as main purpose the transport of only a small number of people or personal materials. Note that we are not looking for photos taken from within the vehicle, but only from the outside.

90 *transport_truckbus* (69,44)

The picture shows a truck or bus. In principle, vehicles that belong to this category have six wheels or more and have as main purpose the transport of many people or goods. Note that we are not looking for photos taken from within the vehicle, but only from the outside.

91 *transport_rail* (93,61)

The picture shows a rail vehicle, such as a train, tram and metro. Note that we are not looking for photos taken from within the vehicle, but only from the outside.

92 *transport_water* (187,127)

The picture shows a water vehicle, such as a rowing boat, sailing boat, yacht, navy ship and freighter. Note that we are not looking for photos taken from within the vehicle, but only from the outside.

93 *transport_air* (89,50)

The picture shows an air vehicle, such as an airplane and helicopter. Note that we are not looking for photos taken from within the vehicle, but only from the outside.

B Query descriptions

In this appendix we list all query definitions, which were part of the instructions we gave to the crowdsourcing workers.

0 flying airplane

The user is looking for photos showing one or more airplanes flying in the sky. He is not looking for photos that show airplanes on the ground, taking off or landing, nor for pictures of airplanes from the inside.

1 horse riding

The user is looking for photos showing one or more persons riding horses, which includes people riding horses at a rodeo. He is not interested in pictures of people on the ground next to a horse.

2 mountain coast

The user is looking for photos showing mountains or hills right besides the coast or lake.

3 single performer live music

The user is looking for photos showing a single person performing live on stage, so only one person should be visible that is singing, mixing, or playing an instrument.

4 snowy trees

The user is looking for photos of trees that are covered in snow or frost, where the scene contains more than one tree.

5 hot air balloon

The user is looking for photos showing one or more hot air balloons.

6 beach sunset sunrise

The user is looking for photos taken at the beach during sunset or sunrise.

7 old men

The user is looking for photos showing only one or more elderly men, so no other people should be additionally visible.

8 train station

The user is looking for photos showing a train stopping at or traveling through a station. He is not looking for photos showing the train from the inside. Pictures of metros and trams are also fine.

9 sad dogs

The user is looking for photos showing one or more dogs that look sad. He is particularly looking for pictures exhibiting selective focus, where the dogs are in focus and the background is out of focus.

10 silence before the storm

The user is looking for photos of an overcast or fogged over day at sea, where it seems it could start storming any moment.

11 smooth water flow

The user is looking for photos showing a stream of water. The picture should be taken with a long exposure time, so that the flowing water has become motion blurred.

12 birds in a tree

The user is looking for photos of one or more birds sitting on a tree branch.

13 person silhouette

The user is looking for photos showing the silhouette of a person, where the silhouette captures all, or almost all, of the entire person.

14 traffic light trails

The user is looking for photos showing light trails made by traffic at night. The picture effectively is taken with a long exposure time and thus the trails form straight or smoothly curving lines following the flow of traffic.

15 city reflections by day

The user is looking for photos taken during the day showing part of a city and its reflection in a large body of still water. The water surface ideally is like a mirror to give a crystal clear reflection, although some diffusion in the reflection is acceptable.

16 close-up red roses

The user is looking for photos showing one or more red roses. The picture should have selective focus, where the focus is on the roses and the background is out of focus. The roses shown should only be red and not have other colors.

17 double rainbow

The user is looking for photos showing a double rainbow. Pictures where the second rainbow is faintly but still distinctly visible are fine.

18 fast car

The user is looking for photos of a fast road car, such as a Porsche, Lambourghini, Ferrari etc. He is also interested in images of sports cars on a circuit. The picture may show motion blur, although this does not necessarily have to be the case.

19 autumn park leaves

The user is looking for photos of a park in autumn, where the trees have yellow /reddish leaves. Ideally the ground is covered in those leaves as well, although this is not necessary.

20 close-up cupcakes

The user is looking for close-up photos of one or more cupcakes.

21 halloween costumes

The user is looking for inspiration what to wear for Halloween and wants to find photos showing one or more people wearing a costume.

22 surf swim

The user is looking for photos showing a sea or lake where one or more persons are doing sports, such as surfing or swimming, or are relaxing, such as simply floating or bathing.

23 flower field

The user is looking for photos showing a field of flowers.

24 foggy forest

The user is looking for photos showing a park or forest where fog is present. He is not interested in seeing smoke.

25 grass field recreation

The user is looking for photos showing one or more people on a grass field.

26 woman short hairstyles

The user is looking for portraits showing a single teenage or adult woman with short hair.

27 skyline fireworks

The user is looking for night time photos showing a city skyline with exploding fireworks.

28 full moon

The user is looking for photos showing a full moon. He is interested in both pictures showing a close-up of the moon as well as the moon being in the background of a scene, as long as the moon is clearly visible and practically in the full moon stage.

29 sleeping baby

The user is looking for photos showing a sleeping (calm, quiet) baby.

30 graffiti artist

The user is looking for photos showing a person and a graffiti artwork.

31 fish tank

The user is looking for photos showing one or more fish in a tank or bowl.

32 people dancing at party

The user is looking for photos showing a large group of people dancing to music.

33 beautiful sceneries

The user is looking for photos showing a scenery where no people are visible. The scenery, for example a landscape, cityscape, seascape or mountainscape, should be in focus and not a close-up of anything. To ensure a good photographer took the picture, there should be a small copyright notice visible in one of the corners of the image.

34 underwater sea life no divers

The user is looking for photos showing sea life under the surface, so at least one aquatic animal should be visible. He is not interested in seeing photos containing divers.

35 euphoric people

The user is looking for photos showing one or more persons that are exceptionally happy or euphoric.

36 fire without smoke

The user is looking for photos showing a fire or explosion where no smoke is visible.

37 high speed cycling

The user is looking for photos showing one or more people cycling through the city at high speeds, so the picture should show high amounts of motion blur.

38 water drops

The user is looking for photos showing one or more water drops in selective focus, where the water drops may be accompanied by a splash or stream of water, as long as the water drops are clearly recognizable. The picture does not necessarily have to be a close-up.

39 dark clothing

The user is looking for photos showing one or more persons wearing mostly black or dark gray clothes, so there should be no people visible wearing other colored clothing. Small parts may be colored, such as shoes for instance, but the persons outer layers should be made of dark fabric.

40 above the clouds

The user is looking for photos that were taken at high altitude, above cloud level. These pictures can be taken from space or from airplanes for instance. The picture should clearly show that the clouds were below the photographer when he or she took the photo. If there are multiple layers of clouds present in the image the user is satisfied if at least one layer of clouds is shown below the vantage point. The photo may contain other objects such as the engine or wings of a plane.

41 bride

The user is looking for photos showing a bride. No other people should be visible in the picture, unless it is the groom.