Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask

Mauricio Villegas,[†] Roberto Paredes[†] and Bart Thomee[‡]

[†] ITI/DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mvillegas,rparedes}@iti.upv.es

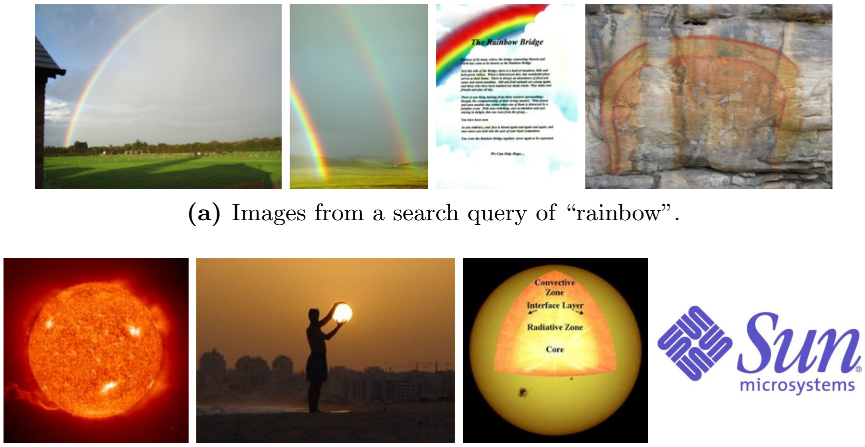
[‡] Yahoo! Research
Avinguda Diagonal 177, 08018 Barcelona, Spain
bthomee@yahoo-inc.com

Abstract. The ImageCLEF 2013 Scalable Concept Image Annotation Subtask was the second edition of a challenge aimed at developing more scalable image annotation systems. Unlike traditional image annotation challenges, which rely on a set of manually annotated images as training data for each concept, the participants were only allowed to use automatically gathered web data instead. The main objective of the challenge was to focus not only on the image annotation algorithms developed by the participants, where given an input image and a set of concepts they were asked to decide which of them were present in the image and which ones were not, but also on the scalability of their systems, such that the concepts to detect were not exactly the same between the development and test sets. The participants were provided with web data consisting of 250,000 images, which included textual features obtained from the web pages on which the images appeared, as well as various visual features extracted from the images themselves. To evaluate the performance of the submitted systems a development set was provided containing 1,000 images that were manually annotated for 95 concepts and a test set containing 2,000 images that were annotated for 116 concepts. In total 13 teams participated, submitting a total of 58 runs, most of which significantly outperformed the baseline system for both the development and test sets, including for the test concepts not present in the development set and thus clearly demonstrating potential for scalability.

1 Introduction

Automatic concept detection within images is a challenging and as of yet unsolved research problem. Over the past decades impressive improvements have been achieved, albeit admittedly not yet successfully solving the problem. Yet, these improvements have been typically obtained on datasets for which all images have been manually, and thus reliably, labeled. For instance, it has become common in past image annotation benchmark campaigns [10,16] to use crowdsourcing approaches, such as the Amazon Mechanical Turk[1], in order to let mul-

[1] www.mturk.com



(a) Images from a search query of "rainbow".

(b) Images from a search query of "sun".

Fig. 1: Example of images retrieved by a commercial image search engine.

tiple annotators label a large collection of images. Nonetheless, crowdsourcing is expensive and difficult to scale to a very large amount of concepts. The image annotation datasets furthermore usually include exactly the same concepts in the training and test sets, which may mean that the evaluated visual concept detection algorithms are not necessarily able to cope with detecting additional concepts beyond what they were trained on. To address these shortcomings a novel image annotation task [20] was proposed last year for which automatically gathered web data was to be used for concept detection, where the concepts varied between the evaluation sets. The aim of that task was to reduce the reliance of cleanly annotated data for concept detection and rather focus on uncovering structure from noisy data, emphasizing the importance of the need for scalable annotation algorithms able to determine for any given concept whether or not it is present in an image. The rationale behind the scalable image annotation task was that there are billions of images available online appearing on webpages, where the text surrounding the image may be directly or indirectly related to its content, thus providing clues as to what is actually depicted in the image. Moreover, images and the webpages on which they appear can be easily obtained for virtually any topic using a web crawler. In existing work such noisy data has indeed proven useful, e.g. [17,22,21].

The second edition of the scalable image annotation task is what is presented in this overview paper, which is one of several ImageCLEF benchmark campaigns [3]. The paper is organized as follows. In Section 2 we describe the task in more detail, which includes introducing the dataset that was created specifically for this challenge, the baseline system and the evaluation measures. In Section 3 we then present and discuss the results submitted by the participants. Finally, we conclude the paper with final remarks and future outlooks in Section 4.

2 Overview of the Subtask

2.1 Motivation and Objectives

Image concept detection generally has relied on training data that has been manually, and thus reliably, annotated, which is an expensive and laborious endeavor that cannot easily scale. To address this issue, the ImageCLEF 2013 scalable annotation subtask concentrated exclusively on developing annotation systems that rely only on automatically obtained data. A very large amount of images can be easily gathered from the web, and furthermore, from the webpages that contain the images, text associated with them can be obtained. However, the degree of relationship between the surrounding text and the image varies greatly. Moreover, the webpages can be of any language or even a mixture of languages, and they tend to have many writing mistakes. Overall the data can be considered to be very noisy.

To illustrate the objective of the evaluation, consider for example that someone searches for the word “rainbow” in a popular image search engine. It would be expected that many results be of landscapes in which in the sky a rainbow is visible. However, other types of images will also appear, see Figure 1a. The images will be related to the query in different senses, and there might even be images that do not have any apparent relationship. In the example of Figure 1a, one image is a text page of a poem about a rainbow, and another is a photograph of an old cave painting of a rainbow serpent. See Figure 1b for a similar example on the query “sun”. As can be observed, the data is noisy, although it does have the advantage that this data can also handle the possible different senses that a word can have, or the different types of images that exist, such as natural photographs, paintings and computer-generated imagery.

In order to handle the web data, there are several resources that could be employed in the development of scalable annotation systems. Many resources can be used to help match general text to given concepts, amongst which some examples are stemmers, word disambiguators, definition dictionaries, ontologies and encyclopedia articles. There are also tools that can help to deal with noisy text commonly found on webpages, such as language models, stop word lists and spell checkers. And last but not least, language detectors and statistical machine translation systems are able to process webpage data written in various languages.

In summary, the goal of the scalable image annotation subtask was to evaluate different strategies to deal with noisy data, so that the unsupervised web data can be reliably used for annotating images for practically any topic.

2.2 Challenge Description

The subtask[2] consisted of the development of an image annotation system given training data that only included images crawled from the Internet, the corresponding webpages on which they appeared, as well as precomputed visual and

[2] Subtask website at http://imageclef.org/2013/photo/annotation

textual features. As mentioned in the previous section, the aim of the subtask was for the annotation systems to be able to easily change or scale the list of concepts used for image annotation. Apart from the image and webpage data, the participants were also permitted and encouraged to use any other automatically obtainable resources to help in the processing and usage of the training data. However, the most important rule was that the systems were not permitted to use any kind of data that had been explicitly and manually labeled for the concepts to detect.

For the development of the annotation systems, the participants were provided with the following:

- A training dataset of images and corresponding webpages compiled specifically for the subtask, including precomputed visual and textual features (see Section 2.3).
- Source code of a simple baseline annotation system (see Section 2.4).
- Tools for computing the appropriate performance measures (see Section 2.5).
- A development set of images with ground truth annotations (including precomputed visual features) for estimating the system performance.

After a period of two months, a test set of images was released that did not include any ground truth labels. The participants had to use their developed systems to predict the concepts for each of the input images and submit these results to the subtask organizers. A maximum of 6 submissions (also referred to as *runs*) were allowed per participating group. Since one of the objectives was that the annotation systems be able to scale or change the list of concepts for annotation, the list of concepts for the test set was not exactly the same as those for the development set. The development set consisted of 1,000 images labeled for 95 concepts, and the test set consisted of 2,000 images labeled for 116 concepts (the same 95 concepts for development and 21 more).

To observe the possible overfitting of the development set and the difference of performance with respect to the test set, the participants were also required to submit the concept predictions of the development set, using exactly the same system and parameters as for the test set.

The concepts to be used for annotation were defined as one or more WordNet synsets [4]. So, for each concept there was a concept name, the type (either noun or adjective), the synset offset(s), and the sense number(s). Defining the concepts this way, made it straightforward to obtain the concept definition, synonyms, hyponyms, etc. Additionally, for most of the concepts, a link to a Wikipedia article about the respective concept was provided. The complete list of concepts, as well as the number of images in both the development and test sets, is included in Appendix A.

2.3 Dataset

The dataset[3] used was mostly the same as the one in ImageCLEF 2012 for the first edition of this task [20]. To create the dataset, initially a database of

[3] Dataset available at http://risenet.iti.upv.es/webupv250k

over 31 million images was created by querying Google, Bing and Yahoo! using words from the Aspell English dictionary [19]. The images and corresponding webpages were downloaded, taking care to avoid data duplication. Then, a subset of 250,000 images (to be used as the training set) was selected from this database by choosing the top images from a ranked list. The motivation for selecting a subset was to provide smaller data files that would not be so prohibitive for the participants to download/handle, and because a limited amount of concepts had to be chosen for evaluation. The ranked list was generated by retrieving images from our database using a manually defined list of concepts, in essence more or less as if the search engines had only been queried for these concepts. From this ranked list, some types of problematic images were removed, and it was guaranteed that each image had at least one webpage in which they appeared. Unlike the training set, the development (1,000 images) and test (2,000 images) sets were manually selected and labeled for the concepts being evaluated. For further details on how the dataset was created, please refer to [20].

The 250,000 training set images were exactly the same as the ones for ImageCLEF 2012. However, some images from the development and test sets had been changed. To guaranty that the visual features were the same for the new images, due to changes in software versions, the features were recalculated and therefore are different from those supplied in the previous edition of this subtask. Also this year the original images and webpages were provided. The most significant change of the dataset with respect to 2012 was the labeling of the development and test sets, where the images have now been labeled and linked to concepts in WordNet [4], thus making it much easier to automatically obtain more information for each concept. Moreover, for most of the concepts a corresponding Wikipedia article was additionally supplied, which may prove to be a useful resource.

Textual Data: Since the textual data was to be used only during training, it was only provided for the training set. Four sets of data were made available to the participants. The first one [4] was the list of words used to find the image when querying the search engines, along with the rank position of the image in the respective query and search engine it was found on. The second set of textual data[4] contained the image URLs as referenced in the webpages they appeared in. In many cases the image URLs tend to be formed with words that relate to the content of the image, which is why they can also be useful as textual features. The third set of data were the webpages in which the images appeared, for which the only preprocessing was a conversion to valid XML just to make any subsequent processing simpler. The final set of data[4] were features obtained from the text extracted near the position(s) of the image in each webpage it appeared in.

To extract the text near the image, after conversion to valid XML, the script and style elements were removed. The extracted text were the webpage title and all the terms closer than 600 in word distance to the image, not including the

[4] This textual data was identical to the 2012 edition [20].

HTML tags and attributes. Then a weight $\mathrm{s}(t_n)$ was assigned to each of the words near the image, defined as

$$\mathrm{s}(t_n) = \frac{1}{\sum_{\forall t \in \mathcal{T}} \mathrm{s}(t)} \sum_{\forall t_{n,m} \in \mathcal{T}} F_{n,m} \, \mathrm{sigm}(\mathrm{d}_{n,m}) \ , \tag{1}$$

where $t_{n,m}$ are each of the appearances of the term t_n in the document $\mathcal{T}$, $F_{n,m}$ is a factor depending on the DOM (e.g. title, alt, etc.) similar to what is done in the work of La Cascia et al. [7], and $\mathrm{d}_{n,m}$ is the word distance from $t_{n,m}$ to the image. The sigmoid function was centered at 35, had a slope of 0.15 and minimum and maximum values of 1 and 10 respectively. The resulting features include for each image at most the 100 word-score pairs with the highest scores.

Visual Features: Seven types of visual features were made available to the participants. Before feature extraction, images were filtered and resized so that the width and height had at most 240 pixels while preserving the original aspect ratio. The first feature set *Colorhist* consisted of 576-dimensional color histograms extracted using our own implementation. These features correspond to dividing the image in 3 × 3 regions and for each region obtaining a color histogram quantified to 6 bits. The second feature set *GETLF* contained 256-dimensional histogram based features. First, local color-histograms were extracted in a dense grid every 21 pixels for windows of size 41 × 41. Second, these local color-histograms were randomly projected to a binary space using 8 random vectors and considering the sign of the resulting projection to produce the bit. Thus, obtaining a 8-bit representation of each local color-histogram that can be considered as a *word*. Finally, the image is represented as a bag-of-words, leading to a 256-dimensional histogram representation. The third set of features consisted of *GIST* [11] descriptors. The other four feature types were obtained using the colorDescriptors software [15]. Features were computed for *SIFT*, *C-SIFT*, *RGB-SIFT* and *OPPONENT-SIFT*. The configuration was dense sampling with default parameters and a hard assignment 1,000 codebook using a spatial pyramid of 1×1 and 2×2 [8]. Since the vectors of the spatial pyramid were concatenated, this resulted in 5,000-dimensional feature vectors. Keeping only the first fifth of the dimensions would be like not using the spatial pyramid. The codebooks were generated using 1.25 million randomly selected features and the k-means algorithm.

2.4 Baseline Systems

A toolkit was supplied to the participants as a performance reference for the evaluation, as well as to serve as a starting point. This toolkit included software that computed the evaluation measures (see Section 2.5) and the implementations of two baselines. The first baseline was a simple random, which is important since any system that gets worse performance than random is useless. The other baseline, referred to as Co-occurrence Baseline, was a basic technique that gives better performance than random, although it was simple enough to give the participants a wide margin for improvement. In the latter technique, when given an input image, obtains its nearest $K = 32$ images from the training set using only the 1,000 bag-of-words C-SIFT visual features and the L1 norm. Then, the textual features corresponding to these K nearest images are used to derive a score for each of the concepts. This is done by using a concept-word co-occurrence matrix estimated from all of the training set textual features. In order to make the vocabulary size more manageable, the textual features are first processed keeping only English words. Finally, the annotations assigned to the image are always the top 6 ranked concepts.

2.5 Performance Measures

Ultimately the goal of an image annotation system is to make decisions about which concepts to assign to given image from a predefined list of concepts. Thus to measure annotation performance what should be considered is how good are those decisions. On the other hand, in practice many annotations systems are based on estimating a score for each of the concepts and then a second technique uses these scores to finally decide which concepts are chosen. For systems of this type a measure of performance can be based only on the concept scores, which considers all aspects of the system except for the technique used for concept decisions, making it an interesting characteristic to measure.

For this task, two basic performance measures have been used for comparing the results of the different submissions. The first one is the F-measure (F_1), which takes into account the final annotation decisions, and the other is the Average Precision (AP), which considers the concept scores.

The F-measure is defined as

$$\mathrm{F}_1 = \frac{2PR}{P + R} \ , \tag{2}$$

where P is the precision and R is the recall. In the context of image annotation, the F_1 can be estimated from two different perspectives, one being concept-based and the other sample-based. In the former, one F_1 is computed for each concept, and in the latter one F_1 is computed for each image to annotate. In both cases, the arithmetic mean is used as a global measure of performance, and will be referenced as MF_1-concepts and MF_1-samples, respectively.

The AP is algebraically defined as

$$\mathrm{AP} = \frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} \frac{k}{\mathrm{rank}(k)} \ , \tag{3}$$

where $\mathcal{K}$ is the ordered set of the ground truth annotations, being the order induced by the annotation scores, and $\mathrm{rank}(k)$ is the order position of the k-th ground truth annotation. The fraction $k/\,\mathrm{rank}(k)$ is actually the precision at the k-th ground truth annotation, and has been written like this to be explicit on the way it is computed. In the cases that there are ties in the scores, a random permutation is applied within the ties. The AP can also be estimated for both the concept-based and sample-based perspectives, however, the concept-based AP is not a suitable measure of annotation performance (it is more adequate for a retrieval scenario), so only the sample-based AP has been considered in this evaluation. As a global measure of performance, also the arithmetic mean is used, which will be referred to as MAP-samples.

A bit of care must be taken when comparing systems using the MAP-samples measure. What the MAP-samples turns out saying is that if for a given image the scores are used to sort the concepts, how good would it rank the true concepts for the image. Depending on the system, its scores could or could not be optimal for ranking the concepts. Thus a system with a relatively low MAP-samples, could still have a good annotation performance if the method used to select the concepts is adequate for its concept scores. Because of this, as well as the fact that there can be systems that do not rely on scores, it was optional for the participants of the task to provide scores.

3 Evaluation Results

3.1 Participation

The participation was excellent, especially considering that this was the second edition of the task and last year there was only one participant. In total 13 groups took part, submitting 58 runs overall. The following teams participated:

- **CEA LIST:** The team from the Vision & Content Engineering group of CEA LIST (Gif-sur-Yvettes, France) was represented by Hervé Le Borgne, Adrian Popescu and Amel Znaidia.
- **INAOE:** The team from the Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, Mexico) was represented by Hugo Jair Escalante.
- **KDEVIR:** The team from the Computer Science and Engineering department of the Toyohashi University of Technology (Aichi, Japan), was represented by Ismat Ara Reshma, Md Zia Ullah and Masaki Aono.
- **LMCHFUT:** The team from Hefei University of Technology (Hefei, China) was represented by Yan Zigeng.
- **MICC:** The team from the Media Integration and Communication Center of the Università degli Studi di Firenze (Florence, Italy) was represented by Tiberio Uricchio, Marco Bertini, Lamberto Ballan and Alberto Del Bimbo.
- **MIL:** The team from the Machine Intelligence Lab of the University of Tokyo (Tokyo, Japan) was represented by Masatoshi Hidaka, Naoyuki Gunji and Tatsuya Harada.
- **RUC:** The team from the School of Information of the Renmin University of China (Beijing, China) was represented by Xirong Li, Shuai Liao, Binbin Liu, Gang Yang, Qin Jin, Jieping Xu and Xiaoyong Du.
- **SZTAKI:** The team from the Datamining and Search Research Group of the Hungarian Academy of Sciences (Budapest, Hungary) was represented by Bálint Daróczy.

Table 1: Comparison of the systems for the best submission of each group.

System	Visual Features [Total Dim.]	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
TPT [13] **#6**	Provided by organizers (All 7) [Tot. Dim. = 21312]	* Morphological expansions	Manual morphological expansions of the concepts (plural forms). Training images selected by appearance of concept in supplied textual features.	Multiple SVMs per concept, with context dependent kernels. Annotation based on threshold (the same for all concepts).
MIL [6] **#4**	Fisher Vectors (SIFT, C-SIFT, LBP, GIST) [Tot. Dim. = 262144]	* WordNet * ActiveSupport library for word singularization	Extract webpage title, image attributes, surrounding text, and singularize nouns. Label training images by appearance of concept, defined by WordNet synonyms and hyponyms with a single meaning.	Linear multilabel classifier learned by PAAPL. Annotation of the top 5 concepts.
UNIMORE [5] **#2**	Multiv. Gauss. Distrib. of local desc. (HSV-SIFT, OPP-SIFT, RGB-SIFT) [Tot. Dim. = 201216]	* WordNet * NLTK (stopwords and stemmer) * +100k training images	Stopword removal and stemming of supplied preprocessed features and webpage title. Label training images by appearance of concept, defined by WordNet synonyms and hyponyms with a single meaning. Disambiguation by negative context from other senses of concept word.	Linear SVMs learned by stochastic gradient descent. Annotation based on threshold (the same for all concepts).
RUC [9] **#4**	Provided by organizers (All 7) [Tot. Dim. = 21312]	* Search engine keywords * Flickr tags dataset	Positive training images selected by a combination of supplied textual features and search engine keywords weighted by a tag co-occurrence measure derived from a Flickr dataset. Negative examples selected by Negative Bootstrap.	Multiple staked hikSVMs and kNNs (L_1 distance). Annotation of the top 6 concepts.
UNED&UV [1] **#3**	—[a]	* Webpage of test images * WordNet * Lucene	Concept indexing (using Lucene) of WordNet *definition*, *forms*, *hypernyms*, *hyponyms* and *related* components.	Text retrieval of concepts using webpage img field. Annotation by a cut-off percentage of the maximum scored concept.
CEA LIST [2] **#4**	Bag of Visterms (SIFT) [Tot. Dim. = 8192]	* Wikipedia * Flickr tags dataset	Training images selected by ranking the images using tag models learned from Flickr and Wikipedia data. The first 100 ranked images as positive and the last 500 images as negative.	Linear SVM. Annotation of concepts with score above $\mu + \sigma$ (μ, σ are mean and standard deviation of all concept scores).
KDEVIR [12] **#1**	Provided by organizers (colorhist, C-SIFT, OPP-SIFT, RGB-SIFT) [Tot. Dim. = 15576]	* WordNet * Lucene stemmer	Stopwords and non-English words removal and stemming of supplied textual features. Matching of features with concepts defined by WordNet synonyms and application of bm25 to obtain concept scores per image.	kNN (IDsim) and aggregating concept scores (bm25). Annotation of top 10 concepts.
URJC& UNED [14] **#3**	HSV histograms, LBP and provided by organizers (C-SIFT) [Tot. Dim. = 5384]	* Search engine keywords * WordNet * English Stopwords list * Porter stemmer	Stopword removal and stemming of supplied textual features, and enriched by WordNet synonyms and hyperonyms. Generation of keywords-concepts co-occurrence matrix.	kNN (Bhattacharyya, χ^2, and L_2 distances) and aggregating concept scores (co-occurrence). Annotation based on threshold (the same for all concepts).
MICC [18] **#5**	Provided by organizers (All 7) [Tot. Dim. = 21312]	* WordNet * Wikipedia * Search engine keywords * Training image URLs	Stopword removal of supplied textual features, search engine keywords and URL extracted words. Enriched textual features with WordNet synonyms and Wikipedia link structure.	kNN (Gaussian kernel distance) and rank concepts by tagRelevance. Annotation of the top 7 concepts.
SZTAKI **#1**	Fisher Vectors [Tot. Dim. = Unknown]	* Wikipedia	Fisher vector-based learning of visual model given training images per category.	Textual ranking of images based on Wikipedia concept descriptions. Prediction via visual models.
INAOE **#3**	SIFT [Tot. Dim. = Unknown]	Unknown	Documents are represented by a distribution of occurrences over other documents in the corpus, so that documents are represented by their context, yielding a prototype per concept.	Ensemble of linear classifiers per concept.
THSSMPAM **#2**	Global: CEDD, Color, Bag of visterms (SIFT) Local: SIFT, SURF [Tot. Dim. = Unknown]	* WordNet	Unknown	Tags of NN image ranked by TF-IDF. Similarity between tags and concepts using WordNet. Annotation by bipartite graph algorithm.
LMCHFUT **#1**	Provided by organizers (SIFT) [Tot. Dim. = 5000]	Unknown	Training images selected by appearance of concept in supplied textual features.	Single SVM learned per concept given visual features of positive and negative training examples.

[a] Unlike the other systems that take as input image visual features, the UNED&UV system receives as input the image webpage.

- **THSSMPAM:** The team from Beijing, China was represented by Jile Zhou.
- **TPT:** The team of CNRS TELECOM ParisTech (Paris, France) was represented by Hichem Sahbi.
- **UNED&UV:** The team from the Universidad Nacional de Educación a Distancia (Madrid, Spain) and the Universitat de València was represented by Xaro Benavent, Angél Castellanos Gonzáles, Esther de Ves, D. Hernández-Aranda, Ruben Granados and Ana Garcia-Serrano.
- **UNIMORE:** The team from the University of Modena and Reggio Emilia (Modena, Italy) was represented by Costantino Grana, Giuseppe Serra, Marco Manfredi, Rita Cucchiara, Riccardo Martoglia and Federica Mandreoli.
- **URJC&UNED:** The team of the Universidad Rey Juan Carlos (Móstoles, Spain) and the Universidad Nacional de Educación a Distancia (Madrid, Spain) was represented by Jesús Sánchez-Oro, Soto Montalvo, Antonio Montemayor, Juan Pantrigo, Abraham Duarte, Víctor Fresno and Raquel Martínez.

In Table 1 we provide a comparison of a the key details of the best submission of each group. For a more in depth look of the annotation systems of each team, please refer to their corresponding paper listed in the table. Note that there were four groups that did not submit a working notes paper describing their system, so for those submissions less information could be listed.

3.2 Results

Table 2 presents the performance measures (mentioned in 2.5) for the baseline techniques and all of the submitted runs by the participants.The last column of the table corresponds to the MF$_1$-concepts measure which was only computed for the 21 concepts that did not appear in the development set. The systems are ordered by performance, beginning at the top with the best performing one. This order of the systems has been derived by considering for the test set the average rank when comparing all of the systems, using the MF$_1$-samples, the MF$_1$-concepts and the MF$_1$-concepts unseen in dev. measures, while breaking ties by the average of the same three performance measures.

For an easier comparison and a more intuitive visualization, the same results of Table 2 are presented as graphs in Figure 2 (only for the test set). These graphs include for each result the 95% confidence intervals. These intervals have been estimated by Wilson's method, employing the standard deviation for the individual measures (for the samples or concepts, and for the average precisions (AP) or F-measures (F$_1$), depending on the case).

Finally, in Figure 3 there is for each of the 116 test set concepts, a boxplot (or also known as box-and-whisker plot) for the F$_1$-measures when combining all runs. In order to fit all of the concepts in the same graph, for multiple outliers with the same value, only one is shown. The concepts have been sorted by the median performance of all submissions, which in a way orders them by difficulty.

Table 2: Performance measures (in %) for the baseline techniques and all submissions. The best submission for each team is highlighted in bold font.

System	MAP-samples		$\mathbf{MF_1}$-samples		$\mathbf{MF_1}$-concepts		$\mathbf{MF_1}$-concepts unseen in dev.
	dev.	test	dev.	test	dev.	test	test
Baseline OPP-SIFT	24.6	21.4	19.2	16.4	13.8	11.8	10.3
Baseline C-SIFT	24.2	21.2	18.6	16.2	10.7	10.5	10.8
Baseline RGB-SIFT	24.3	21.2	18.5	15.8	13.0	11.7	10.5
Baseline SIFT	24.0	21.0	17.8	15.9	11.0	11.0	10.1
Baseline Colorhist	22.1	19.0	16.1	13.9	8.0	8.0	9.6
Baseline GIST	20.9	17.8	14.5	12.5	6.1	6.9	7.3
Baseline GETLF	21.0	17.7	14.9	12.5	6.6	5.4	5.9
Baseline Random	10.9	8.7	6.2	4.6	4.8	3.6	2.3
TPT #6	**50.4**	**44.4**	**51.3**	**42.6**	**45.0**	**34.1**	**45.1**
TPT #4	48.9	43.2	50.7	41.8	42.5	33.7	45.3
MIL #4	**43.8**	**41.4**	**34.0**	**32.4**	**34.7**	**32.3**	**35.8**
MIL #1	44.5	42.1	34.6	33.2	35.2	32.6	33.8
MIL #2	43.1	40.7	34.3	32.7	33.9	31.8	31.4
UNIMORE #2	**46.0**	**44.1**	**27.3**	**27.5**	**34.2**	**33.1**	**34.8**
UNIMORE #5	47.9	45.6	33.3	31.5	33.7	31.9	31.9
UNIMORE #1	39.2	36.7	33.0	31.1	34.1	32.0	31.3
TPT #2	38.5	37.0	41.4	38.1	30.9	30.0	30.9
UNIMORE #6	46.0	44.1	33.0	31.1	34.1	32.0	31.3
RUC #4	**41.2**	**38.0**	**31.6**	**29.0**	**33.4**	**30.4**	**32.8**
MIL #5	42.2	39.7	34.0	31.7	33.4	30.9	30.2
MIL #3	42.5	39.6	34.2	31.8	33.4	30.2	29.5
RUC #5	40.5	37.6	31.0	28.3	32.7	29.6	31.5
UNED&UV #3	**27.1**	**26.6**	**22.5**	**23.1**	**31.5**	**31.3**	**43.2**
UNIMORE #3	43.7	41.9	23.1	23.1	32.4	31.5	35.5
UNED&UV #5	35.5	33.2	27.6	24.4	31.7	29.2	35.4
TPT #5	49.8	44.3	38.7	32.5	33.0	26.7	27.3
RUC #3	39.4	36.9	29.8	27.8	31.4	29.2	30.2
TPT #3	49.0	43.6	38.8	31.9	30.2	24.8	24.7
RUC #2	38.2	35.5	28.8	26.5	30.8	28.5	29.9
UNIMORE #4	39.7	36.2	26.8	24.1	31.7	29.5	28.0
UNED&UV #4	31.0	29.8	29.9	30.0	26.3	22.8	24.6
UNED&UV #1	32.8	30.3	25.0	23.0	27.5	25.0	31.7
RUC #1	36.1	32.4	28.8	25.4	26.6	23.9	22.7
UNED&UV #2	32.4	30.6	24.4	22.9	26.1	24.0	30.6
CEA LIST #4	**40.3**	**34.2**	**32.2**	**26.0**	**26.1**	**21.2**	**20.1**
CEA LIST #5	39.2	33.6	31.6	25.7	25.4	21.0	20.0
CEA LIST #3	40.4	34.1	31.8	25.2	25.3	20.2	20.5
CEA LIST #2	39.6	33.6	30.2	24.2	24.6	20.1	20.1
CEA LIST #1	34.6	29.4	28.7	23.0	23.6	19.0	19.8
KDEVIR #1	**28.7**	**26.1**	**25.3**	**22.2**	**21.1**	**18.0**	**17.3**
URJC&UNED #3	**32.6**	**28.1**	**27.9**	**24.1**	**19.8**	**17.3**	**14.8**
MICC #5	**29.1**	**26.2**	**22.7**	**20.0**	**21.4**	**18.0**	**18.6**
MICC #4	29.2	26.1	22.4	20.0	21.0	18.0	18.6
MICC #3	29.0	26.1	22.3	20.0	21.0	18.1	18.5
URJC&UNED #2	32.2	27.6	27.7	23.8	19.7	17.2	14.6
URJC&UNED #1	32.0	27.6	27.4	23.7	19.2	17.1	14.6
MICC #2	29.0	26.1	23.3	20.4	20.7	17.5	17.0
MICC #1	28.7	25.9	20.4	18.7	20.3	17.3	17.6
KDEVIR #3	28.6	24.8	24.8	21.1	18.7	15.9	15.6
TPT #1	38.6	36.8	30.2	23.0	24.2	19.2	8.2
KDEVIR #6	28.3	24.3	24.5	20.8	18.4	15.7	15.0
KDEVIR #4	29.2	26.4	24.7	20.5	18.5	15.4	15.3
KDEVIR #5	29.0	25.6	24.6	20.2	18.5	15.1	14.5
KDEVIR #2	26.4	23.5	25.0	20.7	19.2	14.8	12.6
SZTAKI #1	**32.9**	**28.2**	**10.4**	**9.5**	**17.7**	**16.4**	**16.7**
INAOE #3	**24.0**	**19.1**	**19.7**	**15.4**	**17.7**	**15.2**	**11.1**
SZTAKI #2	32.7	28.0	9.8	8.8	17.1	15.1	16.0
THSSMPAM #3	**20.9**	**15.9**	**17.0**	**14.8**	**13.0**	**12.7**	**11.1**
THSSMPAM #2	21.7	16.1	17.0	14.8	13.0	12.7	11.1
LMCHFUT #1	N/A[a]	N/A[a]	**12.2**	**11.0**	**13.6**	**12.1**	**11.3**
INAOE #1	21.5	17.5	21.3	16.9	9.0	6.9	5.1
THSSMPAM #1	16.3	12.0	18.2	11.8	13.7	10.0	6.6
INAOE #2	23.6	19.0	24.8	16.7	6.3	4.8	4.7
THSSMPAM #4	15.9	11.9	15.5	11.8	12.2	10.0	6.6
THSSMPAM #5	15.8	11.9	15.5	11.8	12.2	10.0	6.6
INAOE #4	17.9	8.3	15.9	6.2	11.7	3.4	2.3

[a] Concept scores not provided, only annotation decisions.

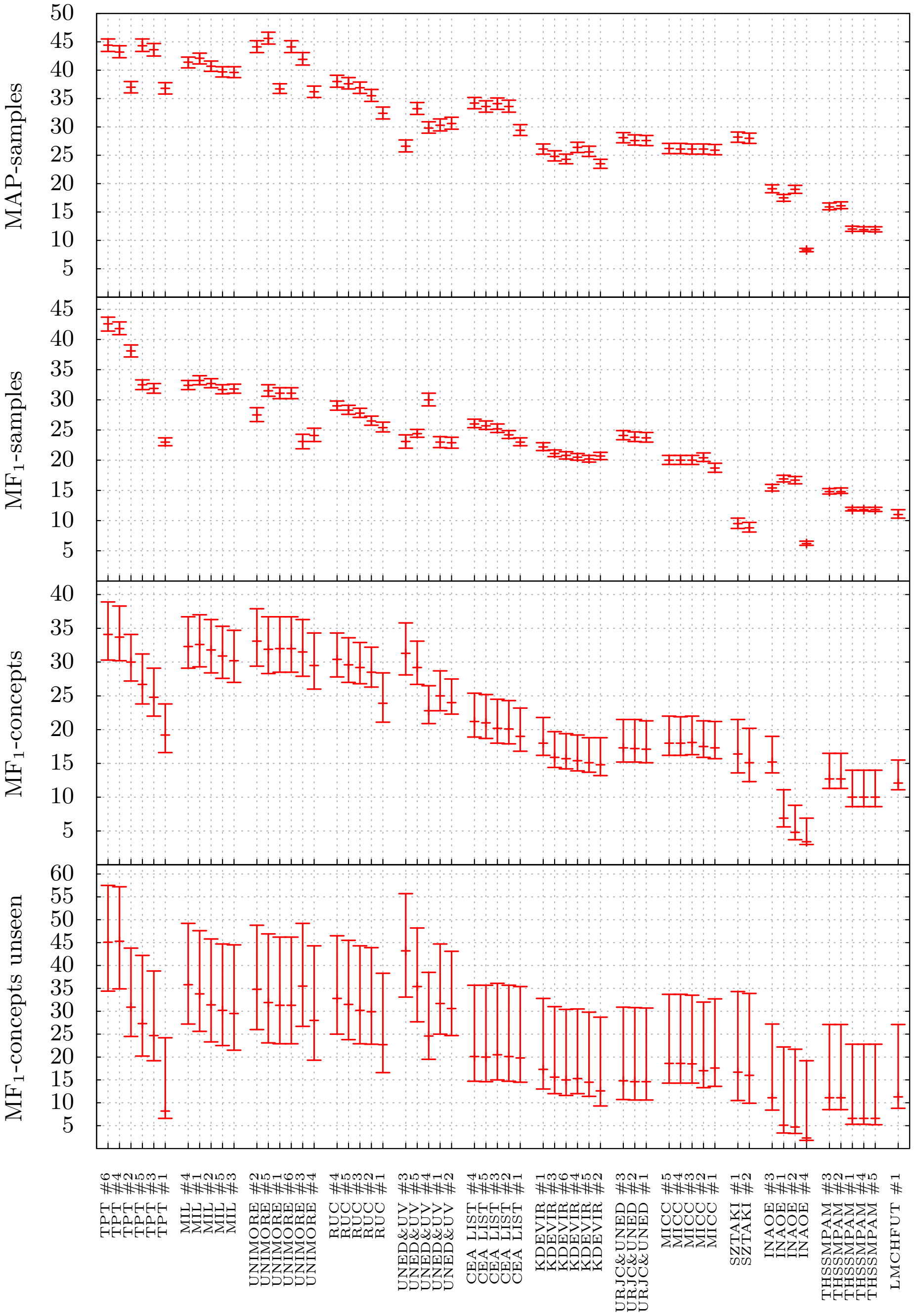


Fig. 2: Graphs showing the test set performance measures (in %) for all the submissions. The error bars correspond to the 95% confidence intervals computed using Wilson's method.

3.3 Discussion

Due to the considerable participation in this evaluation very interesting results have been obtained. As can be observed in Table 2 and Figure 2, most of the submitted runs significantly outperformed the baseline system for both the development and test sets. When analyzing the sample based performances, very large differences can be observed amongst the systems. For both MAP-samples and MF_1-samples the improvement has been from below 10% to over 40%. Moreover, the confidence intervals are relatively narrow, making the improvements quite significant. An interesting detail to note is that for MAP-samples there are several top performing systems, however, when comparing to the respective MF_1-samples measures, three of the TPT submissions clearly outperform the rest. The key difference between these is the method for deciding which concepts are selected for a given image. This leads to believe that that many of the systems could improve greatly by changing that last step of their systems. As a side note, many of the participants chose to use the same scheme as the baseline system for selecting the concepts, the top N and fixed for all images. The number of concepts per image is expected to be variable, thus making this strategy less than optimal. Future work should be addressed in this direction.

The MF_1-concepts results in Figure 2, in contrast to the sample based performances, present much wider confidence intervals. This is due to two reasons, there are fewer concepts than sample images and the performance for different concepts varies greatly (see Figure 3). This effect is even greater for the MF_1-concepts unseen, since these were only 21. Nevertheless, for MF_1-concepts unseen, the top performing systems are statistically significantly better than the baselines and some of the lower performance systems. Moreover, in Figure 3 it can be observed that the unseen concepts do not tend to perform worse. The difficulty of each particular concept affects more the performance than the fact that these have not been seen during development, or from another perspective the systems have been able to generalize rather well to the new concepts. Thus, this demonstrates potential for scalability of the systems. It would be desired for future benchmarking campaigns of this type to have more labeled data available for the evaluation, or find an alternative more automatic analysis, to be able to compare better the systems in this scalability performance aspect.

In contrast to usual image annotation evaluations with labeled training data, this challenge required work in more fronts, such as handling the noisy data, textual processing and multilabel annotations. This has given considerable freedom to the participants to concentrate their efforts in different aspects. Several teams extracted their own visual features, for which they did observe improvements with respect to the features provided by the organizers. On the other hand, for the textual processing, several different approaches were tried by the participants. Some of these teams (namely MIL, UNIMORE, CEA LIST, and URJC&UNED) reported in their working notes papers and/or as observed in the results in this paper that as more information and additional resources are used (e.g. synonyms, plus hyponyms, etc.) the performance of the systems improved. Curiously, the best performing system, TPT, only used the provided visual features and did a very simple expansion of the concepts. Overall it seems that several of the proposed ideas by the participants are complementary, and thus considerable improvements could be expected in future works.

4 Conclusions

This paper presented an overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask, the second edition of a challenge aimed at developing more scalable image annotation systems. The goal was to develop annotation systems that for training only rely on unsupervised web data and other automatically obtainable resources, thus making it easy to add or change the concepts for annotation.

Considering that it is a relatively new challenge, the participation was excellent, 13 teams submitted in total 58 system runs. The performance of the submitted systems was considerably superior to the provided baselines, improving from below 10% to over 40% for both MAP-samples and MF$_1$-samples measures. With respect to the performance of the systems when analyzed per concept, it was observed that the concepts vary greatly in difficulty. An important result was that for the concepts that were not seen during the development, the improvement was also significant, thus showing that the systems are capable of successfully using the noisy web data and generalizing well to new concepts. This clearly demonstrates potential for scalability of the systems. Finally, the participating teams presented several interesting approaches to address the proposed challenge, concentrating their efforts in different aspects of the problem. Many of these approaches are complementary, thus considerable improvements could be expected in future works.

Due to the success of this year's campaign and the very interesting results obtained, it would be important to continue organizing future editions. To be able to derive better conclusions about the performance generalization to unseen concepts, it would be desirable to have more labeled data available and/or find an alternative more automatic analysis which can help in giving more insight in this respect. Also, related challenges could be organized, for instance it could be assumed that for some concepts there is labeled data available, and find out how to take advantage of both the supervised and unsupervised data.

Acknowledgments

The authors are very grateful of the support of the CLEF campaign for the ImageCLEF initiative. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under the tranScriptorium project (#600707), the LiMoSINe project (#288024), and from the Spanish MEC under the STraDA project (TIN2012-37475-C02-01).

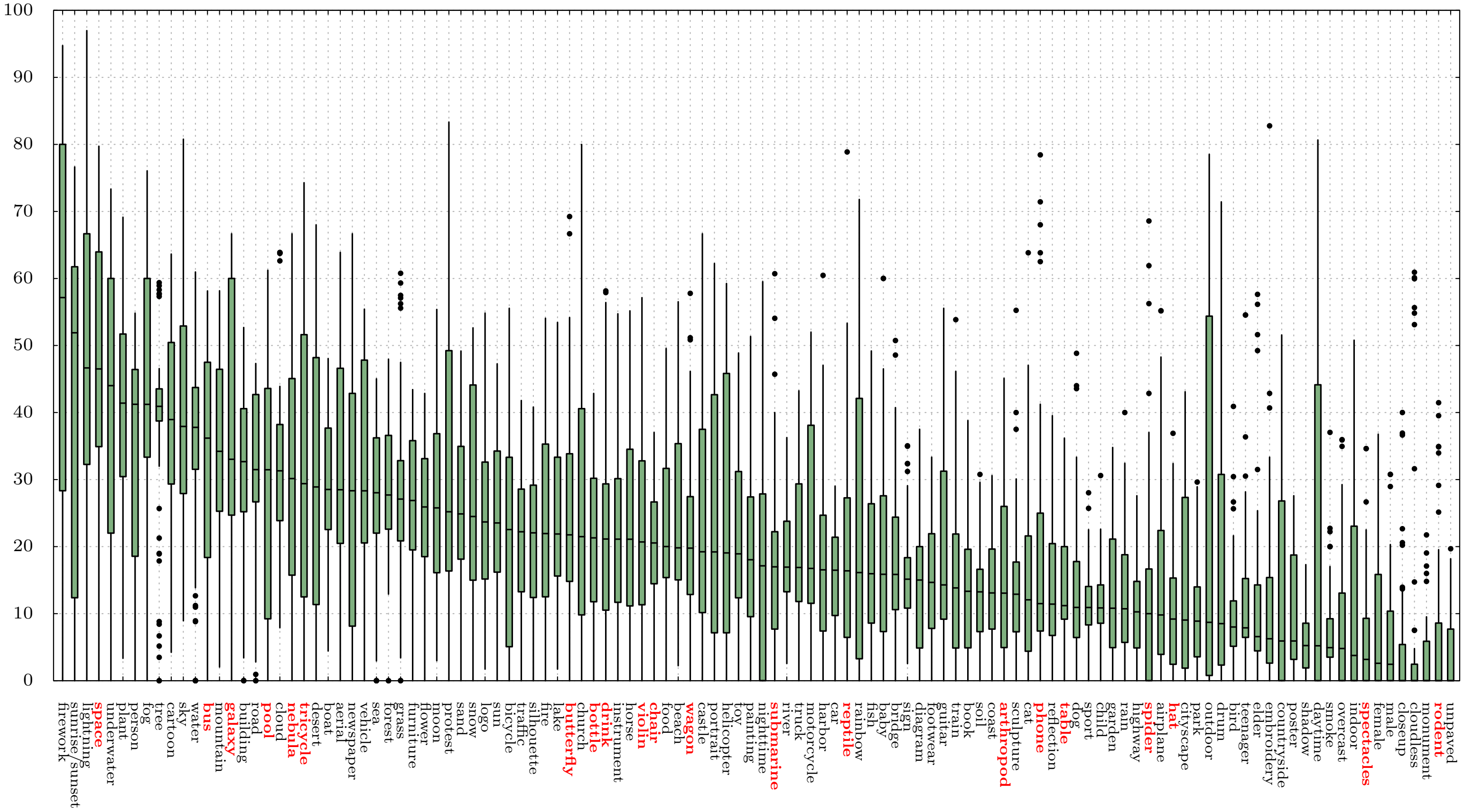


Fig. 3: Boxplots (also known as box-and-whiskers) for the test set of the per concept annotation F_1-measures (in %) for all runs combined. The plots are ordered by the median performance. Concepts in red font are the ones not seen in development.

References

1. Benavent, X., Castellanos, A., de Ves, E., Hernández-Aranda, D., Granados, R., Garcia-Serrano, A.: A multimedia IR-based system for the Photo Annotation Task at ImageCLEF2013. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
2. Borgne, H.L., Popescu, A., Znaidia, A.: CEA LIST@imageCLEF 2013: Scalable Concept Image Annotation. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
3. Caputo, B., Müller, H., Thomee, B., Villegas, M., Paredes, R., Zellhöfer, D., Goeau, H., Joly, A., Bonnet, P., Martínez-Gómez, J., García-Varea, I., Cazorla, M.: ImageCLEF 2013: the vision, the data and the open challenges. In: CLEF. Lecture Notes in Computer Science, Springer, Valencia, Spain (September 23-26 2013) [Cited on page 2]
4. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA; London (May 1998) [Cited on pages 4 and 5]
5. Grana, C., Serra, G., Manfredi, M., Cucchiara, R., Martoglia, R., Mandreoli, F.: UNIMORE at ImageCLEF 2013: Scalable Concept Image Annotation. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
6. Hidaka, M., Gunji, N., Harada, T.: MIL at ImageCLEF 2013: Scalable System for Image Annotation. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
7. La Cascia, M., Sethi, S., Sclaroff, S.: Combining textual and visual cues for content-based image retrieval on the World Wide Web. In: Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on. pp. 24–28 (1998) [Cited on page 6]
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. pp. 2169–2178. CVPR '06, IEEE Computer Society, Washington, DC, USA (2006), `http://dx.doi.org/10.1109/CVPR.2006.68` [Cited on page 6]
9. Li, X., Liao, S., Liu, B., Yang, G., Jin, Q., Xu, J., Du, X.: Renmin University of China at ImageCLEF 2013 Scalable Concept Image Annotation. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
10. Nowak, S., Nagel, K., Liebetrau, J.: The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands (2011) [Cited on page 1]
11. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision 42(3), 145–175 (May 2001), `http://dx.doi.org/10.1023/A:1011139631724` [Cited on page 6]
12. Reshma, I.A., Ullah, M.Z., Aono, M.: KDEVIR at ImageCLEF 2013 Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
13. Sahbi, H.: CNRS - TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]

14. Sánchez-Oro, J., Montalvo, S., Montemayor, A.S., Pantrigo, J.J., Duarte, A., Fresno, V., Martínez, R.: URJC&UNED at ImageCLEF 2013 Photo Annotation Task. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
15. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1582–1596 (2010) [Cited on page 6]
16. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: CLEF 2012 working notes. Rome, Italy (2012) [Cited on page 1]
17. Torralba, A., Fergus, R., Freeman, W.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30(11), 1958–1970 (nov 2008) [Cited on page 2]
18. Uricchio, T., Bertini, M., Ballan, L., Bimbo, A.D.: KDEVIR at ImageCLEF 2013 Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) [Cited on page 9]
19. Villegas, M., Paredes, R.: Image-Text Dataset Generation for Image Annotation and Retrieval. In: Berlanga, R., Rosso, P. (eds.) II Congreso Español de Recuperación de Información, CERI 2012. pp. 115–120. Universidad Politécnica de Valencia, Valencia, Spain (June 18-19 2012) [Cited on page 5]
20. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop, Online Working Notes. Rome, Italy (September 17-20 2012) [Cited on pages 2, 4, and 5]
21. Wang, X.J., Zhang, L., Liu, M., Li, Y., Ma, W.Y.: ARISTA - image search to annotation on billions of web photos. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 0, 2987–2994 (2010) [Cited on page 2]
22. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning 81, 21–35 (2010) [Cited on page 2]

A Concept List

Concept	Type	WN 3.0	WN 3.0	Wikipedia article	#images	
		sense#	offset		dev.	test
aerial	adj.	1	01380267	Aerial_photography	39	72
airplane	noun	1	02691156	Airplane	13	20
baby	noun	1	09827683	Baby	9	29
beach	noun	1	09217230	Beach	36	51
bicycle	noun	1	02834778	Bicycle	17	15
bird	noun	1	01503061	Bird	25	30
boat	noun	1	02858304	Boat	59	83
book	noun	2, 1	02870092, 06410904	Book	22	22
bridge	noun	1	02898711	Bridge	35	44
building	noun	1	02913152	Building	188	288
car	noun	1	02958343	Car	47	86
cartoon	noun	1	06780678	Cartoon	31	73
castle	noun	2	02980441	Castle	18	20
cat	noun	1	02121620	Cat	12	21
child	noun	1	09917593	Child	23	60
church	noun	2	03028079	Church_(building)	12	16
cityscape	noun	1	06209770	Cityscape	67	95
closeup	noun	1	03049695	Closeup	15	56
cloud	noun	2	09247410	Cloud	239	340
cloudless	adj.	1	00460946	-	99	159
coast	noun	1	09428293	Coast	46	64
countryside	noun	1	08645033	Countryside	43	74
daytime	noun	1	15164957	Daytime_(astronomy)	587	989
desert	noun	1	08505573	Desert	19	27
diagram	noun	1	03186399	Diagram	11	23
dog	noun	1	02084071	Dog	28	34
drum	noun	1	03249569	Drum	12	9
elder	noun	1	10048218	Elderly	12	37
embroidery	noun	2	03282933	Embroidery	10	14
female	noun	2	09619168	Female	41	149
fire	noun	3, 1	13480848, 07302836	Fire	28	34
firework	noun	1	03348454	Firework	11	20
fish	noun	1	02512053	Fish	17	33
flower	noun	2	11669335	Flower	46	111
fog	noun	2	14521648	Fog	17	39
food	noun	2, 1	07555863, 00021265	Food	20	59
footwear	noun	1, 2	03381126, 03380867	Footwear	19	40
forest	noun	1, 2	08438533, 09284015	Forest	96	129
furniture	noun	1	03405725	Furniture	52	120
garden	noun	1	03417345	Garden	14	21
grass	noun	1	12102133	Grass	162	253
guitar	noun	1	03467517	Guitar	7	13
harbor	noun	1	08639058	Harbor	20	35
helicopter	noun	1	03512147	Helicopter	8	14
highway	noun	1	03519981	Highway	15	16
horse	noun	1	02374451	Horse	18	44
indoor	adj.	1	01692786	-	87	218
instrument	noun	6	03800933	Musical_instrument	34	58
lake	noun	1	09328904	Lake	43	65
lightning	noun	1, 2	11475279, 07412993	Lightning	10	16
logo	noun	1	07272084	Logo	15	35
male	noun	2	09624168	Male	53	115
monument	noun	1	03743902	Monument	8	19
moon	noun	1	09358358	Moon	7	31
motorcycle	noun	1	03790512	Motorcycle	12	20
mountain	noun	1	09359803	Mountain	100	181
newspaper	noun	3, 1	03822171, 06267145	Newspaper	9	9

continues in next page

Concept	Type	WN 3.0	WN 3.0	Wikipedia article	#images	
		sense#	offset		dev.	test
nighttime	noun	1	15167027	Nighttime	54	90
outdoor	adj.	1, 2	01692222, 03095372	-	615	1023
overcast	noun	1	14524198	Overcast	64	71
painting	noun	1	03876519	Painting	41	82
park	noun	2	08615374	Park	19	28
person	noun	1	00007846	Person	233	538
plant	noun	2	00017222	Plant	393	694
portrait	noun	1	07202391	Portrait	10	26
poster	noun	1	06793426	Poster	6	16
protest	noun	2	01177033	Protest	9	19
rain	noun	1	11501381	Rain	12	29
rainbow	noun	1	09403427	Rainbow	9	15
reflection	noun	4, 5	04747115, 04068976	Mirror_image	63	78
river	noun	1	09411430	River	79	77
road	noun	1	04096066	Road	122	212
sand	noun	1	15019030	Sand	47	79
sculpture	noun	2	00937656	Sculpture	23	55
sea	noun	1	09426788	Sea	97	133
shadow	noun	2	08646306	Shadow	50	98
sign	noun	2	06793231	Sign	55	78
silhouette	noun	1	08613345	Silhouette	26	39
sky	noun	1	09436708	Sky	440	678
smoke	noun	1	11508092	Smoke	26	17
snow	noun	2	15043763	Snow	52	80
soil	noun	2	14844693	Soil	40	79
sport	noun	1	00523513	Sport	31	87
sun	noun	1	09450163	Sun	30	62
sunrise/sunset	noun	1, 1	15168790, 15169248	Sunrise/Sunset	36	52
teenager	noun	1	09772029	Teenager	18	27
toy	noun	1	03964744	Toy	24	31
traffic	noun	1	08425303	Traffic	24	39
train	noun	1	04468005	Train	32	24
tree	noun	1	13104059	Tree	272	451
truck	noun	1	04490091	Truck	23	38
underwater	adj.	1, 2	02472252, 00124685	Underwater	23	55
unpaved	adj.	1	01739987	-	13	21
vehicle	noun	1	04524313	Vehicle	203	374
water	noun	6	07935504	Water	288	430
arthropod	noun	1	01767661	Arthropod	-	78
bottle	noun	1	02876657	Bottle	-	32
bus	noun	1	02924116	Bus	-	45
butterfly	noun	1	02274259	Butterfly	-	16
chair	noun	1	03001627	Chair	-	63
drink	noun	1	07885223	Drink	-	53
galaxy	noun	3	08271042	Galaxy	-	21
hat	noun	1	03497657	Hat	-	78
nebula	noun	3	09366940	Nebula	-	16
phone	noun	1	04401088	Phone	-	26
pool	noun	1	03982060	Swimming_pool	-	30
reptile	noun	1	01661091	Reptile	-	34
rodent	noun	1	02329401	Rodent	-	44
space	noun	4	08500433	Outer_space	-	84
spectacles	noun	1	04272054	Spectacles	-	62
spider	noun	1	01772222	Spider	-	19
submarine	noun	1	04347754	Submarine	-	24
table	noun	2	04379243	Table_(furniture)	-	49
tricycle	noun	1	04482393	Tricycle	-	15
violin	noun	1	04536866	Violin	-	23
wagon	noun	1	04543158	Wagon	-	29