Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task

Mauricio Villegas and Roberto Paredes

PRHLT, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mauvilsa,rparedes}@prhlt.upv.es

Abstract. The ImageCLEF 2014 Scalable Concept Image Annotation task was the third 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, the participants were only allowed to use data and/or resources that as new concepts to detect are introduced do not require significant human effort (such as hand labeling). The participants were provided with web data consisting of 500,000 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 optimize their systems, the participants were provided with a development set of 1,940 samples and its corresponding hand labeled ground truth for 107 concepts. The performance of the submissions was measured using a test set of 7,291 samples which was hand labeled for 207 concepts among which 100 were new concepts unseen during development. In total 11 teams participated in the task submitting overall 58 system runs. Thanks to the larger amount of unseen concepts in the results the generalization of the systems has been more clearly observed and thus demonstrating the 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 [9,15,3] to use crowdsourcing approaches, such as the Amazon Mechanical Turk[1], in order to let multiple annotators label a large collection of images. Nevertheless, 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

[1] www.mturk.com

concepts beyond what they were trained on. To address these shortcomings a novel image annotation task [19] was proposed in 2012 for which automatically gathered web data had 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. [16,22,21].

This paper presents the overview of the third edition of the Scalable Concept Image Annotation task [19,20], one of the four benchmark campaigns organized by ImageCLEF [2] in 2014 under the CLEF initiative[2]. Section 2 describes the task in detail, including the participation rules and the provided data and resources. Followed by this, Section 3 presents and discusses the results of the submissions. Finally, Section 4 concludes the paper with final remarks and future outlooks.

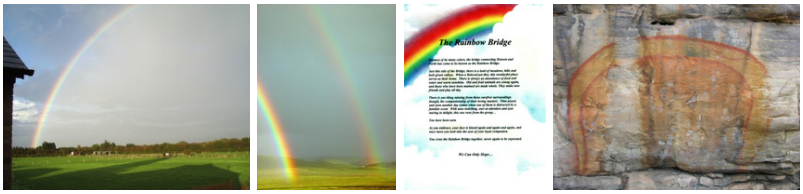
2 Overview of the Task

2.1 Motivation and Objectives

Image concept detection research generally has relied on training data that has been manually, and thus reliably annotated, an expensive and laborious endeavor that cannot easily scale as the number of concepts is increased. As an alternative to clean labeled 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. Motivated by this need for scalability and the possibility of cheaply obtaining useful data, the ImageCLEF 2014 Scalable Concept Image Annotation task concentrated exclusively on developing annotation systems that rely only on automatically obtained data.

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

[2] http://www.clef-initiative.eu



(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.

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 task 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 challenge[3] 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

[3] Challenge website at http://imageclef.org/2014/annotation

and textual features. As mentioned in the previous section, the aim of the task 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 similar datasets and 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 concept detection learning.

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 task, 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 three and a half months to work on the development set, a test set of images was released which 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 task organizers. About one month was given to work on the test data and a maximum of 10 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. Moreover, each test image had its own list of concepts to detect, so not all images had to be annotated for all the possible concepts. The development set consisted of 1,940 samples labeled for 107 concepts, and the test set consisted of 7,291 samples labeled for 207 concepts (the same 107 concepts from development and 100 additional ones).

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), 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 samples in the test sets, is included in Appendix A.

2.3 Dataset

The dataset[4] used was very similar to the one of the first two editions of the task [19,20]. To create the dataset, initially a database of over 31 million images was created by querying Google, Bing and Yahoo! using words from the Aspell

[4] Dataset available at http://risenet.prhlt.upv.es/webupv-datasets

English dictionary [18]. The images and corresponding webpages were downloaded, taking care to avoid data duplication. Then, a subset of 500,000 images (to be used as the training set) was selected from this database by choosing the top images from a ranked list. Half of this data was exactly the same as the training set from last year, the additional data was merely intended to supply images for the new concepts that were introduced. The motivation for selecting a subset was to provide smaller data files that would not be so prohibitive for the participants to download and handle. The ranked list was generated by retrieving images from our database using the list of concepts, in essence more or less as if the search engines had only been queried for these. From the 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 and test sets were manually selected and labeled for the concepts being evaluated. For further details on how the dataset was created, please refer to [19].

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 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 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 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 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. [6], 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: Before visual feature extraction, images were filtered and resized so that the width and height had at most 240 pixels while preserving the original aspect ratio. These raw resized images were provided to the participants but also seven types of precomputed visual features. 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. Then, 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* [10] descriptors. The other four feature types were obtained using the colorDescriptors software [13], namely *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 [7]. Since the vectors of the spatial pyramid were concatenated, this resulted in 5,000-dimensional feature vectors. 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. 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 amount of concepts assigned to the image is variable, the concepts selected are the ones with a score higher than the sum of the mean and the standard deviation for all the concept scores of that image. Since there were seven visual features provided, each one was considered separately for a baseline.

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_1 is defined as

$$F_1 = \frac{2PR}{P+R} \ , \tag{2}$$

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, although there was a slight decrease in participation with respect to last year. In total, 11 groups took part in the task and submitted overall 58 system runs. Among the 11 participating groups, only 7 of them submitted a corresponding paper describing their system, thus only for these there were specific details available. Last year the participation was 13 groups, 58 runs and 9 papers. The following 11 teams were the ones that participated:

- **DISA:** [1] The team from the Laboratory of Data Intensive Systems and Applications of the Masaryk University (Brno, Czech Republic) was represented by Petra Budikova, Jan Botorek, Michal Batko and Pavel Zezula.
- **IPL:** [14] The team from the Information Processing Laboraroty of the Athens University of Economics and Business (Athens, Greece) was represented by Spyridon Stathopoulos and Theodore Kalamboukis.
- **KDEVIR:** [11] 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.
- **MIL:** [5] The team from the Machine Intelligence Lab of the University of Tokyo (Tokyo, Japan) was represented by Atsushi Kanehira, Masatoshi Hidaka, Yusuke Mukuta, Yuichiro Tsuchiya, Tetsuaki Mano and Tatsuya Harada.
- **MindLab:** [17] The team from the Machine learning, perception and discovery Lab from the Universidad Nacional de Colombia (Bogotá, Colombia) was represented by Jorge A. Vanegas, John Arevalo, Sebastian Otálora, Fabián Páez, Santiago A. Pérez-Rubiano and Fabio A. González.
- **MLIA:** [23] The team from the Department of Advanced Information Technology of the Kyushu University (Fukuoka, Japan) was represented by Xing Xu, Atsushi Shimada and Rin-ichiro Taniguchi.
- **RUC:** [8] The team from the School of Information of the Renmin University of China (Beijing, China) was represented by Xirong Li, Xixi He, Gang Yang, Qin Jin and Jieping Xu.
- **FINKI:**[5] The team from the Faculty of Computer Science and Engineering of the Ss. Cyril and Methodius University (Skopje, Republic of Macedonia) was represented by Ivica Dimitrovski.
- **IMC:**[5] The team from the Institute of Media Computing of the Fudan University (Shanghai, China) was represented by Yong Cheng.

[5] No paper describing their system submitted.

– **INAOE:**[5] The team from the Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, Mexico) was represented by Hugo Jair Escalante and Luis Pellegrin.
– **NII:**[5] The team from the National Institute of Informatics (Tokyo, Japan) was represented by Duy Dinh Le.

Table 1 provides the main key details for the best submission of each group that submitted a paper describing their system. This table serves as a summary of the systems, and also is quite illustrative for quick comparisons. For a more in depth look of the annotation systems of each team, please refer to their corresponding paper.

3.2 Scalability Analysis

Since the objective of this task was to compare annotation systems that are scalable, a very important aspect to evaluate is precisely their scalability. However, unlike the annotation performance, it is difficult to quantify the scalability of a system so that the submissions can be compared in this respect. Therefore, instead of attempting to give a measure for scalability, in this section we make a few comments about possible aspects of the proposed systems in which the scalability could be compromised.

One characteristic observed in this year's task was that three teams based their system on Convolutional Neural Networks (CNN) pre-trained using ImageNet, a dataset which was manually hand labeled for 1,000 WordNet synsets. Two of the teams, MIL and MindLab, used the CNN output of an intermediate layer as a visual feature. There are works in which it has been observed that CNN features perform well in new problems different from the one that was trained for, so in some sense their use does not violate the competition rule of no hand labeled data usage. However, a minor detail is that the ImageNet synsets overlap considerably with the current task's concepts, so the annotation performance for these systems might be a bit optimistic in comparison to the others. The third team that used CNN was MLIA, which employed the synsets predicted by the CNN to clean the concepts automatically assigned using the webpage data. In this case the performance of the system could be greatly affected if the concepts for annotation differ significantly from the ones of ImageNet.

Also this year most of the teams proposed approaches based on classifiers that need to be learned. In the case of the MIL team, the classifier is multilabel. A multilabel classifier could be problematic since each time the list of concepts to detect changes, the classifier would have to be relearned. However, the PAAPL algorithm of MIL is designed with special consideration of scalability, so in their case it does not seem an issue. The alternative of multilabel is having one classifier per concept, which are learned one concept at a time using positive and negative samples. For scalability, the learning should be based on a selection of negative images so that this process is independent of how many concepts there are. It seems that all of the teams consider this adequately. However, with respect to a multilabel classifier this might not be the optimal approach. When

new concepts are introduced it could be advisable to learn new classifiers to consider the relationships between the concepts. However, this relationship could be taken into account in a step after classification which is what the KDEVIR team has done with their constructed ontologies.

3.3 Annotation Performance Results

The test set for this year was composed of 4 subsets of samples, each of which had a different list of concepts for annotation. The first subset contained 3,000 images which were exactly the same as last year's development and test sets, and the list of concepts for annotation were also the same 116. The second subset had 1,747 images and the list of concepts were 52 related to the topic *animals*. The third subset had 479 images and the list of concepts were 41 related to the topic *foods*. For both the animals and foods subsets all of the concepts for annotation were not among the ones seen in development. The final subset had 2065 images and the list of concepts for annotation were all the 207.

Table 2 presents the performance measures (mentioned in 2.5) for the baseline techniques and all of the submitted runs by the participants. The table includes the results for the complete test set (referenced as *all*) and for three of the mentioned subsets: animals, foods and the one annotated for all the 207 concepts, referenced as *ani.*, *food* and *207*, respectively. Also for the MF_1-concepts measure the *unseen* column presents the results for the complete test set, but only considering the 100 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 was derived by considering for the test set the average rank when comparing all of the systems, using the complete test set for the three performance measures and also MF_1-concepts unseen. Ties were broken by the average of the same measures. Considering only the performance measures, this ranking indicates that the best system this year was the one developed by KDEVIR.

For an easier comparison and a more intuitive visualization, the results for the complete test set and all the submissions are presented as graphs in Figure 2. In the graphs the error bars correspond to the 95% confidence intervals 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). For the MF_1-concepts measure two results for each submission is presented, one that includes all concepts and another that considers only the unseen concepts. Similarly Figure 3 presents the results for all the submissions, but in each case depicting the performance for three of the subsets of the test set: animals, foods and the 207 concepts subset. The fourth subset of the test is intended for comparison of the systems with respect to the previous edition of the task, so this is presented in a separate graph, Figure 4, although for space reasons and make the comparison more illustrative, only the best submission of each group is included.

Finally, in Figure 5 there is for each of the 207 test set concepts, a boxplot for the F_1 when combining all runs. In order to fit all of the concepts in the

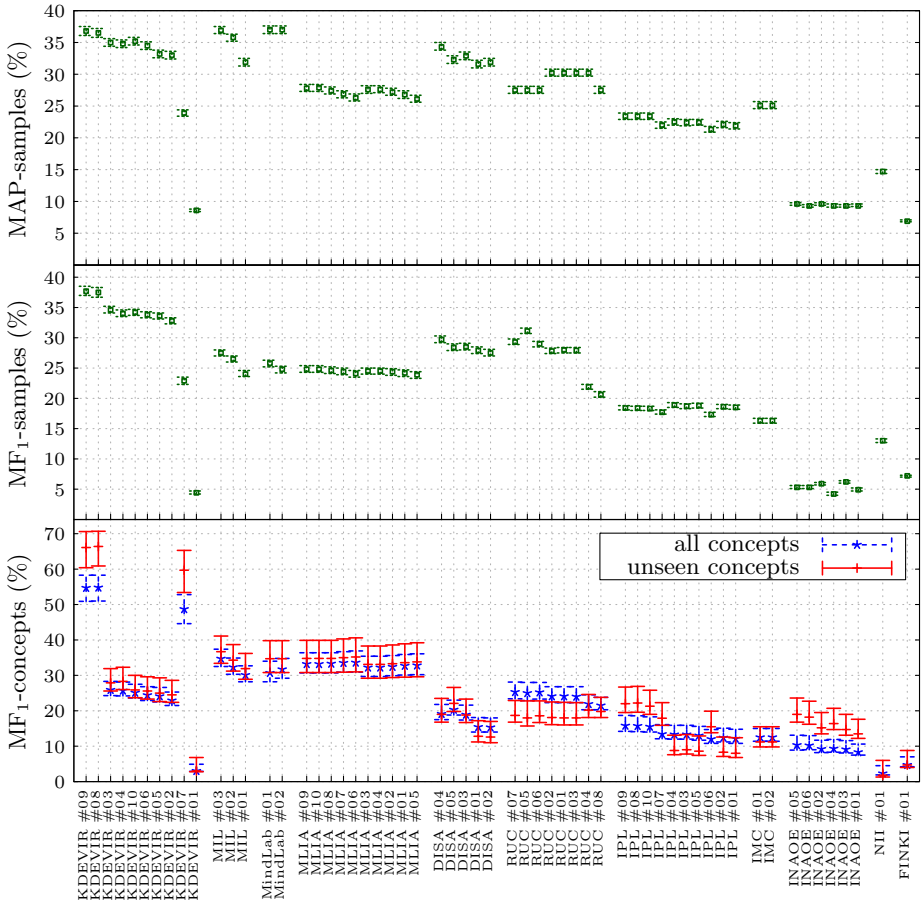


Fig. 2: Performance measures for the complete test set for all the submissions.

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.

3.4 Discussion

As can be observed in Figure 4, the performance of the systems has improved somewhat with respect to what was obtained in the previous edition of the task. This year 5 teams obtained all of the performance measures over 30% in contrast to just 3 from last year. An interesting detail is that it seams that the improvements for the MF_1 measures are greater than for the MAP-samples. Thus it can be observed that this year better approaches have been developed for making the final concept annotations decisions.

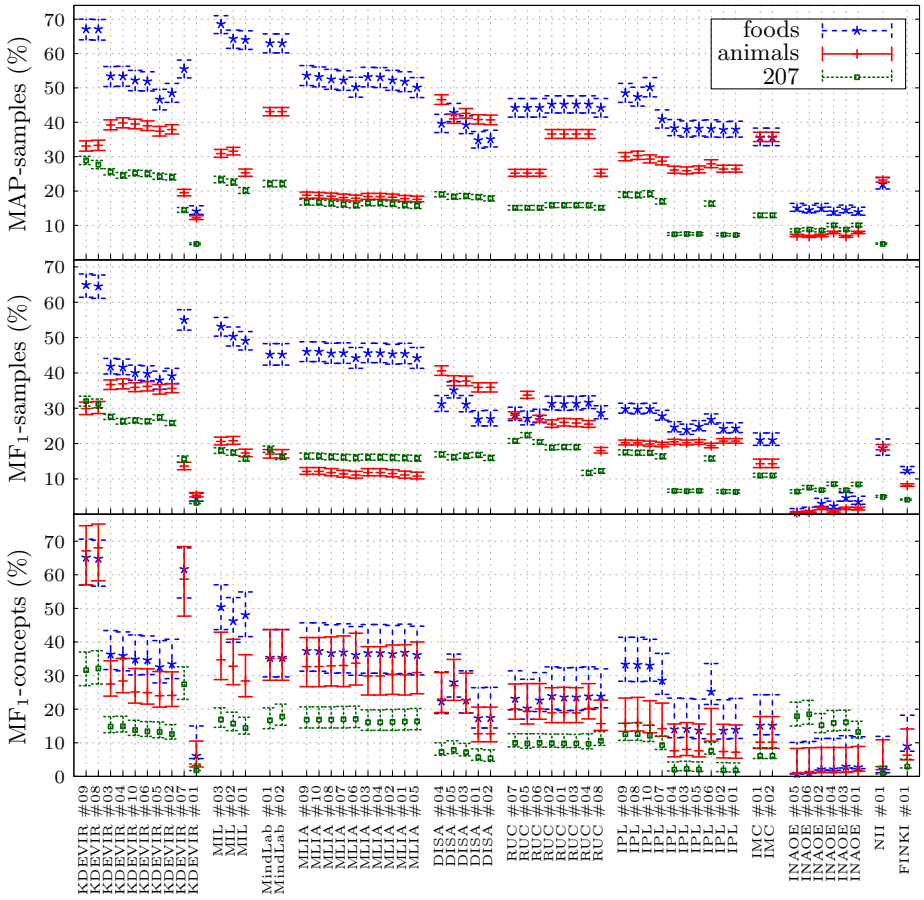


Fig. 3: Performance measures for three of the subsets in the test set for all the submissions.

Observing the results for the complete test set in Figure 2, the best MAP-samples and MF_1-samples values are somewhat lower than for last year's test set. This could be related to the fact that this year the list of concepts was larger, thus making the problem a bit more difficult. With respect to the MF_1-concepts measure, last year the results were characterized by having relatively large confidence intervals, which made drawing conclusions a bit difficult specially for the unseen concepts. The increase in the number of unseen concepts has made the results clearer. For the three measures the performance for two of the systems submitted by KDEVIR significantly outperforms all of the others. Most impressive is the advantage obtained for the MF_1-concepts measure and even more if only the unseen concepts are considered, obtaining a performance over 65%. Note that the good performance for the unseen concepts is due to the fact that many of the new concepts were used in the animals and foods subsets

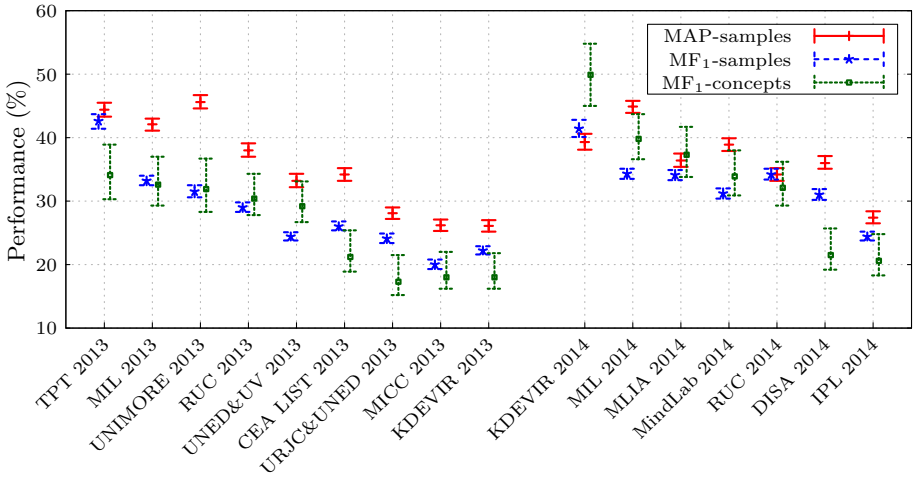


Fig. 4: Performance measures for the best submission of each group for both this and last year's edition of the task. The results for both years are for exactly the same test set.

which had smaller concept lists, making the problem a bit easier. Analyzing the key details of the systems presented in Table 1, it can be noted that the success of the KDEVIR system is most probably due to the usage of concept ontologies both in the training phase for better selecting the images used for optimizing the classifiers and in the testing phase for taking into account the relationships between the concepts. Moreover, the KDEVIR system also employed the technique of last year's winner TPT [12], which uses a learning technique that takes into account context, effectively finding a way to exploit the information available in the noisy webpage data.

Even though the foods and animals subsets consisted purely of unseen concepts, in the results in Figure 3 it can be seen that the performance for these is in general much better, mostly because of the known relationship between the size of the list of concepts for annotation and the performance. The smaller the list of concepts, the easier the problem becomes. Moreover, similar to last year in Figure 5 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 are able to generalize rather well to new concepts.

Considering both the annotation performance measures and the scalability analysis, it can be declared that this year's winner is the KDEVIR system. The fact that KDEVIR only used the provided visual features shows the characteristic of this evaluation, which in contrast to usual image annotation tasks with labeled training data, this challenge requires work in more fronts in order to get important improvements.

4 Conclusions

This paper presented an overview of the ImageCLEF 2014 Scalable Concept Image Annotation task, the third 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 cheaply obtainable resources, thus making it easy to add or change the concepts for annotation.

The participation was similar to last year although with a slight decrease, 11 teams submitted in total 58 system runs. The performance of the submitted systems was somewhat superior to last year's results, in particular improving more for the MF_1 measures, which indicate a greater success in the developed techniques for choosing the final annotated concepts. Thanks to the larger amount of concepts in the test set that were not seen during development, the results for the MF_1-concepts measure had narrower confidence intervals, so it made the comparison the systems more conclusive. Moreover, by having subsets in the test set which had to be annotated using only unseen concepts, it has been observed that the systems are able to generalize well. The clear winner of this year's evaluation was the KDEVIR [11] team, which after analyzing the key components of the system it can be observed that most of the success is due to the usage of a classifier learning technique that takes into account context, effectively finding a way to exploit the information available in the noisy webpage data; and the usage of automatically generated concept ontologies both in the training phase for better selecting the images used for optimizing the classifiers and in the testing phase for taking into account the relationships between the concepts.

The results of the task have been very interesting and show that useful annotation systems can be built using noisy web crawled data. Since the problem requires to cover many fronts, there is still a lot of work that can be done, so it would be interesting to continue this line of research. Papers on this topic should be published, demonstration systems based on these ideas be built and more evaluation of this sort be organized. Also it remains to see how this can be used to complement systems that are based on clean hand labeled data and find ways to take advantage of both the supervised and unsupervised data.

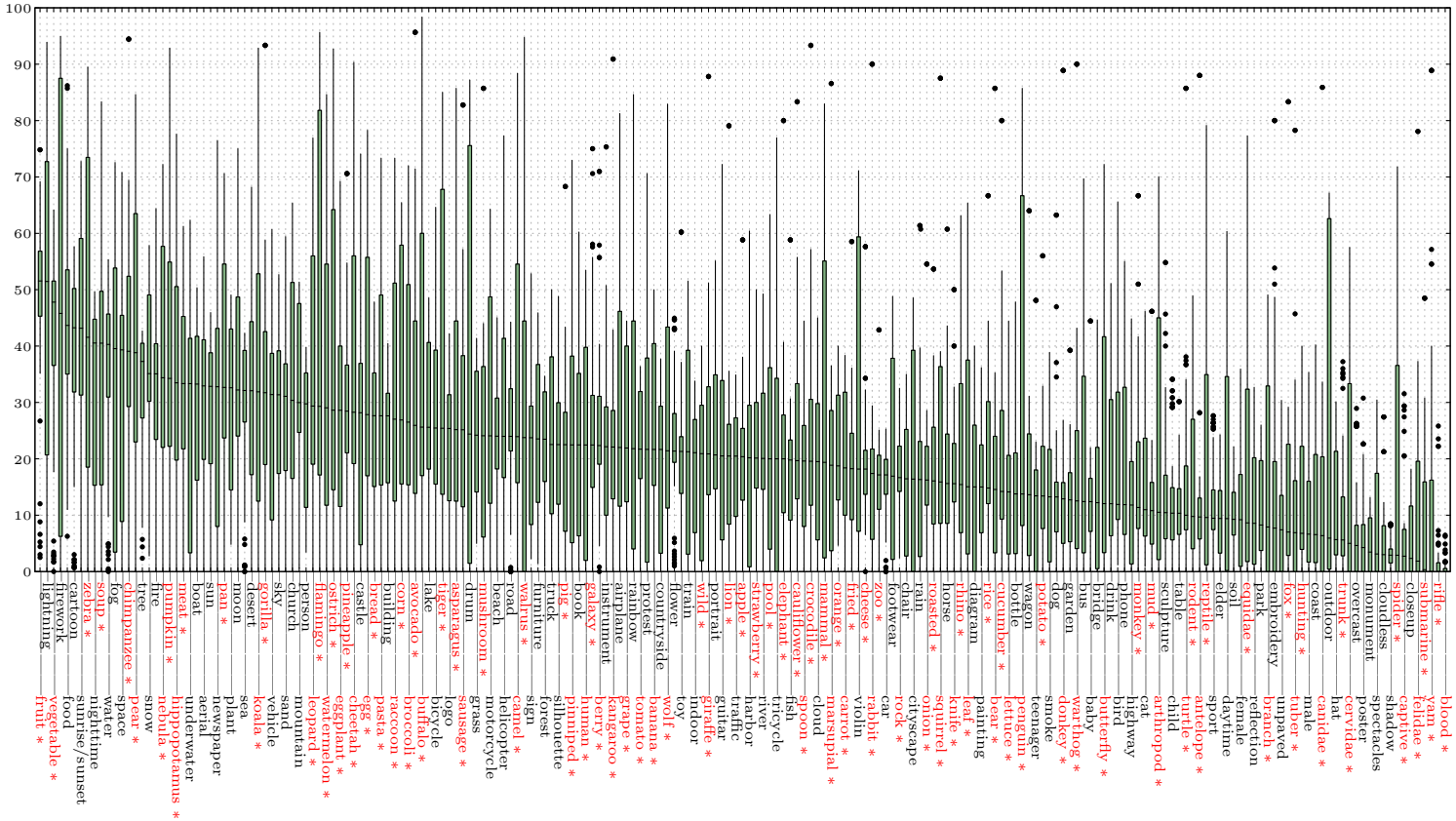


Fig. 5: Boxplots for the test set of the per concept annotation F_1 (in %) for all runs combined. The plots are ordered by the median performance. Concepts in red font and with an asterisk (∗) are the ones not seen in development.

Table 1: Key details of the best system for each of the groups that submitted a paper describing their system.

System	Visual Features [Total Dim.]	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
KDEVIR run #9 [11]	Provided by organizers (All 7) [T.Dim. = 21312]	* WordNet * Wikipedia * Pling stemmer	Ontology built per concept using WordNet and Wikipedia. Provided webpage features processed by noun and adjective detection, singularization and concept weighting based on appearance. Top-m weighted images selected and merged with predecessor concepts of highest semantic confidence according to built ontologies.	Multiple SVMs per concept with context dependent kernel. Positive and negative samples selected by exploiting constructed ontologies. Annotation of top-k weighted concepts along with predecessors of highest semantic confidence according to built ontologies.
MIL run #3 [5]	Fisher Vectors & ImageNet CNN [T.Dim. = 266240]	* WordNet * ImageNet pre-trained CNN (DeCAF)	Extract webpage title, image tag attributes and singularize nouns. Label training images by appearance of concept, defined by WordNet synonyms and hyponyms.	Linear multilabel classifier learned by PAAPL. Annotation of the 4% top scored concepts.
MindLab run #1 [17]	ImageNet CNN [T.Dim. = 4096]	* ImageNet pre-trained CNN (Caffe)	Extract words from webpages with stopword removal and stemming. Concept list is stemmed and training samples are assigned labels by concept word appearance.	A logistic regression (soft-max) model is trained. Annotation based on threshold (the same for all concepts) optimized using development set.
MLIA run #9 [23]	Provided by organizers (All 7) [T.Dim. = 21312]	* WordNet * ImageNet pre-trained CNN (Overfeat) * Lucene stemmer	Provided webpage features processed by stopword removal and stemming. Initial list of concepts assigned to training images by appearance of WordNet synonyms, then list filtered by considering the nouns assigned by Overfeat.	One SVM per concept with parameters selected by cross-validating F-measure on development set. Annotation based on SVM classification decision.
DISA run #4 [1]	Five MPEG7 global visual descriptors [T.Dim. = 256]	* WordNet * Profiset dataset * MUFIN indexing system * Visual Concept Ontology (VCO)	Provided webpage dataset and Profiset indexed for efficient image similarity search using MUFIN system.	Retrieval of the 25 most similar images over both collections, matching their descriptions to concepts using WordNet (hypernymy, hyponymy, holonymy, meronymy) and VCO. Annotation of at most 7 of the top concepts.
RUC run #7 [8]	SIFT BoW with codebook of 4000 [T.Dim. = 16000]	* Search engine keywords * Bing user-clicked dataset (1M) * Flickr tags dataset (4M)	Positive samples selected by: search engine keywords (provided dataset), click count (Bing) and a semantic based relevance measurement (Flickr). Negative samples selected by Negative Bootstrap.	Ensemble of fikSVMs per concept. Annotation of the top k concepts with k adapted per image derived from an estimation the value that would have been selected for the development set concepts.
IPL run #9 [14]	CEDD, FCTH and provided Opp-SIFT [T.Dim. = 8024]	* WordNet	Provided webpage features processed by stopword removal. Concepts assigned to training images by appearance of WordNet synset synonyms.	Retrieval of the 800 most similar images by Latent Semantic Analysis and used for estimation of the posterior probability of each concept. Annotation of at most 8 of the top concepts.

Table 2: Test set performance measures (%) for the baselines and all submissions. The best submission for each team is highlighted with a gray background.

System	MAP-samples				MF_1-samples				MF_1-concepts				
	all	**ani.**	**food**	**207**	**all**	**ani.**	**food**	**207**	**all**	**ani.**	**food**	**207**	unseen
OPP-SIFT	20.2	28.2	34.8	11.4	16.7	25.3	21.6	8.1	9.8	8.1	7.7	5.7	9.6
RGB-SIFT	20.0	27.9	34.3	11.0	16.6	25.4	21.5	8.0	9.7	8.0	7.6	6.0	9.6
C-SIFT	20.4	29.4	35.1	11.8	16.7	25.1	22.4	8.6	8.5	5.6	8.6	5.6	8.9
SIFT	19.9	27.6	34.8	11.0	16.5	25.2	21.6	7.8	9.2	8.0	8.0	5.5	9.3
Colorhist	19.2	28.0	34.2	10.9	15.7	25.0	20.9	7.6	7.0	5.2	7.7	5.0	8.2
GIST	17.8	26.1	32.8	10.1	15.0	24.7	20.2	7.0	5.7	4.9	7.0	4.3	6.4
GETLF	18.2	26.5	32.4	10.3	14.9	24.5	20.2	7.1	5.3	4.3	6.6	4.0	6.4
Random	8.8	12.4	14.3	4.8	3.5	5.2	6.9	1.8	2.6	3.2	6.6	0.8	3.0
KDEVIR 9	36.8	33.1	67.1	28.9	37.7	29.9	64.9	32.0	54.7	67.1	65.1	31.6	66.1
KDEVIR 8	36.5	33.3	67.1	27.7	37.5	30.2	64.5	31.3	54.8	68.0	64.8	32.1	66.4
MIL 3	36.9	30.9	68.6	23.3	27.5	20.6	53.1	18.0	34.7	34.7	50.4	16.9	36.7
MindLab 1	37.0	43.1	63.0	22.1	25.8	17.0	45.2	18.3	30.7	35.1	35.3	16.7	34.7
KDEVIR 3	35.0	39.2	53.4	25.6	34.6	36.7	41.8	27.5	25.9	27.5	36.3	14.8	27.9
MIL 2	35.8	31.6	64.3	22.5	26.5	20.8	50.3	17.4	32.3	32.8	46.2	15.7	34.3
KDEVIR 4	34.8	39.8	53.4	24.5	34.0	36.9	41.6	26.3	25.7	28.4	35.9	14.9	28.3
MindLab 2	37.0	43.1	63.0	22.1	24.8	17.0	45.2	16.2	31.7	35.1	35.3	17.8	34.7
KDEVIR 10	35.2	39.5	52.2	25.2	34.2	35.9	40.0	26.5	25.1	25.1	34.8	13.8	25.9
MLIA 9	27.8	18.8	53.6	16.7	24.8	12.1	46.0	16.4	33.2	32.7	37.3	16.9	34.8
KDEVIR 6	34.5	39.0	51.9	25.0	33.8	36.2	39.8	26.3	24.4	24.9	34.6	13.4	25.6
MLIA 10	27.9	18.7	53.2	16.7	24.8	12.1	46.0	16.4	33.2	32.7	37.3	16.9	34.8
KDEVIR 5	33.2	37.4	46.6	24.2	33.6	35.3	37.9	27.4	24.1	24.0	32.5	13.2	25.0
MLIA 8	27.4	18.5	52.5	16.4	24.6	11.8	45.5	16.2	33.3	32.9	36.7	16.9	34.8
MLIA 7	26.9	18.1	52.2	16.1	24.4	11.4	45.6	16.1	33.5	33.0	36.9	17.0	35.0
KDEVIR 2	33.0	37.9	48.5	24.0	32.8	35.6	39.1	25.8	22.9	24.1	33.4	12.6	24.5
MLIA 6	26.3	17.9	50.2	15.8	24.1	11.1	44.2	16.0	33.6	33.7	36.1	17.1	35.2
KDEVIR 7	23.9	19.5	55.5	14.4	22.9	13.6	55.0	15.6	48.7	58.7	61.7	27.4	59.7
MLIA 3	27.6	18.4	53.2	16.5	24.5	11.8	45.6	16.1	32.2	29.9	36.7	16.1	33.1
MIL 1	31.9	25.3	64.0	20.1	24.0	17.3	49.1	15.7	30.1	28.4	48.0	14.4	31.9
MLIA 4	27.6	18.4	53.1	16.5	24.5	11.8	45.6	16.1	32.2	29.9	36.7	16.1	33.1
DISA 4	34.3	46.6	39.6	19.0	29.7	40.6	31.2	16.9	19.1	23.0	22.3	7.3	19.0
DISA 5	32.3	41.0	42.8	18.3	28.4	37.8	35.1	16.1	20.3	27.0	28.0	7.9	22.1
MLIA 2	27.2	18.2	52.2	16.2	24.4	11.5	45.3	16.0	32.4	30.3	36.4	16.2	33.3
MLIA 1	26.8	17.8	51.8	15.9	24.2	11.1	45.4	15.9	32.7	30.2	36.8	16.3	33.6
MLIA 5	26.1	17.6	50.1	15.7	23.9	10.8	44.2	15.8	32.8	30.9	36.1	16.4	33.8
DISA 3	32.9	42.6	39.2	18.5	28.5	37.7	31.1	16.4	18.9	22.8	22.6	7.2	18.8
RUC 7	27.5	25.2	44.2	15.1	29.3	28.0	28.2	20.7	25.3	20.1	23.1	10.0	18.7
RUC 5	27.5	25.2	44.2	15.1	31.1	33.7	27.1	22.3	25.0	19.4	20.2	9.8	18.0
RUC 6	27.5	25.2	44.2	15.1	29.0	27.0	27.6	20.4	25.2	20.1	22.6	10.0	18.6
RUC 2	30.2	36.6	45.2	15.8	27.8	25.5	31.3	18.8	24.1	18.9	23.9	9.9	18.1
RUC 1	30.2	36.6	45.2	15.8	28.0	26.0	31.2	19.0	24.1	19.0	23.5	9.8	18.0
RUC 3	30.2	36.6	45.2	15.8	27.9	25.8	31.3	18.9	24.0	18.9	23.5	9.8	18.0
RUC 4	30.2	36.6	45.2	15.8	21.9	25.5	31.6	11.6	21.9	20.2	23.8	9.7	20.2
DISA 1	31.6	40.8	34.8	18.2	27.9	35.9	26.9	16.7	15.4	12.7	17.3	5.7	12.8
DISA 2	31.9	40.7	35.1	17.8	27.5	35.9	27.0	15.9	15.3	12.7	17.4	5.4	12.6
RUC 8	27.5	25.2	44.2	15.1	20.6	18.0	28.7	12.2	21.5	15.7	23.8	10.6	19.8
IPL 9	23.4	30.0	48.5	18.9	18.4	20.2	29.8	17.5	15.8	15.8	33.3	12.5	22.0
IPL 8	23.4	30.3	47.4	18.8	18.4	20.2	29.7	17.4	15.7	16.0	33.2	12.6	22.2
IPL 10	23.4	29.3	50.2	19.1	18.3	19.9	29.8	17.3	15.5	15.2	33.0	12.1	21.3
IPL 7	22.0	28.7	40.9	16.9	17.7	19.7	27.8	16.3	13.4	14.2	28.5	9.3	17.9
IPL 4	22.5	26.1	38.3	7.4	18.9	20.5	24.6	6.6	13.3	7.6	14.0	2.0	8.8
IPL 3	22.4	25.9	38.0	7.5	18.7	20.2	23.9	6.5	13.3	8.0	14.1	2.2	9.0
IMC 1	25.1	35.7	35.6	12.9	16.3	14.3	21.0	10.9	12.5	10.2	15.1	6.1	11.2
IPL 5	22.4	26.3	38.2	7.5	18.8	20.4	24.8	6.6	13.0	7.6	13.7	2.0	8.6
IMC 2	25.1	35.7	35.6	12.9	16.3	14.3	21.0	10.9	12.5	10.2	15.1	6.1	11.2
IPL 6	21.3	27.9	38.1	16.3	17.3	19.4	26.8	15.7	12.0	12.6	25.2	7.6	15.5
IPL 2	22.1	26.4	37.8	7.3	18.6	20.7	24.2	6.4	12.4	7.4	13.6	1.8	8.3
IPL 1	21.9	26.4	37.9	7.2	18.5	20.7	24.3	6.3	12.1	7.2	13.9	1.8	8.0
INAOE 5	9.6	6.9	15.0	8.5	5.3	0.4	0.5	6.4	10.3	1.0	0.8	17.9	19.0
INAOE 6	9.3	6.7	14.5	8.8	5.3	0.5	0.8	7.5	10.2	1.2	1.1	18.5	18.2
INAOE 2	9.6	7.0	15.0	8.5	5.9	1.5	3.0	6.8	9.2	1.3	2.1	15.2	15.2
INAOE 4	9.3	7.8	13.9	10.0	4.2	0.8	2.3	8.5	9.3	1.3	2.1	15.9	16.4
INAOE 3	9.3	6.7	14.5	8.8	6.2	1.5	4.6	6.8	9.1	1.3	2.9	16.1	14.7
INAOE 1	9.3	7.9	13.9	10.0	4.9	1.5	3.5	8.4	8.3	1.7	2.7	13.2	13.5
NII 1	14.7	23.2	22.0	4.6	13.0	18.9	18.7	4.9	2.3	3.0	2.1	0.9	1.8
FINKI 1	6.9	N/A	N/A	N/A	7.2	8.1	12.3	4.1	4.7	6.3	9.0	2.9	4.7
KDEVIR 1	8.6	12.2	14.1	4.6	4.4	5.4	4.5	3.2	3.0	3.2	6.1	1.8	3.0

Acknowledgments

The authors are very grateful with the CLEF initiative for supporting ImageCLEF. 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) and from the Spanish MEC under the STraDA project (TIN2012-37475-C02-01).

References

1. Budikova, P., Botorek, J., Batko, M., Zezula, P.: DISA at ImageCLEF 2014: The search-based solution for scalable image annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)
2. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2014)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255 (june 2009), doi:10.1109/CVPR.2009.5206848
4. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA; London (May 1998)
5. Kanehira, A., Hidaka, M., Mukuta, Y., Tsuchiya, Y., Mano, T., Harada, T.: MIL at ImageCLEF 2014: Scalable System for Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)
6. 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), doi:10.1109/IVL.1998.694480
7. 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), doi:10.1109/CVPR.2006.68
8. Li, X., He, X., Yang, G., Jin, Q., Xu, J.: Renmin University of China at ImageCLEF 2014 Scalable Concept Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)
9. 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)
10. 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), doi:10.1023/A:1011139631724
11. Reshma, I.A., Ullah, M.Z., Aono, M.: KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation Task: Ontology based Automatic Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)

12. 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)
13. 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), doi:10.1109/TPAMI.2009.154
14. Stathopoulos, S., Kalamboukis, T.: IPL at ImageCLEF 2014: Scalable Concept Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)
15. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: CLEF 2012 working notes. Rome, Italy (2012)
16. 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), doi:10.1109/TPAMI.2008.128
17. Vanegas, J.A., Arevalo, J., Otálora, S., Páez, F., Pérez-Rubiano, S.A., González, F.A.: MindLab at ImageCLEF 2014: Scalable Concept Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)
18. 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)
19. 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), http://mvillegas.info/pub/Villegas12_CLEF_Annotation-Overview.pdf
20. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013), http://mvillegas.info/pub/Villegas13_CLEF_Annotation-Overview.pdf
21. Wang, X.J., Zhang, L., Liu, M., Li, Y., Ma, W.Y.: ARISTA - image search to annotation on billions of web photos. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 2987–2994 (June 2010), doi:10.1109/CVPR.2010.5540046
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), doi:10.1007/s10994-010-5198-3
23. Xu, X., Shimada, A., ichiro Taniguchi, R.: MLIA at ImageCLFE 2014 Scalable Concept Image Annotation Challenge. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)

A Concept List 2014

The following tables present the 207 concepts used in the ImageCLEF 2014 Scalable Concept Image Annotation task. In the electronic version of this document, each concept name and Wikipedia article name are hyperlinks to webpages of the corresponding WordNet synset and the Wikipedia article, respectively.

Concepts seen in both development and test sets:

Concept	WordNet 3 type	WordNet 3 sense#	Wikipedia article	#test
aerial	adj.	1	Aerial_photography	112
airplane	noun	1	Airplane	33
baby	noun	1	Baby	38
beach	noun	1	Beach	87
bicycle	noun	1	Bicycle	33
bird	noun	1	Bird	110
boat	noun	1	Boat	144
book	noun	2, 1	Book	44
bottle	noun	1	Bottle	32
bridge	noun	1	Bridge	79
building	noun	1	Building	478
bus	noun	1	Bus	45
car	noun	1	Car	135
cartoon	noun	1	Cartoon	104
castle	noun	2	Castle	38
cat	noun	1	Cat	38
chair	noun	1	Chair	64
child	noun	1	Child	88
church	noun	2	Church_(building)	28
cityscape	noun	1	Cityscape	163
closeup	noun	1	Closeup	348
cloudless	adj.	1	-	274
cloud	noun	2	Cloud	609
coast	noun	1	Coast	113
countryside	noun	1	Countryside	117
daytime	noun	1	Daytime_(astronomy)	2186
desert	noun	1	Desert	49
diagram	noun	1	Diagram	35
dog	noun	1	Dog	66
drink	noun	1	Drink	59
drum	noun	1	Drum	21
elder	noun	1	Elderly	49
embroidery	noun	2	Embroidery	24
female	noun	2	Female	211
fire	noun	3, 1	Fire	62
firework	noun	1	Firework	31
fish	noun	1	Fish	54
flower	noun	2	Flower	160
fog	noun	2	Fog	57
food	noun	2, 1	Food	490
footwear	noun	1, 2	Footwear	62
forest	noun	1, 2	Forest	235
furniture	noun	1	Furniture	177
garden	noun	1	Garden	37
grass	noun	1	Grass	654
guitar	noun	1	Guitar	20
harbor	noun	1	Harbor	55
hat	noun	1	Hat	104
helicopter	noun	1	Helicopter	22
highway	noun	1	Highway	31
horse	noun	1	Horse	67
indoor	adj.	1	-	357
instrument	noun	6	Musical_instrument	92
lake	noun	1	Lake	110
lightning	noun	1, 2	Lightning	26
logo	noun	1	Logo	51
male	noun	2	Male	207
monument	noun	1	Monument	28
moon	noun	1	Moon	38
motorcycle	noun	1	Motorcycle	33
mountain	noun	1	Mountain	287
newspaper	noun	3, 1	Newspaper	19
nighttime	noun	1	Nighttime	154
outdoor	adj.	1, 2	-	2255
overcast	noun	1	Overcast	137
painting	noun	1	Painting	125
park	noun	2	Park	47
person	noun	1	Person	856
phone	noun	1	Phone	26
plant	noun	2	Plant	1261
portrait	noun	1	Portrait	36
poster	noun	1	Poster	23
protest	noun	2	Protest	28
rainbow	noun	1	Rainbow	24
rain	noun	1	Rain	41
reflection	noun	4, 5	Mirror_image	149
river	noun	1	River	159
road	noun	1	Road	344
sand	noun	1	Sand	141
sculpture	noun	2	Sculpture	79
sea	noun	1	Sea	233
shadow	noun	2	Shadow	203
sign	noun	2	Sign	133
silhouette	noun	1	Silhouette	68
sky	noun	1	Sky	1230
smoke	noun	1	Smoke	43
snow	noun	2	Snow	175
soil	noun	2	Soil	247
space	noun	4	Outer_space	84
spectacles	noun	1	Spectacles	71
sport	noun	1	Sport	118
sun	noun	1	Sun	92
sunrise/sunset	noun	1, 1	Sunrise/Sunset	90
table	noun	2	Table_(furniture)	49
teenager	noun	1	Teenager	45
toy	noun	1	Toy	56
traffic	noun	1	Traffic	63
train	noun	1	Train	57
tree	noun	1	Tree	906
tricycle	noun	1	Tricycle	15
truck	noun	1	Truck	61
underwater	adj.	1, 2	Underwater	84
unpaved	adj.	1	-	40
vehicle	noun	1	Vehicle	583
violin	noun	1	Violin	23
wagon	noun	1	Wagon	30
water	noun	6	Water	807

continues in next page

Concepts seen only in the test set:

Concept	WordNet 3 type	WordNet 3 sense#	Wikipedia article	#test
antelope	noun	1	Antelope	28
apple	noun	1	Apple	70
arthropod	noun	1	Arthropod	78
asparagus	noun	2	Asparagus	24
avocado	noun	1	Avocado	24
banana	noun	2	Banana	46
bear	noun	1	Bear	34
berry	noun	1, 2	Berry	38
blood	noun	1	Blood	22
branch	noun	2	Branch	994
bread	noun	1	Bread	70
broccoli	noun	1	Broccoli	32
buffalo	noun	1	African_buffalo	62
butterfly	noun	1	Butterfly	16
camel	noun	1	Camel	46
canidae	noun	1	Canidae	188
captive	noun	2	Captivity_(animal)	332
carrot	noun	1	Carrot	52
cauliflower	noun	1	Cauliflower	28
cervidae	noun	1	Cervidae	114
cheese	noun	1	Cheese	76
cheetah	noun	1	Cheetah	32
chimpanzee	noun	1	Chimpanzee	38
corn	noun	3, 1	Maize	50
crocodile	noun	1	Crocodile	32
cucumber	noun	2	Cucumber	36
donkey	noun	2	Donkey	20
egg	noun	2	Egg_(food)	54
eggplant	noun	1	Eggplant	34
elephant	noun	1	Elephant	40
equidae	noun	1	Equidae	156
felidae	noun	1	Felidae	208
flamingo	noun	1	Flamingo	22
fox	noun	1	Fox	26
fried	adj.	1	Frying	58
fruit	noun	1	Fruit	390
galaxy	noun	3	Galaxy	21
giraffe	noun	1	Giraffe	46
gorilla	noun	1	Gorilla	32
grape	noun	1	Grape	78
hippopotamus	noun	1	Hippopotamus	60
human	noun	1	Human	998
hunting	noun	1	Hunting	40
kangaroo	noun	1	Kangaroo	24
knife	noun	1	Knife	30
koala	noun	1	Koala	30
leaf	noun	1	Leaf	2012
leopard	noun	2	Leopard	44
lettuce	noun	3	Lettuce	56
lion	noun	1	Lion	40

Concept	WordNet 3 type	WordNet 3 sense#	Wikipedia article	#test
mammal	noun	1	Mammal	2264
marsupial	noun	1	Marsupial	72
meat	noun	1	Meat	152
monkey	noun	1	Monkey	36
mud	noun	1	Mud	60
mushroom	noun	5, 1	Edible_mushroom	32
nebula	noun	3	Nebula	16
onion	noun	1, 3	Onion	48
orange	noun	1	Orange_(fruit)	58
ostrich	noun	2	Ostrich	44
pan	noun	1	Frying_pan	28
pasta	noun	2	Pasta	38
pear	noun	1	Pear	30
penguin	noun	1	Penguin	18
pig	noun	1	Pig	52
pineapple	noun	2	Pineapple	44
pinniped	noun	1	Pinniped	76
pool	noun	1	Swimming_pool	31
potato	noun	1	Potato	32
pumpkin	noun	2	Pumpkin	30
rabbit	noun	1	Rabbit	22
raccoon	noun	2	Raccoon	38
reptile	noun	1	Reptile	62
rhino	noun	1	Rhinoceros	26
rice	noun	1	Rice	26
rifle	noun	1	Rifle	22
roasted	adj.	1	Roasting	54
rock	noun	1, 2	Rock_(geology)	320
rodent	noun	1	Rodent	53
sausage	noun	1	Sausage	32
soup	noun	1	Soup	50
spider	noun	1	Spider	19
spoon	noun	1	Spoon	56
squirrel	noun	1	Squirrel	32
strawberry	noun	1	Strawberry	54
submarine	noun	1	Submarine	24
tiger	noun	2	Tiger	42
tomato	noun	1	Tomato	80
trunk	noun	1	Trunk_(botany)	706
tuber	noun	1	Tuber	54
turtle	noun	2	Turtle	38
vegetable	noun	1	Vegetable	338
walrus	noun	1	Walrus	18
warthog	noun	1	Warthog	22
watermelon	noun	2	Watermelon	30
wild	noun	2	Wilderness	588
wolf	noun	1	Wolf	46
yam	noun	1, 4	Yam_(vegetable)	10
zebra	noun	1	Zebra	36
zoo	noun	1	Zoo	140