Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas and Krystian Mikolajczyk

Abstract. Since 2010, ImageCLEF has run a scalable image annotation task, to promote research into the annotation of images using noisy web page data. It aims to develop techniques to allow computers to describe images reliably, localise different concepts depicted and generate descriptions of the scenes. The primary goal of the challenge is to encourage creative ideas of using web page data to improve image annotation. Three subtasks and two pilot teaser tasks were available to participants; all tasks use a single mixed modality data source of 510,123 web page items for both training and test. The dataset included raw images, textual features obtained from the web pages on which the images appeared, as well as extracted visual features. Extracted from the Web by querying popular image search engines, the dataset was formed. For the main subtasks, the development and test sets were both taken from the "training set". For the teaser tasks, 200,000 web page items were reserved for testing, and a separate development set was provided. The 251 concepts were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of the visual content of images and were mined from the texts of our extensive database of image-webpage pairs. This year seven groups participated in the task, submitting over 50 runs across all subtasks, and all participants also provided working notes papers. In general, the groups' performance is impressive across the tasks, and there are interesting insights into these very relevant challenges.

1 Introduction

How can you use large-scale noisy data to improve image classification, caption generation and text illustration? This challenging question is the basis of this year's image annotation challenge. Every day, users struggle with the ever-increasing quantity of data available to them. Trying to find "that" photo they took on holiday last year, the image on Google of their favourite actress or band, or the images of the news article someone mentioned at work. There are a huge number of images that can be cheaply found and gathered from the Internet. However, more valuable is mixed-modality data, for example, web pages containing both images and text. A significant amount of information about the image is present on these web pages and vice-versa. However, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and unrelated. Despite the obvious benefits of using such information in automatic learning, the weak supervision it provides means that it remains a challenging problem. Fig. 1 illustrates the expected results of the task.

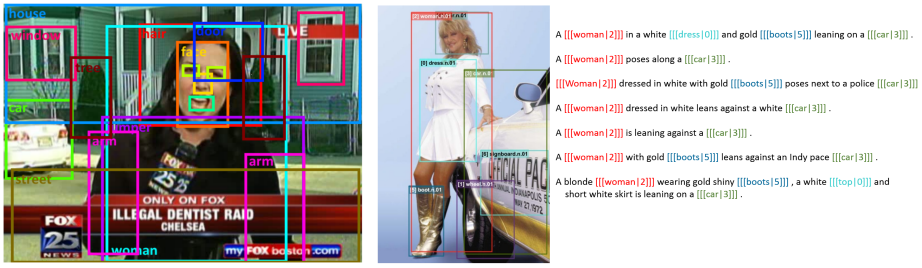


Fig. 1: Image annotation and localization of concepts and natural language caption generation.

The Scalable Concept Image Annotation task is a continuation of the general image annotation and retrieval task that has been part of ImageCLEF since its very first edition in 2003. In the early years the focus was on retrieving relevant images from a web collection given (multilingual) queries, from 2006 onwards annotation tasks were also held, initially aimed at object detection, but more recently also covering semantic concepts. In its current form, the 2016 Scalable Concept Image Annotation task is its fifth edition, having been organized in 2012 [24], 2013 [26], 2014 [25], and 2015 [8]. In the 2015 edition [8], the image annotation task was expanded to concept localization and also natural language sentential description of images. In this year's edition, we further introduced a text illustration 'teaser' task, to evaluate systems that analyse a text document and select the best illustration for the text from a large collection of images provided. As there is an increased interest in recent years in research combining text and vision, the new tasks introduced in both the 2015 and 2016 editions aim at further stimulating and encouraging multimodal research that uses both text and visual data for image annotation and retrieval.

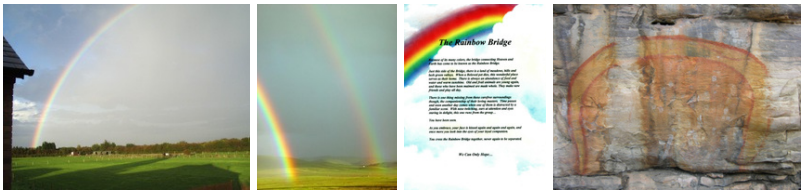
This paper presents the overview of the fifth edition of the Scalable Concept Image Annotation task [24,26,25,8], one of the three benchmark campaigns organized by ImageCLEF [22] in 2016 under the CLEF initiative[1]. Section 2 describes the task in detail, including the participation rules and the provided data and resources. Section 3 presents and discusses the results of the submissions received for the task. Finally, Section 4 concludes the paper with final remarks and future outlooks.

2 Overview of the Task

2.1 Motivation and Objectives

Image annotation has relied on training data that has been manually, and thus reliably annotated. Annotating training data is an expensive and laborious endeavour that cannot be easily scaled, particularly as the number of concepts

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



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

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

Fig. 2: Example of images retrieved by a commercial image search engine.

grows. However, images for any topic can be cheaply gathered from the Web, along with associated text from the web pages that contain the images. The degree of relationship between these web images and the surrounding text varies considerably, i.e., the data are very noisy, but overall these data contain useful information that can be exploited to develop annotation systems. Figure 2 shows examples of typical images found by querying search engines. As can be seen, the data obtained are useful and furthermore a wider variety of images is expected, not only photographs but also drawings and computer generated graphics. This diversity has the advantage that this data can also handle the different possible senses that a word can have or the various types of images that exist. Likewise, there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies. There are also tools that can contribute to deal with noisy text commonly found on web pages, such as language models, stop word lists and spell checkers.

Motivated by the need for exploiting this useful (albeit noisy) data, the ImageCLEF 2016 Scalable Concept Image Annotation task aims to develop techniques to allow computers to describe images reliably, localise the different concepts depicted in the images, generate a description of the scene and select images to illustrate texts. The primary objective of the 2016 edition is to encourage creative ideas of using noisy, web page data so that it can be used to improve various image annotation tasks – concept annotation and localization, selecting important concepts to be described, generating natural language descriptions, and retrieving images to illustrate a text document.

2.2 Challenge Description

This year the challenge[2] consisted of 3 subtasks and a teaser task.[3]

1. **Subtask 1** (*Image Annotation and Localization*): The image annotation task remains the same as the 2015 edition. Participants are required to develop a system that receives as input an image and produces as output a prediction of which concepts are present in that image, selected from a predefined list of concepts. Like the 2015 edition, they should also output bounding boxes indicating where the concepts are located within the image.
2. **Subtask 2** (*Natural Language Caption Generation*): This subtask was geared towards participants interested in developing systems that generate textual descriptions directly with an image as input. For example, by using visual detectors to identify concepts and generating textual descriptions from the detected concepts, or by learning neural sequence models in a joint fashion to create descriptions conditioned directly on the image. Participants used their own image analysis methods, for example by using the output of their image annotation systems developed for Subtask 1. They are also encouraged to augment their training data with the noisy content of the web page.
3. **Subtask 3** (*Content Selection*): This subtask was primarily designed for those interested in the Natural Language Generation aspects of Subtask 2 while avoiding visual processing of images. It concentrated on the content selection phase when generating image descriptions, i.e. which concepts (from all possible concepts depicted) should be selected, and mentioned in the corresponding description? Gold standard input, bounding boxes labelled with concepts for each test image was provided, and participants were expected to develop systems that predict the bounding box instances most likely to be mentioned in the corresponding image descriptions. Unlike the 2015 edition, participants were not required to generate complete sentences but were only requested to provide a list of bounding box instances per image.
4. **Teaser task** (*Text Illustration*): This pilot task is designed to evaluate the performance of methods for text-to-image matching. Participants were asked to develop a system to analyse a given text document and find the best illustration for it from a set of all available images. At test time, participants were provided as input a selection of text documents as queries, and the goal was to select the best illustration for each text from a collection of 200,000 images.

As a common dataset, participants were provided with 510,123 web images, the corresponding web pages on which they appeared, as well as precomputed visual and textual features (see Sect. 2.4). As in the 2015 task, external training

[2] Challenge website at http://imageclef.org/2016/annotation

[3] A Second teaser task was also introduced, aimed at evaluating systems that identify the GPS coordinates of a text document's topic based on its text and image data. However, we had no participants for this task, and thus will not discuss this second teaser task in this paper.

data such as ImageNet ILSVRC2015 and MSCOCO is also allowed, and participants were also encouraged to use other resources such as ontologies, word disambiguators, language models, language detectors, spell checkers, and automatic translation systems.

We observed in the 2015 edition that this large-scale noisy web data was not used as much as we anticipated – participants mainly used external training data. To encourage participants to utilise the provided data for training, in this edition participants were expected to produce two sets of related results:

1. using only external training data;
2. using both external data and the noisy web data of 510,123 web pages.

The aim is for participants to improve the performance of externally trained systems, using the provided noisy web data. However none of the participants submitted results, *only* on the supplied noisy training; this is probably due to the fact the groups are chasing the optimal image annotation results, and not actively attempting to research into using the noisy training data.

Development datasets: In addition to the training dataset and visual/textual features mentioned above, the participants were provided with the following for the development of their systems:

- A development set of images (a small subset of the training data) with ground truth labelled bounding box annotations and precomputed visual features for estimating the system performance for **Subtask 1**.
- A development set of images with at least five textual descriptions per image for **Subtask 2**.
- A subset of the development set above for **Subtask 3**, with gold standard inputs (bounding boxes labelled with concepts) and correspondence annotation between bounding box inputs and terms in textual descriptions.
- A development set for the **Teaser task**, with approximately 3,000 image-web page pairs. This set is disjoint from the 510,123 noisy dataset.

2.3 Concepts

For the three subtasks, the 251 concepts were retained from the 2015 edition. They were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of the visual content of images. They include animate objects such as people, dogs and cats, inanimate objects such as houses, cars and balls, and scenes such as city, sea and mountains. With the concepts mined from the texts of our database of 31 million image-webpage pairs [23]. Nouns that are subjects or objects of sentences are extracted and mapped onto WordNet synsets [7]. In addition, filtered to 'natural', basic-level categories (*dog* rather than a *Yorkshire terrier*), based on the WordNet hierarchy and heuristics from a large-scale text corpora [28]. The organisers manually shortlisted the final list of concepts such that they were (i) visually concrete and localizable;

(ii) suitable for use in image descriptions; (iii) at an appropriate 'every day' level of specificity that was neither too general nor too specific. The complete list of concepts, as well as the number of samples in the test sets, is included in Appendix A.

2.4 Dataset

The dataset this year[4] was built on the 500,000 image-webpage pairs from the 2015 edition. The 2015 dataset used was very similar to previous three editions of the task [24,26,25]. To create the dataset, a database of over 31 million images was created by querying Google, Bing and Yahoo! using words from the Aspell English dictionary [23]. The images and corresponding web pages were downloaded, taking care to avoid data duplication. Then, a subset of 500,000 images was selected from this database by choosing the top images from a ranked list. For further details on the dataset creation, please refer to [24]. By retrieving images from our database using the list of concepts, the ranked list was generated, in essence, more or less as if the search engines was queried. From the ranked list, some types of problematic images were removed, and each image had at least one web page in which they appeared.

To incorporate the teaser task this year, the 500,000 image dataset from 2015 was augmented with 10,123 new image-webpage pairs, taken from a subset of the BreakingNews dataset [16] which we developed, expanding the size of the dataset to 510,123 image-webpage pairs. The aim of generating the BreakingNews dataset was to further research into image and text annotation, where the textual descriptions are loosely related to their corresponding images. Unlike the main subtasks, the textual descriptions in this dataset do not describe real image content but provide connotative and ambiguous relations that may not be directly inferred from images. More specifically, the documents and images were obtained from various online news sources such as BBC News and The Guardian. For the teaser task, a random subset of image-article pairs was selected from the original dataset, and we ensured that each image corresponds to only one text article. The reports were converted to a 'web page' via a generic template.

Like last year, for the main subtasks the development and test sets were both taken from the "training set". Both sets were retained from last year, making the evaluation for the three subtasks comparable across both 2015 and 2016 editions. To generate these sets, a set of 5,520 images was selected using a CNN trained to identify images suitable for sentence generation. Crowd-sourcing, annotated the images in three stages: (i) image level annotation for the 251 concepts; (ii) bounding box annotation; (iii) textual description annotation. A subset of these samples was then selected for subtask 3 and further annotated by the organisers with correspondence annotations between bounding box instances and terms in textual descriptions.

The development set for the main subtask contained 2,000 samples, out of which 500 samples were further annotated and used as the development set for

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

subtask 3. Only 1,979 samples from the development set include at least one bounding box annotation. The number of textual descriptions for the development set ranged from 5 to 51 per image (with a mean of 9.5 and a median of 8 descriptions). The test set for subtasks 1 and 2 contains 3,070 samples, while the test set for subtask 3 comprises 450 samples which are disjoint from the test set of subtasks 1 and 2.

For the teaser task, 3,337 random image-article pairs were selected from the BreakingNews dataset as the development set; these are disjoint from the 10,123 selected in the main dataset. Again, each image corresponds to only one article.

Like last year, the training and the test images were all contained within the 510,123 images. In the case of the teaser task, we divided the dataset into 310,123 for training and 200,000 for testing, where all 10,123 documents from the BreakingNews dataset were contained within the 200,000 test set. Participants of the teaser task were thus not allowed to explore the data for these 200,000 test documents.

The training and development sets for all tasks were released approximately three months before the submission deadline. For subtasks 1 and 2, participants were expected to provide classification/generate a description for all 510,123 images. The test data for subtask 3 was released one week before the submission deadline. While the train/test split for the teaser tasks was provided right from the beginning, the test input was only released 1.5 months before the deadline. The test data were 180,000 text documents extracted from a subset of the web pages in the 200,000 test split. Text extraction was performed using the *get_text*() method of the Beautiful Soup library[5], after removal of unwanted elements (and their content) such as *script* or *style*. A maximum of 10 submissions per subtask (also referred to as *runs*) was allowed per participating group.

Textual Data: 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 used. The second set of textual data contained the image URLs as referenced in the web pages 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 was the web pages 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 web page it appeared in.

To extract the text near the image, after conversion to valid XML, the script and style elements were removed. The extracted texts were the web page title, and all the terms closer than 600 in word distance to the image, not including the HTML tags and attributes. Then a weight $s(t_n)$ was assigned to each of the

[5] https://www.crummy.com/software/BeautifulSoup/

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. [10], 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 eight types of precomputed visual features. The first feature set *Colorhist* consisted of 576-dimensional colour histograms extracted using our implementation. These features correspond to dividing the image in 3×3 regions and for each region obtaining a colour 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 eight random vectors and considering the sign of the resulting projection to produce the bit. Thus, obtaining an 8-bit representation of each local color-histogram that can be regarded 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* [13] descriptors. The following four feature types were obtained using the colorDescriptors software [19], namely *SIFT*, *C-SIFT*, *RGB-SIFT* and *OPPONENT-SIFT*. The configuration was dense sampling with default parameters and a hard assignment 1,000 dimension codebook using a spatial pyramid of 1×1 and 2×2 [11]. Concatenation of the vectors of the spatial pyramid resulted in 5,000-dimensional feature vectors. The codebooks were generated using 1.25 million randomly selected features and the k-means algorithm. Moreover, finally, *CNN* feature vectors have been provided computed as the seventh layer feature representations extracted from a deep CNN model pre-trained with the ImageNet dataset [17] using the Berkeley Caffe library[6].

2.5 Performance Measures

Subtask 1 Ultimately the goal of an image annotation system is to make decisions about which concepts to assign and localise to a given image from a predefined list of concepts. Consideration on how to measure annotation performance should be how good and accurate are those decisions. Ideally, a recall

[6] More details can be found at https://github.com/BVLC/caffe/wiki/Model-Zoo

measure would also be used to penalise a system that has additional false positive output. However given difficulties and unreliability of the hand labelling of the concepts for the test images it was not possible to guarantee all concepts were labelled. However, the labels present are assumed to be accurate and of a high quality.

The annotation and localization of Subtask 1 were evaluated using the PASCAL VOC [6] style metric of intersection over union (IoU), IoU is defined as

$$IoU = \frac{|BB_{fg} \cap BB_{gt}|}{|BB_{fg} \cup BB_{gt}|} \tag{2}$$

Where BB is a rectangle bounding box, fg is a foreground proposed annotation label, gt is the ground truth label of the concept. It calculates the area of intersection between the foreground in the proposed output localization and the ground-truth bounding box localization, divided by the area of their union. IoU is superior to a more simple measure of the percentage of correctly labelled pixels as IoU is normalised by the size of the object automatically and penalises segmentation's that include the background. Causing small changes in the percentage of correctly labelled pixels to correspond to large differences in IoU, and as the dataset has a wide variation in object size, the performance increases from our approach are more reliably measured. The evaluation of the ground truth and proposed output overlap was recorded from 0% to 90%. At 0%, this is equivalent to an image level annotation output, and 50% is the standard PASCAL VOC style metric used. The localised IoU is then used to compute the mean average precision (MAP) of each concept independently. The MAP is reported both per concept and averaged over all concepts. In comparison to previous years, the MAP was averaged over all possible concept labels in the test data, instead of just the concepts the participant used. This was to penalise correctly approaches that only contained a subset of a full approach such as a face detector, as these were producing unrepresentative performances overall MAP, however, registering on only a few concepts.

Subtask 2 Subtask 2 was evaluated using the Meteor evaluation metric [4], which is an F-measure of word overlaps taking into account stemmed words, synonyms, and paraphrases, with a fragmentation penalty to penalise gaps and word order differences. This measure was chosen as it was shown to correlate well with human judgments in evaluating image descriptions [5]. Please refer to Denkowski and Lavie [4] for details about this measure.

Subtask 3 Subtask 3 was evaluated with the fine-grained metric for content selection which we introduced in last year's edition. Please see [8] or [27] for a detailed description. The *content selection* metric is the F_1 score averaged across all 450 test images, where each F_1 score is computed from the precision and recall averaged over all gold standard descriptions for the image. Intuitively, this measure evaluates how well the sentence generation system selects the correct concepts to be described against gold standard image descriptions. Formally, let

$I = \{I_1, I_2, ...I_N\}$ be the set of test images. Let $G^{I_i} = \{G_1^{I_i}, G_2^{I_i}, ..., G_M^{I_i}\}$ be the set of gold standard descriptions for image I_i, where each $G_m^{I_i}$ represents the set of unique bounding box instances referenced in gold standard description m of image I_i. Let S^{I_i} be the set of unique bounding box instances referenced by the participant's generated sentence for image I_i. The precision P^{I_i} for test image I_i is computed as:

$$P^{I_i} = \frac{1}{M} \sum_{m}^{M} \frac{|G_m^{I_i} \cap S^{I_i}|}{|S^{I_i}|} \tag{3}$$

where $|G_m^{I_i} \cap S^{I_i}|$ is the number of unique bounding box instances referenced in both the gold standard description and the generated sentence, and M is the number of gold standard descriptions for image I_i.

Similarly, the recall R^{I_i} for test image I_i is computed as:

$$R^{I_i} = \frac{1}{M} \sum_{m}^{M} \frac{|G_m^{I_i} \cap S^{I_i}|}{|G_m^{I_i}|} \tag{4}$$

The content selection score for image I_i, F^{I_i}, is computed as the harmonic mean of P^{I_i} and R^{I_i}:

$$F^{I_i} = 2 \times \frac{P^{I_i} \times R^{I_i}}{P^{I_i} + R^{I_i}} \tag{5}$$

The final P, R and F scores are computed as the mean P, R and F scores across all test images.

The advantage of the macro-averaging process in equations (3) and (4) is that it implicitly captures the relative importance of the bounding box instances based on how frequently to which they are referred across the gold standard descriptions.

Teaser task For the teaser task, participants are requested to rank the 200,000 test images according to their distance to each input text document. Recall at the k-th rank position ($R@k$) of the ground truth image were used as the performance metrics. The testing of several values of k was performed, and participants were asked to submit the top 100 ranked images. Please refer to Hodosh et al. [9] for more details about the metrics.

3 Evaluation Results

3.1 Participation

This year the participation was not so good as 2015 where it increased considerably in previous years. In total seven groups took part in the task and submitted overall 50 system runs. All seven participating groups submitted a working paper describing their system, thus for these there were specific details available:

– **CEA LIST:** [2] The team from CEA, LIST, Laboratory of Vision and Content Engineering, France, represented by Herve Le Borgne, Etienne Gadeski, Ines Chami, Thi Quynh Nhi Tran, Youssef Tamaazousti, Alexandru Lucian Gînscă and Adrian Popescu.
– **CNRS TPT:** [18] The team from CNRS TELECOM ParisTech, France, represented by Hichem Sahbi.
– **DUTh:** [1] The team from Democritus University of Thrace, DUTh, Greece, was represented by Georgios Barlas, Maria Ntonti and Avi Arampatzis.
– **ICTisia:** [29] The team from Key Laboratory of Intelligent Information Processing, Institute of Computing Technology Chinese Academy of Sciences, China, represented by Yongqing Zhu, Xiangyang Li, Xue Li, Jian Sun, Xinhang Song and Shuqiang Jiang.
– **INAOE:** [14] The team from Instituto Nacional de Astrofisica, Optica y Electronica (INAOE), Mexico was represented by Luis Pellegrin, A. Pastor López-Monroy, Hugo Jair Escalante and Manuel Montes-Y-Gómez.
– **MRIM-LIG:** [15] The team from LIG - Laboratoire d'Informatique de Grenoble, and CNRS Grenoble, France, was represented by Maxime Portaz, Mateusz Budnik, Philippe Mulhem and Johann Poignant.
– **UAIC:** [3] The team from UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania, represented by Alexandru Cristea and Adrian Iftene.

Tables 7, 8, 9 and 10 provide the main key details for some the top groups submission describing their system for each subtask. These tables serve as a summary of the systems, and are also quite illustrative for quick comparisons. For a more in-depth look at the systems of each team, please refer to their corresponding paper.

3.2 Results for Subtask 1: Image Annotation and Localization

Unfortunately subtask 1 had a lower participation than last year, however there were some excellent results showing improvements over previous years. All submissions were able to provide results on all 510,123 images, indicating that all groups have developed systems that are scalable enough to annotate large amounts of images. However one group only processed 180K image (**MRIM-LIG** [15]) due to computational constraints. Final results are presented in Table 1 in terms of mean average precision (MAP) over all images of all concepts, with both 0% overlap (i.e. no localization) and 50% overlap.

Three of the four groups have achieved good performance across the dataset, in particular, the approach of **CEA LIST**. An excellent result given the challenging nature of the images used and the wide range of concepts provided. The graph in Figure 3 shows the performance of each submission for an increasing amount of overlap of the ground truth labels. All the approaches show a steady drop off in performance which is encouraging, illustrating that the approaches do not fail to detect some concepts correctly even with a high degree of accuracy. Even 90% overlap with the ground truth the MAP for **CEA LIST** was 0.20,

Table 1: Subtask 1 results.

Group	0% Overlap	50% Overlap
CEA LIST	0.54	0.378
MRIM-LIG	0.21	0.14
CNRS	0.25	0.11
UAIC	0.003	0.002

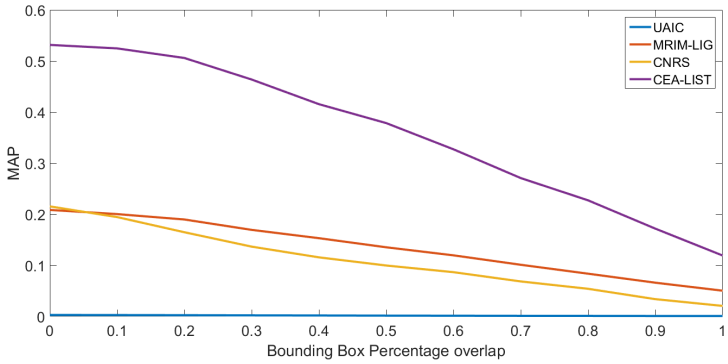


Fig. 3: Increasing percentage of ground truth bounding box overlap of submissions for sub task 1

which is impressive. The results from the groups seem encouraging, and the approaches use a now standard CNN as their foundation. Improved neural network structures such as the proposed approach from VGG [20], have provided much of this improvement.

CEA LIST used a recent deep learning framework [20], however, focused on improving the localisation of the concepts. They attempted to use a face body part detector, boosted by last year's results. However, the use of a face detector was oversold in the previous years results and didn't improve the performance. They used *EdgeBoxes* a generic *objectness* object detector, however the performance also didn't increase as expected in the test runs. They hypothesise that this could be due to the generation of many more candidate bounding boxes, and a significant number estimate the concept incorrectly. **MRIM-LIG** also used a classical deep learning framework and the object localisation of [21], where an apriori set of bounding boxes are defined which are expected to contain a single concept each. They also investigated the false lead on performance improvement through face detection, with a similar lack of performance increase. Finally **CNRS** focused on concept detection and used label enrichment to increase the training data quantity in conjunction with an SVM and VGG [20] deep network. As each group could submit ten different approaches, in general,

the best-submitted approaches contained a fusion of all the various components of their proposed approaches.

Some of the test images have nearly 100 ground truth labelled concepts, and due to limited resources, some of the submitted groups might not have labelled all possible concepts in each image. However, Fig. 4 shows a similar performance between groups as previously in Fig. 3. An improvement over previous years where groups struggled to annotate the 500K images fully.

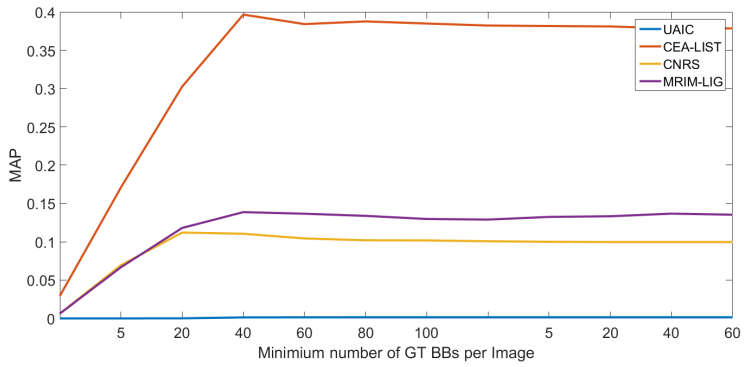


Fig. 4: MAP performance with a minimum number of ground truth bounding boxes per Image

Much of the difference between the groups, is their ability to localise the concepts effectively. The ten concepts with the highest average MAP across the groups, with 0% overlap with the bounding box are in general human-centric: face, hair, arm, woman, tree, man, car, ship, dress and airplane. These are non-rigid classes that are being detected on the image, however not yet successfully localised in the picture as well. With the constraint of 50% overlap with the ground truth bounding box, the list becomes more based around defined objects: car, aeroplane, hair, park, floor, boot, sea, street, face and tree. These are objects that have a rigid shape that can be learnt. Table 2 shows numerical examples of the most successfully localised concepts, together with the percentage of concept occurrence per image in the test data. No method managed to localise 38 concepts, these include the concepts: nut, mushroom, banana, ribbon, planet, milk, orange fruit and strawberry. These are smaller and less represented concepts, in both the test and validation data, in generally occurring in less that 2% of the test images. In fact, many of these concepts were poorly localised in the previous years challenge too, making this an area to direct the challenge objectives in future years.

Discussion for subtask 1 From a computer vision perspective, we would argue that the ImageCLEF challenge has two key differences in its dataset construc-

Table 2: Successfully localised Concepts ranked by 0.5 BB Overlap

Concept	Ave MAP across all Groups		% of Occurrence
	0.5 BB Overlap	0.5 BB Overlap	in test images
Ship	0.61	0.57	28.0%
Car	0.62	0.55	25.3%
Airplane	0.60	0.55	3.2%
Hair	0.74	0.52	93.0%
Park	0.41	0.52	13.9%
Floor	0.41	0.51	13.4%
Boot	0.43	0.59	4.2%
Sea	0.45	0.49	8.8%
Street	0.54	0.47	18.0%
Face	0.75	0.47	95.7%
Street	0.64	0.45	59.9%

tion to that of the other popular data sets ImageNet [17] and MSCOCO [12]. All three are working on detection and classification of concepts within images. However, the ImageCLEF dataset is created from Internet web pages, providing a fundamental difference to the other popular datasets. The web pages are unsorted and unconstrained meaning the relationship or quality of the text and image about a concept can be very variable. Therefore, instead of a high-quality Flickr style photo of a car from ImageNet, the image in the ImageCLEF dataset could be a fuzzy abstract car shape in the corner of the image. Allowing the ImageCLEF image annotation challenge to provide additional opportunities to test proposed approaches on. Another important difference is that in addition to the image, text data from web pages can be used to train and generate the output description of the image in a natural language form.

3.3 Results for Subtask 2: Natural Language Caption Generation

For subtask 2, participants were asked to generate sentence-level textual descriptions for all 510,123 training images. Two teams, **ICTisia** and **UAIC**, participated in this subtask. Table 3 shows the Meteor scores, for all submitted runs by both participants. The Meteor score for the human upper-bound was estimated to be 0.3385 via leave-one-out cross validation, i.e. by evaluating one description against the other descriptions for the same image and repeating the process for all descriptions.

ICTisia achieved the better Meteor score of 0.1837, by building on the state-of-the-art joint CNN-LSTM image captioning system, but fine-tuning the parameters of the image CNN as well as the LSTM. On the other hand, **UAIC**, who also participated last year, improved on their Meteor score with 0.0934 compared to their best performance from last year (0.0813). They generated image descriptions using a template-based approach and leveraged external ontologies and CNNs to improve their results compared to their submissions from last year.

Table 3: Results for subtask 2, showing the Meteor scores for all runs from both participants. We consider the mean Meteor score as the primary measure, but for completeness, we also present the median, min and max scores.

Team	Run	Meteor			
		Mean ± Std	**Median**	**Min**	**Max**
Human	-	0.3385 ± 0.1556	0.3355	0.0000	1.0000
ICTisia	1	0.1826 ± 0.0834	0.1710	0.0180	0.5842
	2	0.1803 ± 0.0823	0.1676	0.0205	0.5635
	3	0.1803 ± 0.0823	0.1676	0.0205	0.5635
	4	0.1837 ± 0.0847	0.1711	0.0180	0.5934
UAIC	1	0.0896 ± 0.0297	0.0870	0.0161	0.2230
	2	0.0934 ± 0.0249	0.0915	0.0194	0.2514

Neither teams have managed to bridge the gap between system performance and the human upper-bound this year, showing that there is still scope for further improvement on the task of generating image descriptions.

3.4 Results for Subtask 3: Content Selection

For subtask 3 on content selection, participants were provided with gold standard labelled bounding box inputs for 450 test images, released one week before the submission deadline. Participants were expected to develop systems capable of predicting, for each image, the bounding box instances (among the gold standard input) that will be mentioned in the gold standard human-authored textual descriptions.

Two teams, **DUTh** and **UAIC**, participated in this task. Table 4 shows the F-score, Precision and Recall across 450 test images for each participant, both of whom submitted only a single run. The generation of a random per image baseline by selecting at most three bounding boxes from the gold standard input at random was perfomed. Like subtask 2, a human upper-bound was computed via leave-one-out cross validation. The results for these are also shown in Table 4. As observed, both participants performed significantly better than the random baseline. Compared against the human upper-bound, like subtask 2, much work can still be done to improve further the performance on the task.

Unlike the previous two subtasks, neither team used neural networks directly for content selection. **DUTh** achieved a higher F-score compared to the best performing team from last year (0.5459 vs. 0.5310), by training SVM classifiers to predict whether a bounding box instance is important or not, using various image descriptors. **UAIC** used the same system as subtask 2, and while they did not significantly improve on their F-score from last year, their recall score showed a slight increase. An interesting note is that both teams this year seem to have concentrated on recall R at the expense of a lower precision P, in contrast

Table 4: Results for subtask 3, showing the content selection scores for all runs from all participants.

Team	Content Selection Score		
	Mean F	**Mean** P	**Mean** R
Human	0.7445 ± 0.1174	0.7690 ± 0.1090	0.7690 ± 0.1090
DUTh	0.5459 ± 0.1533	0.4451 ± 0.1695	0.7914 ± 0.1960
UAIC	0.4982 ± 0.1782	0.4597 ± 0.1553	0.5951 ± 0.2592
Baseline	0.1800 ± 0.1973	0.1983 ± 0.2003	0.1817 ± 0.2227

to last year's best performing team who used an LSTM to achieve high precision but with a much lower recall.

3.5 Results for Teaser task: Text Illustration

Two teams, **CEA LIST** and **INAOE**, participated in the teaser task on text illustration. Participants were provided with 180,000 text documents as input, and for each document were asked to provide the top 100 ranked images that correspond to the document (from a collection of 200,000 images). Table 5 shows the recall at different ranks k ($R@k$), for a selected subset of 10,112 input documents comprised of news articles from the BreakingNews dataset (see Sect. 2.4). Table 6 shows the same results, but on the full 180,000 test documents. Because the full set of test documents were extracted from generic web pages, the domain of the text varies. As such, they may consist of noisy documents such as text from navigational links or advertisements.

Table 5: Results for Teaser 1: Text Illustration. Recall@k for a selected subset of test set

Team	Run	Recall (%)						
		R@1	**R@5**	**R@10**	**R@25**	**R@50**	**R@75**	**R@100**
Random Chance	-	0.00	0.00	0.01	0.01	0.03	0.04	0.05
CEA LIST	1	0.02	0.05	0.11	0.26	0.46	0.67	0.80
	2	0.00	0.04	0.12	0.34	0.71	0.92	1.17
	3	0.01	0.07	0.12	0.38	0.84	1.26	1.61
	4	0.01	0.09	0.16	0.41	0.77	1.24	1.55
	5	0.02	0.06	0.14	0.43	0.78	1.17	1.55
	6	0.00	0.07	0.09	0.17	0.31	0.40	0.55
	7	0.02	0.07	0.18	0.48	0.88	1.32	1.60
INAOE	1	37.05	73.12	78.06	79.55	79.74	79.77	79.77
	2	0.03	0.30	1.22	4.99	11.91	17.33	22.32
	3	0.19	1.55	3.91	9.98	18.43	24.76	29.59

Table 6: Results for Teaser 1: Text Illustration. Recall@k for full 180K test set

Team	Run	Recall (%)						
		R@1	**R@5**	**R@10**	**R@25**	**R@50**	**R@75**	**R@100**
Random Chance	-	0.00	0.00	0.01	0.01	0.03	0.04	0.05
CEA LIST	1	0.02	0.10	0.22	0.48	0.84	1.16	1.44
	2	0.03	0.12	0.23	0.53	0.97	1.38	1.74
	3	0.14	0.56	0.97	1.90	2.98	3.82	4.47
	4	0.18	0.63	1.05	1.97	3.00	3.87	4.51
	5	0.18	0.62	1.04	1.95	2.99	3.85	4.50
	6	0.11	0.36	0.62	1.11	1.68	2.11	2.47
	7	0.18	0.63	1.07	1.93	2.93	3.69	4.33
INAOE	1	28.75	63.50	75.48	84.39	86.79	87.36	87.59
	2	2.57	5.65	7.71	11.76	16.69	20.34	23.40
	3	3.68	7.73	10.46	15.62	21.36	25.48	28.78

This task yielded some interesting results. Bearing in mind the difficulty of the task (selecting one correct image from 200,000 images), **CEA LIST** yielded a respectable score that is clearly better than chance performance. The recall also increased as the rank k is increased. **CEA LIST**'s approach involves mapping visual and textual modalities onto a common space and combining this method with a semantic signature. **INAOE** on the other hand produced excellent results with run 1, which is a retrieval approach based on a bag-of-words representation weighted with tf-idf, achieving a recall of 37% even at rank 1 and almost 80% at rank 100 (in Table 5). In contrast, their runs based on a neural network trained word2vec representation achieved a much lower recall, although it did increase to 29.59% at rank 100. Comparing Tables 5 and 6, both teams performed better on the larger test set of 180,000 generic (and noisy) web text than the smaller test set of 10,112 restricted to news articles. Although interestingly **INAOE**'s bag-of-words approach performed worse at smaller ranks (1-10) for the full test set compared to the news article test set, although still significantly better than their word2vec representation. This increase in overall scores, despite the significant increase in the size of the test set, suggests that there may be some slight overfitting to the training data with most of the methods.

It should be noted that the results of both teams are not directly comparable, as **INAOE** based their submission on the assumption that the webpages for test images are available at test time while **CEA LIST** did not. This assumption made the text illustration problem significantly less challenging since the test documents were extracted directly from these webpages, hence the superior performance by **INAOE**. On hindsight, this should have been specified more clearly in our task description for a level playing field. As such we do not consider one method being superior over the other, but instead concentrate on the technical contributions of each team.

3.6 Limitations of the challenge

There are two major limitations that we have identified with the challenge this year. Very few of the groups used the provided data set and features, we found this surprising, considering the state of the art CNN features and many others were included. However, this is likely to be due to the complexity and challenge of the 510,123 web page based images. Given they were from the Internet with little, a large number of the images are poor representations of the concept. In fact, some participants annotated a significant amount of their more comprehensive training data, as their learning process assumes perfect or near perfect training examples, it will fail. As the number of classes increases and become more varied annotating all comprehensive data will be made more difficult.

Another shortcoming of the overall challenge is the difficulty of ensuring the ground truth has 100% of concepts labelled, thus allowing a recall measure to be used. Especially problematic as the concepts selected include fine-grained categories such as *eyes* and *hands* that are small but frequently occur in the dataset. Also, it was difficult for annotators to reach a consensus in annotating bounding boxes for less well-defined categories such as *trees* and *field*. Given the current crowd-source based hand-labelling of the ground truth, the concepts have missed annotations. Thus, in this edition, a recall measure is not evaluated for subtask 1.

4 Conclusions

This paper presented an overview of the ImageCLEF 2016 Scalable Concept Image Annotation task, the fifth edition of a challenge aimed at developing more scalable image annotation systems. The focus of the three subtasks and teaser task available to participants had the goal to develop techniques to allow computers to annotate the images reliably, localise the different concepts depicted in the images, select important concepts to be described, generate a description of the scene, and retrieve a relevant image to illustrate a text document.

The participation was lower than the previous year, however, in general, the performance of the submitted systems was somewhat superior to last year's results for subtask 1. In part probably due to the increased CNN usage as the feature representation had improved localisation techniques. The clear winner of this year's subtask 1 evaluation was the **CEA LIST** [2] team, which focused on using a state of the art CNN architecture and then also investigated improved localisation of the concepts which helped provide a good performance increase. In contrast to subtask 1, the participants for subtask 2 did not significantly improve the results from last year. The approaches used were very similar to those of last year. For subtask 3, both participating teams concentrated on achieving high recall with traditional approaches like SVM's, compared to last year's winning team which focused on obtaining high precision with a neural network approach. For the pilot teaser task of text illustration, both participating teams performed respectably, with different techniques proposed with varied results. Because of

the ambiguity surrounding one aspect of the task description, the results of the teams are not directly comparable.

The results of the task have been 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 much work, 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 organised. Also, it remains to see how this can be used to complement systems that are based on clean hand-labelled data and find ways to take advantage of both the supervised and unsupervised data.

Table 7: Key details of the best system for top performing groups (subtask 1).

System	Visual Features	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
CEA LIST [2]	16-layer CNN 50-layer ResNet	* Bing Image Search	They collected a set of roughly 251,000 images (1,000 images per concept) from the Bing Images search engine. For each concept they used its name and its synonyms (if present) to query the search engine. They used 90% of the dataset for training and 10% for validation.	They used EdgeBoxes, a generic objectness object detector, extracting a maximum of 100 regions per image then feeding each one to the CNN models. The concept that had the highest probability among the 251 concepts it has been kept.
MRIM-LIG [15]	152-layer ResNet	* Bing Image Search	Two-step learning process using two validation sets. First set of training images,learn the last layer of CNN. Retrain using 200 additional training images defined by the authors according to the low quality recognition concepts	An apriori set of bounding boxes which are expected to contain a single concept each is defined. Each of these boxes have been used as an input image on which the CNN has been applied to detect objects. Localization of parts of faces is achieved through the Viola and Jones approach and facial landmarks detection.
CNRS [18]	VGG deep network	* Google Image Search	2,000 images of the dev set have been used in order to enrich the labels of all the training set transferring the knowledge about the co-occurrence of some labels.	For each concept it has been trained “one-versus-all” SVM classifier.

Acknowledgments

The Scalable Concept Image Annotation Task was co-organized by the VisualSense (ViSen) consortium under the ERA-NET CHIST-ERA D2K 2011 Programme, jointly

Table 8: Key details of the best system for top performing groups (subtask 2).

System	Visual Representation	Textual Representation	Other Used Resources	Summary
ICTisia [29]	VGGNet (FC7)	LSTM	* MSCOCO * Flickr8K/Flickr30K * Manually selected image-caption pairs, captions generated from training set	CNN-LSTM caption generator, but fine-tuning *both* CNN and LSTM parameters. Also fine-tune on different datasets.
UAIC [3]	TensorFlow CNN (architecture unknown)	Text labels	* Face recognition module * WordNet * DuckDuckGo	Concept detection using textual features and visual feature (subtask 1), and generate descriptions using templates (with backoff).

Table 9: Key details of the best system for top performing groups (subtask 3).

System	Representation	Content Selection Algorithm	Summary
DUTh [1]	* Bounding box (Position, size) * Local descriptors (Canny, Harris, BRISK, SURF, FAST) * Entropy	Nonlinear, binary SVM classifier (RBF, Polynomial kernels)	SVM classifier to classify whether a bounding box is important/not important, using combinations of local features.
UAIC [3]	Text labels	Selection by generating descriptions (subtask 2).	Bounding box selection by selecting up to three tuples (concept1, verb, concept2).

Table 10: Key details of the best system for top performing groups (teaser task).

System	Visual Representation	Textual Representation	Other Used Resources	Summary
CEA LIST [2]	VGGNet (FC7)	word2vec (TF-IDF weighted average)	* WordNet * Flickr Groups * NLTK Pos Tagger	Two methods: (i) Semantic signature - fixed sized vector, each element corresponding to a semantic concept. Inverse indexing for retrieval. (ii) Projection onto common, bimodal latent space via kCCA.
INAOE [14]	_	* TF-IDF weighted bag of words * word2vec (simple average)	_	IR queries using bag-of-words or word2vec.

supported by UK EPSRC Grants EP/K01904X/1 and EP/K019082/1, French ANR Grant ANR-12-CHRI-0002-04 and Spanish MINECO Grant PCIN-2013-047. The task was also supported by the European Union (EU) Horizon 2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943).

References

1. Barlas, G., Ntonti, M., Arampatzis, A.: DUTh at the ImageCLEF 2016 Image Annotation Task: Content Selection. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
2. Borgne, H.L., Gadeski, E., Chami, I., Tran, T.Q.N., Tamaazousti, Y., Gînscă, A.L., Popescu, A.: Image annotation and two paths to text illustration. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
3. Cristea, A., Iftene, A.: Using Machine Learning Techniques, Textual and Visual Processing in Scalable Concept Image Annotation Challenge. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
4. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
5. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 452–457. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
6. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111(1), 98–136 (Jan 2015)
7. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA; London (May 1998)
8. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandréa, E., Gaizauskas, R.J., Villegas, M., Mikolajczyk, K.: Overview of the imageclef 2015 scalable image annotation, localization and sentence generation task. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015), http://ceur-ws.org/Vol-1391/inv-pap6-CR.pdf
9. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research (JAIR) 47(1), 853–899 (May 2013)
10. 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
11. 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
12. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014), http://arxiv.org/abs/1405.0312

13. 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
14. Pellegrin, L., López-Monroy, A.P., Escalante, H.J., Montes-Y-Gómez, M.: INAOE's participation at ImageCLEF 2016: Text Illustration Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
15. Portaz, M., Budnik, M., Mulhem, P., Poignant, J.: MRIM-LIG at ImageCLEF 2016 Scalable Concept Image Annotation Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
16. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: Article annotation by image and text processing. CoRR abs/1603.07141 (2016), http://arxiv.org/abs/1603.07141
17. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) pp. 1–42 (April 2015)
18. Sahbi, H.: CNRS TELECOM ParisTech at ImageCLEF 2016 Scalable Concept Image Annotation Task: Overcoming the Scarcity of Training Data. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)
19. 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
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint arXiv:1409.1556 (2014)
21. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2), 154–171 (2013)
22. Villegas, M., Müller, H., García Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science, Springer International Publishing (2016)
23. 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)
24. 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
25. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF2014 Working Notes. CEUR Workshop Proceedings, vol. 1180, pp. 308–328. CEUR-WS.org, Sheffield, UK (September 15-18 2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Image-VillegasEt2014.pdf
26. 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

27. Wang, J., Gaizauskas, R.: Generating image descriptions with gold standard visual inputs: Motivation, evaluation and baselines. In: Proceedings of the 15th European Workshop on Natural Language Generation (ENLG). pp. 117–126. Association for Computational Linguistics, Brighton, UK (September 2015), http://www.aclweb.org/anthology/W15-4722
28. Wang, J.K., Yan, F., Aker, A., Gaizauskas, R.: A poodle or a dog? Evaluating automatic image annotation using human descriptions at different levels of granularity. In: Proceedings of the Third Workshop on Vision and Language. pp. 38–45. Dublin City University and the Association for Computational Linguistics, Dublin, Ireland (August 2014)
29. Zhu, Y., Li, X., Li, X., Sun, J., Song, X., Jiang, S.: Joint Learning of CNN and LSTM for Image Captioning. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 2016)

A Concept List 2016

The following tables present the 251 concepts used in the ImageCLEF 2016 Scalable Concept Image Annotation task. In the electronic version of this document, each concept name is a hyperlink to the corresponding WordNet synset webpage.

Concept	WordNet type	WordNet sense#	#dev.	#test
airplane	noun	1	22	76
anchor	noun	1	-	7
apple	noun	1	3	8
apron	noun	1	2	28
arm	noun	1	83	4352
bag	noun	1	37	150
bag	noun	4	28	88
ball	noun	1	36	63
balloon	noun	1	7	12
banana	noun	1	2	2
barn	noun	1	6	4
baseball_glove	noun	1	10	27
basin	noun	1	2	20
basket	noun	1	12	7
bat	noun	1	-	-
bathroom	noun	1	5	8
bathtub	noun	1	2	1
beach	noun	1	27	5
bear	noun	1	7	20
beard	noun	1	22	178
bed	noun	1	32	31
bee	noun	1	1	5
beer	noun	1	3	10
bell	noun	1	1	-
bench	noun	1	36	81
bicycle	noun	1	30	56
bin	noun	1	22	49
bird	noun	1	14	48
blackberry	noun	1	-	1
blanket	noun	1	17	55
boat	noun	1	76	104
bomb	noun	1	1	5
book	noun	1	30	45
boot	noun	1	19	101
bottle	noun	1	42	81
bouquet	noun	1	-	-
bowl	noun	1	12	24
box	noun	1	28	86
bread	noun	1	8	6
brick	noun	1	21	116
bridge	noun	1	34	80
bucket	noun	1	9	19
bullet	noun	1	2	2
bus	noun	1	25	94
butter	noun	1	2	-
butterfly	noun	1	1	1
cabinet	noun	1	29	89
camera	noun	1	18	37
can	noun	1	8	4
canal	noun	1	5	13
candle	noun	1	7	9
candy	noun	1	2	30
cannon	noun	1	4	13
cap	noun	1	67	223
car	noun	1	181	603
cat	noun	1	5	20
cathedral	noun	1	15	58
cave	noun	1	4	5
ceiling	noun	1	21	124
chair	noun	1	111	448

Concept	WordNet type	WordNet sense#	#dev.	#test
cheese	noun	1	1	1
city	noun	1	37	36
cliff	noun	1	9	22
clock	noun	1	5	3
computer	noun	1	14	41
corn	noun	1	-	-
cow	noun	1	19	66
crab	noun	1	3	3
cross	noun	1	4	30
cup	noun	1	20	96
curtain	noun	1	41	127
dam	noun	1	2	2
deer	noun	1	13	57
dish	noun	1	13	71
dog	noun	1	49	76
doll	noun	1	8	11
door	noun	1	87	429
dress	noun	1	100	384
drill	noun	1	3	-
drum	noun	1	13	25
dryer	noun	1	2	-
ear	noun	1	27	1803
egg	noun	1	-	1
elephant	noun	1	9	23
eye	noun	1	39	2783
face	noun	1	43	3205
fan	noun	1	4	2
farm	noun	1	3	3
feather	noun	1	2	3
female_child	noun	1	72	206
fence	noun	1	94	423
field	noun	1	185	163
fireplace	noun	1	9	8
fish	noun	1	9	36
flag	noun	1	35	131
flashlight	noun	1	1	2
floor	noun	1	69	327
flower	noun	1	96	359
foot	noun	1	14	1291
fork	noun	1	7	5
fountain	noun	1	10	7
fox	noun	1	-	5
frog	noun	1	1	2
fruit	noun	1	6	17
garden	noun	1	35	142
gate	noun	1	12	58
goat	noun	1	12	7
grape	noun	1	-	7
guitar	noun	1	26	42
gun	noun	1	20	34
hair	noun	1	121	2644
hallway	noun	1	13	82
hammer	noun	1	3	2
hand	noun	1	170	3455
hat	noun	1	92	391
head	noun	1	30	3861
helicopter	noun	1	8	16
helmet	noun	1	51	186
hill	noun	1	19	85

continues in next page

Concept	WordNet type	WordNet sense#	#dev.	#test
hog	noun	3	1	24
hole	noun	1	1	6
hook	noun	1	1	11
horse	noun	1	58	83
hospital	noun	1	1	2
house	noun	1	135	725
jacket	noun	1	60	654
jean	noun	1	51	370
key	noun	1	1	5
keyboard	noun	1	10	6
kitchen	noun	1	9	8
knife	noun	1	5	8
ladder	noun	1	14	32
lake	noun	1	28	74
leaf	noun	1	116	134
leg	noun	1	30	3185
letter	noun	1	13	46
library	noun	1	2	1
lighter	noun	2	1	537
lion	noun	1	9	5
lotion	noun	1	-	4
magazine	noun	1	7	20
male_child	noun	1	89	260
man	noun	1	681	2962
mask	noun	1	12	15
mat	noun	1	6	5
mattress	noun	1	3	10
microphone	noun	1	27	67
milk	noun	1	1	1
mirror	noun	1	19	75
monkey	noun	1	4	7
motorcycle	noun	1	22	61
mountain	noun	1	85	77
mouse	noun	1	1	1
mouth	noun	1	48	1568
mushroom	noun	1	-	6
neck	noun	1	14	1400
necklace	noun	1	50	37
necktie	noun	1	33	210
nest	noun	1	1	2
newspaper	noun	1	16	26
nose	noun	1	16	1970
nut	noun	1	1	2
office	noun	1	9	3
onion	noun	1	-	-
orange	noun	1	1	9
oven	noun	1	1	6
painting	noun	1	45	156
pan	noun	1	2	4
park	noun	1	27	344
pen	noun	1	11	14
pencil	noun	1	4	5
piano	noun	1	9	9
picture	noun	1	25	158
pillow	noun	1	19	48
planet	noun	1	-	1
pool	noun	1	23	20
pot	noun	1	4	17
potato	noun	1	3	2
prison	noun	1	-	-
pumpkin	noun	1	1	9
rabbit	noun	1	5	11
rack	noun	1	10	1
radio	noun	1	1	14
ramp	noun	1	3	3
ribbon	noun	1	11	45

Concept	WordNet type	WordNet sense#	#dev.	#test
rice	noun	1	-	-
river	noun	1	51	82
rock	noun	1	94	239
rocket	noun	1	4	9
rod	noun	1	7	31
rug	noun	1	35	52
salad	noun	1	1	2
sandwich	noun	1	3	5
scarf	noun	1	23	67
sea	noun	1	107	215
sheep	noun	1	7	10
ship	noun	1	50	183
shirt	noun	1	153	1946
shoe	noun	1	59	1145
shore	noun	1	41	93
short_pants	noun	1	39	368
signboard	noun	1	91	624
skirt	noun	1	16	120
snake	noun	1	9	6
sock	noun	1	7	185
sofa	noun	1	36	62
spear	noun	1	1	-
spider	noun	1	1	-
stadium	noun	1	27	99
star	noun	1	2	1
statue	noun	1	35	84
stick	noun	1	17	156
strawberry	noun	1	-	1
street	noun	1	143	440
suit	noun	1	77	199
sunglasses	noun	1	45	144
sweater	noun	1	33	107
sword	noun	1	5	5
table	noun	2	125	320
tank	noun	1	7	10
telephone	noun	1	6	20
telescope	noun	1	4	1
television	noun	1	10	29
temple	noun	1	14	26
tent	noun	1	10	57
theater	noun	1	2	19
toilet	noun	1	5	5
tongue	noun	1	4	17
towel	noun	1	6	20
tower	noun	1	32	93
town	noun	1	10	199
tractor	noun	1	7	7
train	noun	1	13	27
tray	noun	1	3	28
tree	noun	1	460	1444
truck	noun	1	44	86
tunnel	noun	1	3	3
valley	noun	1	13	29
vase	noun	1	14	26
vest	noun	1	10	113
wagon	noun	1	6	14
wall	noun	1	104	855
watch	noun	1	29	93
waterfall	noun	1	1	4
well	noun	1	-	1
wheel	noun	1	52	331
wicket	noun	1	-	5
window	noun	1	134	1308
wine	noun	1	10	25
wolf	noun	1	2	1
woman	noun	1	474	1491