Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task

Mauricio Villegas and Roberto Paredes

Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mvillegas,rparedes}@iti.upv.es

Abstract. The ImageCLEF 2012 Scalable Image Annotation Using General Web Data Task proposed a challenge, in which as training data instead of relying only on a set of manually annotated images, the objective was to make use of automatically gathered Web data, with the aim of developing more scalable image annotation systems. To this end, the participants were provided with a new dataset, composed of 250,000 images for training, which included various visual feature types, and textual features obtained from the websites in which the images appeared. Two subtasks were defined. The first subtask employed the same test set as the ImageCLEF 2012 Flickr Photo Annotation subtask, with the particularity that both the Flickr and Web training sets had to be used. The idea was to determine if the Web data could help to enhance the annotation performance in comparison to using only manually annotated data. The second subtask consisted in using only automatically gathered Web data to develop an image annotation system. For this, we provided a development and test sets of 1,000 and 2,000 images, respectively, both manually annotated for 95 and 105 concepts, respectively. The participants of the first subtask were not able to take advantage of the Web data to enhance the annotation performance. On the contrary, in the second subtask interesting results were obtained. As expected, the overall performance of the systems is worse than using manually annotated data, nonetheless, the results are promising when analyzing per concept. For some concepts the performance is relatively good, confirming that the Web data can in fact be quite useful. Moreover, due to the low participation and the relatively simple techniques used, it is believed that there is considerable room for improvement on both subtasks.

1 Introduction

The rapidly increasing amount of digital information that people have to deal with every day, has created huge interest in developing automatic indexing systems, so that information needs can be easily and efficiently fulfilled. In the case of images and video, this indexing can be addressed by means of an automatic image annotation system, in which images are associated with one or more concepts. The research on image concept detection has generally relied on

training data that have been manually, and thus reliably labeled, an expensive and laborious endeavor that cannot easily scale. Because of this, it has become common in past image annotation benchmark campaigns [6,10] to use crowdsoursing approaches such as the Amazon Mechanical Turk[1] (MTurk), in order to label a large amount of images. Still, crowdsoursing is expensive and difficult to scale to a very large amount of concepts, thus it is advisable to explore possible alternatives.

With the advance of multimedia technology and the Internet, we have at our disposal billions of images available online. Furthermore, the images are found on webpages surrounded with text which might have a direct relationship with the content of the image. Even though this surrounding unsupervised text is noisy and sometimes unrelated to the image, it potentially has useful information, and furthermore it can be cheaply gathered and be obtained for practically any topic. Thus, determining whether this kind of data can be used for reliably annotating images is important. Previous research indicates that this information is useful, being the work of Torralba et al. [11] an example of this, in which almost 80 million tiny images were effectively used for several tasks such as person detection. More closely related, in the work of Weston et al. [16] an image annotation learning method is proposed that scales to millions of images and thousands of possible annotations. Another related work is the Arista project [15] in which accurate tags can be generated for popular Web images that have near-duplicates included in their Web image database of billions of images.

This paper presents an overview of the ImageCLEF 2012 Scalable Image Annotation Using General Web Data Task, a benchmark campaign oriented at using automatically gathered Web data for image annotation. The paper is organized as follows. Section 2 describes the generation of the dataset that was created specifically for this evaluation. Followed by this, the two subtasks that were defined are presented in section 3. Then, section 4 presents the results submitted by the participants and a discussion of these. Finally, section 5 is the conclusion of the paper.

2 Creation of the Dataset

2.1 Web Crawling

Among the objectives for the dataset being created [14], was to have a wide variety of images with a relatively small amount of images (not billions). Thus in order to obtain a good set of image URLs, we opted to use the same crawling strategy as in [11], where the image URLs (and the corresponding URLs of the webpages that contain the images) are obtained by querying popular image search engines. We selected Google, Bing and Yahoo, and queried them using words from the English dictionary that comes with the aspell spelling checker.

The next step in the crawling process was to download the images and corresponding webpages, and store a snapshot of these. At the end, in total we

[1] www.mturk.com

obtained over 31 million of both images and webpages. In order to avoid duplicate images, several precautions were taken. First the URLs were normalized to prevent different versions of the same URL to be downloaded several times. However, there was also the possibility that the same image was found under different URLs. To account for this, the images were stored using a unique code or image identifier, composed of: part of the MD5 checksum of an 864-bit image signature (in some aspects similar to the one presented in [17]) and, part of the MD5 checksum of the file. This scheme guaranties storing exactly the same file only once and easily identifying duplicates or near duplicates (accounting for images in various formats, at different resolutions and with minor modifications such as some watermarks). The final image identifiers are 16 base64url digits.

2.2 Image Subset Selection

Even though the set of downloaded images was obtained using all of the words of the English dictionary, and therefore it contains images from practically any topic, a subset of the images was selected for practical reasons. Basically selecting a subset permitted to provide smaller data files that would not be so prohibitive for the participants to download and handle. Furthermore, since the test sets had to be manually labeled and this could be done only for a relatively small list of concepts, we could select only the images indexed with words related to the list of concepts. The size of the training set was chosen to be of 250,000 images, which results in feature vector sets of moderate size that can be easily handled on current personal computers.

Another reason for selecting a subset, was to discard some types of images. Even though the URLs were obtained from trustworthy search engines, inevitably there is a certain amount of problematic images that we decided to remove. Among the problematic images are for instance a message saying "Image removed", or dummy images some servers send specifically to web crawlers. Removing this type images is in itself a difficult problem, however we noticed that most of these tended to have many different URLs linking to them or be images that appeared in a large amount of webpages. So the approach to remove most of these was simply not to include images that had more than N URLs linking to them or that appeared in more than M webpages. The values of N and M were set manually from a quick look at the images being considered for removal.

When crawling the Web, another problem encountered was that an image could be still reachable, although the webpage where it appeared has changed and no longer includes the image, or has been removed and does not supply the proper HTTP 404 code. Resolving this issue was simple due to the requirements of the dataset. For each image in the dataset there was supposed to be at least one webpage that contained the image, thus we verified the location of the images within the webpages. Any image not having a corresponding webpage was obviously not considered for inclusion.

The image identifier codes guarantied storing exactly the same image file only once. However, for this subset selection we also employed a very simple near im-

age duplicate removal scheme. This was done using the same image signature mentioned in the previous section. To reduce the amount computation required for duplicate detection, the 864-bit image signatures were first reduced to 128 bits using PCA followed by a random rotation and thresholding [3]. The duplicate removal scheme was not including images that had a normalized hamming distance lower than 0.1 to the other images already in the subset.

The final selection of the 250,000 images was based on a list of 158 concepts that were manually defined. This list included all of the concepts of the test sets of both subtasks described on section 3. A small set of 3,000 images was first manually labeled using 115 of the concepts. These are the same images and ground truth labels used as development and test sets in subtask 2. Then for each concept (defined by the concept words and synonyms of these) and co-occurrences of concepts in the labeled set, we retrieved ranked lists of images, using the query results from the search engines, and also querying our own image index generated from the downloaded webpages. The lists were sorted by rank and the first 250,000 images were the ones selected for the final dataset.

2.3 Available Data

This dataset was made available under a Creative Commons license, however, since the data was gathered from the Internet, and their original copyright conditions are difficult to determine automatically, only the feature vectors were distributed[2]. Nonetheless, as it is commonly done on image search engines, thumbnails of the images could be obtained from a web server by using the image identifiers[3].

For each of the 250,000 training images, both the textual and visual features described next were available. For the additional 3,000 images used in subtask 2 that were manually labeled, and the 25,000 flickr images used in subtask 1, only the visual features were included.

Textual Features: Four sets of textual features were extracted. First is the list of the words used to find the image when querying the search engines, along with the rank position of the image in the respective query and the search engine it was found on. The second textual features were 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, this is why they can also be useful as textual features. The other two textual features available correspond to text extracted from the webpages near the position of the image. The difference between these two feature sets was the amount of preprocessing.

To extract the text near the image, we first converted the webpages to valid XML to ease processing and removed the script and style elements. The text considered close, was the webpage title and all the terms that are closer than 600 in word distance to the image, not including the HTML tags and attributes.

[2] http://risenet.iti.upv.es/webupv250k
[3] http://risenet.iti.upv.es/db/img/{IID}.jpg

The first level of processing of the features included this raw text, although some types of terms were converted to a special symbol, such as words with non-latin characters. In these features, the position of the image was also indicated and the words replaced by a special symbol serve to preserve word distances.

For the next level of processing, 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. [4], 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: As features extracted from the images, we made available seven types. As preprocessing, we filtered the images and resized them so that the width and height had at most 240 pixels while preserving the original aspect ratio. The first feature set were 576-dimensional color histograms extracted using our own implementation. The second set of features were the GIST [7]. The other four features were obtained using the colorDescriptors software [8]. We computed features for SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT. As configuration we used dense sampling with default parameters, and a hard assignment 1,000 and 10,000 codebooks using a spatial pyramid of 1×1 and 2×2 [5]. Since the vectors of the spatial pyramid were concatenated, this resulted in 5,000-dimensional and 50,000-dimensional feature vectors, respectively. Keeping only the first fifth of the dimensions would be like not using the spatial pyramid. The codebooks were generated using 1.25 million randomly selected features and the k-means algorithm.

Individual features (i.e. without using a codebook) were also made available for SIFT, C-SIFT, RGB-SIFT, OPPONENT-SIFT, and the seventh feature type SURF, extracted using the TOP-SURF software [9]. In this case the preprocessing was a little different since these were extracted exactly the same as for the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task [10] to ease participation in both tasks. The images were filtered with the catrom filter and resized to 256×256 pixels ignoring the original aspect ratio.

3 Task Description

As commented before, a very large amount of images can be cheaply 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, thus this data can be considered to be very noisy. Moreover, the webpages can be of any language

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

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

Fig. 1: Example of images retrieved by a Web search engine for different queries.

or even a mixture of languages, and they tend to have many writing mistakes. The goal of this task is to evaluate different strategies to deal with noisy data so that it can be reliably used for annotating images from practically any topic.

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

Based on these observations, an interesting research topic would be: how to use and handle the automatically retrieved noisy Web data to complement the manually labeled training data and obtain a better performing annotation system than when using the manually labeled data alone. On the other hand, since the Web data can easily be obtained for any topic, another research topic would be: how to use the noisy Web data to develop an annotation system with a somewhat unbounded list of concepts, using only automatically retrieved image and textual Web data.

Both of the research topics just mentioned have been addressed in two separate subtasks.

3.1 Subtask 1: Complementing Manually Annotated Data

In this subtask the list of concepts and the test samples were exactly the same as the ones used in the ImageCLEF 2012 Flickr Photo Annotation subtask [10]. The ImageCLEF 2012 Flickr dataset consisted of a training and test sets of 15,000 and 10,000 images, respectively, that were manually labeled by means of crowdsoursing using a list of 94 concepts. For further details on this dataset, the reader should refer to the overview paper of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task [10].

In this subtask, the participants had available for developing their annotation systems, both the Flickr and Web training datasets. The objective was to develop techniques to take advantage of the Web data, trying to obtain better concept annotation performance in comparison to using only the Flickr manually annotated data. The participants had to submit results using as training only the Flickr dataset, and using both the Flickr and Web datasets.

3.2 Subtask 2: Scalable Concept Image Annotation

In this subtask, the objective was to develop systems that could easily change or scale the list of concepts used for image annotation. In other words, the list of concepts is also considered to be an input to the system. Thus, the system when given an input image and a list of concepts, its job is to give a score to each of the concepts in the list and decide how many and which of them assign as annotations. To observe this scalable characteristic of the systems, the list of concepts was different for the development and test sets, and the participants only had available the ground truth annotations for the development set.

The idea was that the participants use the 250,000 images of the Web training set, including the visual and textual features (see Section 2), to develop and estimate the models for image annotation. It was not permitted to use any manually annotated data, such as the Flickr training set. However, the use of other additional language resources, such as language models, language detectors, stemmers, WordNet [2], spell checkers, etc., was permitted and encouraged.

The development set consisted of 1,000 images annotated for 95 concepts, and the test set consisted of 2,000 images for 105 concepts, among which 85 were common to the development set, i.e. 10 concepts were removed and 20 were added. The list of concepts and the number of images for each can be observed in Table 1. So that the Web training would be the same as for subtask 1, for this first edition of the task, the list of concepts overlapped considerably with the concepts of the Flickr annotation task.

For this subtask, so that there could be a reference performance and also serve as a starting point, a toolkit was supplied to the participants. This toolkit included software that computed the evaluation measures (see Section 4.1), and the implementations of two baselines. The first baseline was a simple random, which is important since any system which gets worse performance than the random baseline means that this system is doing nothing.

Table 1: The number of images for the ground truth annotations per concept for the development and test sets of subtask 2.

Concept	Dev.	Test	Concept	Dev.	Test	Concept	Dev.	Test
aerial	30	56	indoor	37	104	traffic	24	40
airplane/helicopter	21	34	lake	26	45	train/tram/metro	31	27
baby	9	31	lightning	9	16	tree	183	287
beach	34	45	logo	12	30	truck	19	33
bicycle/tricycle	20	24	moon	6	29	underwater	22	54
bird	21	24	motorcycle	10	17	unpaved	12	22
boat	50	75	mountain	85	148	water	177	280
book	20	20	music+instrument	31	56	clouds	138	-
bridge	32	43	newspaper	7	10	computer+generated	18	-
building	124	199	nighttime	25	36	drums	9	-
car	30	71	outdoor	135	254	fog	13	-
cartoon	21	51	overcast	24	27	highway	14	-
castle	17	21	painting	23	59	lying	5	-
cat	11	20	person/people	178	427	portrait	8	-
child	18	48	plant	64	110	standing	4	-
church	12	14	poster	6	12	stream	18	-
cityscape	58	79	protest	9	19	vehicle	12	-
daytime	77	191	rain	10	26	bottle	-	26
desert	16	23	rainbow	9	14	bus	-	38
dirt	16	46	reflection	52	60	chair	-	39
dog	27	31	river	55	77	drink	-	38
drawing/diagram	71	155	road	108	189	galaxy	-	16
droplets	14	25	rural	36	61	glass	-	75
elder	10	28	sand	29	65	glasses	-	31
embroidery	9	13	sculpture	22	54	hat	-	39
fire	28	31	sea	72	94	insect	-	51
fireworks	11	20	shadow	29	48	nebula	-	13
fish	16	31	sign	51	76	pencil	-	34
flower	43	98	silhouette	12	25	phone	-	23
food	19	55	sitting	8	18	pool	-	30
footwear	10	23	sky	197	325	reptile	-	32
forest	62	84	smoke	15	14	rodent	-	44
furniture	39	97	snow	43	74	space	-	72
garden/park	39	62	sports	25	86	submarine	-	24
graffiti	14	10	stars	3	47	table	-	33
grass	99	175	sun	26	68	violin	-	21
guitar	6	12	sunrise/sunset	33	45	wagon	-	24
harbor/port	19	33	teenager	12	22			
horse	18	46	toy	21	27			

The other baseline, referred to as Co-occurrence Baseline, was a very basic technique for this image annotation task, which obviously gives better performance than random, although it was simple enough to give the participants a wide margin for improvement. In this technique when given an input image, its nearest $K = 32$ images from the training set are obtained, using only the 1,000 bag-of-words C-SIFT visual features and the L1 norm. Then, the textual features corresponding to these K nearest images are used to derive a score for each of the concepts. This is done by using a concept-word co-occurrence matrix estimated from all of the training set textual features. In order to make the vocabulary size more manageable, the textual features are first processed keeping only the words from the English dictionary. Finally for the selection of concepts for annotation, for all input images, the first 5 ranked concepts are always chosen as annotations.

4 Evaluation Results

4.1 Performance Measures

The participants were asked to submit the results in the following way. For each image to annotate, a score had to be given for every one of the concepts in the list and also indicate which concepts had finally been selected as annotations.

Two basic performance measures have been used for comparing the results of the different submissions. These basic measures are the Average Precision (AP) and the F-measure (F_1). The AP only takes into account the scores assigned to the concepts and ignores the decisions of the selected annotations. On the other hand, the F_1 only considers the the selected annotations.

The AP is algebraically defined as

$$\mathrm{AP} = \frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} \frac{k}{\mathrm{rank}(k)} \ , \tag{2}$$

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.

In the context of image annotation, the AP can be estimated from two different perspectives, one being concept-based and the other example-based. In the former, one AP is computed for each concept, and in the latter one AP is computed for each image to annotate. Which of these is more correct to use actually depends on exactly what the scores are. If the scores for example relate to the probability that the concept is present for a given image, and the comparison between scores for different images is not clearly defined, then the concept-based AP does not make sense and will probably not be a good indicator of the performance of the system. On the other hand, in this case the example-based AP will be a good indicator of the performance of the system. In the instructions given to the participants, this was not clearly explained, however, for all of the submissions, the scores seemed to be image based. Therefore, in this paper we present results only for the example-based AP. Finally, to obtain a global performance measure of the systems, we have taken the arithmetic mean, in which case it is known as the Mean Average Precision (MAP).

The other performance measure used, the F_1, is defined as

$$\mathrm{F}_1 = \frac{2PR}{P+R} \ , \tag{3}$$

where P is the precision and R is the recall. Again this measure can also be estimated from the concept-based and the example-based perspectives. In this case both approaches are adequate and serve to analyze different aspects. For the example-based F_1, as a global system performance measure the arithmetic

mean is used, thus obtaining a mean F-measure (MF_1). On the other hand, the concept-based F_1 is used to analyze the behavior for different concepts.

Other performance measures were computed and analyzed, however, for the received submissions they do not give any important details that are not already observed with the previously mentioned measures. Therefore for simplicity, we are not including them in this paper. These other measures were, the AP using the geometric mean and the interpolated versions, i.e. the Geometric Mean Average Precision (GMAP) the Interpolated Average Precision (IAP), the Mean Interpolated Average Precision (MIAP) and the Geometric Mean Interpolated Average Precision (GMIAP).

4.2 Participation

In total, 47 groups registered for the task and signed the license agreement, and therefore had access for downloading the datasets. Unfortunately in the end, the participation was considerably low. For subtask 1 we received 15 runs from three groups, and for subtask 2 we received 10 runs from one group. Also, one of the groups that submitted results for subtask 1 said that they made a mistake and did not intend to participate in the task.

KIDS-NUTN: The Knowledge, Information, and Database System Laboratory (KIDS-NUTN), from the National University of Tainan [1] submitted in total 9 runs. All of the runs were for subtask 1, being 5 using only the Flickr training set and the other 4 using both the Flickr and Web training sets. They used a combination of several visual feature types, namely AutoColorCorrelogram, ColorLayout, FCTH, Gabor, GIST, and ROI background, and as textual features they used the EXIF data. For the annotation, they used Random Forests and for comparison the also tried as a baseline the Multiple Bernoulli Relevance Models (MBRM). For further details, please refer to [1].

ISI: The Intelligent Systems and Informatics Laboratory (ISI), from the University of Tokyo [13] submitted in total 20 runs. Half of the runs were for subtask 1, being 6 using only the Flickr training set and the other 4 using both the Flickr and Web training sets. For the other 10 runs for subtask 2, half correspond to the development set, and the other half to the test set. Their effort was targeted at making the system scalable, so for annotation they used the Passive-Aggressive with Averaged Pairwise Loss (PAAPL) [12], which is an online learning method they propose for multiclass multilabel classification using a linear model. As visual features, they used the provided *SIFT features, and to tackle the Web data, they artificially labeled it by looking at the textual features and if a word that defined a concept appeared, then that concept was assumed to be present. The images that did not have any concept were simply discarded. In subtask 1, they tried first to learn models for the Flickr and Web data separately, and combine the results, and second they tried learning the models by merging all of the data. All of their submissions were using the latter approach, since during development it was the one that performed best. The difference between submissions is simply the combination of visual features. In subtask 2, again

Table 2: Results for subtask 1, (2a) best result for each group for the submissions that only used Flickr training data, and the random baseline, and (2b) all of the submissions for each group using both Flickr and Web training data.

(a)

	MAP	**MF$_1$**
Random Baseline	0.103	0.100
ISI 1424	0.708	0.553
KIDS-NUTN 1451	0.579	0.454

(b)

	MAP	**MF$_1$**
ISI 1393	0.250	0.182
ISI 1398	0.247	0.181
ISI 1399	0.245	0.178
ISI 1400	0.241	0.175
KIDS-NUTN 1369	0.521	0.399
KIDS-NUTN 1370	0.538	0.397
KIDS-NUTN 1371	0.493	0.331
KIDS-NUTN 1372	0.528	0.400

they artificially labeled the training data for learning, and the submissions differ in the combination of visual features. For further details, please refer to [13].

4.3 Results and Discussion for Subtask 1

The results for subtask 1, given by the example-based MAP and the MF$_1$, are presented in Tables 2a and 2b. The first table includes the best result of each group when using as training only the Flickr manually annotated data, and a baseline, which is randomly assigning scores to the concepts and selecting randomly the top N as annotations. The second table includes results for all of the submissions using both the Flickr and Web training data.

As can be observed in the tables, all of the results using both the Flickr and Web training data have a worse performance than when using only Flickr data. In the case of the KIDS-NUTN, the difference between using or not using Web data is not so high. When we inquired them about these results, they answered that the textual data of the Web dataset did not help, but they did not give us a clearer explanation on how they arrived to this conclusion or exactly to what the submissions correspond. Moreover, they said that they were not able to dedicate much time on the problem.

Regarding the results of ISI, the difference between using or not using Web data is considerably high. In fact it seems that during development [13] they obtained better results, and these did not generalize to the test set. For the Flickr task, they obtained an MF$_1$ higher than 0.5 both during development and during test. However, using both Flickr and Web training data they obtained

Table 3: Annotation results for all of the submissions for subtask 2 and the Random and Co-occurrence baselines for (3a) the development and (3b) the test sets.

(a)

	MAP	**MF_1**
Random Baseline	0.084	0.063
Co-occurrence Baseline	0.222	0.168
ISI 1406	0.331	0.262
ISI 1409	0.336	0.260
ISI 1410	0.340	0.267
ISI 1413	0.338	0.266
ISI 1414	0.333	0.264

(b)

	MAP	**MF_1**
Random Baseline	0.067	0.055
Co-occurrence Baseline	0.221	0.171
ISI 1407	0.315	0.246
ISI 1408	0.322	0.251
ISI 1411	0.324	0.252
ISI 1412	0.323	0.254
ISI 1415	0.321	0.249

an MF_1 in the order of 0.48 during development in contrast to the 0.18 they obtain for the test set. This suggests that possibly there was some mistake or something was different in the models used for annotating the test set. In fact in subtask 2 (see Section 4.4), they obtain better results even though they have used the same technique, and the problem is harder since only Web data can be used as training and the random baseline is lower.

4.4 Results and Discussion for Subtask 2

Tables 3a and 3b present the results for the example-based MAP and MF_1 for the development and test sets, respectively. The tables include the results for the submitted runs and the two baselines, assigning random scores and selecting the random top N concepts per image, and the co-occurrence baseline as described in Section 3.2. The first thing to note is that the results for the development set, generalizes well to the test set, unlike what was observed in the ISI results for subtask 1. The second thing to note is that the submitted runs have a considerably better performance than the supplied co-occurrence baseline. This is a great achievement, even though the co-occurrence baseline is considerably simple. Unfortunately, there was only one participant, so there was no much competition, and definitely it is not possible to say that this level of performance is more or less what can be achieved using Web data as training, so that it could be compared to using labeled data as training. Furthermore, the ISI system can also be

Table 4: F_1 ranges per concept when training with (4a) manually annotated data (ISI 1424), and (4b) automatically gathered Web data (ISI 1411).

(a)

Concepts	F_1 range
none, noblur, dog, fireworks, flower, partialblur, fooddrink, adult	$0.6 \leq F_1 < 1.0$
female, outdoor, tree, bird, coast, citylife, day	$0.5 \leq F_1 < 0.6$
male, stars, moon, car, graffiti, baby	$0.4 \leq F_1 < 0.5$
insect, closeupmacro, cycle, fish, indoor, silhouette, cat	$0.3 \leq F_1 < 0.4$
clearsky, rainbow, flames, lenseffect, sun, circularwarp, partylife, calm, homelife, grass, sunrisesunset, air	$0.2 \leq F_1 < 0.3$
child, forestpark, underwater, portrait, fogmist, horse, inactive, smoke, overlay	$0.1 \leq F_1 < 0.2$
rail, reflection, overcastsky, other, motionblur, shadow, seaocean, water, biggroup, familyfriends, plant, rural, pictureinpicture, artifacts, sportsrecreation, completeblur, happy, riverstream, lightning, mountainhill, graycolor, melancholic, active, unpleasant, elderly, teenager, amphibianreptile, spider, smallgroup, three, two, coworkers, strangers, desert, euphoric, scary, truckbus, lake, snowice	$0.0 \leq F_1 < 0.1$

(b)

Concepts	F_1 range
fireworks, pencil, stars	$0.6 \leq F_1 < 1.0$
drawing/diagram, galaxy	$0.5 \leq F_1 < 0.6$
newspaper, lightning, forest, pool, fire, aerial, horse, bicycle/tricycle, protest	$0.4 \leq F_1 < 0.5$
sky, building, nebula, mountain, cartoon, church, footwear, logo, lake, moon, grass, road, underwater, tree, snow, painting	$0.3 \leq F_1 < 0.4$
water, plant, desert, furniture, airplane/helicopter, beach, sun, food, guitar, flower, train/tram/metro, boat, rainbow, silhouette, sand, glass, harbor/port	$0.2 \leq F_1 < 0.3$
river, bus, truck, car, cat, dog, castle, fish, baby, book, chair, embroidery, sports, child, phone, toy, garden/park	$0.1 \leq F_1 < 0.2$
motorcycle, wagon, bottle, poster, bird, rain, sculpture, table, outdoor, cityscape, daytime, dirt, drink, droplets, elder, glasses, graffiti, hat, indoor, insect, music+instrument, nighttime, overcast, reflection, reptile, rodent, rural, shadow, sitting, smoke, submarine, teenager, traffic, unpaved, violin	$0.0 \leq F_1 < 0.1$

considered to be a relatively simple technique. It does seem that their proposed PAAPL technique is able to learn from the data despite the large amount of noise that it has. However, their use of the textual features is extremely simple, only searching exactly for the words that define the concepts. They have not used synonym information, stemming, WordNet, or any other resources that could be quite useful, and surely the performance could be improved.

Even though the test sets of subtask 1 and 2, differ in difficulty, image quality, number of concepts, etc., if we dare compare the results of image annotation using manually labeled data (see Table 2a) with using automatically gathered data (see Table 3b), the performance is lower for the latter. This is something expected since learning with Web data is considerably more challenging. The real objective is to observe how much can be achieved using the Web data. Ultimately in practice one or the other or a combination of both approaches will be better in

Table 5: Top performing concepts according to F_1 relative improvement with respect to random when training with (5a) manually annotated data (ISI 1424), and (5b) automatically gathered Web data (ISI 1411).

(a)

Concept	F_1	R. Imp. (%)
none	0.872	85.8
noblur	0.827	80.8
dog	0.722	70.7
fireworks	0.667	66.6
flower	0.662	64.9
partialblur	0.648	61.5
fooddrink	0.623	59.9
adult	0.612	57.7
one	0.593	55.5
female	0.589	55.4
outdoor	0.588	55.2
bird	0.561	54.8

(b)

Concept	F_1	R. Imp. (%)
fireworks	0.703	69.7
pencil	0.692	68.0
stars	0.646	62.8
sunrise/sunset	0.565	54.4
galaxy	0.500	49.0
drawing/diagram	0.564	48.5
newspaper	0.462	45.8
space	0.483	44.5
lightning	0.452	44.3
pool	0.426	40.8
fire	0.424	40.6
protest	0.400	39.1

a certain circumstance. Thinking about the problem, it would be understandable that an annotation system will work better for some concepts than for others. In Tables 4a and 4b, there are lists of concepts categorized by the range of their concept-based F_1, for the best system in subtask 1 and 2, respectively. Here, it can be observed that for some concepts, the Web data performs rather well, and in general it does not look too bad with respect to the results using manually labeled data. The same thing can be observed in Tables 5a and 5b, which show the top performing concepts according to the relative improvement[4] with respect to the random baseline.

5 Conclusions

The ImageCLEF 2012 Scalable Image Annotation Using General Web Data proposed two subtasks. The overall objective was to take advantage of automatically gathered image and textual Web data for training, in order to develop more scalable image annotation systems. In the first subtask, the participants could use for developing their annotation systems, both manually labeled data, and automatically gathered Web data. In this subtask, none of the participants were able to use the Web data to obtain a better performance than when using only manually labeled data. The participation was extremely low, there being only three groups, and it seemed that they were not able to invest much time in the problem. Due to this, few conclusions can be drawn from the results. Although it certainly cannot be stated that the Web data is simply not useful, since in

[4] Relative improvement defined as the absolute improvement divided by the difference between the baseline performance and perfect performance which for F_1 is 1.

subtask 2 the results were somewhat positive, suggesting that in subtask 1 also good results could be achieved.

Subtask 2 consisted in using only automatically gathered Web data, and possibly additional external language resources, to develop a more scalable image annotation system. A special characteristic was that the list of concepts was different for development than for test. In this subtask the participation also low, having participated only one group. However, the obtained results were specially interesting. The submissions obtained a considerably better performance than the two provided baselines, and the results generalized well to the test set, despite the change of concept list. Furthermore, the system of the participant was specially targeted at scalability, by using an online learning method adequate for this type of problem, thus it fulfills the initial objective. On the other hand, the processing of the textual data could only be considered to be very basic, thus suggesting that a much better performance could be achieved.

Another interesting aspect of the results of subtask 2 was that when analyzing on a per concept basis, in some cases the performance was comparable to good annotation systems learned using manually labeled data. Therefore, for some concepts the Web data is considerably effective. This also suggests that one possible way to address what was proposed in subtask 1, is to do some type of fusion per concept.

Since the participation was low, and there were positive results, it would be interesting to repeat this benchmark, but making a greater effort to get more groups to participate. However, even though it is believed that in subtask 1, good results could be achieved, it is not so interesting from a scalability point of view. For a future edition it could be changed slightly. For example it could be that for the concepts where there is manually labeled data available, the annotation systems would use a combination of manual and automatically gathered data, otherwise only automatically gathered data is used.

Acknowledgments

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. Work supported by the Spanish MICINN under the MIPRCV Consolider Ingenio 2010 program (CSD2007-00018) and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

1. Chien, B.C., Chen, G.B., Gaou, L.J., Ku, C.W., Huang, R.S., Wang, S.E.: KIDS-NUTN at ImageCLEF 2012 Photo Annotation and Retrieval Task. In: CLEF 2012 working notes. Rome, Italy (2012) 10
2. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA; London (May 1998) 7
3. Gong, Y., Lazebnik, S.: Iterative quantization: A procrustean approach to learning binary codes. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 817 –824 (june 2011) 4

4. 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) 5
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. pp. 2169–2178. CVPR '06, IEEE Computer Society, Washington, DC, USA (2006), http://dx.doi.org/10.1109/CVPR.2006.68 5
6. 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) 2
7. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision 42(3), 145–175 (May 2001), http://dx.doi.org/10.1023/A:1011139631724 5
8. 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) 5
9. Thomee, B., Bakker, E.M., Lew, M.S.: TOP-SURF: a visual words toolkit. In: Proceedings of the 18th International Conference on Multimedia 2010. pp. 1473–1476. ACM, Firenze, Italy (October 25-29 2010) 5
10. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: CLEF 2012 working notes. Rome, Italy (2012) 2, 5, 7
11. 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) 2
12. Ushiku, Y., Harada, T., Kuniyoshi, Y.: Efficient Image Annotation for Automatic Sentence Generation. In: ACM Multimedia. Nara, Japan (2012), accepted 10
13. Ushiku, Y., Muraoka, H., Inaba, S., Fujisawa, T., Yasumoto, K., Gunji, N., Higuchi, T., Hara, Y., Harada, T., Kuniyoshi, Y.: ISI at ImageCLEF 2012: Scalable System for Image Annotation. In: CLEF 2012 working notes. Rome, Italy (2012) 10, 11
14. 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) 2
15. Wang, X.J., Zhang, L., Liu, M., Li, Y., Ma, W.Y.: ARISTA - image search to annotation on billions of web photos. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 0, 2987–2994 (2010) 2
16. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learningtorank withjoint word-image embeddings. Machine Learning 81, 21–35 (2010), http://dx.doi.org/10.1007/s10994-010-5198-3 2
17. Wong, H.C., Bern, M., Goldberg, D.: An image signature for any kind of image. In: Proc. of International Conference on Image Processing 2002. pp. 409–412 (2002) 3