Semantic-Analysis Object Recognition: Automatic Training Set Generation Using Textual Tags

Sami Abduljalil Abdulhak[1], Walter Riviera[1], Nicola Zeni[2],
Matteo Cristani[1], Roberta Ferrario[2], Marco Cristani[1]

[1] Department of Computer Science, Cá Vignal 2, Verona, Italy
{sami.naji,walter.riviera,matteo.cristani,marco.cristani}@univr.it
[2] Laboratory for Applied Ontology, Consiglio Nazionale delle Ricerche (CNR), Via alla Cascata 56/C, Trento, Italy
{nicola.zeni,roberta.ferrario}@loa.istc.cnr.it

Abstract. Training sets of images for object recognition are the pillars on which classifiers base their performances. We have built a framework to support the entire process of image and textual retrieval from search engines such that, giving an input keyword, calculate the statistical analysis, the semantic analysis and automatically build a training set. We have focused our attention on textual information and we have explored, with several experiments, three different approaches to automatically discriminate between positive and negative images: keyword position, tag frequency and semantic analysis. We present the best results for each approach.

Keywords: training set, semantic, ontology, semantic similarity, image retrieval, object recognition

1 Introduction

The process of automatically building a training set of images for object recognition giving a class name, is a recent challenge originated from the Semantic Robot Vision Challenge [1]. The idea is to mine on-line repository of images and use them to support the object recognition of image classifiers [2]. Given this strategy, the goal is to mine search engine and retrieve images that can be used to feed a training set for a specific class.

The problem falls under the field of Image Retrieval(IR) task where given a certain query in a form of keyword or image, the system must present images relevant to the query. There are two main strategies to tackle this problem: the content-based image retrieval (CBIR) [3] and the tag/keyword-based image retrieval (TBIR) [4].

CBIR leverages the concept of visual similarity between the query image and the retrieved ones using low-level visual features (e.g., color, shape, etc) to perform matching, while TBIR tries to overcome the limitations of CBIR by using textual information conveyed with images through applying document retrieval techniques to boost the retrieval performances. Nevertheless TBIR performances are influenced by the availability and quality of the textual information users supplied with images, in fact during manual annotation process of images they often misuse tags or give incomplete textual description of the image content [5–7].

The use of information conveyed with images in process of image retrieval or image classification is not a novelty, there are several works that explore how textual information can be used, among them [8–11]. Recent approaches try to solve these issues by performing a tag completion process either mining extra textual information from Internet either using a content image analysis to fill the gap [6, 12].

In the present work we propose a framework that helps to automate the entire process. The main idea is to use textual information that comes along with images on the web to fully automate the training set construction. To do this, we assume that user annotation process is not always reliable since the users are not experts. Even though users upload images in a social context where other users can use collaborative tagging to annotate images, tags are not validated and so the subjectivity elements are not removed. Moreover since users are non expert they tend to use ambiguous and inappropriate tags to describe images content. The main idea is to explore how statistical and semantic analysis of textual information can help to fully automate the training set construction. In particular, we employ statistical and semantic analysis to filter the textual information, pruning noisy tags and retain only those that are highly correlated with the content of an image in such a way discriminating positive images from negative ones [1]. We use statistical measures such as frequency and tags distribution, as well as WordNet and semantic distances to find tags correlation and explore their contribution in the discriminate process. Our starting assumptions are that injecting incrementally semantic analysis techniques into textual annotation, performances rises and to validate our assumptions, a set of experiments are presented.

The rest of the paper is structured as follows: Section 2 describes the challenges of image retrieval task and provides an overview of work in the area. The method we propose is introduced in Section 3. Sections 4 discuss the experimental setup and evaluation method, while the evaluation results are presented in Section 5. Finally, conclusions and directions for future work are presented in Section 6.

2 Related work

Annotation is a widely used technique of characterizing objects portrayed in images by adding textual tags. The textual tags associated with images have been

[1] We consider positive images such images in which the prominence of the object presented in the image indicates the image fully represents the object and classified as positive. On the contrary we consider as negative, images where targeted object is absence.

shown to be useful, improving the access to photo repositories both using temporal [13] and geographical information [14]. One of the popular online tag-based photo sharing repositories is Flickr, allowing users to have the ability to freely assign one or more chosen keywords for an image for personal organization or retrieval purposes. In other words, it permits users to perform tagging that is the act of adding words to images, describing the semantics of the visual contents. This attribute increases the motivations for adding more keywords, creating relatively large amount of rich descriptions of objects presented in images. However, the textual tags associated with images are often noisy and unreliable, posing a number of difficulties when dealing with IR.

A number of approaches have been proposed to measure the reliability of the textual tags accompanying images [15–17]. In [17], the authors present a Flickr distance to measure the correlation between different concepts obtained from Flickr. Given a pair concepts (e.g., car-dog), the algorithm tries to calculate the semantic distance between them using square root of Jensen-Shennon divergence. The authors rely on the scores by considering the higher score distance as an indication of high relatedness of a pair concepts. Related researches have been also focused on investigating what objects do people observe most in an image, what do they annotate or tag first, and what implications influence them to choose words to describe objects depicted in images.

Spain and Perona [15] study the idea of "*importance*" of objects in an image and conclude that important objects are most likely to be tagged first by human when asked to describe the contents of an image. The authors develop a statistical model validating the notion of prominence of objects in images, demonstrating that one can foresee a set of prominent keywords based on the visual cues through regression. A closely related work to ours is presented by Hwang and Grauman [18]. They introduce an unsupervised learning method for IR that uncovers the implicit information about the object importance in an image, exploiting a list of keyword tags provided by humans. The proposed method is able to disclose the relationship between how humans tend to tag images (e.g., words order in the tag list) and the relative importance of objects in an image.

Traditional techniques are relying on features extracted from visual contents when learned visual category models directly from image repositories that require no manual supervision [8–11]. The intuition behind the approach proposed in [9] is to learn object categories from just a few training images in an incremental manner, using a generative probabilistic model. Similarly, Li-Jia Li and Fei-Fei Li [10] propose an incremental learning framework, capable of automatically collecting large image dataset. The authors build a database from a sample of seed images and use the database to filter out newly crawling images by eliminating irrelevant examples.

Fergus et al. [11] introduce a method able to learn object category by its name, exploiting the raw images automatically downloaded from Google image search engine. The introduced approach is able to incorporate spatial information in translation and scale invariant style, possessing the ability to tackle the high intra-category variability and isolate irrelevant images produced by the search engine.

Vijayanarasimhan and Grauman [19] propose an unsupervised approach to learn visual categories by their names using a collection of images pooled from keyword-based search engines. The main goal underneath the proposed approach is to harvest multiple images, by translating the query names into several languages and crawling the search engines for images using those translated queries. The false positive categories are collected from random sample images found in categories that have different names from the category of interest.

We are working on a different challenge of the one stated previously: given the textual tags provided by humans and associated with images, we want to automatically build a good training set by discriminating images as either relevant to the queried object or otherwise.

3 Method

In this paper our goal is to take advantage of the textual tags available with images to automatically select the most representative of an object category for training a classifier, without looking at the nature of the objects therein. To do so, we exploit both semantic analysis and pure statistical approaches. These considerations lead us to focus on three main features:

- **keyword position**, to capture an image as related or unrelated on the basis of a keyword(i.e., object class name) position in a tag list;
- **semantic analysis**, to measure the semantic relatedness by means of semantic distance measures;
- **tag frequency**, to count the frequency usage of each single tag of all tag list of the object class.

Figure 1 presents a schematic representation of our framework. Detailed description of the procedure is provided in the subsequent subsections.

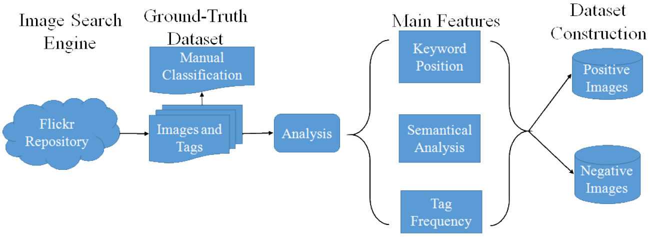

Fig. 1. A schematic representation of our framework.

3.1 Keyword Position

The textual tags given in a tag list and associated with an image describing its contents, could reasonably help us to derive important and valuable information about the nature of the depicted objects. However, the order in which the textual tags are placed in a tag list is most likely to reference to the objects position and size in the visual content [20]. Therefore, it is reasonable to claim that the first textual tags in the list are mostly representing the central entities of an image. Taking this keypoint into account, we use this feature to develop 5 different strategies which follows the same algorithmic structure:

Algorithm 1: Keyword Position

```
Data: a Keyword (i.e.,the object class name) and
      $T = \{t_i| \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
Result: A partition of the Images ∈ Keyword in:
Image-$P = \{i_p|i \in$ Images which are usable to build a training dataset$\}$
Image-$N = \{i_n|i \in$ Images which are outliers$\}$
Initialization;
foreach $i \in Images$ do
    $tags \leftarrow$ load $t_i$;
    clean$(tags) tags_n \leftarrow$ extract the first $n$ tag from $tags$;
    if "$keyword$" $\in tags_n$ then
        Image-$P \leftarrow i$;
    else
        Image-$N \leftarrow i$;
```

Algorithm 1 is designed to demonstrate the systematic workflow of the keyword position feature. Given a tag list that comprises of a number of textual tags and corresponds to a particular image, the algorithm tries to search for the keyword through the list in the first n positions. The algorithm then labels the image as positive (reliable) if it is related to the class name or negative (outlier) It is noteworthy that the clean operation provided in the algorithm is used to remove words with less than 3 characters, empty strings and non-alphabetic texts. It also splits long sentences in single words, when they are separated by the “_” symbol.

3.2 Semantic Analysis

To typically define the semantic relatedness or its inverse of the object class to its belonging textual tags, semantic distance must be measured. Therefore we propose to apply two different semantic distance measures: WordNet and Jiang and Conrath [21]. First we adopt the WordNet distance [22]. WordNet is a large-scale lexical database that organizes English terms and their syntactic roles into synsets. Synsets are interlinked by means of conceptual-semantic and variety of lexical relations. We choose WordNet due to the fact that it is the first to be richly developed and has been used widely as an ontology. Since WordNet provides a lexical relationship between concepts, it is beneficial to semanticly measure relatedness of the object class to its belonging tags by their lexical relationship such as meronymy (bus-wheels).

Second we apply the distance measure proposed by Jiang and Conrath in [21]. They formulate their approach in the form of conditional probability of coinciding an item of a child synset given an item of a parent sysnset. The formula of measuring the semantic distance is given as follows:

$$Dist(w_1, w_2) = IC(c_1) + IC(c_2) - 2 \times IC(LSuper(c_1, c_2))$$

We use this feature to develop 12 different strategies which follows the same algorithmic structure:

Algorithm 2: Semantic Analysis

```
Data: a Keyword (i.e., the object class name) and
      $T = \{t_i| \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
Result: A partition of the Images ∈ Keyword in:
Image-$P$ = $\{i_p | i \in$ Images which are usable to build a training dataset $\}$
Image-$N$ = $\{i_n | i \in$ Images which are outliers$\}$
Initialization;
foreach $i \in Images$ do
    $tags \leftarrow$ load $t_i$;
    clean($tags$);
    $score_i \leftarrow$ sum or mean of the distance values of the $tags$;
    if $if\ score_i \geq a\ Threshold\ \tau$ then
        Image-$P \leftarrow i$;
    else
        Image-$N \leftarrow i$;
```

Algorithm 2 is developed to clearly illustrate how we apply the semantic analysis feature to measure the semantic relatedness or its inverse of the object class to its textual tags. As already mentioned above, we adopt two difference distance measures: WordNet and Jiang and Conrath. The algorithm takes the object class (represented by a keyword) and each image's tag list, then computes the distance of the keyword to every single textual tag in the tag list, yielding a score for each. If the algorithm finds no semantic distance between the keyword and a textual tag, it discards the tag. The algorithm therefore labels an image as positive (reliable) if its score is equal or above a threshold τ; otherwise it labels it as negative (outlier). The threshold value τ changes with respect to the experiment (see Section 4).

3.3 Tag Frequency

To understand which are the most frequently used tags (words) which describe images related to a certain object class, we calculate the frequency values of all the single tag$_{(i,j)}$ as their occurrences probability. The idea is to perform a selection based on the utility of the words used to describe the object depicted in an image. The frequency value of a single tag$_{(i,j)}$ is calculated as follows:

$$Feq(tag_{(i,j)}) = \frac{O - tag_{(i,j)}}{\sum_{i=1}^{N_{images}} length(tag_i)},$$

where tag$_{(i,j)}$ is the j^{th} tag of the tag list associated to image i, and $O - tag_{(i,j)}$ is the total number of a tag$_{(i,j)}$ occurrences. In particular, if a given frequency value of a single tag$_{(i,j)}$ is relatively high, it means that many images of the considered object class requires it into their descriptions. In other words, it is natural to think that if we are looking at an image of a "car", we highly expect to observe higher frequency values for tags like "*wheel*" or "*driver*" than "*pizza*" or "*pencil*".

We use this feature to develop 12 different strategies which follows the same algorithmic structure:

```
Algorithm 3: Tag Frequency
  Data: a Keyword (i.e., the object class name) and
        $T = \{t_i| \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
  Result: A partition of the Images ∈ Keyword in:
  Image-$P$ = $\{i_p | i \in$ Images which are usable to build a training dataset $\}$
  Image-$N$ = $\{i_n | i \in$ Images which are outliers$\}$
  Initialization;
  foreach $i \in Images$ do
      $tags \leftarrow$ load $t_i$;
      clean($tags$);
      $score_i \leftarrow$ sum or mean of the frequency values of the $tags$;
      if $if\ score_i \geq a\ Threshold\ \tau$ then
          Image-$P \leftarrow i$;
      else
          Image-$N \leftarrow i$;
```

Algorithm 3 uses the frequency values to determine if a given image is related to the object class. To do this, it combines the frequency values of each tag$_{(i,j)}$ to produce a score. Then, it labels an image i as positive (reliable) if its score is equal or above a threshold τ; otherwise it labels it as negative (outlier). The threshold value τ changes with respect to the experiment (see Section 4).

4 Experiments

We devote this section to demonstrate the systematic workflow of our framework. Firstly, we pool images for a set of 21 object classes taken from the standard Caltech101 [2], using Flickr online photo sharing[3]. Each class contains 400 images

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101

[3] https://www.flickr.com/

as well as their corresponding tag lists (tag_i). For simplicity, the number of crawled images has been defined in order to minimize the computational time of downloading images and managing their tags during the experiments. The effective number of classes have been normalized to 16, avoiding the classes that are composed by a bi-gram (i.e., two words). The remaining classes are: *accordian, bonsai, euphonium, face, laptop, menorah, nautilus, pagoda, panda, piano, pyramid, revolver, starfish, sunflower, umbrella, watch.* Since there are 400 images and 400 tag lists per class, the dataset composes of 6400 images and 6400 tag lists.

To generate the ground-truth for our experiment in a more effective and efficient way, we build a graphics user interface (GUI) that allows us to manually label an image as positive or as negative to the object class. For reliable manual classification, certain guidelines are defined and adopted. If the following guidelines are satisfied, then an image is labeled as negative; otherwise as positive:

- an image is completely unrelated with the object specified by the category it belongs
- an image contains irrelevant parts of the object, that is, parts that alone are not sufficient to make the category object identifiable
- an image contains only internal parts of the category object (like a cockpit of an aeroplane or an engine of a car)
- an image is drawings and caricatures of the category object

For each single features we run several different experiments based on different strategies. Each strategy differs from another one, with regard to the method used to calculate the threshold. This produces different results in determining if a given tag list is associated to a positive or negative image.

Referring to the algorithms described in 3.1, 3.2, 3.3, we give a brief explanation of the strategies associated to the threshold which produces the best discrimination results:

Feature 1: Based on experiments performances, we obtain the best result when searching if keyword is found in the first three positions in the tag list. Surprisingly, this feature does not involve any cleaning mechanism of textual tags in tag list (it avoids the step number 4 of algorithm 1). However, the feature takes the textual tags as they are provided by Flickr. At this point one may ask why using contaminated textual tags in tag list is, unexpectedly, producing better results than the cleaned version? The answer lays in the “filtering” mechanism of the textual tags.

Cleaning the tag list tag_i implies producing more single words (tag_j)since the tag sentences are split. This increases the probabilities of finding the right match with the keyword, therefore a higher number of tag list labeled as positive. This has been confirmed by the number of false positives generated using the others strategies, which is widely higher than the number of false positives produced by the strategy just described. To better provide a deep understanding of what happen if we do not perform any tag cleaning on the tag list, we provide the following example: Given the tag list relative to a negative image of the *panda* class: “*zoo_atlanta*”, “*taishan*”, “*giant_panda*”, the keyword would not be matched since the substring matching is not performed. Therefore, the image is labeled as negative. This results change if we clean the tag list by splitting the sentences in single words. The cleaned tag list is become: *zoo, atlanta, taishan, giant, panda.* In this case, the keyword would match with the 5^{th} tag and therefore the image is now labeled as positive.

Feature 2: As already explained in Section 3, for this feature we use two different measurements: the standard semantic distance provided by WordNet, and the distance proposed by Jiang and Conrath in [21]. To select the one which produce the best results, we use both the metrics to run the 12 strategies. The comparison results are showed in figure 2 .

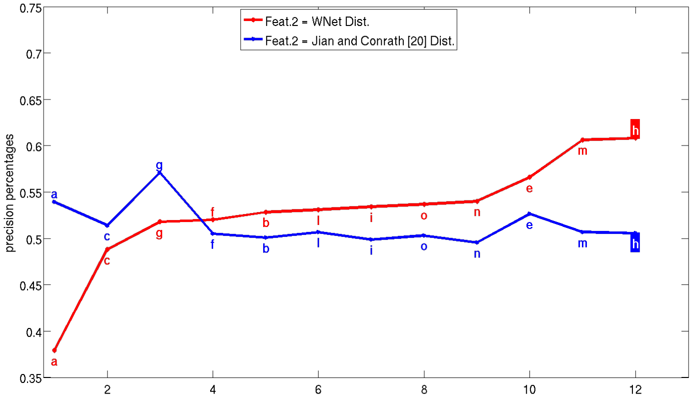

Fig. 2. Summary results obtained by using WordNet and Jiang and Conrath distances in [21] in all the strategies. WordNet distance is outperforming in average in all of the strategies. We calculate the precision rate for each strategy $(a, b, \ldots, o)$ as: $\#TruePositive/(\#TruePositive + \#FalsePositive)$.

Using WordNet distance as shown in figure 2, we observe constant increase in the average performances of all strategies. Therefore, in the following description we are mainly referring to the WordNet distance. The strategy based on the WordNet distance, which gives the best results use the following criteria to split the images set: Defining the $scores_i$ as the mean of the distances between the considered tags and the keyword:

$$scores_i = mean(Distance(tag_{(i,j)} - keyword))$$

$$Feat2(scores_i) = \begin{cases} positive & \text{if } mean(scores_i) \geq \tau \\ negative & \text{otherwise} \end{cases}$$

The best result is obtained using this strategy when the threshold is set to $\tau = median(scores_I)$, where the $scores_I$ is the vector of all the $scores_i$.

Feature 3: The strategy based on the tag frequency feature, which produces the best results, comparing with the other strategies, use the following criteria to split the images set: Defining the $scores_i$ as the sum of the frequencies values of the considered tags with respect to the keyword:

$$scores_i = \sum_{i=1}^{N_{images}} Fq(tag_{(i,j)})$$

$$Feat3(scores_i) = \begin{cases} positive & \text{if } mean(scores_i) \geq \tau \\ negative & \text{Otherwise} \end{cases}$$

We reach the best results when the threshold is set to $\tau = mean(scores_I)$, where the $scores_I$ is the vector of all the $scores_i$.

5 Performances Evaluations

To assess the reliability the experimental performances of the features described beforehand, we select n images labeled as positives from all the strategies and from Flickr. Hence, we count the true positives and the false positives that have been generated by the strategies and by Flickr (in this case, the false positives are the ones we manually label as negatives). Since the main goal of this framework is to generate a reliable dataset of images, for this reason, all our strategies tends to produce more negative labels than the positive. This behavior allows to minimize the number of the false positive labels generated during the experiments. Since not all the strategies produces the same number of positive labels, to avoid the problem of getting some Null values, we fix $n = min(P - labelsofeachfeature)$. The selection of the n labels, has been done randomly for Flickr while for our strategies the firsts n are considered. To ensure the consistency of Flickr performances, we average the results produced after 10 random selections.

Table 5 displays the percentage values of the performances obtained using Flickr and our best strategies. The column $\#P - labels$ contains the different n values used for each class. The column $GT - Positives$ presents the number of true positive within the ground-truth.

To make the performances reported in the table more comparable, we recalculate the precision percentages by fixing $n = 50$[4] positives labels per class. Also in this case, the selection of the 50 labels, has been done randomly for Flickr while for our strategies refers to the firsts n. In figure 3, we provide the averages values of each strategy for all the class with $n = 50$.

In this last case, an exception is done for the "*euphonium*" category since it is composed by just 9 positive images also in the ground-truth.

[4] this parameter has been set by considering the lowest common number of labels

Table 1. Precision results obtained using all the features for 16 classes. Flickr donates the number of correct positive labels from the n images downloaded from Flickr repository. Feat is an abbreviation for feature where Feat.1 donates the keyword position, Feat.2 donates semantic analysis, and Feat.3 donates tag frequency.

Classes	# P- labels	GT-Positives	Flickr	Feat.1	Feat.2	Feat.3
watch	218	386 / 400	94.95	95.87	96.79	96.79
sunflower	178	379 / 400	93.26	97.19	96.63	96.63
bonsai	119	362 / 400	90.76	90.76	92.44	88.24
panda	182	359 / 400	89.56	90.11	32.31	97.25
laptop	171	359 / 400	88.30	92.98	93.57	87.72
pyramid	203	250 / 400	65.02	60.10	64.04	64.04
starfish	170	211 / 400	49.41	60.00	56.47	53.53
piano	50	105 / 400	37.50	58.33	37.50	70.83
umbrella	175	164 / 400	37.14	41.14	41.71	44.00
menorah	148	146 / 400	34.46	33.78	29.73	35.81
accordion	158	118 / 400	31.01	29.75	31.65	28.48
pagoda	167	114 / 400	29.94	32.34	34.13	38.32
face	135	120 / 400	28.15	31.11	25.19	27.41
revolver	127	110 / 400	26.77	38.58	42.52	31.50
nautilus	163	67 / 400	17.79	22.09	25.15	17.79
euphonium	8	9 / 400	0	62.5	0	0

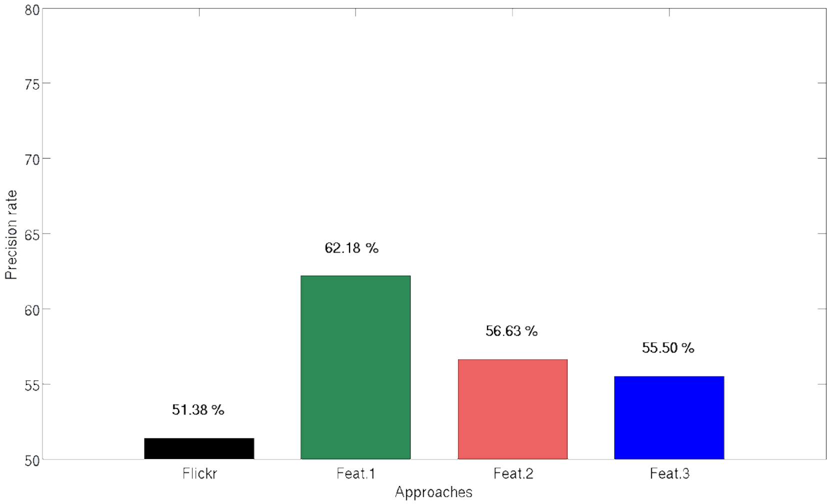

Fig. 3. Summary of results of the all the features by fixing $n = 50$. The highest precision is given using feat.1 (i.e., keyword position).

At this point, one may be skeptical about the reliability of our strategies since we are estimating their performances by considering only 50 images against the 400 downloaded. Therefore, if we observe how the performances change when we consider all the available positives labels showed in 5, we are more confident of our results. Indeed, if we calculate the average of the positives labels considered in the last case, we can observe (see table 5) that the performances remain constant when setting $n \neq 50$. The overall performance of our strategies still outperforms Flickr. In particular, using keyword position, the average performance obtained is encouragingly well (about 11% higher than Flickr). This information is further enriched since it provides us reliable percentage value than the ones provided by the results of $n = 50$.

Table 2. The average performance of all the features when $n = 50$ and $n \neq 50$

# P- labels	Flickr	Feat.1	Feat.2	Feat.3
$\neq 50$	50.87	61.18	56.62	55.50
$= 50$	50.87	62.18	56.62	55.50

6 Conclusions

We have presented a framework to support the entire process of image and textual retrieval from search engines such that, giving an input keyword, calculate the statistical analysis, the semantic analysis and automatically build a training set. We have conducted several experiments to validate our assumptions about the analysis of textual information and the evaluation that we did on three methods investigated have shown that the position of tags, their order, is pretty relevant. We have investigated the semantical aspects by using semantic distance. Unfortunately, the results achieved does not benefit at all from the adopted semantic features. However, the methods suggested are currently under continuous experimentation and need to bee further investigated. In particular we take into account for future work to explore the use of different search engines such as Google[5], ImageNet[6], InstaGram[7] or Pinterest [8] to check if they are interchangeable or can be combined to improve performances. We plan also to extend and investigate other semantic features related to ontological relationship of textual information and combine them with the aim of creating a waterfall model which combines different strategies.

Acknowledgment This research was supported by the VISCOSO project financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

[5] http://www.google.com
[6] http://www.image-net.org
[7] http://instagram.com/
[8] http://www.pinterest.com

References

1. Helmer, S., Meger, D., Viswanathan, P., McCann, S., Dockrey, M., Fazli, P., Southey, T., Muja, M., Joya, M., Jim, L., Lowe, D.G., Mackworth, A.K.: Semantic robot vision challenge: Current state and future directions. In: IJCAI workshop. (2009)
2. Setti, F., Cheng, D.S., Zeni, N., Ferrario, R., Cristani, M.: Semantically-driven automatic creation of training sets for object recognition. Computer Vision and Image Understanding **Submitted** (2014) x–x
3. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1) (February 2006) 1–19
4. Liu, Y., Xu, D., Tsang, I.W., Luo, J.: Textual query of personal photos facilitated by large-scale web data. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5) (2011) 1022–1036
5. Heymann, P., Paepcke, A., Garcia-Molina, H.: Tagging human knowledge. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. WSDM '10, New York, NY, USA, ACM (2010) 51–60
6. Lin, Z., Ding, G., Hu, M., Wang, J., Ye, X.: Image tag completion via image-specific and tag-specific linear sparse reconstructions. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. (June 2013) 1618–1625
7. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3) (March 2013) 716–727
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2. (2005) 524–531 vol. 2
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding **106**(1) (2007) 59 – 70
10. Li, L.J., Li, F.F.: Optimol: Automatic online picture collection via incremental model learning. International Journal of Computer Vision **88**(2) (2010) 147–168
11. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 2. (2005) 1816–1823
12. Gilbert, A., Bowden, R.: A picture is worth a thousand tags: Automatic web based image tag expansion. In Lee, K., Matsushita, Y., Rehg, J., Hu, Z., eds.: Computer Vision ? ACCV 2012. Volume 7725 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 447–460
13. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: Proceedings of the 15th International Conference on World Wide Web. WWW '06, New York, NY, USA, ACM (2006) 193–202
14. Ahern, S., Naaman, M., Nair, R., Yang, J.: World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In: In Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Press (2007) 1–10
15. Spain, M., Perona, P.: Some objects are more equal than others: Measuring and predicting importance. In: Proceedings of the 10th European Conference on Computer Vision: Part I. ECCV '08, Berlin, Heidelberg, Springer-Verlag (2008) 523–536
16. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press (2007) 971–980

17. Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr distance: A relationship measure for visual concepts. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(5) (May 2012) 863–875
18. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int. J. Comput. Vision **100**(2) (November 2012) 134–153
19. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (June 2008) 1–8
20. Spain, M., Perona, P.: Some objects are more equal than others: Measuring and predicting importance. In: Proceedings of the 10th European Conference on Computer Vision: Part I. ECCV '08, Berlin, Heidelberg, Springer-Verlag (2008) 523–536
21. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR **cmp-lg/9709008** (1997)
22. Fellbaum, C.: WordNet: An Electronic Lexical Database. Language, Speech and Communication. Mit Press (1998)