Multi-Modal Region Selection Approach for Training Object Detectors

Elisavet Chatzilari[1,2], Spiros Nikolopoulos[1,3], Yiannis Kompatsiaris[1],
Josef Kittler[2]
[1] Centre for Research & Technology Hellas - Informatics and Telematics Institute
[2] Centre for Vision, Speech and Signal Processing University of Surrey Guildford, UK
[3] School of Electronic Engineering and Computer Science, QueenMary University of London
ehatzi@iti.gr, nikolopo@iti.gr, ikom@iti.gr, j.kittler@surrey.ac.uk

ABSTRACT

Our purpose in this work is to boost the performance of object classifiers learned using the self-training paradigm. We exploit the multi-modal nature of tagged images found in social networks, to optimize the process of region selection when retraining the initial model. More specifically, the proposed approach uses a small number of manually labelled regions to train the initial object detection classifiers. Then, a large number of loosely tagged images, pre-segmented by an automatic segmentation algorithm, is used to enhance the initial training set with additional image regions. However, in contrast to the typical case of self-training where the image regions are selected based solely on how well they fit to the original classification model, our approach aims at optimizing this selection by making combined use of both visual and textual information. The experimental results show that the object detection classifiers generated using the proposed approach outperform the classifiers generated using the typical self-training paradigm.

Categories and Subject Descriptors

I.4.6 [**Image Processing and Computer Vision**]: Segmentation—*Pixel classification*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Selection process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Experimentation

Keywords

multi-modal, region selection, large scale dataset, self-training, object detection, MIRFLICKR

ICMR '12, June 5-8, Hong Kong, China

1. INTRODUCTION

Machine learning algorithms have been widely used for automatic object detection, a challenging task for the research community in the last decade. The efficient estimation of model parameters mainly depends on two factors, the quality and the quantity of the training samples. The high quality of the training samples is usually accomplished through manual annotation, which is a laborious and time consuming task and typically results into a limited number of training samples. This has a direct impact on the second factor affecting the performance of the generated models and formulates the typical trade-off in machine learning. In an effort to balance this trade-off, the self-training paradigm [1] [24] has been adopted to combine the benefits of both labelled data in terms of effectiveness, and unlabelled data in terms of scalability. Its basic idea is to gradually improve the effectiveness of the classification model by iteratively enhancing the utilized training set with samples that are deemed most confident by the model of the previous iteration. In the case of region based classification, a typical self-training framework consists of an initial classification model trained by labelled regions and a set of unlabelled image regions some of which will be selected to enhance the initial set of training samples. These regions are represented by feature vectors, which are expected to express their visual information. However the effectiveness of self-training heavily relies on the quality of the initial model, which is based solely on the visual information of the initial training samples and is prone to all different errors that are inherent to visual analysis.

On the other hand, the excessive use of Web 2.0 has made available large amounts of user tagged images. This type of images can be obtained at almost no cost, while at the same time offering more information than the mere image visual content. This fact has motivated a significant amount of research effort aiming to combine the advantages of manually labeled data (i.e. high quality) with the cost effectiveness of crowdsourcing [12] and social networks (i.e. vast amounts of digital images, along with an indication of their depicted content, are provided at no cost). Driven by the same motivation, our objective is to exploit the tagged images (from now on called loosely tagged images) offered by social sites like flickr, towards optimizing the process of region selection in model retraining.

In this work, we propose a multi-modal region selection strategy that leverages loosely tagged images to obtain the

additional training samples for enhancing the initial classifiers. Considering the imperfection of the available visual analysis techniques, the key is to effectively combine the two types of knowledge carried by the loosely tagged images (i.e. visual features and tags) in order to yield optimal performance in the region selection process. In order to achieve this and eventually boost the performance of the generated models, an extra layer of confidence is added in the region selection process, which is provided by the textual information, i.e. tags. More specifically, for every concept, a set of regions is selected to enhance the initial training set based on: a) the visual similarity of the region with the examined concept as expressed by the initial object detection model, b) the confidence that the examined concept is present in the image the region belongs to, as determined by its textual information and c) the pixel-size of the region being relatively large with respect to the average size of all regions identified in the host image.

The rest of the manuscript is organized as follows. The related literature is reviewed in Section 2. The proposed approach and its components are described in Section 3 and the experimental setup along with the evaluation results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

In the area of object detection, datasets of manual annotated image regions have been widely used [4],[16]. The authors of [4] present a new benchmark for evaluating pixel-based or region-based methods, which consists of 20000 images manually annotated at pixel level. They also apply and evaluate a variety of known machine learning algorithms (e.g. Support Vector Machines, naive bayes classifier, random forests, etc). In [16] semantic segmentation is achieved by learning the conditional distribution of the class labels given an image, using a Conditional Random Field (CRF) model. However, given that manual annotation of image regions is a time consuming task, approaches that operate on annotations at global image level were proposed. In this case, the image level keywords are associated with the image regions by either relying on aspect models like probabilistic Latent Semantic Analysis (pLSA) [17] or by incorporating multiple instance learning [2]. The authors of [21] propose a method that combines aspect models (pLSA) with spatial models (Markov Random Fields) with the aim of labelling image regions. Finally, considering that the performance of pattern recognition systems is highly influenced by the number of the training samples [14] and that manual annotation, even at a global level, is very expensive, the semi supervised approaches have recently become the subject of intense research efforts [24].

Lately there has been also considerable interest on loosely labeled data and their potential to serve as the training samples for various computer vision tasks. Considering that the size of such datasets can grow almost unlimitedly, active learning has been introduced as a special case of the semi supervised algorithms [19]. In the case of active learning, the purpose is to reduce the number of training examples to be labelled, by selectively sampling a subset of informative data-concept pairs for a human to label. A example for this case is [6], where the authors propose an SVM-based algorithm that combines cross-domain, semi-supervised, multi-concept and active learning for video concept detection. Similarly, the authors of [23] propose an active learning algorithm that uses flickr data to effortlessly acquire training samples. More specifically, a model is initially trained by manually labeled data using Support Vector Machines. Then, the samples that are closest to the hyperplane are selected as the most informative ones and are manually annotated to improve the discrimination ability of the enhanced classifier.

From the perspective of optimal region selection from loosely tagged images, our work can be considered closest to [8] and [15]. In [8], loosely-tagged images are sampled to enrich the negative training set of an object classifier. The authors claim that the tags of such images can reliably determine if an image does not include a concept, thus making social sites a reliable pool of negative examples. The selected negative samples are further sampled by a two stage sampling strategy. First, a subset is randomly selected and then, the initial classifier is applied on the remaining negative samples. The examples that are classified closest to the margin are considered as most informative and are finally selected to boost the classifier. The authors of [15] propose a multiple instance learning algorithm that operates on one million flickr images. They incorporate the various ambiguities between classes by constructing an object correlation network that models the inter-object visual similarities and the co-occurrences of the classes.

The proposed approach is essentially a method for object detection that operates on loosely tagged images and uses the associated textual information to optimize the selection of training samples in a modified version of self-training. In contrast to active learning, where the goal is to select the most informative samples so as to minimize the required human effort for annotation, the goal of the proposed approach is to be completely discharged from the laborious task of manual annotation. In order to do this, the human annotator is replaced by an automatic region selection strategy that exploits the textual information carried by the images in social networks. Moreover, we opt to enhance the training set with positive samples, instead of negative as in [8], allowing for a higher performance boost of the final classifiers. For the same reason, a semi-supervised learning algorithm was chosen instead of the multiple instance learning algorithm that is utilized in [15].

3. PROPOSED APPROACH

The proposed approach for extracting training samples from loosely tagged images is depicted in Fig. 1. Given a concept c_k, we start from a set of samples that are labelled with this concept and use Support Vector Machines (SVMs) to train the initial classifier recognizing c_k. Then, we try to identify more samples representing this concept in a pool of unlabelled entities, which in our case are regions of loosely tagged images harvested from Web 2.0 applications. In these images, there is no knowledge of the real objects depicted, or of the exact location of the objects within the image. To overcome this obstacle, the following process takes place. First, the loosely tagged images are automatically segmented into regions. In order to overcome the flawed nature of segmentation, an automatic preprocessing step is proposed to dismiss the over-segmented regions and finally obtain a set of regions that roughly correspond to semantic objects. Afterwards, visual features are extracted to represent each region. Applying the initial classifier to

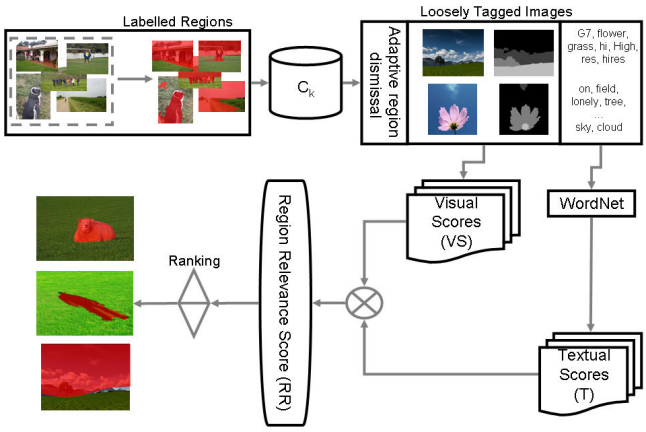

Figure 1: System Overview

the unlabelled regions provides the visual scores, which indicate the confidence that the examined region depicts c_k based on its visual information. Next, the textual scores are extracted by the textual information that accompanies the loosely tagged images. Essentially, the textual information adds an extra layer of confidence in the selection process by enhancing the probability that the examined concept c_k is expressed visually in the image when it is tagged with c_k. This is derived by the reasonable assumption that an image tagged with c_k is more likely to depict the concept c_k than an image that does not include a tag related to c_k.

3.1 Adaptive Region Dismissal and Representation

Segmentation is applied to all images used by this framework aiming to extract spatial masks of visually meaningful regions. In our work we used the K-means with connectivity constraint (KMCC) segmentation algorithm as described in [11]. The output of this algorithm applied to a single image is a set of segments some of which roughly correspond to meaningful objects. However, considering that most segmentation algorithms are prone to over-segmentation, there is a need to automatically dismiss poorly segmented regions before they are even considered as possible candidates, especially since the number of loosely tagged images can grow to an arbitrary size. The proposed approach works under the assumption that a small, over-segmented region is more likely to confuse the classifier instead of enhancing its performance, even if it depicts the object partially. Consider an image I and its size $|I|$ ($|I| = width \times height$ in pixels) which, after given to the segmentation algorithm, is found to have N_R regions ($R^0 = \{R_1, R_2, \ldots, R_{N_R}\}$). Then the following iterative process is applied on the set of regions:

$$R^i = \{R_j^{i-1} \quad | \quad j = \arg_j \left(|R_j^{i-1}| < \alpha * \overline{|R^{i-1}|}\right)\} \qquad (1)$$

where, $\overline{|R^{i-1}|}$ is the average size of the remaining regions in iteration $i-1$ and $|R_j^{i-1}|$ is the size of the j^{th} region. This practically means that in each iteration the regions with size smaller than a percentage α of the average size of all regions that have survived the previous iteration, are dismissed. The process stops when the following criterion is met:

$$\frac{R^i}{R^{i-1}} > \lambda \qquad (2)$$

The parameters α and λ influence the threshold under which regions are considered unsuitable for enhancing the initial classifiers.

In order to represent visually the remaining regions, we have employed a bag of visual words approach similar to the one described in [20] with the important difference that, in our case, descriptors are extracted to represent each of the identified image regions rather than the whole image. More specifically, for detecting interest points we have applied the Harris-Laplace point detector on the intensity channel, which has shown good performance for object recognition [22]. In addition, we have also applied a dense-sampling approach where interest points are taken every 6^{th} pixel in the image. For each interest point (identified both using the Harris-Laplace and dense sampling approach) the 128-dimensional SIFT descriptor is computed using the version described by Lowe [9]. Then, a Visual Word Vocabulary (Codebook) is created by using the K-Means algorithm to cluster in 500 clusters approximately one million SIFT descriptors that were sub-sampled from the total amount of SIFT descriptors extracted from all training images. The Codebook allows the SIFT descriptors of all interest points enclosed by an image region, to be vector quantized against the set of Visual Words and create a histogram [18]. Thus, for every region, a 500-dimensional feature vector is extracted, that contains information about the presence or absence of the visual words included in the Codebook. Then, all feature vectors are normalized so as the sum of all elements in each feature vector is equal to one.

3.2 Visual and Textual Scores Estimation

For every concept c_k, an object detection model (SVM_{c_k}) is trained using as positive examples the regions that are labelled with c_k while the rest are used as negative examples (One Versus All / OVA approach). For each region extracted from the loosely tagged images, a score is extracted by applying the SVM_{c_k} classifier. This score is based on the distance of the feature vector that represents this region from the margin of the SVM_{c_k} model [7]. The higher the outcome of the model for a specific region the higher the possibility that this region depicts the concept c_k. We will refer to this score as visual score, $VS_{c_k}(r_m^I)$, of region r_m^I of image I with respect to the concept c_k.

In addition, loosely tagged images contain textual information which can guide the training sample selection process. Although these tags describe the images globally and do not provide any information for the location of the objects within an image, they can still be used as an additional criterion besides the visual score of the region. For example, if a region with high visual score for the concept *grass* belongs to an image which is not tagged with the literal *grass*, the region can be disregarded. However, in order to exploit this textual information, we need to overcome the well known problems of social tagging systems (i.e., lack of structure, ambiguity, redundancy, emotional tagging, etc). To this end we use three approaches in order to measure the semantic relatedness between the image tags and the concepts' lexical description. Firstly, an adapted version of the Google Similarity Distance [3] was used. The original

Google Similarity Distance between words x and y is given by the following expression:

$$GD = \frac{\max\{\log f(x), \log f(y)\} - \log f(x;y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (3)$$

where $f(x)$ denotes the number of pages containing x and $f(x;y)$ denotes the number of pages containing both x and y, as reported by Google. N is a normalization factor that is typically equal to the maximum possible value of the function $f(x)$. In our case, where the objective is to measure the distance between image tags, the Google Similarity Distance was modified in order to rely on the co-occurrence of two tags in the space of social networks, rather than the co-occurrence of two words in the general space of web documents. More specifically, the functions $f(x)$ and $f(x;y)$ were substituted by the functions $fl(x)$ and $fl(x;y)$, which denote the total number of images in the entire flickr database that are tagged with x and both x and y, respectively. Then, the distance between tags x and y was calculated by the equivalent of Eq. 3, replacing f with fl, which will be called Google-Flickr Distance (GFD). All extracted distances were normalized to the $[0, 1]$ range and finally the similarity between two tags was calculated to be $1 - norm(GFD)$.

Alternatively, the widely known lexical database WordNet [5], was used in order to measure the semantic relatedness between image tags and concepts. More specifically, we employ the *vector* similarity metric [13] that combines the benefits of using the strict definitions of WordNet along with the knowledge of the concepts' co-occurrence which is derived from a large data corpus. Finally, an extra manual step is taken towards disambiguating the textual information. More specifically, when judging the relatedness score between two words, WordNet considers all different "meanings" for each word and outputs the maximum score among all possible combinations. This is an undesired behavior especially in cases where the examined words, apart from their "meaning" intended during the manual annotation process, happened to have other "meanings" that caused a severe misinterpretation of their semantic relatedness. For example, the word "palm" has five different meanings in the WordNet database. The first meaning of the word is the *inner surface of the hand from the wrist to the base of the fingers* while another one refers to *any plant of the family Palmae having an unbranched trunk crowned by large pinnate or palmate leaves*. In order to tackle this problem, while querying WordNet about the similarity between a concept and a tag, we manually select the intended meaning of the concept resulting in more accurate similarities. In this example, if we intended to search for palm trees we would select manually the second of the two aforementioned meanings of that word. Eventually, the use of any of these three approaches results in a textual similarity score between an image tag tag_j^I and a concept c_k, $TSim(tag_j^I, c_k)$. For every concept, its maximum similarity with the tags of the image I is chosen to gauge the possibility that the concept exists in the specific image:

$$t_{c_k}^I = \max_j \{TSim(tag_j^I, c_k)\} \quad (4)$$

Here, $t_{c_k}^I$ is a number in the $[0, 1]$ range and indicates the possibility that the concept c_k is present in the image I.

Finally, in order to decide how to combine the aforementioned independent scores into a single region relevance score we make the following observations. Our goal is that the relevance score of a region r_m^I corresponds to the system's confidence that this region depicts c_k. Thus, this score should be proportional to the visual score of the region $VS_{c_k}(r_m^I)$ and the textual score of the image it belongs to with respect to the concept c_k, $t_{c_k}^I$. Moreover, we want to select the regions that have both scores high, since a) these regions will come from images that do indeed contain the examined concept based on the image tags, and b) will have high visual similarity with the visual model as it is expressed by the SVM_{c_k} classifier. Based on the above, the region relevance scores are obtained by:

$$RR_{c_k}(r_m^I) = VS_{c_k}(r_m^I) * t_{c_k}^I \quad (5)$$

Finally, the regions of the loosely tagged images are ranked according to their region relevance score, and finally the top N regions with the highest relevance scores are selected to enhance the initial training set.

4. EXPERIMENTAL STUDY

The objective of our experimental setup is to show the benefits of the proposed multi modal region selection approach with respect to the straight forward self-training approach. To accomplish that, we examine two configurations based on the calculation of the RR function (eq. 5). In the first case RR is calculated using only the visual scores, which corresponds to a typical self-training approach, while in the second case RR is calculated according to eq. 5. The first case is essentially used as the baseline for measuring the improvement introduced by the incorporation of textual information.

In this context, in Section 4.2.1 the sample selection process is applied on a fully controlled dataset, where we can directly assess the quality of the samples selected to enhance the initial training set. In Section 4.2.2 the performance of the models retrained based on the proposed sample selection approach are compared to the performance of the initial classifiers which are trained using the manually labelled regions. Different methods for estimating the textual similarity between a concept and the image tags are evaluated in Section 4.2.3, while Section 4.2.4 compares the proposed approach with existing methods in the literature.

4.1 Datasets

The datasets that were used in our experimental study are shown in Table 1. The MIRFLICKR-1M dataset [10] consists of one million loosely tagged images harvested from flickr. The images of MIRFLICKR-1M were tagged with 862115 distinct tags of which 46937 were meaningful (included in WordNet). After the textual preprocessing 131302 images had no meaningful tag, 825365 images were described with 1 to 16 meaningful tags and 43333 images had more than 16 meaningful tags. This dataset constitutes the pool of loosely tagged images. For each concept, 1000 regions were selected from this dataset as positive examples to enhance the initial training set

The SAIAPR TC-12 dataset [4] consists of 20000 images labelled at region detail and was split into 3 parts (70% train, 10% validation and 20% test). To acquire comparable measures over the experiments, the images of the SAIAPR TC-12 dataset were segmented by the segmentation algo-

Name	Source	Size	Annotation Type	Usage	# positive training samples per concept
MIRFLICKR-1M	flickr	1 million	Loose Tags	100% training images	1000
SAIAPR TC-12	imageCLEF 2006	20000	Manual region-level annotations	70-10-20% training-testing-validation images	95 (average)

Table 1: Datasets

rithm described in Section 3.1 and the ground truth label of each segment was taken to be the label of the hand-labeled region that overlapped with the segment by more than the 2/3 of the segment's area. The validation set was used to train the initial classifiers for the proposed approach. After the segmentation step, on average 95 regions per concept were used as positive samples to train the classifiers. The testing set was used for evaluation purposes and the mean average precision (MAP) served as the metric for evaluating the proposed approach.

4.2 Evaluation results

4.2.1 Sample Selection Performance Assessment

The objective of this experiment is to show the impact of employing the textual information to the relevance rank order of the analyzed regions. In other words, we aim to assess the quality of the training samples that are selected to enhance the initial training set, before building the new classification model. In order to do this, the pool of the loosely tagged images need to be annotated at region level. For this reason, we artificially treat the training set of SAIAPR TC-12 as the pool of loosely tagged images by loosening the region labels to become tags for the whole image. Then, we train the initial models using the validation set of SAIAPR TC-12 (2k images) and apply these models to the regions of the SAIAPR TC-12 training set. Finally, the region relevance scores are obtained by using both the baseline (i.e. using only the visual information) and the proposed multi-modal approach configurations.

Each implementation is assessed by calculating the MAP of the regions that are ranked based on their relevance score. The validation set is annotated with 188 distinct concepts, of which 166 are present in the 14k images of the training set. The concepts that had less than 15 instances (22 concepts) were not included in the evaluation procedure to ensure the statistical safety of the conclusions. Table 2 shows the results averaged over all concepts. The particularly low performance (4.56%) exhibited by the "Visual" classifier in Table 2 can be attributed to the fact that having to classify the full set of regions extracted from all images in the SAIAPR TC-12 training set, the "Visual" classifier operates on a particularly dense feature space. In such a space it is extremely difficult for a classifier to perform satisfactorily, even if using a hyperplane like the SVM models employed in our work. Such an explanation is further supported by the fact that the performance of the exact same classification model is dramatically increased when the regions are initially filtered by the textual score before being subject to classification. In this way, the density of the feature space is significantly reduced allowing the classifier to more easily distinguish between the relevant and irrelevant regions. It is evident from the above that the quality of the samples selected to enhance the initial training set is particularly high when the proposed multi-modal approach is employed, which is a positive indication for the performance of the retrained object detection models.

Metric	Visual	Visual*Textual
MAP	4.56%	58.78%

Table 2: Region selection by artificially treating the SAIAPR TC-12 training set as the pool of loosely tagged images (MAP)

4.2.2 Performance comparison of the retrained models

In this experiment the performance of the initial classifiers which were trained using the manually labelled regions is compared to the performance of the enhanced classifiers (i.e. the ones trained by the combination of the labelled and the selected regions). The initial classifiers were enriched by the top N regions ranked using their Relevance Score, and provided that they had survived the adaptive region dismissal approach described in Sec. 3.1. For this experiment, the textual similarities were calculated using the WordNet method. In all experiments N was set at 1k. The validation set of the SAIAPR TC-12 dataset was used for training the initial models and the test set was used to evaluate the performance of all generated models. The concepts that had less than 15 instances in the test set were not included in the evaluation procedure to ensure the statistical soundness of the conclusions. Moreover, the concepts that were not members of the WordNet database were also excluded. The average precision of each concept separately as well as the MAP are shown in Fig. 2. The first bar is the performance of the initial classifiers, second bar is the performance of the enhanced classifiers with the regions that were selected using the typical self-training approach. Finally, for the third bar both visual and textual scores contributed to the region relevance scores. By examining this figure, we can see that the configuration incorporating both visual and textual information exhibits the highest performance in 44 out of the 63 examined concepts, compared to 4 for the typical self-training configuration and 15 for the configuration based on the initial classifiers. In average, the initial classifiers scored 5.7% in terms of MAP while the performance of the retrained models using solely the visual information was worse than initial classifiers (5.1%). This was actually expected given that the quality of the samples that were selected using the typical self-training approach was particularly poor (see Table 2), leading to the degradation of the retrained models. On the other hand, the enhanced classifiers trained using the proposed approach did manage to

	Visual	Visual*Textual
Without Preprocessing	4.9%	6%
Adaptive Dismissal	5.1%	7%

Table 3: Comparing different region dismissal algorithms (MAP)

improve their performance over the baseline by scoring 7% in terms of MAP.

Moreover, in an attempt to estimate the improvement resulting from the employment of the adaptive region dismissal approach, we have also measured the MAP without any pre-processing on the automatically segmented regions. The results (Table 3) show that the proposed approach for adaptive region dismissal greatly increases the performance of the resulting classifiers.

4.2.3 Evaluation of different textual similarity estimation approaches

In order to further investigate the impact of textual analysis in the process of optimizing the region selection process, we have comparatively evaluated the performance of the three methods that were described in Section 3.2 for calculating the textual scores. To this end, the proposed multimodal region selection approach was applied three times, each one using a different textual similarity estimation method. The results for each concept are shown in Fig. 3. The first bar shows the results using WordNet, the second using the manual disambiguation process with WordNet and finally the third bar using the Google-Flickr Distance. In general we can see that for the majority of concepts all three methods perform equivalently, with WordNet and disambiguated WordNet performing slightly higher than Google-Flickr Distance. On average, both WordNet and disambiguated WordNet scored $\sim 7\%$ in terms of MAP, while the Google-Flickr Distance scored 6.8%. This was expected since the Google-Flickr Distance is based solely on the words' co-occurrences while WordNet includes the information that is provided by the WordNet lexical database. However, the benefit of the Google-Flickr Distance is that it is fully automatic and can be estimated for any word as long as it exists in flickr, while on the other hand, WordNet limits the concepts and tags to the words included in its lexical database. Finally, by looking more closely to the results obtained using WordNet and its disambiguated version, it is interesting to note that the performance of some ambiguous concepts like *palm* and *branch* was boosted by the use of this extra disambiguation step. Based on the above, it is evident that the quality of textual scores largely depends on the nature of the considered concept (e.g. ambiguous concepts, concepts with overlapping WordNet glosses, concepts that can be better explained through their co-occurrence than their meaning) and different methods can be used to cover all existing cases.

4.2.4 Comparing with existing methods

In order to compare the proposed approach with existing methods the results of [4] were used. The authors introduce the SAIAPR TC-12 dataset and evaluate seven different classification schemes. In all cases, the manually labelled regions of the training set were used to train the classifiers following the OVA approach. Every test region was classified by all the classifiers and their outputs were merged by selecting the prediction of the classifier with the highest confidence. In order to compare our approach with the various classification schemes, the same merging procedure was applied. The classification accuracy served as the evaluation measure. Table 4 shows the results. We can see that the performance of the proposed approach is higher in three of the seven examined cases, i.e. when using Zarbi, Naive Bayes and SVM classifiers. However, given that our purpose is not to evaluate the performance of different classification schemes but to assess the improvement introduced by optimizing the sample selection process, the only value that can be considered directly comparable with our case is the one obtained using the SVM classification scheme. For this case, the proposed approach outperforms the corresponding SVM classifier that was evaluated in [4], by 13.6% units of accuracy.

Classifier	Classification Accuracy (%)
Zarbi	6,4
Naive Bayes	14,8
Klogistic	35
Neural Net	22,9
SVM	6,2
Kridge	30,3
Random Forest	39,8
Proposed Approach	19.8

Table 4: Comparing the performance of our work with all approaches implemented in [4]

5. CONCLUSIONS

In this work we have shown the benefit of using a multimodal region selection approach for boosting the efficiency of object detection classifiers. In order to demonstrate the functionality of our method we have chosen a large collection of tagged images obtained from flickr. Using these images, we have relied on the self-training paradigm to validate the value of using textual information so that the sample selection process for retraining is optimized. Our experimental results have shown that by using our approach we manage to improve the object detection performance over the baseline, which was not the case when using typical self-training.

Moreover, considering the ill-posed nature of the segmentation problem that often results in small and uninformative regions, we have verified the benefit of using a preprocessing step for adaptively dismissing the regions with a relatively small pixel-size. Finally, we have investigated the use of various methods for estimating the textual similarity between two words, aiming to capture different aspects of relatedness such as literal meaning, co-occurrence in social networks and disambiguated meaning. Our experiments have shown that the quality of the obtained textual score is largely affected by the nature of the considered concept, expressing the need for concept-oriented mechanisms in exploiting more efficiently the available textual information.

In our future work we plan to further optimize the process of region selection, by developing more complex schemes for estimating textual similarity and coping with the imperfections of the employed visual analysis scheme. Moreover, in addition to enhancing the positive training samples, enhancing the set of negative samples could help to further improve the performance of the classifiers. Finally, investigating al-

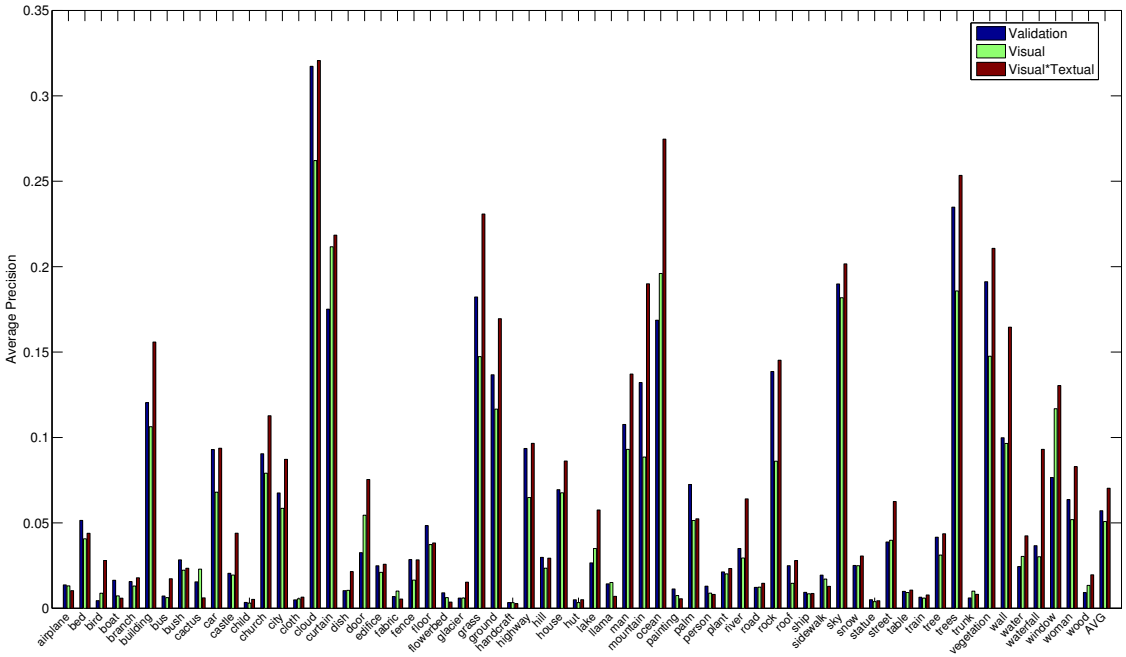

Figure 2: Performance of the initial classifiers (generated by using the labelled regions of the validation subset of the SAIAPR TC-12 dataset) and the enhanced classifiers using Visual and Visual*Textual information.

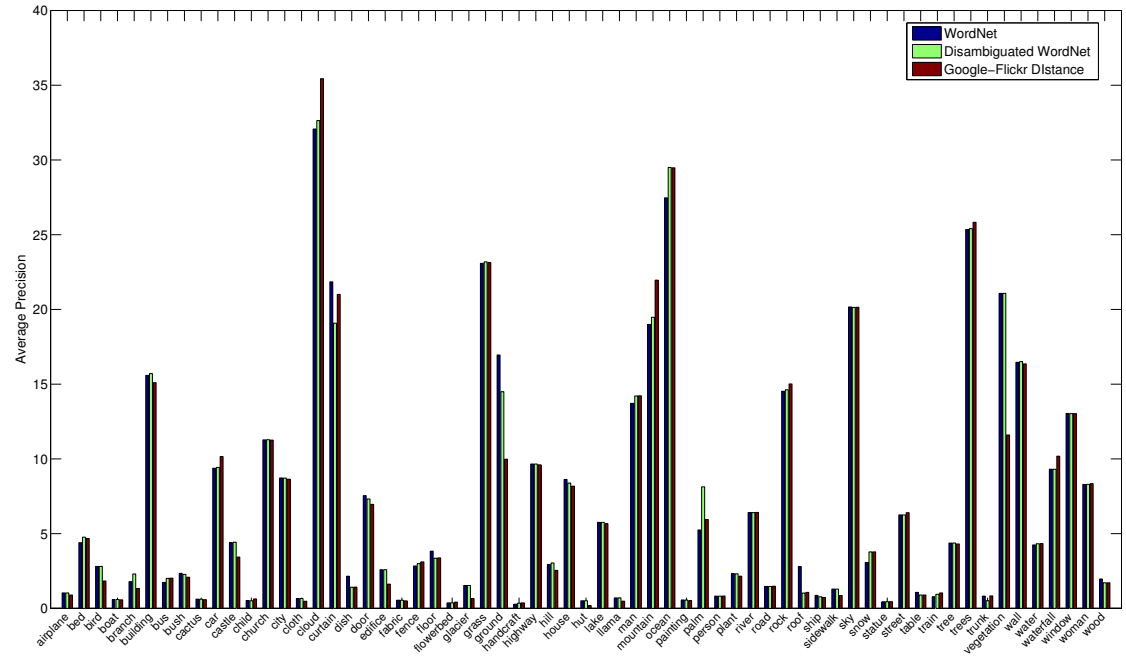

Figure 3: Performance of the three proposed textual similarity estimation approaches.

ternative ways to exploit the multi-modal nature of tagged images by improving the information fusion process, is also within our future plans.

Acknowledgment

This work was supported by the Glocal and SocialSensor FP7 projects, partially funded by the EC, under contract numbers FP7-248984 and 287975 respectively.

6. REFERENCES

[1] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[2] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1931–1947, 2006.

[3] R. Cilibrasi and P. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370 –383, march 2007.

[4] H. J. Escalante, C. A. Hernandez, J. A. Gonzalez, A. Lspez-Lspez, M. Montes, E. F. Morales, L. E. Sucar, L. Villase?or, and M. Grubinger. The segmented and annotated iapr tc-12 benchmark. *CVIU*, 2010.

[5] C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.

[6] W. Jiang and A. Loui. Laplacian adaptive context-based svm for video concept detection. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, WSM '11, pages 15–20, New York, NY, USA, 2011. ACM.

[7] T. Joachims. *Learning to classify text using support vector machines*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers, 2002.

[8] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Social negative bootstrapping for visual categorization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 12:1–12:8, New York, NY, USA, 2011. ACM.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[10] B. T. Mark J. Huiskes and M. S. Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA, 2010. ACM.

[11] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still image segmentation tools for object-based multimedia applications. *IJPRAI*, 18(4):701–725, 2004.

[12] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA, 2010. ACM.

[13] S. Patwardhan. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota, Duluth, August 2003.

[14] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:252–264, 1991.

[15] Y. Shen and J. Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *Proceedings of the international conference on Multimedia*, MM '10, pages 5–14, New York, NY, USA, 2010. ACM.

[16] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV (1)*, pages 1–15, 2006.

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *ICCV*, pages 370–377, 2005.

[18] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society.

[19] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pages 107–118, New York, NY, USA, 2001. ACM.

[20] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2008.

[21] J. Verbeek and W. Triggs. Region Classification with Markov Field Aspect Models. In *IEEE Conference on Computer Vision & Pattern Recognition (CPRV '07)*, pages 1–8, Minneapolis, United States, 2007. IEEE Computer society.

[22] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, 2007.

[23] L. Zhang, J. Ma, C. Cui, and P. Li. Active learning through notes data in flickr: an effortless training data acquisition approach for object localization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 46:1–46:8, New York, NY, USA, 2011. ACM.

[24] X. Zhu. Semi-Supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.