Using Tagged Images of Low Visual Ambiguity to Boost the Learning Efficiency of Object Detectors

Elisavet Chatzilari
Centre for Research & Technology Hellas - Information Technologies Institute
Centre for Vision, Speech and Signal Processing, University of Surrey Guildford, UK
ehatzi@iti.gr

ABSTRACT

Motivated by the abundant availability of user-generated multimedia content, a data augmentation approach that enhances an initial manually labelled training set with regions from user tagged images is presented. Initially, object detection classifiers are trained using a small number of manually labelled regions as the training set. Then, a set of positive regions is automatically selected from a large number of loosely tagged images, pre-segmented by an automatic segmentation algorithm, to enhance the initial training set. In order to overcome the noisy nature of user tagged images and the lack of information about the pixel level annotations, the main contribution of this work is the introduction of the visual ambiguity term. Visual ambiguity is caused by the visual similarity of semantically dissimilar concepts with respect to the employed visual representation and analysis system (i.e. segmentation, feature space, classifier) and, in this work, is modelled so that the images where ambiguous concepts co-exist are penalized. Preliminary experimental results show that the employment of visual ambiguity guides the selection process away from the ambiguous images and, as a result, allows for better separation between the targeted true positive and the undesired negative regions.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

user tagged images, multimedia data augmentation, social bootstrapping, visual ambiguity, semantic segmentation

1. INTRODUCTION

The majority of state-of-the-art methods for automatic object detection rely on the paradigm of pattern recognition through machine learning. Based on this paradigm, a model is parametrized to recognize all different attributes of a concepts' form and appearance, using a set of training examples. The efficient estimation of model parameters mainly depends on two factors, the quality and the quantity of the training examples. High quality is usually accomplished through manual annotation, which is a laborious and time consuming task. This has a direct impact on the second factor since it inevitably leads into a small number of training examples and limits the performance of the generated models. On the other hand, the excessive use of Web 2.0 applications has made available large amounts of user tagged images. Motivated by the above and inspired by semi supervised learning [1], the goal of this work is to combine the advantages of manually labelled data with the cost effectiveness of social networks.

However, the nature of these annotations (i.e., global level) and the noise existing in the associated information disqualifies them from being directly appropriate learning samples. Nevertheless, the tremendous volume of data that is currently hosted in social networks gives us the luxury to disregard a substantial number of candidate examples, provided we can devise a gauging mechanism that could filter out any ambiguous or noisy samples. Towards this goal, the main contribution of this work is to define, model and utilize visual ambiguity, which arises when two semantically different objects share similar visual stimuli under the employed representation system. In the proposed approach, visual ambiguity is modelled through a measure of image trustworthiness and is employed within an adapted self-training technique designed to combine the benefits of both manual annotations in terms of effectiveness, and social sites in terms of scalability. More specifically, for every concept, a set of regions is selected to enhance the initial training set based on three parameters; a) the visual similarity of the region with the examined concept as expressed by the initial object detection model, b) the textual information (tags) of the image the region belongs to and, c) the trustworthiness of the image the region belongs to, as defined by the ambiguity characterizing its content.

The rest of this paper is organized as follows. In Section 2 the related literature is reviewed. The proposed approach is explained in Section 3 while experimental results are shown in Section 4 and conclusions are drawn in Section 5.

2. RELATED WORK

In an attempt to minimize the labelling effort, approaches that rely on active learning (i.e. selectively sampling and annotating examples based on their *informativeness* as they

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502208.

are expected to improve the model performance) have been recently presented [12], [7]. The authors of [12] introduce the concept of *live learning* and propose to replace the human oracle in the typical active learning method with a crowdsourcing service like the MTurk to provide the annotations of the selected informative samples. On the other hand, social networks and user contributed content are leading the recent research efforts, mainly because of their ability to offer more information than the mere image visual content, coupled with the potential to grow almost unlimitedly. In this direction, the authors of [7] propose a solution for actively sampling the most misclassified user tagged images to enrich the negative training set of a concept 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. However, active learning without an expert oracle is feasible in these cases because they either rely on non-expert, but still manual annotations (MTurk) or are applied on image level classifiers, which removes the additional factor of localization. In contrast, the proposed approach utilizes loosely tagged images which are provided at no cost and operates on segmented regions instead of global images.

A few approaches have been proposed towards fully unsupervised object detection exploiting user tagged images ([3], [9]). In [3], a theoretical and experimental study is presented to validate the assumption that if the set of loosely tagged images is properly selected, the most frequently appearing visual object and user contributed tag will coincide. Utilizing this assumption, object detection models are build in an unsupervised manner. In a similar fashion, the authors of [9] 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. Visual ambiguity is also considered in [11], where soft assignment of visual words is proposed by considering the *visual word uncertainty* (i.e. an image feature may have more than one candidates in the visual word vocabulary) and the *visual word plausibility* (i.e. when there is no suitable visual word for the image feature).

A preliminary version of the proposed work was presented in [2], where the approach was inspired by the bootstrapping method. Additionally, in this case, the visual ambiguity between regions is also considered and modelled. This measure, unlike other works, is exploited directly in the classification scheme for discarding the misleading images that contain ambiguous concepts, as in these cases selecting the targeted region would be rather difficult.

3. APPROACH

The proposed approach for extracting training samples from unambiguous loosely tagged images is depicted in Fig. 1. Given a concept c_k, an initial classifier is trained on a set of regions that are labelled with this concept and additional regions representing this concept are chosen from a pool of user tagged images harvested from the web. 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. The loosely tagged images are automatically segmented into regions that roughly correspond to semantic objects and visual features

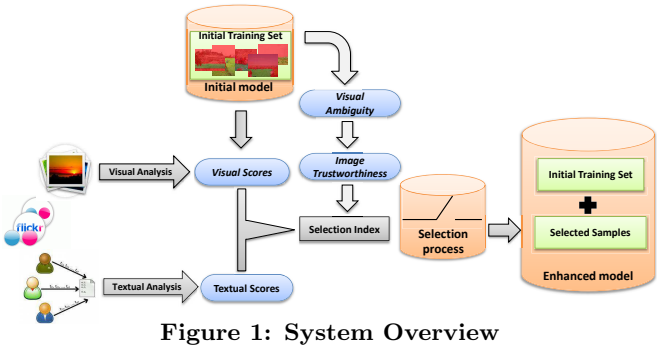


Figure 1: System Overview

are extracted to represent each region. Support Vector Machines (SVMs) are utilized to train initial classifiers using the visual features that were extracted by the labelled regions. Applying these classifiers to the unlabelled regions provides the visual scores. Next, the textual scores are extracted by the textual information that accompanies the loosely tagged images. Finally, visual ambiguity is modelled and transformed into image trustworthiness scores, which practically indicate how much a classifier is trusted to classify the regions that have been extracted from a specific image. In this way, regions are selected so that they represent the concept c_k while at the same time, the ambiguous content is identified and discarded. This luxury is provided by the exuberant amount of the available user contributed content.

Segmentation and feature extraction

In order to segment the images into regions the K-means with connectivity constraint (KMCC) segmentation algorithm [8] is employed. For the visual representation of the regions, a typical bag of visual words (BOW) approach is used. SIFT features [10] are extracted for every key-points detected by the Harris-Laplace and the dense detectors and they are subsequently encoded in a single vector using a vocabulary of 500 words and the soft assignment method [11].

Visual and Textual Scores Estimation

For every concept c_k, an object detection model (SVM_{c_k}) is trained using the one versus all approach. The distance of a region r_m^I, which belongs in image I, from the hyperplane of the SVM_{c_k} model will be referred to as visual score $VS_{c_k}(r_m^I)$ from now on. This score indicates the confidence of the model that the region r_m^I depicts the concept c_k.

In addition, in order to utilize the textual information provided with the user tagged images, the widely known lexical database WordNet [5] is utilized to measure the semantic relatedness between image tags and concepts. More specifically, for a loosely tagged image I with tags $Tag^I = \{tag_1^I, tag_2^I, ..., tag_{N_{tag}}^I\}$ the textual similarity score between its image tags and a concept $TSim(tag_j^I, c_k)$ is calculated using WordNet. For every concept c_k, its maximum similarity with the tags of image I is chosen to gauge the possibility $t_{c_k}^I$ that the concept c_k exists in the specific image resulting in a vector of textual scores for every image:

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

Visual Ambiguity and Image Trustworthiness

In order to model the visual ambiguity that arises between visually similar concepts the visual ambiguity scores are introduced and are estimated using the following process. For a concept c_k, given its model SVM_{c_k}, the visual scores of all the regions that have been used to train this model, are determined. In the ideal case the visual scores of all the

regions depicting c_k should be much higher than the visual scores of all other regions. When regions that do not depict c_k are associated with high visual scores by SVM_{c_k}, the discriminative ability of SVM_{c_k} is low. This is considered as the visual ambiguity between c_k and the concept $c_l, l \neq k$, which is the actual concept depicted by the examined region. The visual ambiguity of c_k and c_l is selected to be the average of the visual scores that the regions belonging to the c_l class received:

$$VA_{c_k,c_l} = \begin{cases} \frac{1}{N_l} \sum_{i=1}^{N_l} VS_{c_k}(r_i^{c_l}) & \text{if } k \neq l \\ 0 & \text{if } k = l \end{cases} \quad (1)$$

where $r_i^{c_l}, i = 1 \dots N_l$ are the regions that depict c_l. The visual ambiguity between two concepts c_k and c_l is high when the model that is trained to detect c_k produces high confidence scores for the $r_i^{c_l}$ regions, which practically means that our system tends to confuse the visual information that depicts c_k with the visual information that depicts c_l. For example, the visual ambiguity scores of the closely related couples of concepts *grass-plant* (0.824) and *grass-bush* (0,874) are higher than the visual ambiguity score of the couple *grass-fence* (0,638).

The visual ambiguity scores indicate how much a specific classifier is trusted to distinguish between two concepts when asked to classify a region. Having this knowledge for every couple of concepts, it could be applied on every image separately if the existent objects in the image were known. This information might not be available explicitly, but the possibility about the existence of an object within an image is available through the textual score of the image. If the textual score of a concept in the image is above a threshold th, we consider that the concept is present in the image. In order to express the trustworthiness of the classifier SVM_{c_k} to classify the regions of an image I, the dynamic visual ambiguity $VA_{c_k}^I$ of an image I with respect to a concept c_k, is calculated as a function of the static information **VA**:

$$\mathbf{VA_{c_k}^I} = \mathbf{T_{th}^I} * \mathbf{VA_{c_k}} \quad (2)$$

The difference between $\mathbf{VA_{c_k}^I}$ and $\mathbf{VA_{c_k}}$ is that $\mathbf{VA_{c_k}^I}$ is calculated for a specific image I and gauges how ambiguous is the specific image, whereas $\mathbf{VA_{c_k}}$ is static information, independent of the image that is based on the visual representation system. Finally, image trustworthiness is defined as the complement of the maximum visual ambiguity score exhibited among the existing concepts with respect to c_k. The trustworthiness score of an image I with respect to c_k gauges how much a classifier can be trusted to classify the regions of the image I with respect to the concept c_k:

$$Trust_{c_k}^I = 1 - \max_l(VA_{c_k,c_l}^I) \quad (3)$$

In the previous example for the concept *grass*, the classifier is trusted more to detect the grass regions within images that contain *fence*, than within images that contain *bush* ($VisualAmbiguity(grass, fence) = 0.638 < VisualAmbiguity(grass, bush) = 0.874$).

Region relevance and selection of training samples

In order to combine the three aforementioned independent scores into a single region relevance score, the geometric mean is chosen over the more typical arithmetic mean due to its robustness when multiplying quantities with different normalizations.

$$RR_{c_k}(r_m^I) = VS_{c_k}(r_m^I) * t_{c_k}^I * Trust_{c_k}^I \quad (4)$$

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. PRELIMINARY RESULTS

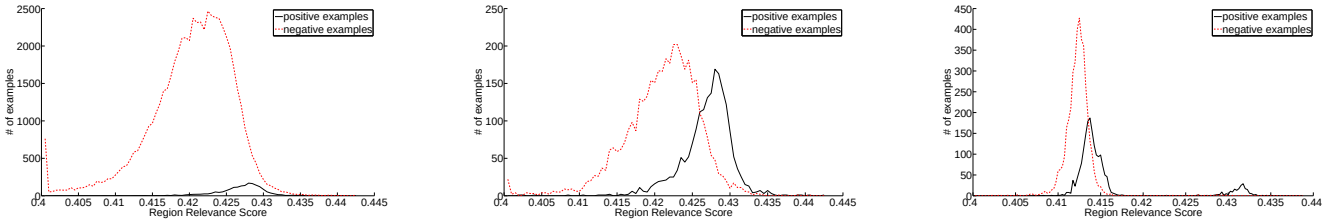
Two datasets were used in the experimental study. The MIRFLICKR-1M dataset [6] consists of one million user tagged images harvested from flickr. This dataset consists the pool of tagged images, from where the training regions were selected to enhance the manually trained models. The second dataset, the 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 and the ground truth label of each segment was taken to be the label of the hand-labelled region that overlapped with the segment by more than the 2/3 of the segment's area. The concepts that had less than 15 instances were removed to ensure statistical safety. The mean average precision (mAP) served as the metric for evaluating the proposed approach.

4.1 Sample Selection Performance

The objective of this experiment is to show the impact of employing the ambiguity and the image trustworthiness scores to the ranking of the regions. In order to be able to evaluate the selection process directly, the user tagged images should be annotated at region level. For this reason, the training set of the SAIAPR TC-12 dataset (14k images) was used by loosening the region labels to image tags-keywords. The initial models were trained using the validation set (2k images) and were applied to the regions of the training set of SAIAPR TC-12. The regions were ranked based on a) the visual scores (**V**), b) the geometric mean of the visual and textual scores (**VT**) and c) the proposed approach, i.e. applying Eq. 4 (**VTA**). In Fig. 2, the distribution of the region relevance scores, calculated as explained by each configuration (i.e. **V**, **VT** and **VTA**), is shown for the concept grass. The black solid line is the distribution of the positive examples, i.e. the targeted regions which we opt to select, and the red dashed line is the distribution of the negative examples. It is obvious, that without the auxiliary information the classifier performs poorly (Fig. 2(a)), since the two distributions overlap significantly. Moreover, we can see that the textual information has eliminated a large number of non-relevant regions (Fig. 2(b)), which was expected since in this case the tags are accurate. Finally the impact of visual ambiguity is clearly shown in Fig. 2(c), where part of the black distribution, i.e. true positives, now stands out receiving much higher region relevance scores compared to the rest. This effect would be ideal in the case of loosely tagged images since it makes more accurate the selection of the top N regions. Additionally, the mAP over all concepts is measured and written in the caption. The numerical results validate the aforementioned conclusions as well.

4.2 Retrained Models Performance

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 $1k$ regions ranked based on the configurations **V**, **VT** and **VTA**. The validation set of the SAIAPR TC-12



(a) **V** (4.56% mAP) (b) **VT** (58.78% mAP) (c) **VTA** (65.55% mAP)

Figure 2: The distribution of the RR scores (Eq. 4) based on the configuration a) V, b) VT and c) VTA.

dataset (2k images) is used for training the initial models and the test set (4k images) is used to evaluate the performance of all generated models. The mAP of the initial models is 5.9%, while adding regions ranked based on the **V** configuration degraded the models, to 4.9% mAP. Using the **VT** and **VTA** configurations, the enhanced models increased their performance to 6 and 6.3% respectively. These results comply with the conclusions reached in the previous section, showing the positive impact of ambiguity to the sample selection process. Examining each concept independently, the configuration incorporating visual ambiguity exhibits the highest performance in 26 out of the 62 examined concepts, compared to 19 for the **VT** configuration, 3 for the **V** configuration and 14 for the configuration based on the initial classifiers.

5. CONCLUSIONS AND FUTURE WORK

In this work we have presented a means to quantify and utilize the visual ambiguity that characterizes the image content, with a view to boost the efficiency of object detection classifiers. More specifically, we have relied on the self-training paradigm to validate the value of using visual ambiguity for the optimization of the sample selection process. Preliminary experimental results show that by using the proposed approach to cope with the existing ambiguities, the improvement in performance is higher than the one achieved using a typical self-training approach, where the sample selection process is based solely on the visual information of the initial models. An interesting observation that came our of our experimental study relates to the use of WordNet and the fact that this similarity metric does not take into account the context of the words to disambiguate their meaning. For example, the words palm and tree would always yield a very high similarity score regardless if the intended meaning for palm was the tree or the hand. In these cases our approach was heavily misled, making impossible the extraction of a reliable score for image trustworthiness. In our future work we plan to investigate ways for alleviating the negative effects of using WordNet and examine other context-based metrics. With respect to visual ambiguity we plan to investigate alternative, more sophisticated ways for fusing the available information from the various modalities towards a better selection strategy. Additionally, we are working on designing experimental set-ups for proving directly that the proposed method for quantifying visual ambiguity is a reasonable measure. Finally, the exploitation of a richer source for positive samples, like flickr groups, is within our future plans. Using better suited content rather than a canned dataset would allow for more iterations and for achieving better performance improvements.

Acknowledgements

This work was supported by the EU 7th Framework Programme under grant number IST-FP7-288815 in project Live+Gov (www.liveandgov.eu).

6. REFERENCES

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

[2] E. Chatzilari, S. Nikolopoulos, Y. Kompatsiaris, and J. Kittler. Multi-modal region selection approach for training object detectors. ICMR. ACM, 2012.

[3] E. Chatzilari, S. Nikolopoulos, I. Patras, and I. Kompatsiaris. Leveraging social media for scalable object detection. *Pattern Recognition*, 45(8):2962–2979, 2012.

[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] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR*. ACM, 2010.

[7] X. Li, C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. Bootstrapping visual categorization with relevant negatives. *IEEE Trans. on Multimedia*, In press, 2013.

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

[9] Y. Shen and J. Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *MM*, pages 5–14. ACM, 2010.

[10] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[11] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[12] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, pages 1449 –1456, 2011.