Towards modelling visual ambiguity for visual object detection

Elisavet Chatzilari
Informatics and Telematics Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
ehatzi@iti.gr

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK
e.chatzilari@surrey.ac.uk

Spiros Nikolopoulos
Informatics and Telematics Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
nikolopo@iti.gr

Yiannis Kompatsiaris
Informatics and Telematics Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
ikom@iti.gr

Josef Kittler
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK
J.Kittler@surrey.ac.uk

ABSTRACT

The widespread adoption of Web 2.0 applications has resulted in the creation of huge amounts of user-generated multimedia content, a fact that motivated the investigation of employing this content for training. However, the nature of these annotations (i.e. global level) and the noise existing in the associated information, as well as the ambiguity that characterizes these examples 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 learning examples, provided we can devise a gauging mechanism that could filter out any ambiguous or noisy samples. Our objective in this work is to define a measure for visual ambiguity, which is caused by the visual similarity of semantically dissimilar concepts, in order to help in the process of selecting positive training regions from user tagged images. This is done by limiting the search space of the potential images to the ones yielding a higher probability to contain the desired regions, while at the same time not including visually ambiguous objects that could confuse the selection algorithm. Experimental results show that the employment of visual ambiguity allows for better separation between the targeted true positive and the undesired negative regions.

i-KNOW '14, September 16 - 19 2014, Graz, Austria
ACM 978-1-4503-2769-5/14/09 $15.00.
http://dx.doi.org/10.1145/2637748.2638431

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

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

1. INTRODUCTION

An important factor that affects the quality of supervised classifiers is the size of the training set. Aiming to improve the performance of the classifiers, the bootstrapping technique was designed to augment the training set with additional training samples [8]. However, within a typical bootstrapping process, the algorithm searches for the targeted true positive samples in a large pool of unlabelled data, the majority of which constitute the undesired negative examples. This fact incommodes the accurate selection of true positive regions, in the case of region level object detection, as the search space is noisy and dense. In order to thin out the search space of the algorithm, multi-modal selection strategies have been proposed by replacing the pool of unlabelled examples with user tagged images and using these tags to refine the pool of candidates into a set of images that are more probable to contain the targeted object [6, 2]. Following the same idea, in this work, we present a multimodal region selection strategy that opts to refine the search space not only by utilizing textual information for selecting the images that are more likely to depict the targeted object, but also by presenting a method that models visual ambiguity in order to disregard the ambiguous content.

Towards devising a gauging mechanism that could filter out the ambiguous samples, 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, which indicates how much the initial object detection model is trusted to find the targeted regions within the examined image. 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 measured by the initial object detection model, b) the textual similarity of the image tags with the examined concept indicating the possibility of its existence in the image and, c) the trustworthiness of the image the region belongs to, as defined by the ambiguity characterizing its content. In this way, the pool of candidates is limited to the most prominent images that will allow the bootstrapping algorithm to select accurately true positive examples.

2. RELATED WORK

A few approaches have been proposed towards fully unsupervised object detection exploiting user tagged images ([3], [10]). 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. In a similar fashion, the authors of [10] propose a multiple instance learning algorithm that incorporates 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). In contrary, in this work, visual ambiguity is exploited directly in the classification scheme for discarding the misleading images that contain ambiguous concepts. This luxury to disregard a substantial number of candidate examples that are deemed ambiguous is afforded because of the vast availability of social media content.

In a similar endeavour but by relying on the basic principles of active learning, the most recent works in this field propose to enhance the training set with the most misclassified negatives [6], or with the most prominent positive examples by jointly considering the oracle's confidence and the sample's informativeness when selecting new samples [2]. Based on the active learning theory as well, 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[1]. 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 (i.e. to detect the exact location of the object in the image). In contrast, the proposed approach utilizes cheaply obtained user tagged images instead of manually annotated regions by a crowdsourcing service and operates on segmented regions instead of global images.

[1]www.mturk.com

3. APPROACH

The proposed approach for extracting training samples from unambiguous user tagged images is depicted in Fig. 1. Given a concept c_k, an initial classifier is trained on a set of regions that have been manually labelled with this concept. This initial training set is enhanced with additional regions representing this concept, which 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 in the image or of their exact location within the image. To overcome this obstacle, the following process takes place. The user tagged images are segmented into regions by an automatic segmentation algorithm and visual features are extracted to represent each region. Support Vector Machines (SVMs) are utilized to train the initial classifiers for each concept c_k. Applying the classifier for c_k to an unlabelled region r_m^I provides the visual scores $VS_{c_k}(r_m^I)$, while the textual scores $t_{c_k}^I$ are extracted by the textual information that accompanies the user tagged image I (Section 3.1). Finally, visual ambiguity is modelled by image trustworthiness scores $Trust_{c_k}^I$, which practically indicate how much the classifier for c_k is trusted to classify the regions that have been extracted from image I (Section 3.2). In order to combine the three aforementioned independent scores into a single region relevance score $RR_{c_k}(r_m^I)$, the geometric mean is chosen over the more typical arithmetic mean due to its robustness when multiplying quantities with different normalizations:

$$RR_{c_k}(r_m^I) = \sqrt[3]{VS_{c_k}(r_m^I) * t_{c_k}^I * Trust_{c_k}^I} \quad (1)$$

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. 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.

3.1 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, we employ the *vector* similarity metric [9] 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. 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 the

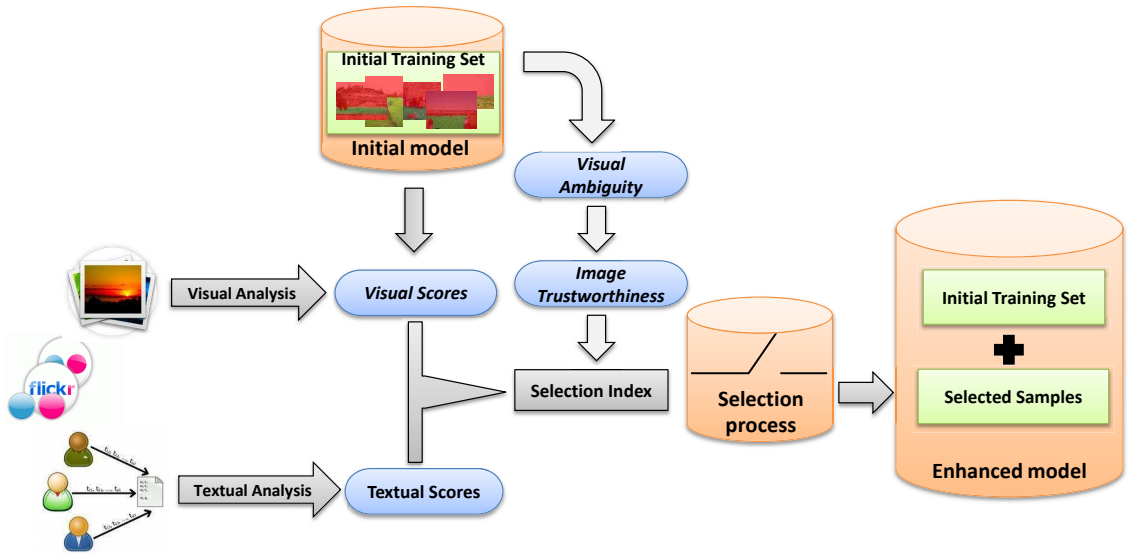


Figure 1: System Overview

linguistic expression of 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:

$$t_{c_k}^I = \max_j \{TSim(tag_j^I, c_k)\} \quad (2)$$

3.2 Visual Ambiguity and Image Trustworthiness

In order to model the visual ambiguity that arises between visually similar concepts the visual ambiguity scores are estimated using the following process. In the ideal case, when applying the model SVM_{c_k}, 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 depict c_l instead 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 the concepts c_k and $c_l, l \neq k$ and is selected to be the average of the visual scores that the regions belonging to the c_l class received:

$$VA(c_k, c_l) = \begin{cases} \frac{1}{N_l} \sum_{i=1}^{N_l} VS_{c_k}(r_i^{c_l}) & \text{if } k \neq l \\ 0 & \text{if } k = l \end{cases} \quad (3)$$

where $r_i^{c_l}, i = 1 \ldots N_l$ 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 the model SVM_{c_k} 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 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. Although this information might not be available explicitly, 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 object referring to this concept is present in the image ($T_{th}^I(c_k) = 1$ if $t_{c_k}^I > th$). The trustworthiness of the classifier SVM_{c_k} to classify the regions of an image I, is defined to be the complement of the visual ambiguity $VA_{c_k}^I$ of a specific image I with respect to a concept c_k, which is calculated as the maximum visual ambiguity of the SVM_{c_k} classifier with respect to the concepts that exist in image I, indicated by the textual scores T_{th}^I

$$Trust_{c_k}^I = 1 - \max_l (T_{th}^I(c_l) * VA(c_k, c_l)) \quad (4)$$

The trustworthiness score of an image I with respect to c_k gauges how much the classifier SVM_{c_k} can be trusted to classify the regions of the image I and depends on the existence of ambiguous concepts (i.e. c_l) in the image I. 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* (i.e. because $VA(grass, fence) = 0.638 < VA(grass, bush) = 0.874$).

4. EXPERIMENTAL RESULTS

Two datasets were used in the experimental study. The MIRFLICKR-1M dataset [7] consists of one million user tagged images harvested from flickr. This dataset constitutes the pool of tagged images, from where the training regions were selected to enhance the manually trained mod-

els. 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 automatically 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. Details about the segmentation process and the extracted visual features can be found in [1]. 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.

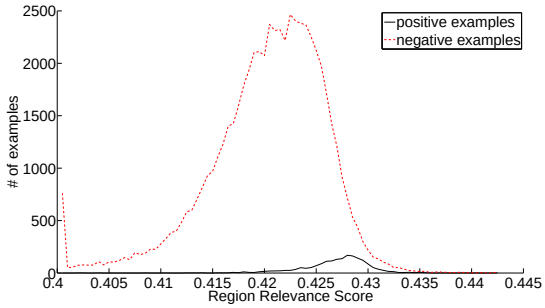
The following configurations of Eq. 1 were tested:

- using only visual scores, i.e. $RR_{c_k}(r_m^I) = VS_{c_k}(r_m^I)$ (**V**)
- using visual and textual scores, i.e. $RR_{c_k}(r_m^I) = VS_{c_k}(r_m^I) * t_{c_k}^I$ (**VT**)
- using Eq. 1 (**VTA**).

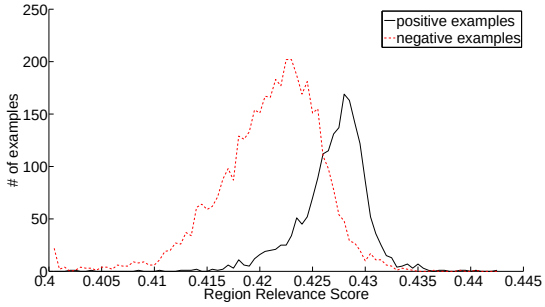
4.1 Sample Selection Performance

The objective of this experiment is to show the impact of employing visual ambiguity on 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 (i.e. if the regions $r1$, $r2$ and $r3$ of an image I are annotated as *sky*, *sea* and *sand* respectively, then we consider that the tags for image I are also *sky*, *sea* and *sand*). 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. In Fig. 2, the distribution of the region relevance scores, calculated as explained for 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 user tagged images since it makes the selection of the top N regions more accurate. Additionally, the mAP over all concepts is measured and written in the caption. The numerical results validate the aforementioned conclusions as well.

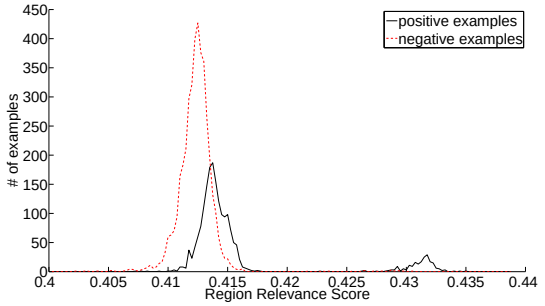
In an attempt to evaluate if the aforementioned conclusions also apply when the pool of candidates consists of user tagged images instead of accurate manual keywords as above, we apply the proposed sample selection approach on the MIRFLICKR-1M dataset and the regions belonging to



(a) **V** (4.56% mAP)



(b) **VT** (58.78% mAP)



(c) **VTA** (65.55% mAP)

Figure 2: The distribution of the RR scores (Eq. 1) based on configurations (a) V, (b) VT and (c) VTA.

the user tagged images are ranked based on the three aforementioned configurations (i.e. **V**, **VT**, **VTA**). Afterwards, we visually examine some of the regions ranked amongst the top N places by each configuration. The examples are shown in figure 3(a), 3(b) and 3(c) based on **V**, **VT** and **VTA** configurations respectively. The images in blue bounding boxes constitute the wrongly selected false positives. It is obvious that in the **V** case, where only the visual scores were used, the performance of the initial model is very poor and the selected regions are very noisy. Adding the textual information allows us to select a number of *grass* regions and the addition of visual ambiguity increases greatly the quality of the selected regions. Note that in the previous experiment (Fig. 2), this specific model of the concept *grass* was ranking the new unlabelled samples with an average precision of 17.14% while when adding the textual information the performance rose to 72.63% and finally to 76.07% when incorporating the visual ambiguity as well. This coincides with the visualization of the selected regions by this model.

4.2 Performance of the enhanced classifiers

Additionally, 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 from the MIRFLICKR-1M dataset). 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 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 previously, showing the positive impact of ambiguity to the sample selection process. Examining each concept independently (Fig. 4), The first bar (black) is the performance of the initial classifiers, second bar (red) is the performance of the enhanced classifiers with the regions that were selected by the baseline configuration **V**. For the third bar (yellow) visual and textual scores contributed to the region relevance scores (**VT**) while for the fourth bar (white) all the scores were used (**VTA**). By examining this figure, we can see that the configuration incorporating visual ambiguity **VTA** 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.

4.3 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. For all the classifiers of [4], the manually labelled regions of the training set were used to train the classifiers, while for the proposed approach the classifier trained in Section 4.2 with the **VTA** configuration was used. Every test region was classified by all classifiers (i.e. one for each concept) and was categorized to the concept with the highest prediction score of its corresponding classifier. The classification accuracy served as the evaluation measure. Table 1 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, it is evident that the proposed approach outperforms greatly the SVM classifier that was evaluated in [4].

Classifier	Classification Accuracy (%)
Zarbi [4]	6.4
Naive Bayes [4]	14.8
Klogistic [4]	35
Neural Net [4]	22.9
SVM [4]	6.2
Kridge [4]	30.3
Random Forest [4]	39.8
Proposed Approach	20.6

Table 1: Comparing Performance of the proposed approach with [4]

5. DISCUSSION OF THE RESULTS

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. Our experimental results have shown 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. In our future work we plan to investigate alternative, more sophisticated ways for fusing the available information from the various modalities towards a better selection strategy. Finally, the exploitation of a richer source for positive samples, which would allow for more iterations and for achieving better performance improvements, is within our future plans.

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] E. Chatzilari, S. Nikolopoulos, Y. Kompatsiaris, and J. Kittler. Multi-modal region selection approach for training object detectors. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pages 5:1–5:8. ACM, 2012.

[2] E. Chatzilari, S. Nikolopoulos, Y. Kompatsiaris, and J. Kittler. Active learning in social context for image classification. In *9th Int. Conference on Computer Vision Theory and Applications (VISAPP)*, Lisbon, Portugal, January 5-8 2014.

[3] E. Chatzilari, S. Nikolopoulos, I. Patras, and I. Kompatsiaris. Leveraging social media for scalable

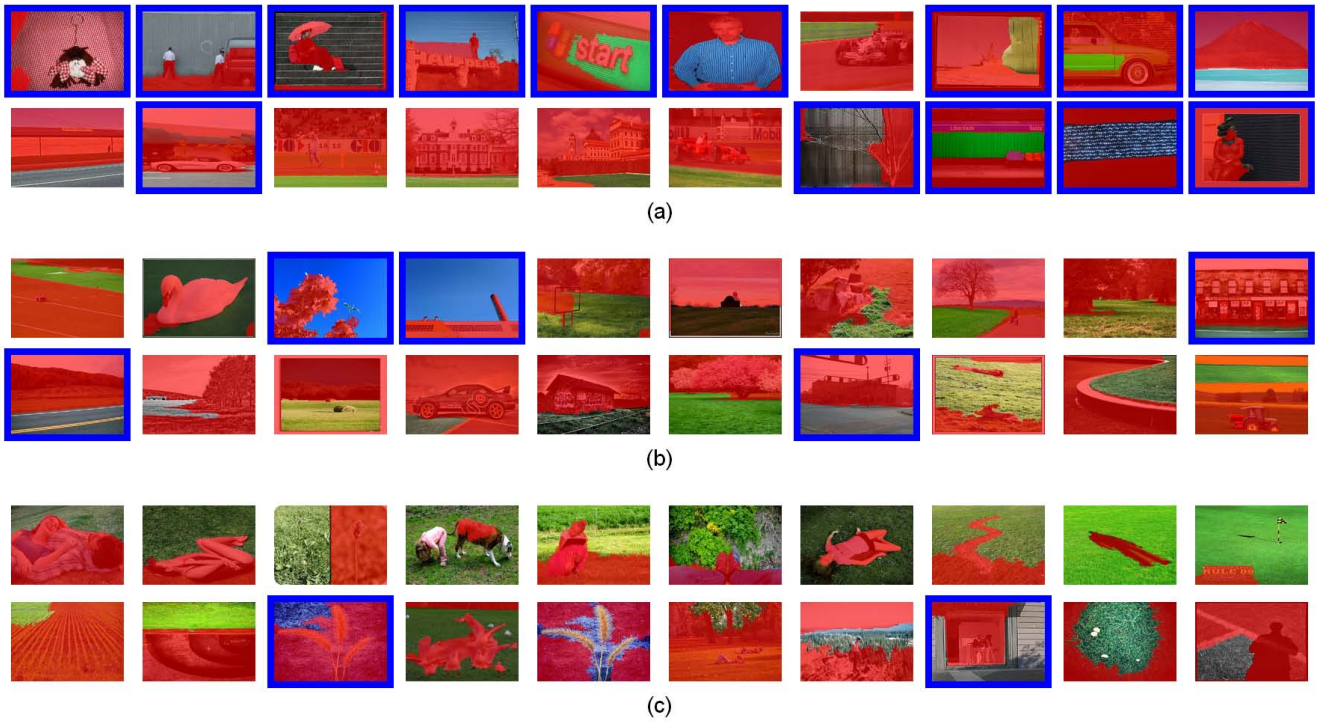


Figure 3: Indicative regions selected using the configurations (a) V, (b) VT and (c) VTA. A blue bounding box indicates a false positive result.

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] 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.

[7] 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.

[8] V. Ng and C. Cardie. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 113–120, 2003.

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

[10] Y. Shen and J. Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *ACM*, MM '10, 2010.

[11] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions 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.

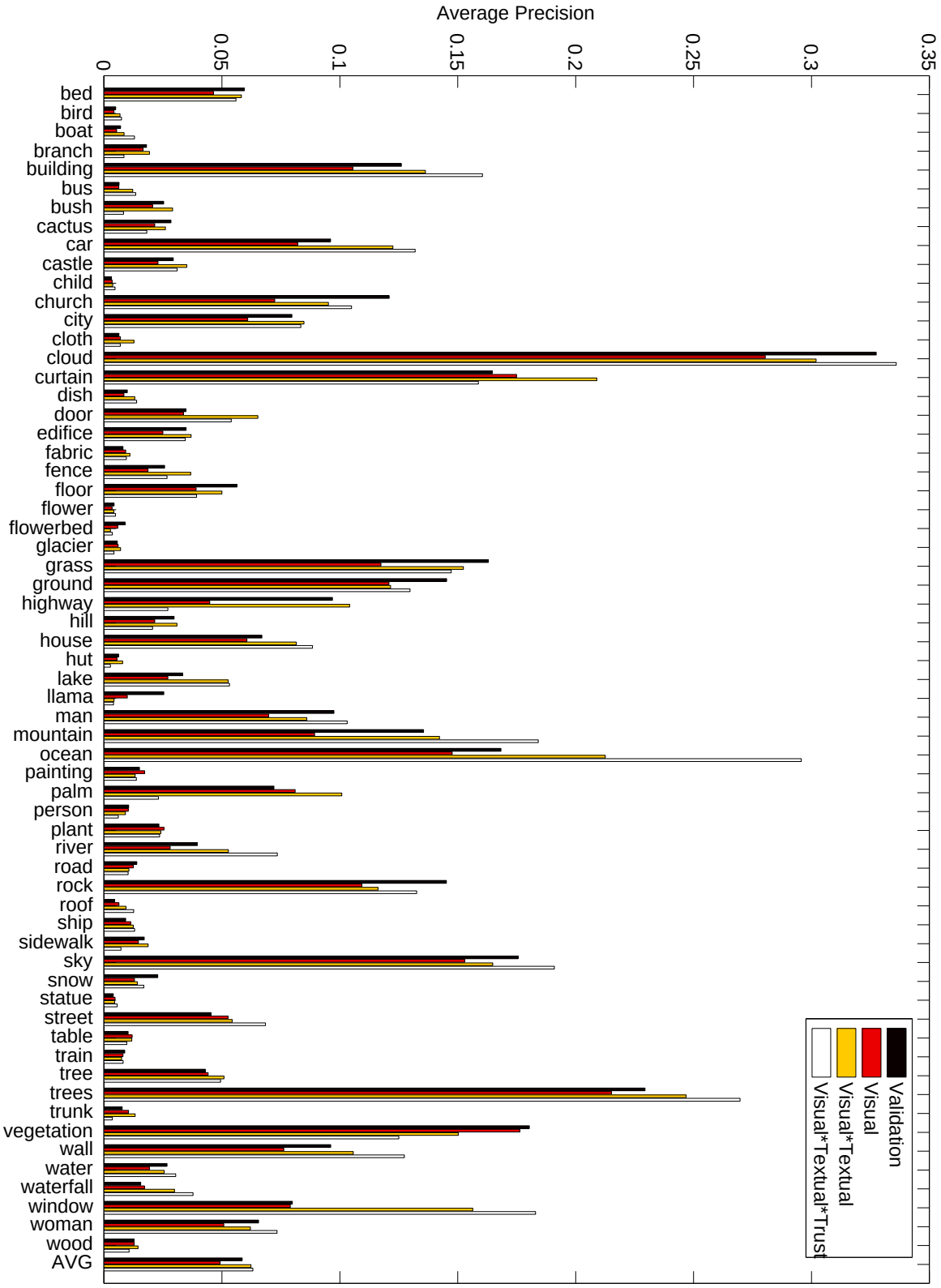


Figure 4: Performance of the initial and the enhanced classifiers using the V, VT and VTA configurations.