Hierarchical Image Annotation Using Semantic Hierarchies

Hichem Bannour
Applied Mathematics and Systems Department,
École Centrale Paris
92 295 Châtenay-Malabry, France
hichem.bannour@ecp.fr

Céline Hudelot
Applied Mathematics and Systems Department,
École Centrale Paris
92 295 Châtenay-Malabry, France
celine.hudelot@ecp.fr

ABSTRACT

Semantic hierarchies have been introduced recently to improve image annotation. They was used as a framework for hierarchical image classification, and thus to improve classifiers accuracy and reduce the complexity of managing large scale data. In this paper, we investigate the contribution of semantic hierarchies for hierarchical image classification. We propose first a new method based on the hierarchy structure to train efficiently hierarchical classifiers. Our method, named One-Versus-Opposite-Nodes, allows decomposing the problem in several independent tasks and therefore scales well with large database. We also propose two methods for computing a hierarchical decision function that serves to annotate new image samples. The former is performed by a top-down classifiers voting, while the second is based on a bottom-up score fusion. The experiments on Pascal VOC'2010 dataset showed that our methods improve well the image annotation results.

Keywords

Image annotation, hierarchical classification.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Content Analysis and Indexing

1. INTRODUCTION

Automatic image annotation is a challenging problem dealing with the textual description of images, i.e. associating tags or even better descriptive text to images. A wide number of approaches have been proposed to address this problem and to narrow the well-known *semantic gap* issue. Most approaches rely on machine learning techniques to provide a mapping function that allows classifying images in semantic classes using their visual features [3]. However, these approaches face the scalability problem when dealing with broad content image databases, i.e. their performances decrease significantly when the concept number is high. This variability may be explained by the huge intra-concept variability and a wide inter-concept similarity on their visual properties that often lead to incoherent annotations. Furthermore, multimedia retrieval systems require an increasing concept classes for annotating images in order to meet user needs. Accordingly, current techniques are struggling to scale up. Therefore, the only use of machine learning seems to be insufficient to solve the image annotation problem. Semantic structures, such as semantic hierarchies, appear to be a good alternative to reduce this problem complexity [1].

Semantic hierarchies have shown to be very useful to narrow the semantic gap. They identify the dependency relationships between concepts and provide valuable information for many problems. Semantic hierarchies can improve image annotation by supplying a hierarchical framework for image classification, and allow for efficiencies in both learning and representation. Three types of hierarchies for image annotation have been recently explored: 1) language-based hierarchies: based on textual information [11, 5], 2) visual hierarchies: based on low-level image features [8], 3) semantic hierarchies: based on both textual and visual features [9, 6, 2]. Semantic hierarchies provide a meaningful semantic structure that helps simplifying the complexity of the classification problem. Thus, this paper proposes a new approach to effectively use semantic hierarchies as a framework for hierarchical image classification.

2. RELATED WORK

Image annotation has been considered in the last decade as a multi-class classification problem. To deal with a large number of concept categories, many approaches proposed to combine hierarchical structure with Support Vector Machines (SVM) classifiers [11, 8, 9, 6, 4]. These approaches can be qualified as top-down methods, i.e. the class hierarchy is built by recursive partitioning of the set of classes [8, 4, 7], or as bottom-up methods, i.e. the class hierarchy is built by agglomerative clustering of the classes [11, 9, 6, 2]. Two directions have been explored for hierarchical image classification: using Decision Directed Acyclic Graphs (DDAGs) [12, 11, 7], and using Binary Hierarchical Decision Trees (BHDTs) [8, 4]. Given $C = \langle c_1, c_2, \cdots, c_{\mathcal{N}} \rangle$ the annotation vocabulary of the database, the DDAG based approaches train $N(N-1)/2$ binary classifiers and use a DAG to decide about the belonging of an image i to a class $c_j \in C$. These methods allow at each node in a distance d from the rooted DAG to eliminate d candidate classes from C, resulting in a $N-1$ decision nodes to be evaluated for labeling a test sample. On the other side, BHDT based ap-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

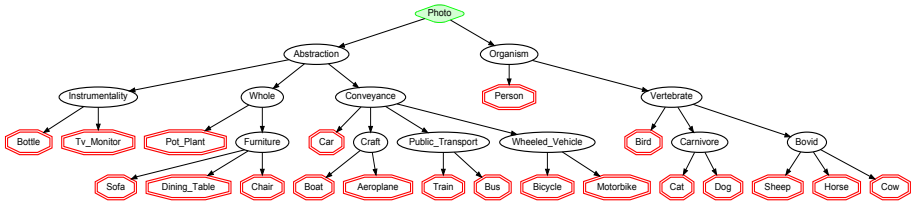


Figure 1: Built semantic hierarchy on VOC'2010 dataset. Double octagon nodes are original concepts.

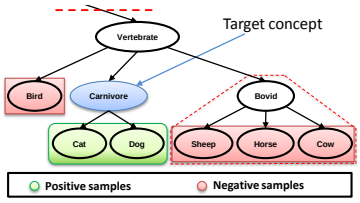


Figure 2: OVON hierarchical classifiers training.

proaches build and use hierarchies as binary trees, i.e. data are divided hierarchically into two subsets until each subset consists of only one class. Data partition is often achieved using a clustering algorithm. Thus, one SVM is trained for each node of the tree, resulting in a $\log_2 N$ SVM runs to label a test sample. BHDT approaches target to optimize the efficiency of SVM classifiers by reducing the unnecessary comparisons while maintaining a high classification accuracy [4]. However, BHDT and DDAG approaches focus on the classification optimization and do not model in anyway image semantics. Although these approaches allow increasing classification accuracy, they constrain hierarchies to binary structures resulting in a significant deadlock when the concepts number is large. For instance, the method of [11] is in a deadlock when the concept number exceeds 30, since intermediate concepts are extracted from WordNet using hypernymy relationships which depth is limited to 15 levels.

Alternative approaches have emerged recently and propose the use of semantic relationships between concepts for the building of hierarchies. Fan et al. [6] proposed to incorporate concept ontology and a multi-task learning algorithm for hierarchical concept learning. The labeling of a new sample is obtained by a voting procedure at all levels of the hierarchy, i.e. $|C + C'|$ SVM runs are necessary for labeling new images (C' is intermediate concept nodes in the hierarchy). In [5], a 'tree-max classifier' based on ImageNet hierarchy is proposed. A classifier at each node of the ImageNet tree is learned. The decision function is computed according to a target class and all its child nodes.

3. HIERARCHICAL CLASSIFICATION

In this paper we propose a new method for learning hierarchical classifiers. Our method relies on the structure of semantic hierarchies to train more accurate classifiers for image classification. Subsequently, we propose two methods for computing the decision function in order to hierarchical image classification. The first one is a *bottom-up* approach for hierarchical image classification by score fusion. Fusion is performed by the spreading of scores starting from leaf nodes until reaching the root node. The second is a *top-down* approach and is performed by classifiers voting. Starting from the root node and according to classifier votes, the hierarchy is traversed until reaching leaf nodes.

For the building of the semantic hierarchy we rely on our previous work [2], where we proposed to compute a semantic similarity between concepts in order to produce a hierarchy faithful to image semantics. This measure, named Semantico-Visual Relatedness of Concepts, integrates: 1) a visual similarity which represents the visual correspondence between concepts, 2) a conceptual similarity which defines a relatedness measure between target concepts based on their definitions in WordNet, and 3) a contextual similarity which measures the distributional similarity between each pair of concepts. The building of the hierarchy is based on a set of rules to link together concepts that are semantically related. Figure 1 illustrates the obtained hierarchy, which is a *N-ary tree* like structure where leaf nodes are initial concepts. Motivation for this method is that the semantics of images and data is much more complex than binary items.

3.1 Learning Hierarchical Classifiers

Based on the hierarchy structure, we propose in the following to train several classifiers that represent the same concept at different levels of abstraction. These classifiers are consistent with each other since they are linked by the subsumption relationship, and will represent the same information with different levels of details. Therefore, these classifiers results can be merged in order to achieve relevant decision on the membership of an image to a class. Concretely, given a class hierarchy, a classifier for each concept node is trained by performing a One-Versus-Opposite-Nodes (OVON) SVM. Indeed, in order to propose a method that scales well with large image databases, a good strategy would be to decompose the problem in several independent tasks based on the hierarchy structure. Thus, instead of considering all database images for training classifiers, we will consider only images of children nodes of a given target concept node. This is similar to cut a target node of a tree from its upper part and treats it independently. Therefore, for training a classifier of a target node, we took as positive samples all images associated with its children leaf nodes. So, if an image is annotated by 'cow' it will also serves to train the classifiers for 'Bovid', 'Vertebrate', etc. Negative samples are all images of sibling nodes - cf. Figure 2.

3.2 Bottom-Up Score Fusion (BUSF)

Starting from leaf concept nodes and following the subsumption relationships, we compute the average confidence scores of all paths in the hierarchy. The decision function is then computed according to the sign of this average score. A practical standpoint is that the classification results of these SVMs are independent. Therefore, it is also possible to run all SVM classifiers to compute the membership degree of an image to all classes. Subsequently, according to the hierarchy structure the decision function can be computed easily for all leaf concepts. Thus, the complexity for labeling a given image is $\leq (2N - 1)$. Let x_i^v be any visual representation of an image i, a classifier is trained for each concept class c_j in the hierarchy. $\mathcal{N} = |\mathcal{C}| + |\mathcal{C}'|$ binary SVM OVON are then used with a decision function $\mathcal{G}(x^v)$:

$$\mathcal{G}(x^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x^v) + b \qquad (1)$$

where $\mathbf{K}(x_i^v, x^v)$ is the value of a kernel function for the training sample x_i^v and the test sample x^v, $y_i \in \{1, -1\}$ the class label of x_i^v, α_i the learned weight of the training sample x_i^v, and b is a learned threshold parameter. RBF kernel is used for SVMs training $\mathbf{K}(x, y) = exp\Big(\frac{\|x-y\|^2}{\sigma^2}\Big)$.

The final decision function to compute the membership degree of an image z to a concept class c_j is:

$$\overline{f_{c_j}(z)} = sign\Big(\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\mathcal{G}_k(z^v)\Big) \quad (2)$$

where $\mathcal{S}$ is the set of subsumer of c_j. $\mathcal{G}_k(z^v)$ is the decision function of the classifier associated with concept k.

From a statistical standpoint, the final decision function $\overline{f_x(z)}$ is computed by achieving n measures of the same event ($n = |\mathcal{S}|$ is the hierarchy depth). Thus, the uncertainty about $\overline{f_x(z)}$ can be computed as the standard deviation $\sigma_{\overline{f_x(z)}} = \frac{\sigma}{\sqrt{n}}$. Therefore, the final decision function is $\sqrt{n}$ times more accurate than the one obtained from a single classifier.

3.3 Top-Down Classifiers Voting (TDCV)

TDCV aims at decomposing the image classification problem into several complementary sub-tasks. It consists in building several classifiers that are able to discriminate one class from the others under a given parent node. Thus, to reach the final decision about the class membership it is essential to descend the hierarchy according to the classifier decisions (votes).TDCV is efficient in terms of complexity since it requires to train less than $2N-1$ classifiers for hierarchical classification, and to evaluate less than $\log_2 N$ decision nodes for labeling a test image - cf. Table 1. However, TDCV is sensitive to the initial classification since classifiers at the subsequent levels cannot recover from the misclassification of a test image that may occur in a higher concept level. Thus, this misclassification can be propagated to the terminal node. Nevertheless, the average precision is strongly high for the nodes in highest levels of the hierarchy, and therefore errors propagation is small - cf. Figure 3.

Image classification is performed top-down as illustrated in Algorithm 1. Starting from the root node, the decision functions of subsequent level nodes are evaluated. The nodes with positive confidence value are recursively explored until reaching leaf nodes. Several paths in the hierarchy can be explored, and thus a test image can be associated to many classes. If a path is explored, but all the leaf classifiers have responded negatively, we keep the concept with higher confidence value.

4. EXPERIMENTAL RESULTS

We used the Bag-of-Features (BoF) representation to describe image features, which is a widely known method. The BoF model has shown excellent performances and became one of the most widely used techniques for image classification and object recognition. In our approach, image features are described as follows: Lowe's DoG Detector [10] is used for detecting a set of salient image regions. A signature of these regions is then computed using SIFT descriptor [10]. Afterwards, given the collection of detected region from the training set of all categories, we generate a codebook of size $K = 1000$ by performing the k-means algorithm. Thus, each detected region in an image is mapped to the most similar visual word in the codebook through a KD-Tree. Each image is then represented by a histogram of K visual words, where each bin in the histogram corresponds to the occurrence number of a visual word in that image.

Algorithm 1: Top-Down Classifiers Voting

Input: A test image, the semantic hierarchy
Result: Image annotation
begin
  $\Omega \leftarrow$ immediate children nodes of the hierarchy root
  while ($|\Omega| > 0$) **do**
    $\Upsilon \leftarrow 0$
    foreach ($k \in \Omega$) **do**
      if $\mathcal{G}_k(x^v) > 0$ **then**
        $\Upsilon \leftarrow \Upsilon + k$
    if $|\Upsilon| = 0$ **then**
      $\Upsilon \leftarrow \underset{j\in\Omega}{\text{argmax}}\ \mathcal{G}_j(x^v)$
    $\Omega \leftarrow$ immediate children of the nodes $\in \Upsilon$

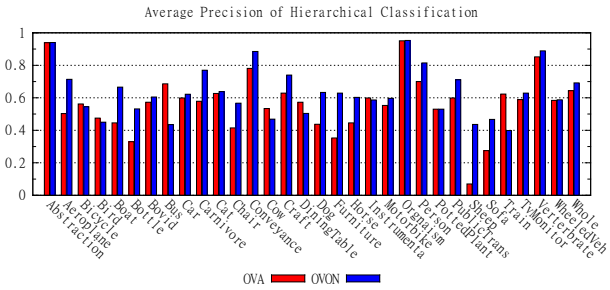
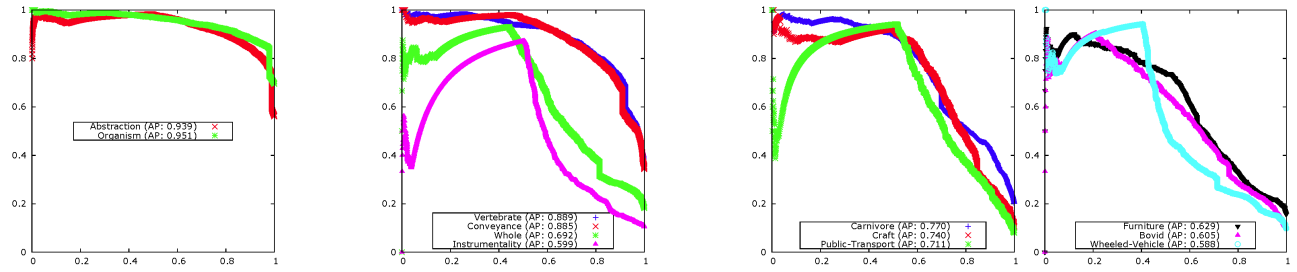


Figure 4: Comparison of the *OVON* and the *OVA* hierarchical classifiers on VOC'2010 dataset.

Experiments are performed on Pascal VOC'2010 dataset. We used 50% of the dataset images for training concept classifiers and the other images for evaluating the proposed approaches. To perform a fair comparison, we used the same visual representation of images for all of these methods, i.e. Bag-of-Features representation. The flat classification is performed by $|C|$ SVM OVA, where the inputs are the BoF representation of images and the outputs are the desired SVM responses for each image (1 or -1). We used cross-validation to overcome the unbalanced data problem, taking at each fold as many positive as negative images. Hierarchical classification with OVA classifiers is performed by training a set of $(|C+C'|)$ hierarchical classifiers consistent with the structure of the hierarchy in Figure 1. The baseline method is built by taking the average submission results to VOC'2010 challenge. In the following, evaluations are performed using the recall/precision curves and Average Precision score(AP).

In Figure 4, we compared our method (OVON) for training hierarchical classifiers to the One-Versus-All one. OVON performs a better result than the OVA classifiers, with an AP of 63.25% while 56.42% for the OVA hierarchical classification. In Figure 5, we compare our methods for hierarchical image classification to the other ones. Our methods achieve a higher AP than the flat classification with a gain of +26.8% for BUSF method and a gain of +16.04% for the TDCV method. Compared to the baseline method our approaches are slightly better. This can be explained by the efficient image features used in the submission of VOC challenge, and



(a) AP 2^{nd} layer concepts. (b) AP 3^{rd} layer concepts. (c) Average Precision for the fourth layer concepts.

Figure 3: Recall/Precision curves for concepts of each level of the hierarchy.

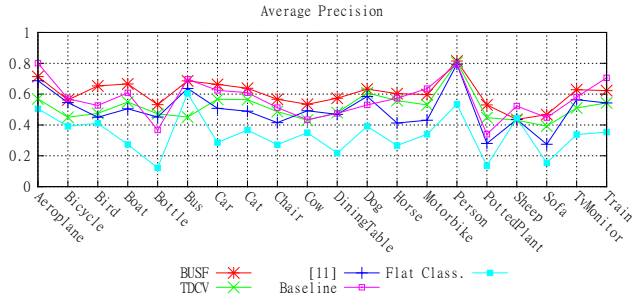


Figure 5: Obtained AP on VOC'2010 dataset for: BUSF, TDCV, [11], Baseline and flat classification.

	Training	Labeling	VOC'10 Training	VOC'10 Labeling
DDAG	$(N^2-N)/2$	$N-1$	190 t	19 t'
BHDT	$2N-1$	$\log_2 N$	39 t	5 t'
TDCV	$\leq 2N-1$	$\leq \log_2 N$	32 t	4 t'
BUSF	$\leq 2N-1$	$\leq 2N-1$	32 t	32 t'

Table 1: Complexity of our methods compared to DDAG and BHDT. t, t': stand for 1 unit of time.

the basic one used in our approach. Moreover, we used only the half of the training set since we do not dispose of the test set used in the challenge. Thus, the obtained results are still promising and can be improved by incorporating more sophisticated image descriptors. For comparison we implemented the method of [11], and we compared it to our methods for hierarchical image classification as illustrated in Figure 5. The BUSF method achieves a higher AP than the others with a gain of +8.99% compared to the TDCV method and a gain of +10.76% compared to the method proposed by [11]. The AP for these methods was as follows: 60.6% for the BUSF method, 51.61% for the TDCV method, and 49.84% for the method of [11].

As illustrated in Figure 3, the classifiers accuracy decreases as we go deeper in the hierarchy. This is because classes in higher level of the hierarchy are sufficiently visually different, i.e. it is easier to find a boundary that separates these classes. They are also more balanced. For instance, the ratio of positive/negative samples in VOC'2010 dataset is about 5%. OVON method allows overcoming this problem as it decomposes image classification into several sub-tasks. The ratio of positive/negative samples is 35.6% for OVON, i.e. these classes are quite balanced and there is no need for techniques as over-sampling or under-sampling to recover the problem of unbalanced data.

5. CONCLUSION

Hierarchical image classification is often considered as a binary classification problem. In this paper, we proposed a new hierarchical classification methodology, based on semantic hierarchies, which performs better on image annotation. Our approach is based on the hierarchy structure to efficiently train hierarchical classifiers, i.e. draws benefits from the hierarchies structure to decompose the image classification problem into several independent and complementary sub-tasks. We also proposed two methods for computing a hierarchical decision function serving to annotate images. The former is achieved by a top-down classifiers voting, while the second is based on a bottom-up score fusion. Compared to existing approaches, our methods achieve higher accuracy on Pascal VOC'2010 dataset.

6. REFERENCES

[1] H. Bannour and C. Hudelot. Towards ontologies for image interpretation and annotation. In *CBMI*, 2011.

[2] H. Bannour and C. Hudelot. Building semantic hierarchies faithful to image semantics. In *MMM*. 2012.

[3] K. Barnard, P. Duygulu, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 2003.

[4] H. Cevikalp. New clustering algorithms for the support vector machine based hierarchical classification. *Pattern Recognition Letters*, 31(11):1285 – 1291, 2010.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE IP.*, 17(3):407–426, 2008.

[7] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[8] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *CVPR*, 2008.

[9] L.-J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *CVPR*, 2010.

[10] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[11] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007.

[12] J. C. Platt, N. Cristianini, and J. Shawe-taylor. Large margin dag for multiclass classification. In *NIPS*, 2000.