Detection of Visual Concepts and Annotation of Images Using Ensembles of Trees for Hierarchical Multi-Label Classification

Ivica Dimitrovski[1,2], Dragi Kocev[1], Suzana Loskovska[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
[2] Department of Computer Science, Faculty of Electrical Engineering and Information Technology
Karpoš bb, 1000 Skopje, Macedonia
ivicad@feit.ukim.edu.mk, Dragi.Kocev@ijs.si, suze@feit.ukim.edu.mk, Saso.Dzeroski@ijs.si

Abstract. In this paper, we present a hierarchical multi-label classification system for visual concepts detection and image annotation. Hierarchical multi-label classification (HMLC) is a variant of classification where an instance may belong to multiple classes at the same time and these classes/labels are organized in a hierarchy. The system is composed of two parts: feature extraction and classification/annotation. The feature extraction part provides global and local descriptions of the images. These descriptions are then used to learn a classifier and to annotate an image with the corresponding concepts. To this end, we use predictive clustering trees (PCTs), which are able to classify target concepts that are organized in a hierarchy. Our approach to HMLC exploits the annotation hierarchy by building a single predictive clustering tree that can simultaneously predict all of the labels used to annotate an image. Moreover, we constructed ensembles (random forests) of PCTs, to improve the predictive performance. We tested our system on the image database from the ImageCLEF@ICPR 2010 photo annotation task. The extensive experiments conducted on the benchmark database show that our system has very high predictive performance and can be easily scaled to large number of visual concepts and large amounts of data.

1 Introduction

An ever increasing amount of visual information is becoming available in digital form in various digital archives. The value of the information obtained from an image depends on how easily it can be found, retrieved, accessed, filtered and managed. Therefore, tools for efficient archiving, browsing, searching and annotation of images are a necessity.

A straightforward approach, used in some existing information retrieval tools for visual materials, is to manually annotate the images by keywords and then

D. Ünay, Z. Çataltepe, and S. Aksoy (Eds.): ICPR 2010, LNCS 6388, pp. 152–161, 2010.

to apply text-based query for retrieval. However, manual image annotation is an expensive and time-consuming task, especially given the large and constantly growing size of image databases.

The image search provided by major search engines, such as Google, Bing, Yahoo! and AltaVista, relies on textual or metadata descriptions of images found on the web pages containing the images and the file names of the images. The results from these search engines are very disappointing when the visual content of the images is not mentioned, or properly reflected, in the associated text.

A more sophisticated approach to image retrieval is automatic image annotation: a computer system assigns metadata in the form of captions or keywords to a digital image [5]. These annotations reflect the visual concepts that are present in the image. This approach begins with the extraction of feature vectors (descriptions) from the images. A machine learning algorithm is then used to learn a classifier, which will then classify/annotate new and unseen images.

Most of the systems for detection of visual concepts learn a separate model for each visual concept [7]. However, the number of visual concepts can be large and there can be mutual connections between the concepts that can be exploited. An image may have different meanings or contain different concepts: if these are organized into a hierarchy (see Fig. 2), hierarchical multi-label classification (HMLC) can be used for obtaining annotations (i.e., labels for the multiple visual concepts present in the image) [7]. The goal of HMLC is to assign to each image multiple labels, which are a subset of a previously defined set (hierarchy) of labels.

In this paper, we present a system for detection of visual concepts and annotation of images, which exploits the semantic knowledge about the inter-class relationships among the image labels organized in hierarchical structure. For the annotation of the images, we propose to exploit the annotation hierarchy in image annotation by using predictive clustering trees (PCTs) for HMLC. PCTs are able to handle target concepts that are organized in a hierarchy, i.e., to perform HMLC. To improve the predictive performance, we use ensembles (random forests) of PCTs for HMLC. For the extraction of features, we use several techniques that are recommended as most suitable for the type of images at hand [7].

We tested the proposed approaches on the image database from the ICPR 2010 photo annotation task [9]. The concepts used in this annotation task are from the personal photo album domain and they are structured in an ontology. Fig. 2 shows a part of the hierarchical organization of the target concepts.

The remainder of this paper is organized as follows. Section 2 presents the proposed large scale visual concept detection system. Section 3 explains the experimental design. Section 4 reports the obtained results. Conclusions and a summary are given in Section 5.

2 System for Detection of Visual Concepts

2.1 Overall Architecture

Fig. 1 presents the architecture of the proposed system for visual concepts detection and image annotation. The system is composed of a feature extraction part and a classification/annotation part. We use two different sets of features to describe the images: visual features extracted from the image pixel values and features extracted from the exchangeable image file format (EXIF) metadata files. We employ different sampling strategies and different spatial pyramids to extract the visual features (both global and local) [4].

As an output of the feature extraction part, we obtain several sets of descriptors of the image content that can be used to learn a classifier to annotate the images with the visual concepts. First, we learn a classifier for each set of descriptors separately. The classifier outputs the probabilities with which an image is annotated with the given visual concepts. To obtain a final prediction, we combine the probabilities output from the classifiers for the different descriptors by averaging them. Depending on the domain, different weights can be used for the predictions of the different descriptors.

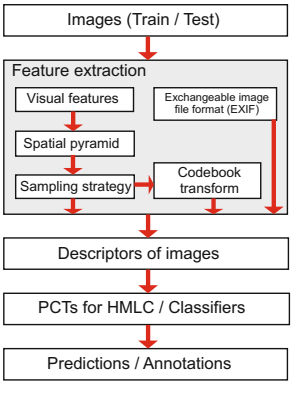

Fig. 1. Architecture of the proposed system for detection of visual concepts and annotation of images

2.2 The Task of HMLC

Hierarchical multi-label classification is a variant of classification were a single example may belong to multiple classes at the same time and these classes are organized in a hierarchy. An example that belongs to some class automatically belongs to all its super-classes, as implied by the hierarchical constraint. Example problems of this kind can be found in several domains including text classification, functional genomics, and object and scene classification. For more detail overview of the possible application areas we refer the reader to [11].

The predefined set of labels can be organized in a semantic hierarchy (see Fig. 2 for an example). Each image is represented with: (1) a set of descriptors (in this example, the descriptors are histograms of five types of edges encountered in the image) and (2) labels/annotations. A single image can be annotated with multiple labels at different levels of the predefined hierarchy. For example, the image in the third row in the Table from Fig. 2 is labeled with clouds and sea. Note that this image is also labeled with the labels: sky, water and landscape because these labels are in the upper levels of the hierarchy.

The data, as presented in the Table from Fig. 2, are used by a machine learning algorithm to train a classifier. The testing set of images contains only the set of descriptors and has no *a priori* annotations.

image	features/descriptors						annotations/labels
	|	—	/	\	S	...	
	48	24	59	66	37	...	flowers
	36	25	53	45	15	...	mountains@sky@plants
	35	25	56	52	19		clouds@sea
...	...	...	...	...	...	...	...

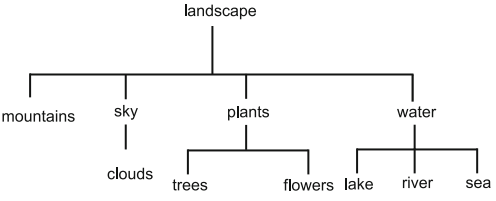

Fig. 2. A fragment of the hierarchy for image annotation. The annotations are part of the hierarchical classification scheme for the ICPR 2010 photo annotation task (right).The table contains set of images with their visual descriptors and annotations (left).

2.3 Ensembles of PCTs for HMLC

In the PCT framework [1], a tree is viewed as a hierarchy of clusters: the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. Note that the hierarchical structure of the PCT does not necessary reflect the hierarchical structure of the annotations.

PCTs are constructed with a standard "top-down induction of decision trees" (TDIDT) algorithm. The heuristic for selecting the tests is the reduction in variance caused by partitioning the instances, where the variance $Var(S)$ is defined by equation (1) below. Maximizing the variance reduction maximizes cluster homogeneity and improves predictive performance.

A leaf of a PCT is labeled with/predicts the prototype of the set of examples belonging to it. With appropriate variance and prototype functions, PCTs can handle different types of data, e.g., multiple targets [3] or time series [12]. A detailed description of the PCT framework can be found in [1]. The PCT framework is implemented in the CLUS system, which is available for download at `http://www.cs.kuleuven.be/~dtai/clus`.

To apply PCTs to the task of HMLC, the example labels are represented as vectors with Boolean components. Components in the vector correspond to

labels in the hierarchy traversed in a depth-first manner. The i-th component of the vector is 1 if the example belongs to class c_i and 0 otherwise. If $v_i = 1$, then $v_j = 1$ for all v_j's on the path from the root to v_i.

The variance of a set of examples (S) is defined as the average squared distance between each example's label v_i and the mean label $\bar{v}$ of the set, i.e.,

$$Var(S) = \frac{\sum_i d(v_i, \bar{v})^2}{|S|} \tag{1}$$

We consider the higher levels of the hierarchy more important: an error at the upper levels costs more than an error at the lower levels. Considering this, a weighted Euclidean distance is used:

$$d(v_1, v_2) = \sqrt{\sum_i w(c_i)(v_{1,i} - v_{2,i})^2} \tag{2}$$

where $v_{k,i}$ is the i'th component of the class vector v_k of an instance x_k, and the class weights $w(c_i)$. The class weights decrease with the depth of the class in the hierarchy, $w(c_i) = w_0 \cdot w(c_j)$, where c_j is the parent of c_i and $0 < w_0 < 1$.

Each leaf in the tree stores the mean $\bar{v}$ of the vectors of the examples that are sorted in that leaf. Each component of $\bar{v}$ is the proportion of examples $\bar{v_i}$ in the leaf that belong to class c_i. An example arriving in the leaf can be predicted to belong to class c_i if $\bar{v_i}$ is above some threshold t_i. The threshold can be chosen by a domain expert.

For a detailed description of PCTs for HMLC the reader is referred to [15]. Next, we explain how PCTs are used in the context of an ensemble classifier, namely ensembles further improve the performance of PCTs.

Random Forests of PCTs. To improve the predictive performance of PCTs, we use ensemble methods. An ensemble classifier is a set of classifiers. Each new example is classified by combining the predictions of each classifier from the ensemble. These predictions can be combined by taking the average (for regression tasks) or the majority vote (for classification tasks) [2]. In our case, the predictions in a leaf are the proportions of examples of different classes that belong to it. We use averaging to combine the predictions of the different trees. As for the base classifiers, a threshold should be specified to make a prediction.

We use random forests as an ensemble learning technique. A random forest [2] is an ensemble of trees, obtained both by bootstrap sampling, and by randomly changing the feature set during learning. More precisely, at each node in the decision tree, a random subset of the input attributes is taken, and the best feature is selected from this subset (instead of the set of all attributes). The number of attributes that are retained is given by a function f of the total number of input attributes x (e.g., $f(x) = x, f(x) = \sqrt{x}, f(x) = \lfloor \log_2 x \rfloor + 1$).

2.4 Feature Extraction

We use different commonly used types of techniques for feature extraction from images. We employ three types of global image descriptors: gist features [10],

local binary patterns (LBP) [13] and a color histogram, with 8 bins in each color channel for the RGB color space. The LBP operator is computed in a spatial arrangement where the image is split into 4x4 sub-regions.

Local features include scale-invariant feature transforms (SIFT) extracted densely on a multi-scale grid or around salient points obtained from a Harris-Laplace detector [6]. The dense sampling and Harris-Laplace detectors give an equal weight to all key-points, independent of their spatial location in the image. To overcome this limitation, one can use spatial pyramids of 1x1, 2x2 and 1x3 regions [14].

We computed six different sets of SIFT descriptors over the following color spaces: RGB, opponent, normalized opponent, gray, HUE and HSV. For each set of SIFT descriptors, we use the codebook approach to avoid using all visual features of an image [14].

The generation of the codebook begins by randomly sampling 50 key-points from each image and extracting SIFT descriptors in each key-point (i.e., each key-point is described by a vector of numerical values). Then, to create the codewords, we employ k-means clustering on the set of all key-points. We set the number of clusters to 4000, thus we define a codebook with 4000 codewords (a codeword corresponds to a single cluster and a codebook to the set of all clusters). Afterwards, we assign the key-points to the discrete codewords predefined in the codebook and obtain a histogram of the occurring visual features. This histogram will contain 4000 bins, one for each codeword. To be independent of the total number of key-points in an image, the histogram bins are normalized to sum to 1.

The number of key-points and codewords (clusters) are user defined parameters for the system. The values used above (50 key-points and 4000 codewords) are recommended for general images [14].

An image can have an associated text file with metadata information in EXIF (EXchangeable Image File) format [16]. The metadata can be used to construct features that describe certain aspects of the imaging technique and the technical specification of the used camera. These describe, for example the image quality (resolution, focal length, exposure time) and when the picture was taken.

3 Experimental Design

3.1 Definition and Parameter Settings

We evaluated our system on the image database from the ImageCLEF@ICPR 2010 photo annotation task. The image database consists of training (5000), validation (3000) and test (10000) images. The images are labeled with 53 visual concepts organized in a tree-like hierarchy [9]. The goal of the task is to predict which of the visual concepts are present in each of the testing images.

We generated 15 sets of visual descriptors for the images: 12 sets of SIFT local descriptors (2 detectors, Harris-Laplace and dense sampling, over 6 different color spaces) with 32000 bins for each set (8 sub-images, from the spatial pyramids: 1x1, 2x2 and 1x3, 4000 bins each). We also generated 3 sets of global descriptors

(LBP histogram with 944 bins, gist features with 960 bins and RGB color histogram with 512 bins). From the EXIF metadata, we selected the most common tags as features, such as: software, exposure time, date and time (original), exposure bias, metering mode, focal length, pixelXDimension, pixelY-Dimension etc. Since the PCTs can handle missing values, these values for the images without EXIF tags were set to 'unknown' or '?'.

The parameter values for the random forests were as follows: we used 100 base classifiers and the size of the feature subset was set to 10% of the number of descriptive attributes. The weights for the PCTs for the HMLC (w_0) were set to 1: each of the classes from the hierarchy has equal influence on the heuristic score.

3.2 Performance Measures

The evaluation of the results is done using three measures of performance suggested by the organizers of the challenge [3]: area under the ROC curve (AUC), equal error rate (EER) and average ontology score (AOS). The first two scores evaluate the performance for each visual concept, while the third evaluates the performance for each testing image.

The ROC curve is widely used evaluation measure (see Fig. 3). It plots the true positive rate (TPR) vs. false positive rate (FPR). The area between the curve and the axis with FPR (AUC) is the probability that a randomly chosen positive example will be ranked higher than a randomly chosen negative example. The EER is the threshold value at which the TPR and FPR are equal. Hence, the EER balances the probability of error with the probability of false rejection. Lower EER means better predictive performance. The hierarchical AOS measure calculates the misclassification cost for each missing or wrongly annotated concept per image. The AOS score is based on structure information (distance between concepts in the hierarchy), relationships from the ontology and the agreement between annotators for a concept [8].

4 Results and Discussion

We present results from two different experiments (see Table 1). In the first experiment, we use just the training images for learning the classifier. For the second experiment, we merge the training and validation set into a single dataset which we then use to learn the classifier. The results show that by using both datasets (training and validation together) we get better scores.

If we focus on the prediction scores for the individual visual concepts, we can note that we predict best the presence of landscape elements (see Table 2); the best predicted concept is 'Sunrise or Sunset' (from the parent-concept 'Time of day'). The worst predicted concepts are from the 'Aesthetics' group of concepts ('Aesthetic Impression', 'Overall quality' and 'Fancy'). But, this is to be expected because the agreement of human annotators on these concepts is only about 75% [8].

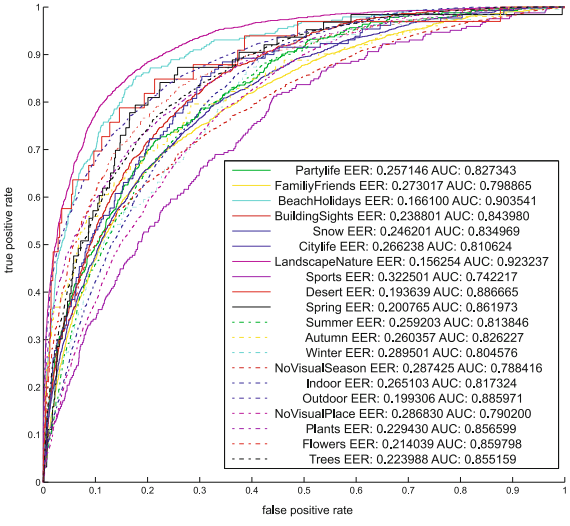

Fig. 3. ROC curves for a subset of the visual concepts

The system has low predictive power when we are predicting the absence of a concept (e.g., 'No persons', 'No visual season' ...). The hierarchy should not include these concepts. These concepts should be assigned after post-processing the results. We also have to predict mutually exclusive concepts (for example: Indoor, Outdoor and No visual place). The notation of HMLC, however, does not account for mutually exclusive concepts. To solve this issues one must re-engineer the hierarchy of the concepts.

Further improvements can be expected if different weighting schemes are used (to combine the predictions of the various descriptors). For instance, the SIFT descriptors are invariant to color changes, and they do not predict well concepts where illumination is important. Thus, the weight of the SIFT descriptors in the combined predictions for those concepts should be decreased.

Let us compare the results of our system with the results from the other participating groups at the ImageCLEF@ICPR 2010 photo annotation task. Our system ranks second by the hierarchical AOS score. By the EER and AUC score it ranks third. Thus, relatively speaking, it performs better under the hierarchical performance measure.

Table 1. Results of the experiments evaluated using Equal Error Rate, Area under Curve and Average Ontology Score

	EER	AUC	AOS
Train and Validation	0.242	0.832	0.706
Train	0.250	0.821	0.703

Table 2. Results per concept for our best run in the Large-Scale Visual Concept Detection Task using the Area Under the Curve. The concepts are ordered by their highest score.

Concept	AUC	Concept	AUC	Concept	AUC
Sunset-Sunrise	0.951	Trees	0.855	Citylife	0.811
Clouds	0.946	Day	0.853	Winter	0.805
Sea	0.939	Portrait	0.849	Out-of-focus	0.803
Sky	0.933	Partly-Blurred	0.848	Animals	0.803
Landscape-Nature	0.923	Building-Sights	0.844	Familiy-Friends	0.799
Night	0.923	No-Visual-Time	0.840	Sunny	0.799
Mountains	0.919	Snow	0.835	No-Persons	0.794
Beach-Holidays	0.904	No-Blur	0.829	Vehicle	0.791
Lake	0.900	Partylife	0.827	No-Visual-Place	0.790
River	0.891	Autumn	0.826	No-Visual-Season	0.788
Food	0.890	Canvas	0.825	Motion-Blur	0.779
Desert	0.887	Indoor	0.817	Single-Person	0.761
Outdoor	0.886	Still-Life	0.817	Small-Group	0.752
Water	0.877	Macro	0.816	Sports	0.742
Underexposed	0.876	Summer	0.814	Aesthetic-Impression	0.661
Spring	0.862	Overexposed	0.812	Overall-Quality	0.657
Flowers	0.860	Big-Group	0.811	Fancy	0.613
Plants	0.857	Neutral-Illumination	0.811	Average	0.832

5 Conclusion

Hierarchical multi-label classification (HMLC) problems are encountered increasingly often in image annotation. However, flat classification machine learning approaches are predominantly applied in this area. In this paper, we propose to exploit the annotation hierarchy in image annotation by using ensembles of trees for HMLC. Our approach to HMLC exploits the annotation hierarchy by building a single classifier that simultaneously predicts all of the labels in the hierarchy.

Applied on the ImageCLEF@ICPR 2010 photo annotation benchmark task our approach was ranked second for the hierarchical performance measure and third for the equal error rate and area the under the curve, out of 12 competing groups. The results were worst for predicting the absence of concepts. This suggests the need for re-engineering the hierarchy or for post processing the predictions to appropriately handle such concepts.

The system we presented is general. It can be easily extended with new feature extraction methods, and it can thus be easily applied to other domains, types of images and other classification schemes. In addition, it can handle arbitrarily sized hierarchies organized as trees or directed acyclic graphs.

References

1. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Proc. of the 15th ICML, pp. 55–63 (1998)
2. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
3. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Ensembles of Multi-Objective Decision Trees. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 624–631. Springer, Heidelberg (2007)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
5. Li, J., Wang, J.Z.: Real-Time Computerized Annotation of Pictures. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(6), 985–1002 (2008)
6. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
7. Nowak, S., Dunker, P.: Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 94–109. Springer, Heidelberg (2010)
8. Nowak, S., Lukashevich, H.: Multilabel classification evaluation using ontology information. In: Workshop on IRMLeS, Heraklion, Greece (2009)
9. Nowak, S.: ImageCLEF@ICPR Contest: Challenges, Methodologies and Results of the PhotoAnnotation Task. In: Ünay, D., Çataltepe, Z., Aksoy, S. (eds.) ICPR 2010. LNCS, vol. 6388, pp. 140–153. Springer, Heidelberg (2010)
10. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
11. Silla, C., Freitas, A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery (in press, 2010)
12. Slavkov, I., Gjorgjioski, V., Struyf, J., Džeroski, S.: Finding explained groups of time-course gene expression profiles with predictive clustering trees. Molecular BioSystems 6(4), 729–740 (2010)
13. Takala, V., Ahonen, T., Pietikainen, M.: Block-Based Methods for Image Retrieval Using Local Binary Patterns. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.) SCIA 2005. LNCS, vol. 3540, pp. 882–891. Springer, Heidelberg (2005)
14. Van de Sande, K., Gevers, T., Snoek., C.: A comparison of color features for visual concept classification. In: CIVR, pp. 141–150 (2008)
15. Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning 73(2), 185–214 (2008)
16. Exchangeable image file format, `http://en.wikipedia.org/wiki/EXIF`