Multimedia autoannotation via hierarchical semantic ensembles

Serhiy Koisnov and Stéphane Marchand-Maillet
Computer Vision and Multimedia Lab, University of Geneva
24 rue du General-Dufour, Geneva, Switzerland
{kosinov,marchand}@cui.unige.ch

Abstract

This paper presents a hierarchical semantic ensemble learning method applied in the context of multimedia autoannotation. In contrast to the standard multiple-category classification setting that assumes independent, non-overlapping and exhaustive set of categories, the proposed approach models explicitly the hierarchical relationships among target classes and estimates their relevance to a query as a trade-off between the goodness of fit to a given category description and its inherent uncertainty. The promising results of the empirical evaluation confirm the viability of the proposed approach, validated in comparison to several techniques of ensemble learning, as well as with different type of baseline classifiers.

1. Introduction

One of the essential challenges in modern information retrieval is to be able to deduce high-level semantics from the low-level perceptual features of multimedia, which the literature sources refer to as the semantic categorization, keyword prediction, autoannotation or automatic linguistic indexing task. The diversity in the problem terminology reflects the variety of contributions from numerous research domains that have been proposed to date. For example, an appealing idea of treating the visual feature data as another language to translate semantic keywords to and from is developed with the aid of generative probabilistic models by Barnard *et al.* [1, 2]. A family of methods [18, 20, 27], related to the cross-language extension of the latent semantic indexing (LSI) technique [5, 13], permit the retrieval of multimedia semantics via low-level feature queries. Yet, the majority of the other approaches consider the multimedia autoannotation problem in the multiple-category classification framework, where unseen documents must be assigned to one or more predefined semantic categories. In [8], for instance, the authors focus on improving several popular ensemble schemes, such as OPC (one per class), PWC (pair-wise coupling) and ECOC (error-correcting output codes). The methods developed in [3, 14, 15] decompose a multiple-category classification task into a collection of binary clasification problems and propose ways of recombining effectively the individual predictions from classifiers as diverse as SVM, BPM, 2D-MHMM. The semantic categories for these and many other classification-based techniques are generally assumed to be independent, non-overlapping and sufficient to cover all of the problem domain.

The approach presented in this paper is also formulated as a classification-based method, but differs from the above work in the important respect that the relationships among the semantic categories derived from the individual keywords of the annotation corpora are explicitly modeled in Bayesian terms, leading to a more consistent autoannotation performance. Furthermore, the proposed method broadens the range of the derived annotation allowing to predict more general notions or semantically-related keyword groups in addition to individual keywords present in the training data vocabulary. Another benefit of the proposed formulation is that it gives an answer to such an important question as how many keywords the system should predict and whether it is reasonable to predict anything at all.

The remainder of this paper is organized as follows. Section 2 presents the problem formulation focused on the autoannotation of digital images as a particular form of multimedia documents, followed by an illustrative example of the proposed method, given in Section 3. An overview of the baseline classifiers used as basic building blocks of the hierarchical ensemble, as well as the technique of fitting posterior probabilities to their raw outputs are discussed in Section 4. Section 5 details the adopted low-level image feature representation, while the experimental results and concluding remarks are provided in Sections 6 and 7, respectively.

2. Problem formulation

We employ a hierarchical ensemble of binary classifiers in order to perform semantic annotation of unseen im-

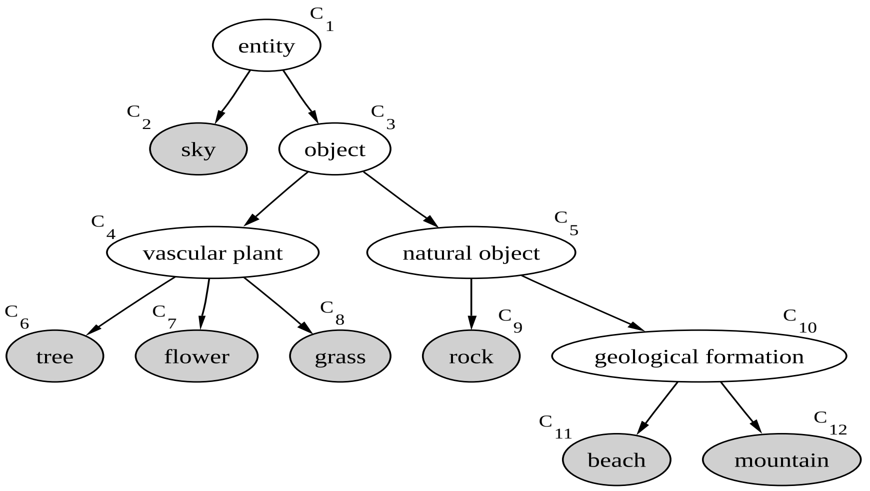


Figure 1. Classifier hierarchy example. Shaded nodes denote $C_i \in V$

ages. Given a training set of annotated images $X^{(T)} = \{I_t, K_t\}_{t=1}^{n}$, where I_t and K_t represent the feature vector of a given image and its associated set of keywords, respectively, the concept hierarchy $H = \{C_i\}_{i=1}^{N}$ is defined by all of the unique nouns comprising the annotation vocabulary $V = \bigcup_{t=1}^{n} K_t$ and their hyponyms derived from WordNet [17]. Every concept C_i occupies a separate node in H, and is associated with a binary classifier Φ_i designed to distinguish the set of leaf concepts subsumed (directly or indirectly) by C_i, denoted as $\mathbf{L}(C_i)$, from all of the others. An example of a hierarchy derived for a simple vocabulary V:{*beach, flower, grass, mountain, rock, sky, tree*} is shown in Figure 1.

In order to perform the autoannotation of an unseen image represented by a low-level feature vector I_U, each concept C_i is assessed as a potential candidate. Thus, the set of possible annotations is no longer restricted to be V, as is the case for the majority of other similar techniques. The relevance of C_i is seen as a trade-off between, on one hand, how well the input data I_U fits the description of C_i from the classification accuracy point of view, and, on the other hand, how specific or non-ambiguous the candidate set of keywords $\mathbf{L}(C_i)$ is. In our method, the first of these two quantites is represented by the posterior probability of a concept given the data, $P(C_i|I_U)$, while the second one is estimated as the posterior probability of a concept given the assumption that a particular keyword k from the set of all homonyms of C_i is chosen correctly, denoted as $P(C_i|k)$.

For a given concept C_i, the estimate of $P(C_i|I_U)$ is determined according to the following theorem, which is a reformulation of a previously established result described in [12]:

Theorem 1, (Kumar *et al.***, 2002).** *The posterior probability $P(C_i|I_U)$ for any input I_U is the product of the posterior probabilities of all the internal classifiers along a unique path from the root node to C_i, i.e.*

$$P(C_i|I_U) = \prod_{l=0}^{\mathscr{D}(C_i)-1} P(C_i^{(l+1)}|I_U, C_i^{(l)}), \qquad (1)$$

where $\mathscr{D}(C_i)$ is the depth of C_i (the depth of the root concept C_1 is 0), $C_i^{(l)}$ is the concept at depth l on the path from the root node to C_i, such that $C_i^{(\mathscr{D}(C_i))} \equiv C_i$ and $C_i^{(0)} \equiv C_1$.

In order to ensure that (1) is applicable in the case of classifiers with non-probabilistic outputs, such as SVM [4], a sigmoid function is fit to the raw classifier output values f_i, as described in detail in Section 4. As for $P(C_i|k)$, the Bayes theorem allows to express this quantity in terms of statistics of the training data as shown in (2):

$$P(C_i|k) = \frac{P(k|C_i)P(C_i)}{\sum_{C_i \in H} P(k|C_i)P(C_i)}, \qquad (2)$$

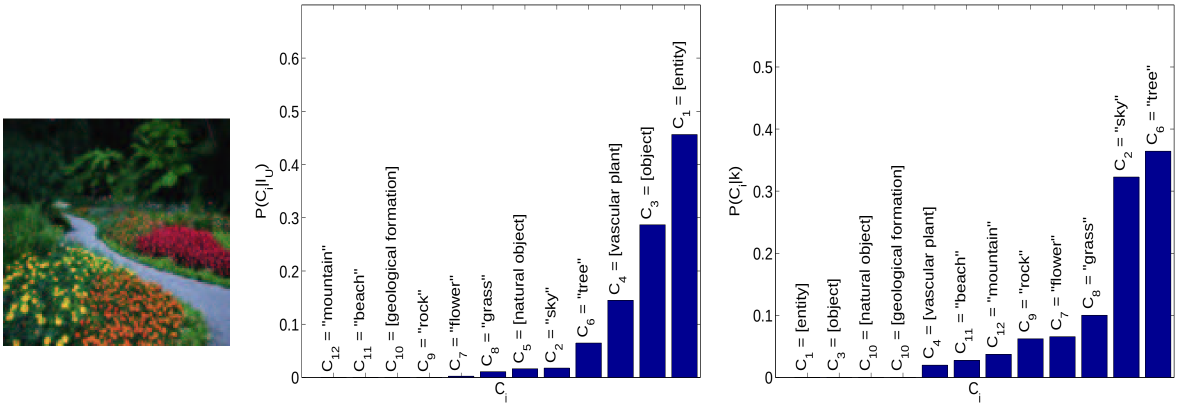
where $P(C_i)$, a prior probability of concept C_i, is estimated from the training data as:

$$P(C_i) = \frac{\sum_{C \in \mathbf{L}(C_i)} freq^{(T)}(C)}{\sum_{C \in V} freq^{(T)}(C)}, \qquad (3)$$

and $P(k|C_i)$, the worst-case estimate of the probability of choosing a correct annotation keyword k given the degree of generality of concept C_i, is deduced from the homonym set cardinality information derived from WordNet:

$$P(k|C_i) = \frac{\min_{C \in \mathbf{L}(C_i)} freq^{(W)}(C)}{freq^{(W)}(C_i)}. \qquad (4)$$

In (3) and (4), the frequency of a given concept in the training data and the cardinality of the WordNet homonym set are denoted as $freq^{(T)}$ and $freq^{(W)}$, respectively.



(a) Query I_U

(b) Goodness of fit $P(C_i|I_U)$

(c) Specificity $P(C_i|k)$

Figure 2. Individual contributions of factors $P(C_i|I_U)$ and $P(C_i|k)$

Finally, assuming that the likelihood of the input data I_U given C_i is not dependent on the correctness of a particular choice of k from the homonym set of C_i, we obtain the following result:

$$P(C_i|I_U, k) \propto P(C_i|I_U)P(C_i|k) = \rho, \qquad (5)$$

which essentially represents a means of comparison of different hypothesis concepts $\{C_i\}$ that takes into account both the goodness of fit of the data I_U to a given concept description and the concept's inherent degree of uncertainty or specificity. The next section illustrates these notions.

3. Illustrative example

Let us come back to the simplified 12-concept classifier hierarchy given in Figure 1. To be able to observe the effect of each of the two factors contributing to the final estimate of the concept relevance, ρ, we plot separately the computed values of $P(C_i|k)$, Figure 2(c), and $P(C_i|I_U)$, Figure 2(b)[1] , for a sample test image query depicted in Figure 2(a). As the diagrams show, there is a natural tendency among the values of $P(C_i|I_U)$ to favor simpler, more general concepts, such as *object*, due to the smaller number of terms to be evaluated in product (1). Quite the opposite trend is noticeable among the estimates of $P(C_i|k)$ that tend to promote very specific, unambiguous concepts, such as *sky*, taking into account their prior probabilities as well. This very trade-off of "Goodness of fit vs. Specificity" is captured by the concept relevance, ρ, leading to the results listed in Table 1 that demonstrate a reasonable degree of coherence between the top ranking concepts C_i and the true keywords of the query $K_U = \{$*flowers, path, grass, trees*$\}$.

Table 1. Candidate concepts ranked by relevance

Rank	$-\log_2 \rho(C_i)$	Concept C_i
1	5.41	C_6 = tree
2	7.46	C_2 = sky
3	8.44	C_4 = vascular plant
4	9.84	C_8 = grass
5	12.64	C_7 = flower
6	**17.26**	$\mathbf{C_1}$ **= entity**
7	17.87	C_3 = object
8	19.42	C_5 = natural object
9	21.00	C_9 = rock
10	44.32	C_{10} = geological formation
11	55.97	C_{12} = mountain
12	56.35	C_{11} = beach

Another important property of the proposed method that the figures from Table 1 help highlight is its ability to determine exactly how many of the top-ranked concepts should be predicted. Many existing approaches [1, 2, 18] resolve this issue by specifying a tunable "refuse-to-predict" parameter that regulates the propensity of image regions to emit concepts or, as some other techniques, by simply considering a fixed number of top-ranked entries. In our case, the relevance of the root node, $\rho_1 = \rho(C_1)$, provides a natural threshold that determines the number of candidate annotation concepts to be selected. An intuitive interpretation of the negative logarithm of this quantity comes from the

[1] One may note that $P(C_1|I_U) \neq 1$ as shown in the figure, contrary to what (1) may imply. This is explained by our use of the global prior of the root concept *entity* computed from overall WordNet statistics.

minimum message length (MML) principle of information theory [26], which interprets $-\log_2 \rho_1$ as the null-model hypothesis test that corresponds to transmitting all the data, since the root concept subsumes all of the other concepts, as is. According to the MML principle, any hypothesis that cannot better the null-model is not acceptable. In our example, this assertion makes us discard all of the candidate concepts ranked 6 or worse (see Table 1).

4. Probabilistic outputs for baseline classifiers

Having discussed the general formulation of the proposed method, we now turn our attention to the basic building blocks of the ensemble, namely, the individual binary classifiers. The main criteria for selecting baseline classifiers Φ_i for each candidate concept C_i were superiority in performance and suitability for the task. Thus, we chose two main types of classifiers: support vector machines [4, 24] due to their exceptional performance record and a great degree of flexibility with various types of kernels, and the recently developed transformational approach of distance-based discriminant analysis [10, 11] that demonstrated very competitive results specifically on the problem of visual object categorization.

The first of the two techniques, support vector machines (SVM), produces an uncalibrated output defined as:

$$f(\mathbf{x}) = h(\mathbf{x}) + b, \qquad (6)$$

where

$$h(\mathbf{x}) = \sum_i y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) \qquad (7)$$

lies in a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{F}$ induced by a kernel k [25]. Training an SVM minimizes an approximation to the training misclassification rate plus a penalty term corresponding to the norm of h in the RHKS:

$$C \sum_i (1 - y_i f_i)_+ + \frac{1}{2} ||h||_{\mathscr{F}} \qquad (8)$$

where $f_i = f(\mathbf{x}_i)$, which also corresponds to the minimization of a bound on the test misclassification rate [24]. The classification decision is made based on the sign of the raw output $f(\mathbf{x})$.

The other method, distance-based discriminant analysis (DDA), finds a data transformation $T \in \mathbb{R}^{m \times n}$ as a solution to the problem of minimization of the following criterion:

$$J(T) = \frac{\left(\prod_{i<j}^{N_X} \Psi\left(d_{ij}^{W}(T)\right) \right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X} \prod_{j=1}^{N_Y} d_{ij}^{B}(T) \right)^{\frac{1}{N_X N_Y}}}, \qquad (9)$$

where the numerator and denominator of (9) represent the geometric means of the within- and between-class distances, and $\Psi(\cdot)$ denotes a Huber robust estimation function [9]. The main advantages of the approach are its non-parametric nature, asymmetric class treatment specifically tailored for unbalanced data sets and the ability to select the dimensionality of the target space automatically. When using DDA, the classification decision is made based on the class label of the nearest neighbor in the T-transformed space.

An important fact that becomes evident even from the above succint overview of the SVM and DDA methods is that neither of the two techniques produces a probabilistic output, as required by (1). It is therefore necessary to fit posterior probabilities to the raw outputs of the classifiers, which is done by implementing the approach from [19] briefly introduced below.

Considering a posterior probability of a given concept C out of its hierarchical context to simplify the notation, we may obtain the following expression via the Bayes theorem:

$$P(C|f) = \frac{p(f|C)P(C)}{p(f|C)P(C) + p(f|\overline{C})P(\overline{C})}, \qquad (10)$$

where $f = f(I_U)$ is the raw output of the classifier given query I_U, which for SVM is defined by (6) whereas for DDA it is a signed nearest neighbor distance, $P(C)$ and $P(\overline{C})$ are prior probabilities of C and its complement, $p(f|C)$ and $p(f|\overline{C})$ are the corresponding class-conditional densities. Assuming exponential behavior of the class-conditional densities, (10) simplifies to a parametrized sigmoid function:

$$P(C|f) = \frac{1}{1 + \exp(Af + B)}. \qquad (11)$$

The parameters A and B of (11) are fit using maximum likelihood estimation from a training data set (f_i, t_i), where t_i is a concept membership indicator such that $t_i = 1$ for C, and zero otherwise. The problem of finding A and B thus becomes that of minimizing the negative log-likelihood of the training data:

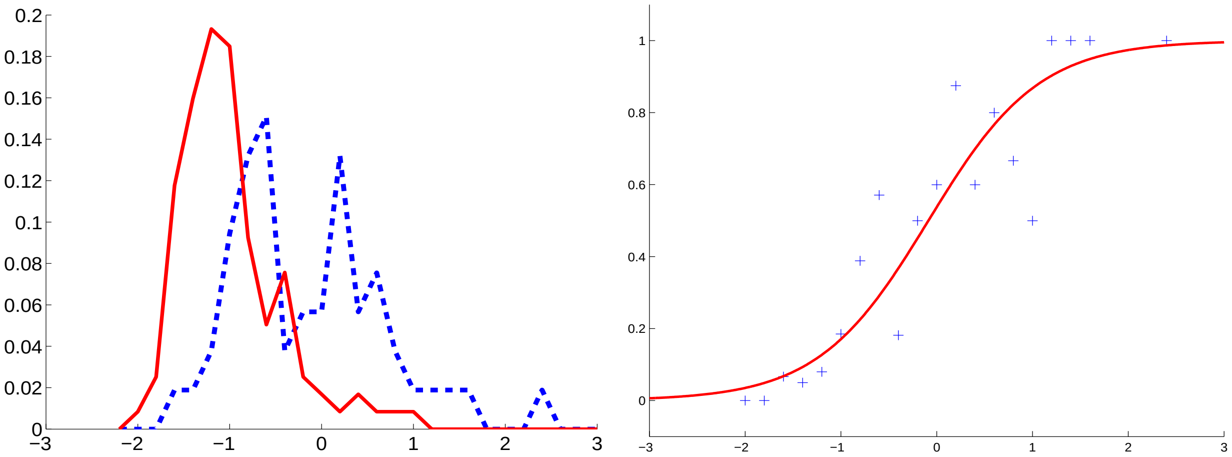
$$\min_{A,B} -\sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i), \qquad (12)$$

where

$$p_i = \frac{1}{1 + \exp(Af_i + B)}. \qquad (13)$$

In order to solve (12) in a robust fashion, the model-trust algorithm [7] suggested by the author is used.

Figure 3 shows an example of applying this posterior probability fitting technique to the output values of a Gaussian kernel SVM classifier for the leaf concept *trees*. Figure 3(a) plots the histograms of class-conditional densities



(a) Class-conditional densities of raw Gaussian SVM outputs for concept C (dashed) and its complement $\overline{C}$ (solid)

(b) Sigmoid function fit

Figure 3. Posterior probability fit for concept "trees"

derived from tenfold cross-validation, while Figure 3(b) demonstrates the fit of the sigmoid function (11) to the posterior probabilities computed from the class-conditional densities via Bayes' rule.

5. Low-level feature representation

As mentioned earlier, the baseline classifiers operate on both query I_U and training set images I_t represented by low-level feature vectors. This representation, extracted by the *Viper* system [22], comprises a subset of features corresponding to global color properties of an image that are concatenated with two-dimensional Gabor filter responses describing texture.

Viper uses a palette of 166 colors, derived by uniformly quantizing the cylindrical *HSV* color space into 18 hues, 3 saturations, and 3 values. These are augmented by 4 gray levels. This choice of quantization means that more tolerance is given to changes in saturation and value, which is desirable since these channels can be affected by lighting conditions and viewpoint. The choice of the *HSV* color space is due to its perceptual uniformity and a relatively low complexity of computation and inversion in comparison to such alternatives as *CIE-LUV* and *CIE-LAB* [21].

As for the texture features, we employ a bank of real, circularly symmetric Gabor filters, defined in the spatial domain by:

$$\begin{aligned} f_{mn}(x,y) \quad &= \quad \frac{1}{2\pi\sigma_m^2} e^{-\frac{x^2+y^2}{2\sigma_m^2}} \cos[2\pi(u_{0_m} x \cos\theta_n \\ &\quad +u_{0_m} y \sin\theta_n)], \end{aligned} \tag{14}$$

where m indexes the scales of the filters, and n their orientations. The center frequency of the filter is specified by u_{0_m}. The half-peak radial bandwidth is given by:

$$B_r = \log_2\left(\frac{2\pi\sigma_m u_{0_m} + \sqrt{2\ln 2}}{2\pi\sigma_m u_{0_m} - \sqrt{2\ln 2}}\right), \tag{15}$$

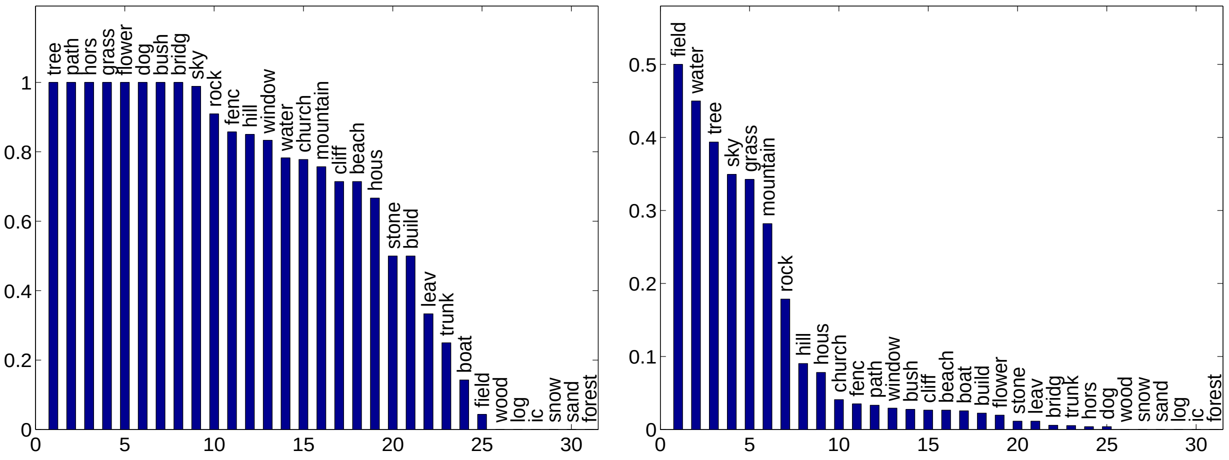
where B_r is chosen to be 1, i.e. a bandwidth of one octave, which then allows us to compute σ_m:

$$\sigma_m = \frac{3\sqrt{2\ln 2}}{2\pi u_{0_m}}. \tag{16}$$

The highest center frequency is $u_{0_1} = \frac{0.5}{1+\tan(1/3)} \approx 0.5$, so that it is within the discrete frequency domain. The center frequency is halved at each change of scale, which implies that σ is doubled (16). The orientation of the filters varies in steps of $\pi/4$, and three scales are used. These choices result in a bank of 12 filters, which renders appropriate coverage of the frequency domain with little overlap between the filters. Given the 10 band energy quantization, this design provides 120 global texture characteristics of the image. Combining this information with the color data, we obtain a common 286-dimensional feature vector representation for every image.

6. Experimental Results

In our experiments we have used data from two separate image collections for training and testing in an attempt to ensure collection-independent learning. The training data



(a) Recall

(b) Precision

Figure 4. Performance indicators on test data vocabulary

was derived from the Washington University annotated image collection [16] with about 600 images, while the testing data constituted a 254 image subset, New Zealand and Ireland sections, from Corel image database. The visual information for each training image was represented by 286-dimensional feature vector as described in the previous section. Annotation keywords appearing only once were eliminated from the target vocaublary V, from which a hierarchical ensemble of 60 concepts was constructed.

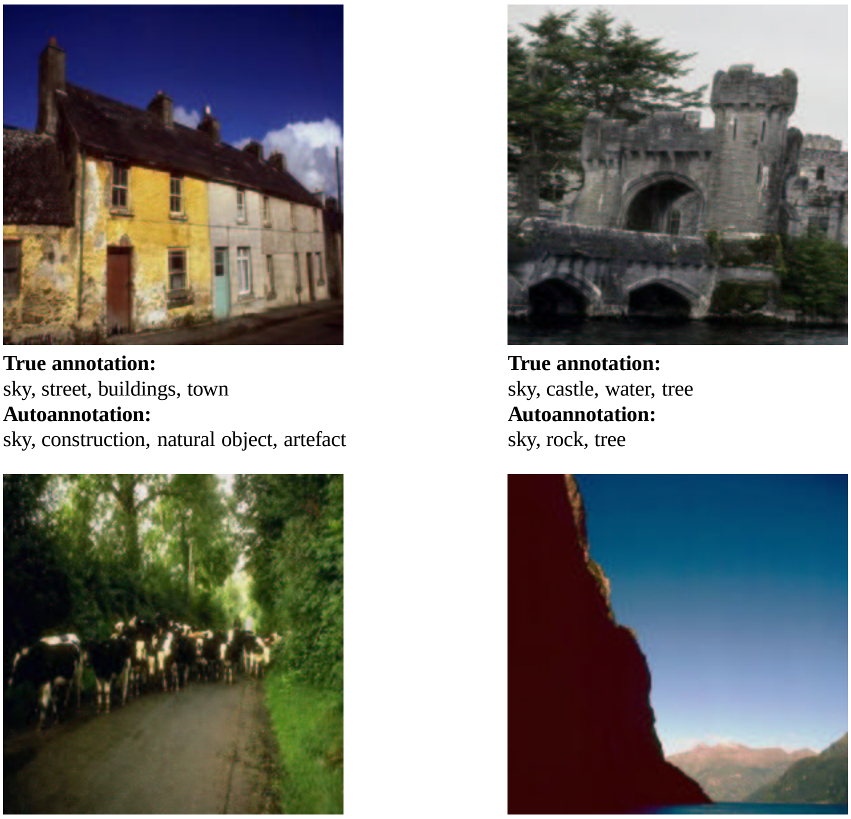
In order to be able to judge the performance of the presented method in terms of the traditional precision and recall indicators, we have adopted the following strategy. Whenever a non-leaf concept, $C_i \notin V$, is predicted, it is evaluated as a union of its underlying keywords, $\mathbf{L}(C_i)$, thus bridging the vocabulary gap between the derived concepts, e.g. *[vessel, watercraft]*, and the actual training data, e.g. *boat*, *sailboat*, *ferryboat*, *rowboat*, at the expense of precision. Using the DDA baseline classifiers [10, 11] for each concept $C_i \in H$, the following precision and recall results on the test set vocabulary were obtained (see Figure 4). As seen from the figure, the naturally high recall results boosted by keyword group retrieval, Figure 4(a) do not necessarily correspond to high frequency common concepts emphasizing the importance of the concept co-occurrence factors, while the significantly lower precision values for complex concepts, such as *church, fence, boat*, Figure 4(b), indicate that these words are much more often retrieved as a group of semantically-related keywords, rather than individually.

An illustration of the automatically derived annotation is provided in Figure 5, showing examples occurrences of out-of-vocabulary words being replaced by a visually similar common concepts $C_i \in V$ (top-right image, *castle* $\rightarrow$ *rock*), members of the vocaulary being predicted as semantically relevant, but more common (and therefore, more likely) concepts C_i (top-left, *buildings* $\rightarrow$ *construction*), as well as other typical predictions.

In addition to the above experiments, we have compared the presented method to several popular classifier ensemble techniques, such as OPC, or one-against-all strategy, and Max Wins algorithms [6] that combined SVM baseline classifiers. As shown in Table 2, the proposed hierarchical semantic ensemble (HSE) approach achieved better results despite the fact that only a fixed number of top-ranked singleton concepts was allowed to be predicted, which was done in order to ensure equal conditions for all of the methods, most of which have no means of determining exactly the number of concepts in the derived annotation. The first

Table 2. Classifier ensemble performance restricted to top 5 keywords

Ensemble	Baseline classifier	% Recall	% Precision
Empirical	none	16.13	5.04
Max Wins	SVM, polyn.	8.14	3.83
Max Wins	SVM, gauss.	10.61	4.47
OPC	SVM, polyn.	20.31	7.85
OPC	SVM, gauss.	21.27	10.19
HSE	DDA	21.22	10.20
HSE+S	DDA	28.42	26.88



True annotation:
sky, street, buildings, town
Autoannotation:
sky, construction, natural object, artefact

True annotation:
sky, castle, water, tree
Autoannotation:
sky, rock, tree

True annotation:
cows, road, trees, grass
Autoannotation:
bush, tree, grass, vascular plant, woody plant, organism

True annotation:
sky, water, mountain, trees,
Autoannotation:
sky, water, geological formation, natural object, artefact

Figure 5. Autoannotation of test images

row of Table 2 represents the reference point performance attained by sampling concepts according to their empirical distribution in the training data annotation, i.e. picking word *tree* first, since it is most likely to occur, then *sky*, and so on, whereas the last row shows an improvement in performance of the presented HSE method when one considers sibling concepts[2] the same, e.g. *sailboat* and *boat*.

We also examined the performance of various types of binary SVM techniques as baseline classifiers in the proposed HSE framework, as illustrated in Table 3. The results of these studies have confirmed earlier findings [23] stating that state-of-the-art individual classifiers do not necessarily always lead to a better performance in ensembles, while the inadequate results for the Max Wins technique, the only scheme to be using raw classifier outputs, emphasize the importance of the role of fitted posterior probabilities in classification ensembles.

Table 3. HSE performance with respect to the choice of baseline classifier

Baseline classifier	% Recall	% Precision
SVM, linear	18.12	5.28
SVM, polynomial	18.34	5.67
SVM, gaussian	18.62	6.05
DDA	21.22	10.20

7. Conclusion

We have presented a hierarchical ensemble learning method applied in the context of multimedia autoannotation. In contrast to the standard multiple-category classification setting that assumes independent, non-overlapping and exhaustive set of categories, the proposed approach models explicitly the hierarchical relationships among target classes using WordNet, and estimates their relevance to

[2] Concept A is considered a sibling of concept B if $A^{(\mathscr{D}(A)-1)} = B^{(\mathscr{D}(B)-1)}$.

a query as a trade-off between the goodness of fit to a given category description and its inherent uncertainty. The latter aspect, formulated in Bayesian terms, brings an additional benefit of allowing to determine exactly the number of categories to be predicted. The promising results of the empirical evaluation confirm the viability of the proposed approach, validated in comparison to several techniques of ensemble learning, as well as with different type of baseline classifiers.

In perspective, we plan to explore the problem of establishing correspondence between individual annotation keywords and low-level feature descriptors, and improve the proposed approach my taking advantage of the meaningful structure of the resulting hierarchical classification ensemble in order to incorporate relevance feedback from the user.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] K. Barnard, P. Duygulu, and D. Forsyth. Recognition as translating images into text. *Internet Imaging IX, Electronic Imaging 2003 (Invited paper)*, 2003.

[3] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: content-based soft annotation for multimodal image retrieval using Bayes point machines. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 13, pages 26–38. 2003.

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analisys. *Journal of the American Society of Information Science*, (41):391–407, 1990.

[6] J. Friedman. Another approach to polychotomous classification. Technical report, Stanford University, 1996.

[7] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London and New York, 1981.

[8] K.-S. Goh, E. Chang, and K.-T. Cheng. Svm binary classifier ensembles for image classification. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 395–402. ACM Press, 2001.

[9] P. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

[10] S. Kosinov, S. Marchand-Maillet, and T. Pun. Iterative majorization approach to the distance-based discriminant analysis. Presented by S. Kosinov at “Conference of the GfKl 2004”, Dortmund, Germany, March 9–11 2004.

[11] S. Kosinov, S. Marchand-Maillet, and T. Pun. Visual object categorization using distance-based discriminant analysis. In *Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering*, Washington, DC, July 2004. to appear.

[12] S. Kumar, J. Ghosh, and M. M. Crawford. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. *Pattern Analysis and Applications*, 5:210–220, 2002.

[13] T. Landauer and M. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, Waterloo, Ontario, 1990. UW Centre for the New OED and Text Research.

[14] B. Li and K. Goh. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 195–206. ACM Press, 2003.

[15] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.

[16] Y. Li and L. G. Shapiro. Object recognition for content-based image retrieval. In *Lecture Notes in Computer Science*. Springer-Verlag, to appear, 2004.

[17] G. A. Miller. Wordnet: a lexical database for English. *Commun. ACM*, 38(11):39–41, 1995.

[18] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, November 2003.

[19] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.

[20] P. Praks, J. Dvorsky, and V. Snasel. Latent semantic indexing for image retrieval systems. In *Proceedings of the SIAM Conference on Applied Linear Algebra (LA03)*, Williamsburg, USA, 2003. The College of William and Mary.

[21] J. R. Smith and S.-F. Chang. Tools and techniques for color image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIF)*, pages 426–437, 1996.

[22] D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *The 11th Scandinavian Conference on Image Analysis*, pages 143–149, Kangerlussuaq, Greenland, june 1999.

[23] V. Tresp. A bayesian committee machine. *Neural Computation*, 12(11):2719–2741, 2000.

[24] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New-York, 1998.

[25] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 69–88. MIT Press, 1998.

[26] C. Wallace and D. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.

[27] R. Zhao and W. Grosky. From features to semantics: Some preliminary results. In *IEEE International Conference on Multimedia and Expo (II)*, pages 679–682, 2000.