Computer Vision and Image Understanding 115 (2011) 787–803

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

An energy-based model for region-labeling

Hugo Jair Escalante [a,*], Manuel Montes-y-Gomé z [b], Luis Enrique Sucar [b]

[a] Universidad Autónoma de Nuevo León, Graduate Program in Systems Engineering San Nicolás de los Garza, NL 66450, Mexico
[b] National Institute of Astrophysics, Optics and Electronics Department of Computational Sciences Luis Enrique Erro #, 1 Tonantzintla, Puebla 72840, Mexico

ARTICLE INFO

Article history:
Received 24 April 2010
Accepted 11 February 2011
Available online 17 February 2011

Keywords:
Region labeling
Energy-based modeling
Random forest
Image annotation
Object recognition

ABSTRACT

This paper introduces an energy-based model (EBM) for region labeling that takes advantage of both context and semantics present in segmented images.The proposed method refines the output of multiclass classification methods that are based on the one-vs-all (OVA) formulation. Intuitively, the EBM maximizes the semantic cohesion among labels assigned to neighboring regions; that is, a tradeoff between label-association information and the predictions from the base classifier. Additionally, we study the suitability of OVA classification for the region labeling task. We report experimental results of our methods in 12 heterogeneous data sets that have been used for the evaluation of different tasks besides region labeling. On the one hand, our results reveal that the OVA approach offers an important potential of improvement in terms of labeling performance that can be exploited by refinement techniques similar to ours. On the other hand, experimental results show that our EBM improves the labeling provided by the base classifier. The EBM is highly efficient and it can be applied without modifications to different data sets. The heterogeneity of the considered databases shows the generality of our approach and its robustness to different scenarios. Our results are superior to other techniques that have been tested in the same collections. Furthermore, results on image retrieval show that the labels generated with our EBM can be helpful for annotation-based image retrieval.

1. Introduction

Region labeling (RL) is the task of assigning textual descriptors to regions in segmented images with to goal of allowing images to be searched by using keywords. This task is a special case of automatic image annotation (AIA) [1], which has been recognized as one of the *"hot topics"* in the new age of multimedia information retrieval [2]. The latter is motivated by the availability of large repositories of images without any textual description associated to them, which limits the way the images can be searched for; as the only way to access such collections is by means of content-based image retrieval (CBIR, i.e., the task of retrieving images by using their visual content) techniques. Whereas CBIR is a mature field in computer vision, CBIR methods still need of a significative amount of user interaction (in terms of specifying query images, sketches and through category browsing and relevance feedback), which is not a desired property for any automatic image retrieval system.

Labels generated with RL methods can also be helpful to improve the performance of text-based image retrieval (TBIR, i.e., image search through text that has been manually assigned to images) methods, in collections where images are accompanied by textual descriptors [3]. This is due to the fact that labels used in AIA correspond to visual information in the image, while manually assigned annotations often refer to high-level semantic information of the image, see Fig. 1; in consequence, both text and labels can be considered complimentary [3]. Thus, RL has an important impact into the multimedia image retrieval domain and therefore the development of effective RL techniques is a crucial task.

There are several factors that make RL a highly challenging problem, including: the selection of the appropriate set of visual attributes (e.g., color, texture or shape attributes) for effectively describing regions; the visual homonymy (i.e., similar regions associated to different objects, like *sky* and *ocean*) and visual synonymy (i.e., different regions associated to a single object, like different regions of *beds*) issues; and the uncertainty of the boundaries of regions (i.e., regions that include parts of different objects) due to inaccurate image segmentation. The combination of these and other factors make of RL a highly noisy and subjective task for which robust methods are required.

This paper introduces an RL method that faces the problem as one of single-label multiclass classification, where a *one-vs-all* (OVA) classifier is built with as many classes as labels are in the vocabulary. The output of the OVA classifier is then refined by an energy-based model (EBM) that takes into account spatial and semantics information. The proposed EBM is a Markov random

* Corresponding author. Fax: +52 222 266 31 52.
E-mail address: hugo.jair@gmail.com (H.J. Escalante).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.02.002

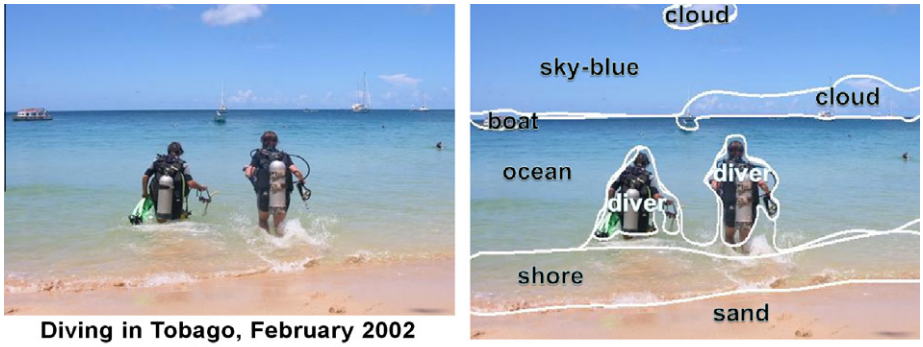

Fig. 1. Illustration of the complementariness of manually assigned text (a) and labels generated with RL methods (b). The images are taken from the SAIAPR TC-12 benchmark [4].

field (MRF) in essence, however, opposed to straightforward random field modeling, the energy-based modeling framework allows us to predefine the energy function without a learning phase; relaxing the strict probability modeling and avoiding intractable partition functions associated with MRFs.

The EBM can be seen as a postprocessing step for correcting the errors of the initial OVA classifier. Intuitively, our approach aims at maximizing the *semantic cohesion* of labels assigned to neighboring regions in an image; that is, a tradeoff between: (1) the evidence we have about a specific label is the correct descriptor for a region (i.e., the output of the OVA classifier) and (2) the relationship among labels assigned to neighboring regions (i.e., label co-occurrence statistics). The configuration of region-label assignments that maximizes this tradeoff is the predicted labeling for the image. We present experimental results in a wide variety of image collections, which include both manually and automatically segmented images. The performance of our EBM is superior to that of current methods that have used the same collections. Furthermore, we show how the labels generated by our RL strategy can be effectively used to search for images under the ABIR framework [5].

The main contributions of this paper are the following:

- We face the RL problem as one of OVA multiclass classification and succeed in improving the performance of other techniques. We present a study on the benefits offered by OVA classification to RL. Our study motivates the development of labeling refinement techniques, similar to ours, that can exploit the large potential of labeling performance provided by OVA.
- We introduce an EBM that refines the output of the OVA classifier for RL. Experimental results give evidence of the efficiency and effectiveness of the EBM. Compared to similar techniques, our method is easier to implement and shows better performance in a variety of collections. Further, our results show that labels generated by the EBM can be helpful for keyword-based image retrieval.
- We conduct an extensive experimentation for evaluating the performance of both OVA and OVA + EBM. Our study considers a suite of data sets that comprise an important diversity in terms of number of labels, number of regions, labeling granularity, image resolution and segmentation methods.

The rest of this paper is organized as follows. The next section reviews related work on RL. Section 3 describes the OVA approach to RL. Section 4 introduces the EBM for labeling refinement. Section 5 presents the databases used for experimentation. Section 6 report experimental results of our methods. Section 7 summarizes the conclusions derived from this work and outlines future work directions.

2. Related work

This section reviews related work on image annotation with region labeling (RL) techniques. Whereas the annotation task has been also faced with image-level methods (e.g., the machine translation approach [1] or the cross-media relevance model [6]) we do not consider such techniques in our review as they are designated to solve a different problem (i.e., assigning labels to the images as a whole); for complete surveys on the latter subject we refer the reader to [2,7,8].

The RL task has been mainly approached with supervised learning techniques, where the usual setting is as follows. Training images are segmented (either manually or by using automatic techniques) and each region is manually associated with a label taken from a predefined set of keywords (i.e., the annotation vocabulary); visual features are extracted from the regions[1]. The pairs of feature vectors and labels form the training set for a usual multiclass classification task, with as many classes as labels in the vocabulary. This scheme has been adopted, for example, in [9–19].

Different multiclass classification formulations have been adopted for RL, including all-vs-all (AVA), *single-machine* and OVA approaches. Vogel et al. adopted an AVA approach for RL in support of CBIR [17]. Such method resulted very effective; however, this approach is impractical for collections with a large number of labels as a total of $\frac{C\times(C-1)}{2}$ classifiers should be build for a problem with C classes. *Single-machine* approaches to RL have been adopted in [9,15,18,19]; the underlying idea is to use multiclass models for the classification task. Winn et al. used Gaussian and KNN classifiers for RL after learning a universal visual dictionary (UVD). Shotton et al. used boosting classifiers coupled with a conditional random field model in their TextonBoost framework for image segmentation and for RL. Escalante et al. and Hernandez et al. proposed Markov random fields for RL based on KNN classifiers [9,15]. Whereas satisfactory performance has been reported with this sort of methods, some *single-machine* techniques are highly complex (because a single model is used for C labels) while others obtain limited classification performance, thus other formulations are preferred. OVA approaches have been adopted for RL as well [10–14], where C binary classifiers are build (one per label) and the outputs of binary classifiers are combined to obtain a multiclass predictions. Acceptable performance has been reported with those techniques, besides OVA is more efficient than AVA and simpler than single-machine techniques.

Because of the difficulty of the RL task, postprocessing techniques have been adopted with the goal of refining the outputs

[1] hereafter we will use the term regions for referring to both, the regions themselves and the feature vectors representing the regions

of multiclass classifiers [9–15]. The underlying idea of these methods is to *perform an initial labeling by multiclass classifiers and then apply different strategies to improve the labeling provided by the preceding classification*. To the best of our knowledge, the first method of this type was proposed by Escalante et al. [9], where a random field model was used to combine the outputs of a classifier with co-occurrence information estimated from an external collection. Hernandez and Sucar [15] proposed a very similar model that incorporated spatial instead of co-occurrence information. Both models resulted very effective for improving the initial classification, however the performance of the initial classifier was very limited. In the rest of this section we review closely related methods for RL, next, in Section 2.1, we highlight the differences of the method proposed in this paper and our previous work [9].

Carsten et al. proposed a fuzzy constraint satisfaction (FCSP) approach [13] and a binary integer programming (BIP) formulation [12] for improving the initial labeling provided by a classifier based on visual features. Both methods use spatial constraint templates obtained from labeled images. Results with FCSP and BIP in the SCEF collection show that their methods can improve the initial labeling provided by the classifier [12,13]. The improvements are more significative for the BIP formulation, which is also much more efficient than the FSCS approach. However, obtaining the background knowledge required for this method (i.e., spatial prototypes) can be expensive and the BIP formulation is moderately complex.

Papadopoulos et al. faced the RL refinement problem as a global optimization problem and used a genetic algorithm (GA) for solving such problem [10,14]. The fitness function of the GA takes as input a confidence value from (OVA) support vector machine classifiers and a set of spatial constraints. Their results show improvements over the initial RL by using the GA. Nonetheless, the extraction of the spatial information is an expensive process and the method is inefficient: it took an average of 24.46 s for refining the labeling for a single image, whereas less than 0.5 s are required for the EBM we propose.

The work by Galleguillos et al. is also closely related to ours in that they consider a conditional random field (CRF) that takes as input the confidence of classifiers over labels and label co-occurrence statistics [11]. They calculate multiple segmentations for each image, each candidate segment is considered a node in the CRF. Each segment is classified by using a bag-of-features probabilistic classifier. Then the probabilities for the labels, co-occurrence information and spatial information are feed into a CRF that selects a single label for each candidate segment in the image. In the learning phase of their model they resort to importance sampling to deal with the intractable partition function of their CRF model (which we avoid with the EBM formulation).

Finally, we would like to mention that nowadays there is a research trend on the development of models that can take advantage of context through adaptive strategies for diverse computer vision tasks (i.e., models that are similar to ours) [20–24]. For example, Llorente et al. developed a Markov random field model for image retrieval using word association information [20]; Jiang et al. proposed a graph diffusion formulation that incorporates contextual information for video annotation [21]; Lee and Grauman proposed a graph based approach to object recognition that incorporates appearance and object-object relationships to discover new object categories [22]; Llorente et al. explored the use of contextual models based on co-occurrence statistics for image annotation at the image-level [23]; Yao et al. developed a random field model that incorporates human poses and object association information for recognition of human-object interaction activities (sports) [24]. This sample of successful contextual methods are evidence that the use of context is beneficial for computer vision and image understanding.

2.1. Discussion

Compared to the above described works, the method we propose is more generic, easier to implement and efficient, besides it compares favorably in performance, see Section 6. Below we summarize the benefits of our approach over related methods:

- By adopting the OVA formulation we avoid the inefficiency of AVA and the limited-performance/high-complexity of some single-machine approaches.
- The proposed EBM requires information that is relatively easy to obtain (i.e., label co-occurrence counts), therefore, our method does not need resources that are only available for particular collections nor information that is difficult to obtain.
- By adopting an EBM, instead of a straight random field method, we avoid computing intractable partition functions and, instead, we focus on inference only, which can be performed with high efficiency.
- The generality of the approach allows us applying the method for any collection of images and using any learning algorithm for the multiclass classifier.

The work described in this paper is built upon our previous work [9], however, we have gone several steps further, in particular: (i) we adopt the RL problem as one of OVA multiclass classification and improved significantly the initial annotation performance; (ii) we carry out a study on the benefits of OVA classification for RL; (iii) we improve, while keeping it simple, the refinement model; (iv) we perform an extensive evaluation of the proposed technique on a number of benchmark collections for RL, previously we only presented results on two data sets; (v) we compare the performance of our proposal with state-of-the-art techniques that face the RL problem; (vi) we present results on image retrieval by using the keywords generated by our method, showing the usefulness of our region-labeling approach.

3. Region labeling as OVA multiclass classification

In RL we are given a training set with N images $D = \{I_1,\ldots,I_N\}$, where each image is of the following form $I_i = \{(\mathbf{r}^i_1, l^i_1), (\mathbf{r}^i_2, l^i_2), \ldots, (\mathbf{r}^i_{N_i}, l^i_{N_i})\}$, with $(\mathbf{r}^i_j, l^i_j)$ being the j^{th} region-label pair of image I_i and N_i is the number of regions in image I_i. Regions $\mathbf{r}^i_j$ are d–dimensional vectors (i.e., $\mathbf{r}^i_j \in \mathbb{R}^d$) of visual features and labels $l^i_j \in W = \{1,\ldots,|V|\}$, where each number in W is associated with a single semantic descriptor in the considered annotation vocabulary $V = \{w_1, w_2,\ldots,w_{|V|}\}$. This way, an RL model is meant to learn the mapping between regions and labels by using the training set D; the performance of the trained model is evaluated on a separate test set D^T.

A natural solution to this problem is approaching it as one of multiclass classification. Several options exists for facing such classification problem, including: OVA, AVA, error-correcting output codes and single-machine approaches. A recent study by Rifkin et al. has shown that OVA (i.e., the simplest approach), is as effective as other strategies that are more complex [25]; this result motivates the use of the OVA formulation in this work. Additionally, the OVA strategy can be naturally combined with the proposed EBM technique, which is able to improve the initial labeling provided by OVA classification.

Fig. 2 illustrates the OVA formulation we consider for RL. $|V|$-binary classifiers are built, where each one is able to discriminate among examples of class l_i (positive class) and the rest $l_{j:j\neq i}$ (negative class); and where each classifier is independent to each other. When a new region $\mathbf{r}^T$ needs to be classified it is passed through the $|V|$-classifiers; each classifier i provides a value $f_i(\mathbf{r}^T)$

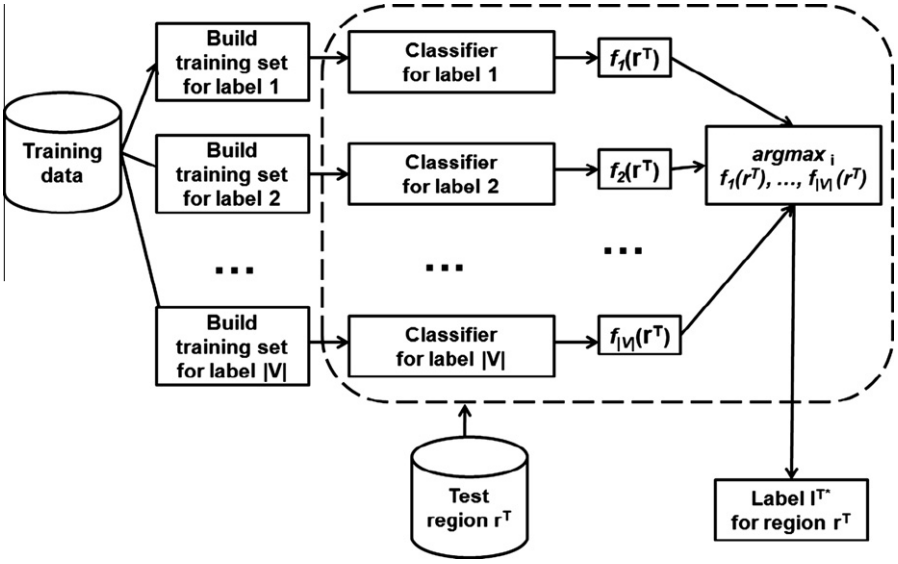

Fig. 2. Illustration of the OVA approach to RL. $|V|$ (independent) binary classifiers are built, each one is able to classify regions of label i against the rest $j{:}j \neq i$. A test region $\mathbf{r}^T$ is assigned the label corresponding to the classifier of the highest confidence $f_i(\mathbf{r}^T)$.

reflecting the confidence it has about that the correct label for $\mathbf{r}^T$ is w_i. In a standard setting, as shown in Fig. 2, the label for $\mathbf{r}^T$ is set by assigning the class of the classifier that obtained the highest confidence value $l^{T*} = argmax_i f_i(\mathbf{r}^T)$.

Despite that very effective binary classification algorithms have been developed so far [28], the application of the OVA approach to RL still remains a challenge, mainly because of the lack of correspondence between low-level features (e.g., color and texture features) and high-level semantics, an issue known as the *semantic gap* [2].

Nevertheless, there is an *"obvious"* fact about OVA classification that here we attempt to exploit for bridging the semantic gap: *"classification performance of OVA classifiers increases if we look for the correct label in the set of the top–k more confident labels instead of the most confident one"* [9,29]. Thus if we rank the labels for $\mathbf{r}^T$ in descending order of $f_i(\mathbf{r}^T)$, the probability of finding the correct class within the top–k labels will increase as we increase k. We call the set of the top–k labels for a test region its k–candidate labels. However, assigning a set of k–labels to each region may cause confusion and deteriorate retrieval performance. Hence, we must adopt additional strategies for assigning a single label to each region, this is the goal of the EBM we propose.

One should note that we have not made any restriction in the form or family of the learning algorithms that can be considered for the binary classifiers in the OVA formulation. Thus, in principle, any learning algorithm can be used for the OVA classifier, even different classification methods can be considered for different labels.

4. Energy-based model for region labeling

EBMs are structured prediction models that capture dependencies between the variables of the model by associating a scalar energy to each configuration [30]. EBMs are defined in a way that correct configurations of the variables obtain low energy values while incorrect configurations are associated high values. Inference consists of fixing the value of the observed variables and finding configurations for the remaining variables that minimize the energy. Energy-based modeling provides a common theoretical framework for several related models, including Markov and conditional random fields. However, unlike Markov and conditional random fields, EBMs do not require computing intractable partition functions and they are not restricted to a strict probabilistic modeling. Hence EBMs offer more modeling flexibility and they can be more efficient than alternative probabilistic models.

In this paper we propose an EBM with a predefined energy function to select the correct label for each region in an image, given: (*i*) the set of candidate labels for the regions in the image together with their respective confidence weights, and (*ii*) an estimate of the semantic association among the considered labels. We now describe the EBM as it is applied to each image I_i with N_i regions.

For each region $\mathbf{r}_j \in I_i$ we consider the set Q_j of its k–candidate labels, $Q_j = \{l_j^1, l_j^2, \ldots, l_j^k\}$; together with the confidence weight for each label, $Z_j = < z_j^1, z_j^2, \ldots, z_j^k >$. Let $a(\mathbf{r}_j) = l_x$ denote the assignment of an arbitrary label l_x to a region $\mathbf{r}_j$ and let $A_i = \{a(\mathbf{r}_1) = l_x, \ldots, a(\mathbf{r}_{N_i}) = l_y\}$ be a labeling for I_i (i.e., a possible annotation for all of the regions in the image), where each region can be only assigned labels from its set of candidate labels. More formally, $a(\mathbf{r}_j)$ is a random variable associated with the region $\mathbf{r}_j$ that can take values $l_x \in Q_j$ (i.e., $a(\mathbf{r}_i)$ is a label). We will use the shorthand a_j to denote $a(\mathbf{r}_j)$ when it is clear from context. The goal of the EBM is to select a labeling A_i^* for I_i such that all of the regions are associated with their correct label. One should note that this is not an easy task, since for each image k^{N_i} different labelings can be generated. Therefore, only a subset of the possible labelings can be explored for most values of k and N_i.

In order to find the labeling configuration A^* for each image I_i, we need a way to characterize the "*goodness*" of any configuration so that we can compare labelings and select the best one. Accordingly, we define the following energy function:

$$E(A) = -\sum_{a_j \in A} \left(\lambda \times \gamma(a_j) + \sum_{a_j \in A} \sum_{a_h \in \eta_{a_j}} \psi(a_j, a_h) \times \gamma(a_h) \right) \quad (1)$$

where $\gamma(a_j)$ is a potential depending on the candidate labels assigned to region $a(\mathbf{r}_j)$ and $\psi(a_{j_1}, a_{j_2})$ is a smoothing term that accounts for the association between labels a_{j_1} and a_{j_2} assigned to neighboring regions $\mathbf{r}_{j_1}$, $\mathbf{r}_{j_2}$; η_{a_j} is the set of neighbors of a_j according to a fixed neighborhood system; λ is a factor that weights the contribution of the observation potential.

This energy function assigns low energy values to labelings that are semantically coherent (in terms of the prediction of the classifier and the association among neighboring labels) and high values

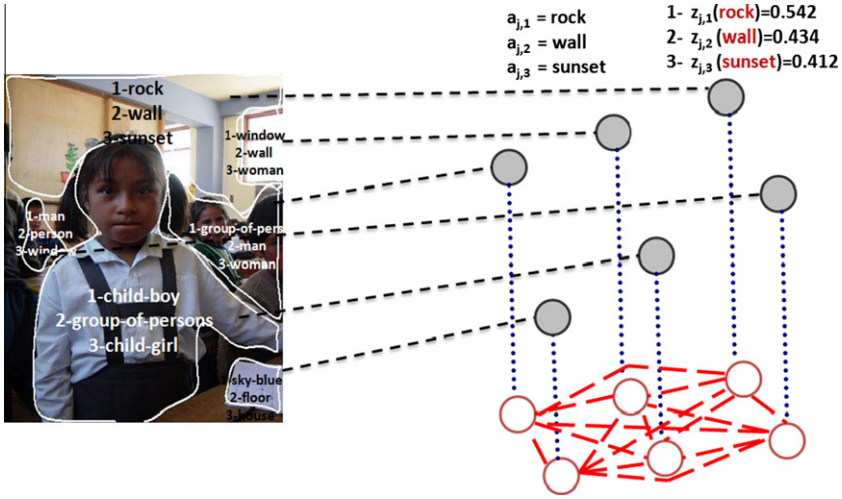

Fig. 3. Illustration of the EBM for RL. Unshaded nodes represent the assignment of labels to regions, while shaded nodes denote the confidence that the OVA classifier has in the candidate labels. We consider dependencies between every pair of regions.

to incorrect configurations. This way, the RL problem is reduced to finding the configuration A^* that minimizes Eq. (1). One should note that we have included the value of $\gamma(a_h)$ in the association potential, this is because we want to give an importance to the association term ψ proportional to the confidence on the neighboring labels.

Fig. 3 illustrates the proposed EBM, it shows candidate labels and their weights for a single region. Unshaded nodes denote assignments (i.e., $a(\mathbf{r}_j) = l_x$); while shaded nodes represent the weights assigned to candidate labels (the observed variable); dashed lines represent the dependency between assignments to neighboring regions according to the next-to relationship.

Potentials γ and ψ should be chosen carefully so that the configuration A^* that minimizes Eq. (1) is the correct annotation for the image. Usually, potentials and energy function are learnt from the data by minimizing the negative log-likelihood of the energy function or a similar expression [29,11,19,26,27,30]. However, the model proposed in this work uses as input the outputs of generic OVA classifiers and simple label co-occurrence statistics; thus, learning in the EBM model is performed offline through the OVA classifier and label occurrence counts. Hence, the generality and easiness of implementation of the proposed approach, which can be used with any OVA classifier and for any image collection, allows us to define the energy function *a priori* in a very simple way and affording efficient inference for the model. Learning the energy function in the EBM would not make sense as we have available *a priori* all of the information we need for the model; this is the main difference of the EBM with related models for RL [29,11,19,26,27], which work with specific classifiers only, needing much more information and relying on expensive training processes. The rest of this section describes the way we defined both potentials.

4.1. Observation potential

Recall that in OVA-based RL every candidate label $l_j^g \in Q_j$ for each region $\mathbf{r}_j$ is accompanied with a relevance weight z_j^g, which accounts for the confidence that classifier g has that the region belongs to its class. Therefore, it is a natural choice to define:

$$\gamma(a(\mathbf{r}_j) = l_g) = z_i^g \tag{2}$$

because we consider classifiers that return outputs in $[-1,1]$ and to ensure the sum of weights over candidate labels equals one, we scale and normalize z_j^g as follows: $z_j^g = \frac{\beta(z_j^g)}{\sum_{ki=1}^{k}\beta(z_k^g i)}$, where $\beta(x)$ is the logistic function $\beta(x) = \frac{1}{1+exp(-x)}$. This way the confidence weights will lie in [0,1] and the weights over the candidate labels will sum to one.

4.2. Association potential

The association potential $\psi(a_{j_1}, a_{j_2})$ accounts for the relationship between labels a_{j_1} and a_{j_2} assigned to regions $\mathbf{r}_{j_1}$ and $\mathbf{r}_{j_2}$, respectively. Because in RL the considered labels are tied with semantic concepts, we consider for ψ an estimate of the semantic association among concepts. Consider the following label association matrix[2]:

$$\mathbf{A}(j_1, j_2) = \frac{\#(a_{j_1}, a_{j_2})}{\#(a_{j_1})\#(a_{j_2})} \tag{3}$$

where $\#(a_j)$ is the number of images in the training set in which label a_j occurs and $\#(a_{j_1}, a_{j_2})$ is the number of images in the training set in which both a_{j_1} and a_{j_2} co-occur[3]. For specific labels a_{j_1} and a_{j_2}, the higher $\mathbf{A}(j_1 j_2)$ the more both labels are related. One should note that the values of $\mathbf{A}$ lie in a scale that may be incompatible with the values obtained from the observation potential. For avoiding the potential effects of this difference in scale, we sort the rows of matrix $\mathbf{A}$ (in descending order) and replace the non-zero values of $\mathbf{A}$ with equally-spaced values in [0.9, 0.05], where the intervals depend on $|V|$; notice that this process implicity smooths the matrix $\mathbf{A}$.

In Eq. (1) the contribution of $\psi(a_{j_1}, a_{j_2})$ is weighted by the confidence value of the neighboring candidate label under consideration $\gamma(a_{j_2})$; this in order to weight the importance of the association value according to the support of candidate labels of neighboring regions. Intuitively, we will be more confident on the significance of $\psi(a_{j_1}, a_{j_2})$ if the confidence weight of a_{j_2} is high;

[2] The ratio $\frac{\widehat{P}(i,j)}{\widehat{P}(i)\widehat{P}(j)}$ is widely used in statistics and information theory, e.g., in the mutual information: $I(X,Y) = \sum_{X,Y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$. However, in this work we do not considered probabilities and instead used (un-normalized) raw counts of word co-occurrence, thus the ratio $\frac{\#(a_{j_1}, a_{j_2})}{\#(j_1)\#(j_2)}$ is just an indicator of association between labels.

[3] In previous work we have investigated the use of different corpora for computing $\mathbf{A}$ [9], (including other image collections and corpus created from the Web), although, as expected, we have found that better results can be obtained with the EBM if co-occurrence statistics are calculated for the collection of images to be annotated.

Table 1
Statistics for the image collections considered for experimentation. We show the total number of images (Imgs.), the number of images used for training (Tr.) and testing (Te.), the total number of labels (Lb.) and regions (Regs.), the average number of regions per label (RxL), the average number of regions per image (RxI), the type of segmentation (Seg.), the references where the collection has been used (Ref., the slash separates the work where the collection was used first from other representative works that have used the collection); $*$ two versions of the SCEF data set were used, see text.

Collection	Imgs	Tr.	Te.	Lb.	Regs	RxL	RxI	Seg.	Refs.
COREL-AN	205	137	68	22	2,008	91.28	9. 79	Ncuts	[29]/[9,15]
COREL-AG	205	137	68	24	4920	205	24	Grid	[29,9]
COREL-BN	299	199	100	44	3068	69.72	10.2	Ncuts	[29,9]
COREL-BG	299	199	100	39	7176	163.9	24	Grid	[29,9]
COREL-CN	504	336	168	56	5076	90.64	10.16	Ncuts	[29]/[9]
COREL-CG	504	336	168	56	12,096	216	24	Grid	[29]/[9]
SCEF*	923	400	523	10	6244	567.63	6.77	KM	[12]/[10,14]
SAIAPR TC-12	20,000	CV	CV	255	99,317	390	4.97	Man	[4]/[35,4]
MSRC-1	240	120	120	10	790	79	3.31	Man	[18]/[19]
MSRC-2	591	335	256	22	2062	85.92	3.49	Man	[19]/[11]
VOGEL	700	CV	CV	10	70,000	7000	100	Grid	[17]/[17]

on the other hand, if the confidence on a_{j_2} is low, the significance of $\psi(a_{j_1}, a_{j_2})$ should have a minor impact in the energy function.

Under the proposed EBM, the problem of labeling refinement is reduced to that of finding the configuration that minimizes Eq. (1). This can be achieved by adopting usual inference techniques for random fields [31]; in this work we use one of the simplest inference methods: iterated conditioned modes (ICM) [32]. We have also tried simulated annealing [33] and graph cuts [34], though we did not found a significant difference in performance but an important increase in computational time.

One should note that the EBM can be easily extended to incorporate further information from the relation among regions (e.g., spatial relationships) or any other external knowledge. This could be done by defining suitable potentials over the nodes of the EBM; we will explore this research direction in future work. Nevertheless, adding more information into the EBM could make the model more complex, which would lessen the simplicity benefit of the proposed EBM.

4.3. Parameter tuning in the EBM

From the Eq. (1) we can see that there is a single tunable parameter for the energy function, namely λ. In our experiments we tune this parameter by trial and error on a validation data set. The value of this parameter depends on the collection under consideration, although generally it is greater than one, see Section 6. A parameter that is implicit in the EBM is k, the number of candidate labels per region. This value depends on the number of labels in the vocabulary and the performance of the OVA classifier. In our experiments we set this parameter by trial an error as well.

We consider two neighborhood systems for the EBM: a fully connected system (assuming every region in an image is connected to each other) and a next-to system (where only neighboring regions are considered to be connected). In preliminary experiments, we found that better results are obtained with the fully connected configuration, thus we use such setting for all of the experiments reported in Section 6.

5. Image databases and experimentation settings

For evaluating the performance of our methods we considered a suite of image collections that have been used by other researchers. We consider the Corel^R subsets[4] provided by Carbonetto et al. [29], the Spatial Context Evaluation Framework[5] (SCEF) considered by Carsten et al. [12,13] and Papadopulos et al. [10,14], the segmented and annotated IAPR TC-12 benchmark[6] introduced by Escalante et al. [4], the Microsoft^R object recognition data sets[7] due to Winn et al. [18] and Shotton et al. [19], and the database of natural scenes of Vogel et al. [17]. All of the data sets in Matlab^R format and Matlab^R code implementing the EBM are available from the website of the first author.

Table 1 describes the considered collections. All of these collections have been segmented and each of the resultant regions were manually labeled using different annotation vocabularies. There is a wide variety of collection sizes, vocabulary lengths, number of regions per image and number of regions per label. Thus, these databases form a suite of heterogeneous RL data sets that will be helpful for showing the generality of our EBM and its robustness to different settings.

Different segmentation methods have been used for the different collections, including grid segmentation (*Grid*), normalized cuts [36] (*Ncuts*), manual segmentation (*Man*) and the K-means-with-connectivity-constraint pixel classification algorithm due to Mezaris et al. (*KM*) [37]. The diversity of segmentation algorithms used for different data sets will provide evidence about the robustness of our approach to different automatic segmentation algorithms.

For all collections, we considered the same data splits for training and testing as used in the references that have considered the same collections (column 10 in Table 1); 10-fold cross validation (*CV*) has been used for VOGEL and SAIAPR TC-12 collections. A total of 23,462 different images have been considered in our experiments comprising about 212,995 different regions. The resolution of the images varies from 192×128 for the Corel^R subsets [29], to 720×480 for the VOGEL data set [17]. The content of the images is very diverse, comprising photos of animals, landscapes, and natural scenes as well as pictures of people under different situations, rooms, buildings and man-made objects among many other contents.

Fig. 4 shows sample images from the considered databases. The collections represent a sample of real world images (in particular those from the SAIAPR TC-12 and SCEF collections) that will be helpful for assessing the performance of the EBM in a realistic scenario. The segmentation quality varies notably, being MSRC-X the collections with the better segmentation, this is obvious as these collections are labeled at a pixel-level [18,19]; images from the VOGEL collection have been segmented into 100 square regions, which makes the segmentation more accurate than that of COREL-XG databases that have been segmented into 24 patches. For the SCEF collection some objects are not considered for classi-

[4] http://www.cs.ubc.ca/ ~ pcarbo/corel.tar.gz
[5] http://mklab.iti.gr/project/scef
[6] http://imageclef.org/SIAPRdata
[7] http://research.microsoft.com/en- us/projects/objectclassrecognition/

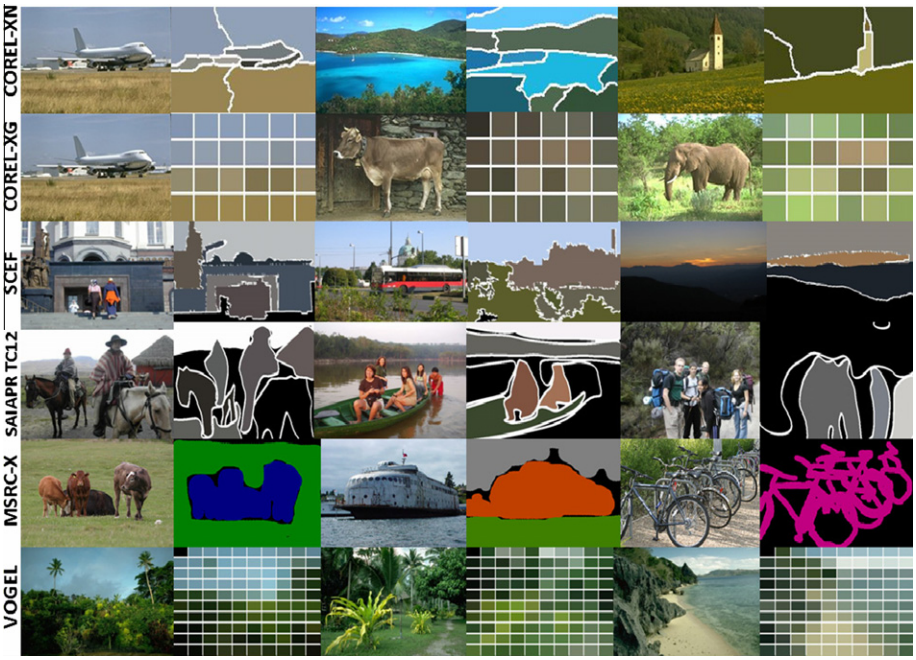

Fig. 4. Sample images from the considered collections. We show the original image followed by its corresponding segmentation mask (showing the average color of each region) for three different images.

Table 2
Summary of the visual features considered for each image collection.

Collection	# Features	Description
COREL-XY	16	Area, and RGB/CIE-Lab color features
SCEF-I	433	MPEG-7 descriptors
SCEF-II	4	Wavelet-based features
SAIAPR TC-12	27	Area, and RGB/CIE-Lab color features
MSRC-X	27	Area, and RGB/CIE-Lab color features
VOGEL	207	Edge and HSI-color histogram, texture features

fication, as these are not represented by the 10 labels the authors predefined (see for example the *'bus'* in the third row of images).

For each data set in Table 1 we considered the visual features that have been used in the corresponding reference. Table 2 summarizes the visual features we considered, for further details we refer the reader to the corresponding references.

5.1. The SAIAPR TC-12 benchmark

Despite all of the collections we consider can be helpful to evaluate the performance of RL techniques, the SAIAPR TC-12 collection is a benchmark specifically designed for RL; therefore, it is worth giving more information about this database [4].

The SAIAPR TC-12 benchmark consists of 20,000 images that have been previously used for the evaluation of CBIR, TBIR and multimedia image retrieval methods [35]. It has been manually segmented and labeled as an effort to increase the scope of the benchmark by evaluating RL methods as well. For the annotation process it was defined a conceptual hierarchy that facilitated the manual labeling process. The hierarchy can also be helpful to evaluate the *soft* RL performance [4]. Sample images from the SAIAPR TC-12 collection are shown in Fig. 1 and row 4 of Fig. 4. Images from this benchmark comprise a wide variety of concepts including: *landscapes, sports, roadways, landmarks, buildings, historical places and people under different situations, among many other*. Hence, most of the concepts covered by the other collections shown in Table 1 are contained in the SAIAPR TC-12 collection.

5.2. Experimentation settings

For our experiments we have used the same evaluation methodology adopted by other authors that have used the same image collections. For each database, we used the training set for training an OVA multiclass classifier; co-occurrence statistics were computed from the training set as well. We used the trained OVA classifier for predicting outputs and obtaining confidence values for the regions in the test set; the top–k labels (according to their confidence values) were kept for each test region. Then, the candidate labels, with associated weights for each region, and the co-occurrence matrix were feeded into the EBM, which used ICM for finding the best labeling for each image in the test set. The output of the EBM is then evaluated.

We considered the accuracy as indicator of RL performance, defined as the percentage of regions that were labeled with the correct class. In addition to RL performance we report the average processing time it took labeling an image with the EBM; all of our experiments were performed on a Laptop with a Core DUO processor (1.8 GHz) and 2 GB in RAM.

When comparing the performance of the EBM with that of related works, we resort to results published by the researchers in their respective papers. We have tried to reproduce exactly the settings under which related works were applied, in such a way that the comparisons are both fair and meaningful.

6. Experimental results

We have divided the results into four sections evaluating different aspects of the OVA RL approach and the EBM we propose. First, we assessed the performance of different binary classifiers in the OVA-based approach to RL. Next, we evaluated the improvement over OVA classification offered by the EBM. Next, we compared the RL performance of OVA and OVA + EBM to that of other approaches that have used the same image collections. Finally, we evaluated the retrieval performance on keyword-based search using the labels generated with the EBM.

6.1. OVA for region labeling

The goal of the experiments described in this subsection is to evaluate the performance of the OVA labeling approach depicted in Fig. 2. For these experiments we considered the SAIAPR TC-12 and SCEF collections, as these are the ones with more realistic images and allowed us to evaluate RL performance with a different number of examples/labels. For the SAIAPR TC-12 collection we considered those labels for which there are at least 200 regions labeled with the concept; this comprises 90 labels and about 91,139 regions out of the 99,317 (i.e., 91.7%). Regarding the SCEF collection, we considered the SCEF-II data set as we have obtained better results with this version than with SCEF-I.

We considered some of the classifiers available in the CLOP[8] machine learning toolbox [38], which are described in Table 3. Rows 2-4 show non-parametric classifiers or classifiers that perform model selection internally; rows 5-8 describe parametric classifiers with default parameters; row 12 shows PSMS, a heuristic search strategy that automatically selects methods[9] for preprocessing, feature selection and classification from the CLOP package, as described by Escalante et al. [39]. One should note that we disregarded some of the classifiers available in CLOP as they are computationally expensive to use them with the SAIAPR TC-12 benchmark (e.g., gkridge); also, note that despite we do not explicitly considered the popular support vector machine (SVM) classifier, the CLOP implementations of both kridge and SVM are equivalent when using default parameters.

Table 4 shows the RL accuracy obtained by the different classifiers averaged over 10 trials for the SCEF-II data set and over a 10-fold CV run for the SAIAPR TC-12 benchmark. Rows 2 and 3 in Table 4 show the performance of two baselines: predicting the majority class, regardless of the region (**Baseline-I**) and making predictions at random (**Baseline-II**).

With exception of the Naive classifier in SCEF-II, all of the considered methods outperformed the baselines in the two collections. Consistently, the best classifier for both databases was random forest (RF). The worst performance for the SAIAPR TC12 collection was that obtained by the Zarbi classifier, while the Naive method gave the worst result on the SCEF-II collection.

Whereas the performance of RF on the SCEF-II collection was, to some extent, satisfactory, its accuracy on the SAIAPR TC-12 collection seemed poor as to support ABIR applications. This is not surprising because: 1) the size of the SAIAPR TC-12 collection is more than 20 times larger than that of the SCEF-II collection; 2) the number of considered labels is about 9-times larger; and 3) the images are more much diverse in the SAIAPR TC-12 benchmark. Thus, despite modest, the performance achieved by RF is rather a strong baseline. Moreover, even with this labeling performance we can obtain reasonably good results on ABIR (See Section 6.4).

Figs. 5 and 6 show the *soft* RL performance of the methods considered in Table 3; for each method, it is shown the accuracy obtained by looking for the correct label at the top–k labels, when sorted by its confidence value, see Section 3. This figure clearly illustrates the potential benefits offered by OVA classification for RL. For the SCEF-II data set, we could achieve up to 97% accuracy if we consider the two most confident labels; while for the SAIAPR TC-12 collection we could obtain above 65% of accuracy if we consider the top–5 labels. Therefore, we could almost duplicate the initial accuracy if we would select the correct label for each region from its set of top–k candidate labels.

Nevertheless, despite the potential advantage brought into play by the OVA formulation, selecting the correct label from a set of k candidate labels is not an easy task. For example, Table 5 shows the performance we would obtain by randomly selecting a single label for each test region from its corresponding set of k candidate labels, as obtained by the RF classifier, for the SCEF-II and SAIAPR TC-12 data sets.

From this table we can see that randomly selecting labels is a rather unreliable way for RL and thus effective techniques are required for solving this problem. As expected, the accuracy of the random selection strategy decreases as k increases, thus as k increases potential accuracy increases but the problem of selecting the correct label out of k candidates is more complicated.

Table 3
Classifiers used in the experiments with default parameters.

Classifier	Description	Parameters
Zarbi	A simple linear classifier	–
Naive	Naive Bayes classifier	–
Klogistic	Kernel logistic regression	–
Neural	Feedforward Neural Network	Units = 10, Shrinkage = 0.1, Epochs = 50
Kridge	Kernel ridge regression	Kernel = *Poly*, Degree = 1, Shrinkage = 0.001
RF	Random Forest	Depth = 1, Shrinkage = 0.3, Units = 100
Logitboost	Logit-boosting with trees	Depth = 1, Shrinkage = 0.3, Units = 50
PSMS	Particle swarm model selection	FMS = 1

Table 4
Average and standard deviation of RL accuracy for different classifiers, under the OVA formulation, on the SAIAPR TC-12 and SCEF-II collections.

Classifier	SAIAPR TC-12	SCEF-II
Baseline-I	$5.75 \pm 0.00\%$	$18.84 \pm 0.00\%$
Baseline-II	$1.09 \pm 0.00\%$	$10.00 \pm 0.00\%$
Zarbi	$7.56 \pm 0.36\%$	$59.03 \pm 0.00\%$
Naive	$9.45 \pm 0.04\%$	$13.64 \pm 0.00\%$
Klogistic	$29.62 \pm 0.28\%$	$67.30 \pm 0.00\%$
Neural	$20.38 \pm 1.14\%$	$48.33 \pm 0.00\%$
Kridge	$22.74 \pm 0.39\%$	$56.00 \pm 0.00\%$
RF	$\mathbf{36.13 \pm 0.71}\%$	$\mathbf{81.45 \pm 0.00}\%$
Logitboost	$27.31 \pm 0.46\%$	$79.48 \pm 0.00\%$
PSMS	$14.38 \pm 2.75\%$	$76.97 \pm 2.88\%$

6.1.1. On random forest for region labeling

The effectiveness of RF to RL can be due to the fact that random forest has been proved to be particulary very effective for classification problems in highly imbalanced data sets [41,42]. The OVA formulation is naturally a severely-imbalanced classification task: when there are the same number of examples per class (i.e., a balanced multiclass classification data set), for each class there are $\frac{N}{|V|}$ positive examples and $(|V|-1)\times\frac{N}{|V|}$ negative ones, with $|V|$ the number of classes and N the number of examples; for instance, for $|V| = 90$ and $N = 91{,}139$, we would have 1,012 positive examples and 90,127 negative ones for each label.

What is more, RF is advantageous over the other classifiers in RL because it is an ensemble method [40]: it well known that ensemble methods offer better generalization performance than individual techniques, furthermore they are more robust to outliers and noisy data in general; the latter is important for RL because it is a highly subjective task (even for humans) and therefore RL is naturally prone to noise. Thus it is not surprising that RF achieved better performance than the other methods we considered in OVA multiclass classification. Additionally, one should note that the random forest implementation that we considered (that from the CLOP toolbox) has proved to be a effective technique in a recent prediction-performance challenge [41,43].

[8] http://clopinet.com/CLOP/

[9] PSMS selects a specific *full model* for each binary classification problem; where a *full model* is the combination of methods for preprocessing, feature selection and classification, see [39,16] for details.

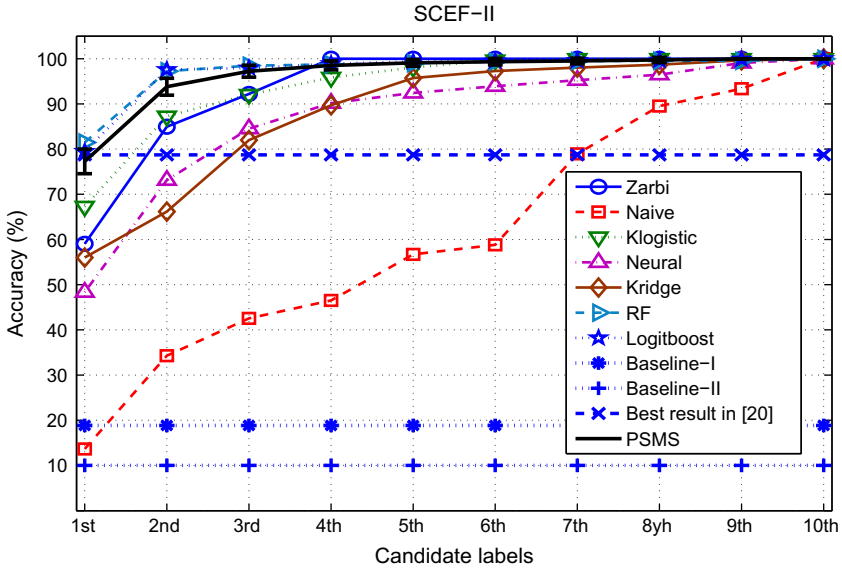

Fig. 5. Soft RL performance of the classifiers shown in Table 3 for the SCEF-II data set. It is shown the accuracy obtained by looking for the correct label at the top–k labels, when sorted by its confidence value.

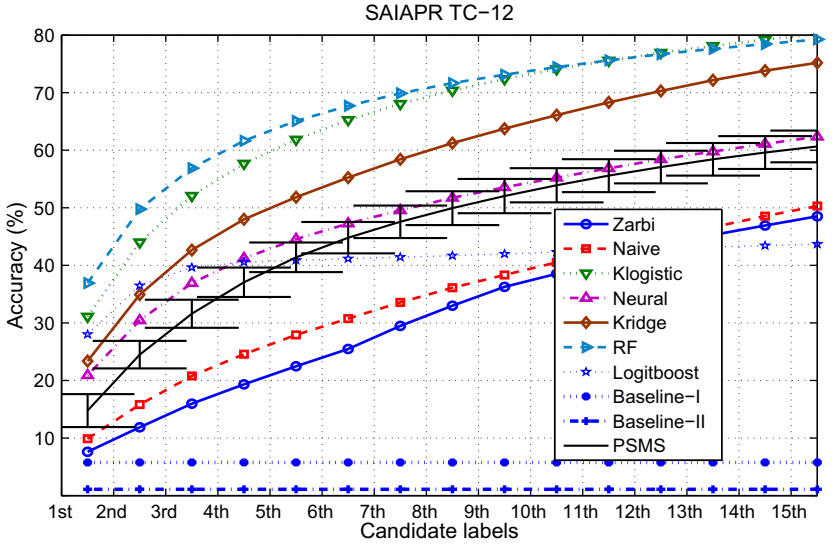

Fig. 6. Soft RL performance of the classifiers shown in Table 3 for the SAIAPR-TC12 data set.

Table 5
RL performance we would obtain by randomly picking a single label for each region from k–candidates, according to the candidate labels generated by a RF classifier for the SAIAPR TC-12 and SCEF-II collections, see Figs. 5 and 6; recall $|V|$ is the number of labels in the annotation vocabulary.

k	SAIAPR TC-12 (%)	SCEF-II (%)
1	36.91	81.45
2	24.88	48.62
3	18.96	32.82
4	15.41	24.68
5	13.02	19.80
10	7.44	10.00
$\|V\|$	1.12	10.00

6.1.2. Random forest vs PSMS

The labeling performance of RF resulted even superior to that of classifiers selected with PSMS (another technique that has proved to be very effective for model selection [39,43]). However, we would like to emphasize that the performance of binary classifiers selected with PSMS was significatively better (one-by-one) than the respective binary RF-classifiers. Fig. 7 compares the individual performance of the $|V|$–binary classifiers selected with PSMS to the RF classifiers for the SCEF-II data set; we show the balanced error rate[10] (BER), which is an adequate quantity given the high class-imbalance inherent in OVA classification.

From Fig. 7, it is evident that PSMS-based classifiers outperformed RF ones. The average BER for the SCEF-II data set was of 15.38% and 7.97% for the RF and PSMS classifiers, respectively; whereas, for the SAIAPR TC-12 the respective BER was of 47.19% and 29.11%, the corresponding figure is not shown for clarity. In average, the difference in BER between RF and PSMS classifiers was of 7.41% and 18.08% for the SCEF-II and SAIAPR TC-12 collections, respectively.

[10] $BER = \frac{E^+ + E^-}{2}$, where E^+ and E^- are the positive and negative error rates.

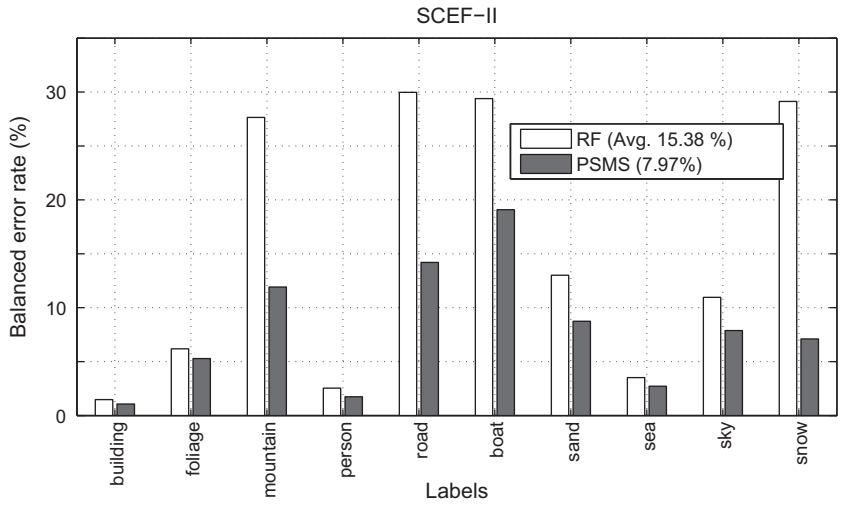

Fig. 7. Balanced error rate of the RF (white) and PSMS (gray) individual classifiers in OVA RL in the SCEF-II collection.

Therefore, better binary classifiers were selected by PSMS; however, when combining the outputs of such classifiers the multiclass performance was rather poor. This is due to the fact that outputs of classifiers selected with PSMS lie in different scales as each classifier may use different methods for preprocessing, feature selection and classification. Thus, the outputs of the different classifiers are not comparable and the combination of such outputs, under OVA, does not make sense, which resulted in lower multiclass classification performance. On the other hand, for RF the $|V|$ classifiers may be different from each other only in terms of the hyperparameters and features considered for the splits. Hence, the outputs of the different RF classifiers are *'more comparable'* and therefore the multiclass performance of RF was better.

Nevertheless, PSMS provides a potential benefit, in terms of the highly accurate individual classifiers one can obtain, which we would like to explore as future work. Finally, one should note that the multiclass classification performance of PSMS was related to the number of classes: the difference in performance, shown in Table 4, between PSMS and RF was larger for the SAIAPR TC-12 benchmark (90 labels) than for the SCEF-II (10 labels) collection.

6.2. Labeling refinement with the EBM

In this section we report experimental results with the EBM method described in Section 4. The goal of these experiments is to show how the EBM improves the labeling provided by an OVA classifier. For the rest of our experiments we consider an RF classifier as this was the method that obtained the best RL performance in the previous section. We compare the RL performance of the RF classifier (OVA-RF) to that obtained after applying the EBM (OVA-RF + EBM); additionally, we show the performance one would obtain by randomly selecting labels. The parameters of the EBM have been fixed by trial and error, see Section 4.3.

Table 6 shows the performance of OVA-RF and OVA-RF+EBM for the data sets described in Section 5; results are averaged over 10 trials. For this experiment we fixed $k = 3$ and ran ICM for 50 iterations, different values of λ were used for the different collections. Column 7 (**S**) in Table 6 indicates whether the difference in performance is statistically significant according to a Wilcoxon signed-rank test, which is the test recommended for the comparison of classifiers across multiple data sets [44]. In the following we will refer to this statistical test with 95% of confidence when mentioning statistical significance.

The EBM method improved the performance of the initial OVA-RF classifier for all of the considered data sets; all of the differences were statistically significant; showing that the proposed method is able to improve the performance of the base classifier. The improvements vary across the data sets, the largest improvement was for the MSRC-2 data set (+7.69%), whereas the smaller improvements were for the SCEF-I (+0.61%) and SAIAPR TC-12 (+0.67%) collections. For all of the data sets, the EBM achieved much more better performance than the baseline (column 8 in Table 6), which proves that the EBM is significantly better than selecting labels at random. The value of λ differs from a data set

Table 6
Comparison of RL accuracy between OVA-RF (column 4) and OVA − RF + EBM (column 5) in the considered data sets. We show the values of λ for the different data sets (column 2), the average processing time required to process a single image with the EBM (column 3), the relative improvement when using the EBM (column 6), the statistical significance of the difference in performance (column 7), and the performance we would obtain by randomly picking labels (column 8).

Collection	λ	time (s)	OVA − RF	OVA − RF + EBM	Impr.	S	Baseline
COREL-AN	3	0.36	$57.90 \pm 0.07\%$	$\mathbf{58.97 \pm 0.00}\%$	1.85%	+	27.11%
COREL-AG	3	0.87	$56.56 \pm 0.23\%$	$\mathbf{57.23 \pm 0.00}\%$	1.18%	+	27.46%
COREL-BN	3	0.35	$46.65 \pm 0.09\%$	$\mathbf{48.74 \pm 0.00}\%$	3.38%	+	22.59%
COREL-BG	3	0.86	$46.08 \pm 0.12\%$	$\mathbf{46.87 \pm 0.00}\%$	1.71%	+	23.51%
COREL-CN	3	0.35	$54.13 \pm 0.61\%$	$\mathbf{55.59 \pm 0.00}\%$	1.46%	+	24.73%
COREL-CG	3	0.88	$52.28 \pm 0.15\%$	$\mathbf{52.71 \pm 0.00}\%$	2.69%	+	25.42%
SCEF-I	7.5	0.42	$59.99 \pm 0.05\%$	$\mathbf{60.35 \pm 0.10}\%$	0.61%	+	25.49%
SCEF-II	3	0.41	$81.55 \pm 0.01\%$	$\mathbf{82.92 \pm 0.01}\%$	1.68%	+	32.16%
SAIAPR TC-12	3	0.16	$34.38 \pm 0.27\%$	$\mathbf{34.61 \pm 0.38}\%$	0.67%	+	18.51%
MSRC-I	3	0.22	$86.60 \pm 3.06\%$	$\mathbf{88.82 \pm 4.77}\%$	2.56%	+	32.74%
MSRC-2	3	0.21	$70.60 \pm 0.61\%$	$\mathbf{76.03 \pm 0.39}\%$	7.69%	+	30.35%
VOGEL	1	3.58	$70.78 \pm 5.91\%$	$\mathbf{72.54 \pm 6.46}\%$	2.49%	+	30.32%

to other, although for most collections $\lambda = 3$ gave the best results. This result reflects the fact that the confidence values of the RF classifier resulted more helpful than the co-occurrence statistics.

Column 3 in Table 6 shows that the processing time required to process a single test image was rather small, giving evidence of the high efficiency of the EBM. Since ICM is a local optimization procedure the processing time of EBM depends on the number of regions per image. In average, for the SAIAPR TC-12 collection were required about 0.16 seconds per image; while for the VOGEL collection were needed about 3.6 seconds per image, which is not surprising as for the latter collection each image has been segmented into 100 regions. For comparison, on the SCEF collection, Papadopulos et al. report that 24.46 seconds were required to process a single image [10], while Carsten et al. report that the required time for the BIP and FCSP methods (see Section 2) is of 1.1 and 40 seconds respectively. EBM is much more efficient than the genetic algorithm of Papadopulos et al. (by a factor of 58) and the FCSP method due to Carsten et al. (by a factor of 96) and it is more than twice as efficient than the BIP technique.

The EBM seems to be robust to different segmentation algorithms, as the improvements were varied across collections that were segmented with different methods. For example, different improvements were obtained for grid-segmented collections and different performance was achieved for collections segmented with the same method (e.g., the MSRC-X) data sets. This suggests our method does not depends on the segmentation algorithm used and hence it is not too sensitive to the inaccuracy of region boundaries.

The diverse improvements reported in Table 6 motivate further analysis on the characteristics of the data sets with large/small improvements. For that reason we consider a nearest-neighbor-based-estimator (HNNE) for illustrating the difficulty of the multi-class classification problem for different data sets. We define HNNE plots as follows. x and y axis represent the regions in the data set, where regions with similar classes are put together. For each region $\mathbf{r}_i$ we plot a square at the intersection of the region (x–axis) with its 1-nearest-neighbor (y–axis), according to the Euclidean distance. This type of plots are helpful for visualizing the difficulty of the task according to the considered features; in particular, they can give us insight about overlap of regions with different labels in the data set. Thus, a large accumulation of dots across the diagonal boxes means that the classification problem is not complicated because most regions are similar to other regions of the same class. Accordingly, a large amount of points outside these boxes means that the data set is difficult as the points outside the boxes represent regions that are similar to regions of different classes. Figs. 8–10 show HNNE plots for the SCEF-II, SCFE-I, and MSRC2 collections, respectively.

It is clear from Fig. 8 that the SCEF-II collection shows very little overlap among regions of different classes, resulting in a high labeling accuracy and a reasonable improvement due to EBM of + 1.68%. On the other hand, as shown in Fig. 9, the SCEF-I collection shows a high overlap among classes, which is reflected in low RL performance and a very small improvement due to EBM (+0.61%). This result gives evidence that EBM can improve the labeling when initial classifiers are more reliable, which is not surprising as the confidence of classifiers contributes to both potentials of the energy function in Eq. (1). One should note that the difference in accuracy between the SCEF-I and SCEF-II data sets was of more than 22%; recall that the difference between SCEF-I and SCEF-II is the set of features considered for representing the regions (see Table 2). Hence, the MPEG-7 descriptors are not useful for describing regions in the SCEF collection, whereas the 4-dimensional Wavelet-based features are powerful descriptors, despite the their low dimensionality.

The overlap of regions with different classes in MSRC-2 seems to lie somewhere between that of the SCEF-I and SCEF-II data sets,

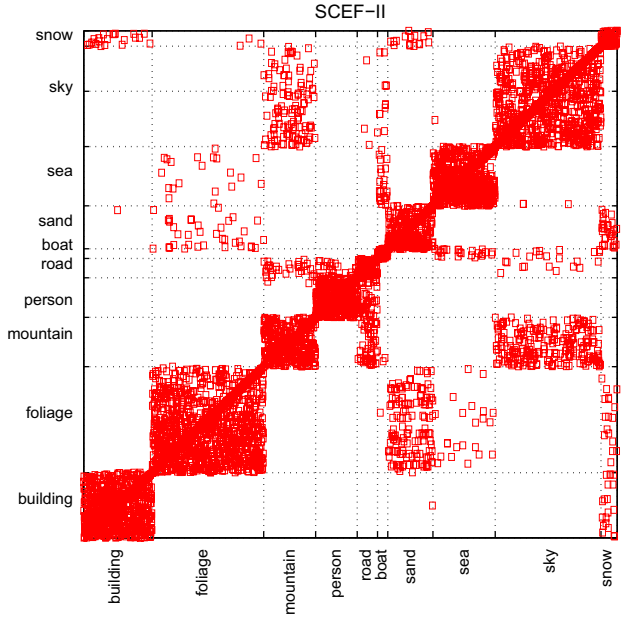

Fig. 8. Visualization of the difficulty of the RL problem for the SCEF-II data set. For each region in the collection, we plot a square where the region and its 1-nearest-neighbor (according to the Euclidean distance) intersect.

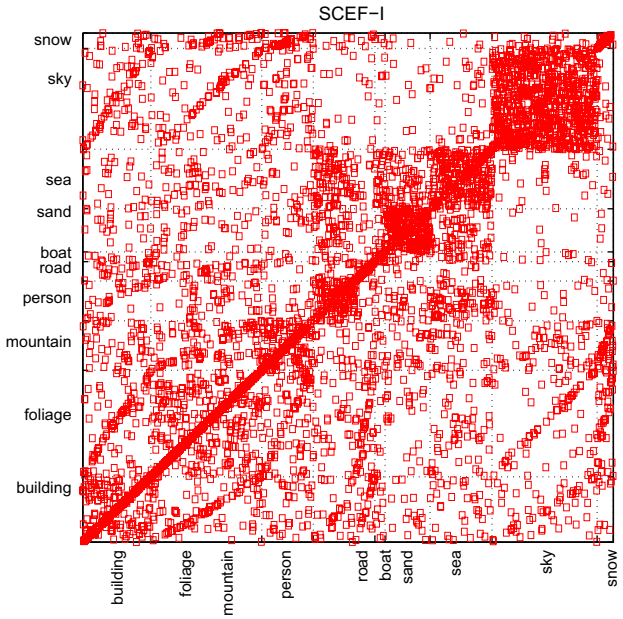

Fig. 9. Visualization of the difficulty of the RL problem for the SCEF-I data set.

see Fig. 10. For this collection EBM achieved the largest improvement (+7.69%), therefore, we can roughly conclude that the EBM is more helpful when there is a tradeoff between the difficulty of the task and the reliability of the OVA classifier. It is interesting to point out that the number of classes of the data sets is not clearly related with the performance of the EBM method, with exception of the SAIAPR TC-12 collection discussed below.

6.2.1. OVA-RF+EBM in the SAIAPR TC-12 benchmark

The small improvement in the SAIAPR TC-12 can be due to several factors, though mostly to the number of classes and the diversity of images. In order to get further insights on the latter aspect,

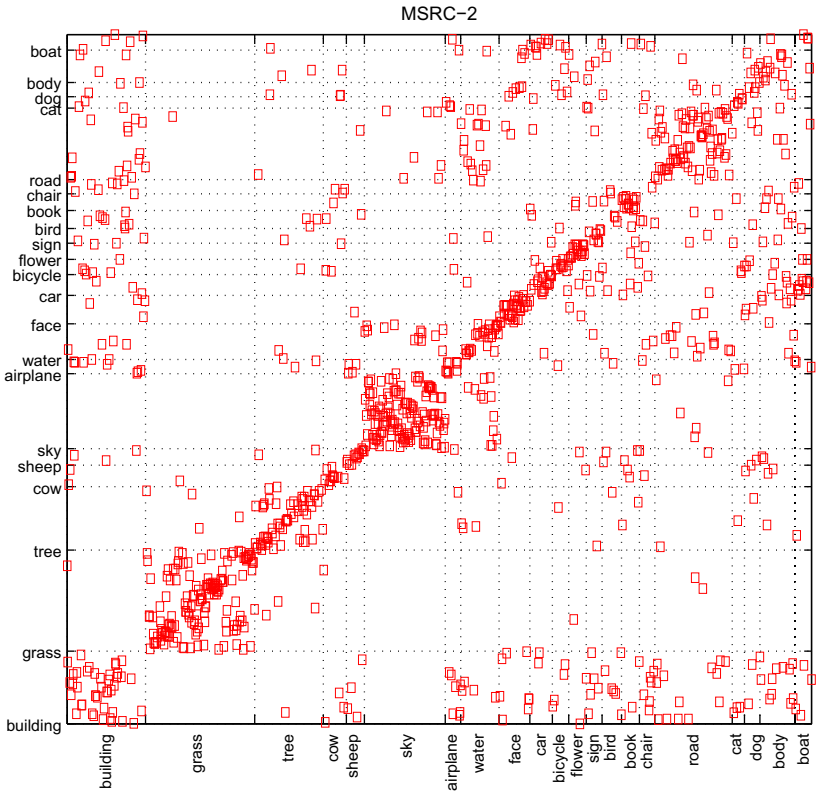

Fig. 10. Visualization of the difficulty of the RL problem for the MSRC-2 data set.

we evaluated the *hierarchical error* for both OVA-RF and OVA-RF+EBM. Specifically, we measured the percentage of test regions that were annotated with a *related* label (i.e., the *parent*, the *child* or the *brother*) of the correct one, according to the conceptual hierarchy defined in [4]. This measure indicates how close was the labeling provided by an RL method to the correct one in terms of the annotation hierarchy.

We found that before applying the EBM method, about 50.25% of the test regions were assigned a related label, after applying the EBM 49.59% of the labels were assigned a related one, thus there was no improvement/loss in terms of visual-semantics. If we just consider those labels that can be improved by the EBM (i.e., those test regions for which the correct label was in its set of top–k candidate labels), then 82.9% and 83.24% of these regions were annotated with a related label, for OVA-RF and OVA-RF+EBM respectively. Hence, only about 17% of the test regions, that can be improved, have been assigned a *unrelated* label; which complicated the job of the EBM because it was attempting to refine the labeling at a very fine granularity. For example, the EBM was trying to correct regions whose correct label is *woman* and the predicted output is *man*; despite being wrong, the confidence of classifiers and, what is more, the co-occurrence statistics are very similar for both labelings.

Fig. 11 shows sample test-set images as annotated with OVA-RF (middle column) and OVA-RF+EBM (rightmost column). We can see that the labeling provided by OVA-RF and OVA-RF+EBM is very similar (semantically) to that of the ground-truth data (leftmost column). EBM improves the initial labeling of the OVA-RF method for some images while it makes it wrong for others. However, we can see that for most of the labels that were changed the new-labels are related to the ground-truth label (either semantically or visually). For example, in the first row both methods misclassify the *woman* region by assigning the label *man*; in the second row, the *child-girl* region is labeled with *child-boy* by the EBM; in row 3 the wrong annotation *man* is substituted by *trees*, that despite being closely related to the true label (*tree*) it is judged as wrong.

6.3. Comparison with related methods

In this section we compare the RL performance of OVA-RF and OVA-RF+EBM to other methods that have used the databases described in Table 1. The goal of these experiments is to show that the performance obtained by our methods is comparable and even better to other approaches that are more complex (in terms of the information they need and the way they face the labeling problem) and more difficult to implement. For each subset we consider for comparison the result of the method that, to the best of our knowledge, has reported highest accuracy on the respective data set. One should note that the results reported in this section should be considered illustrative, in the sense that different refinement methods should be used with the same OVA-RF classifier in order to precisely determine what method performs better than other. A straightforward comparison between the EBM and other labeling refinement methods is presented in [45].

For the Corel^R data sets we considered the results reported by Hernandez and Sucar [15] for the COREL-AN subset and those results of Escalante et al. [16] for the rest. For the SCEF-I and SCEF-II collections we considered the best results reported on a recent comparison of region-labeling refinement methods [14]. For the MSRC-1 collection we considered the best results reported by Winn et al. [18]. For the MSRC-2 data set we considered the region-labeling results reported by Shotton et al. [19]. Finally, for the VOGEL collection we considered the results published by Vogel et al. [17]. We do not report a comparison for the SAIAPR TC-12 collection because this is the first time this collection is used. Table 7 shows the results of the comparison.

As we can see the performance of the OVA-RF approach was quite competitive itself, outperforming the state-of-the-art results

Fig. 11. Sample test set images. Left: ground truth labeling, middle: OVA-RF annotation and right: OVA – RF + EBM labeling.

in 8 out of the 11 data sets, however, the difference was not statistically significant. The EBM outperformed the considered methods in 9 out of 11 data sets, this global difference was statistically significant (recall the difference between OVA-RF and OVA-RF+EBM was statistically significant as well, see Section 6.2). The average improvement of the EBM over the OVA-RF method was of 8.6%, whereas the average improvement of the EBM over the considered related methods was of 11.26%.

Table 7
Comparison of RL accuracy of our methods against related works. The second column shows the performance reported in related works, together with the corresponding reference. The third and fourth show the RL accuracy obtained by OVA – RF and OVA – RF + EBM, respectively. Further, in columns 3 and 4 we show between parentheses the relative improvement obtained by OVA – RF and OVA – RF + EBM over the reference accuracy (column 2).

Collection	Reference	OVA – RF	OVA – RF + EBM
COREL-AN	45.64% [15]	57.90% (26.86%)	**58.97% (29.21%)**
COREL-AG	50.50% [16]	56.56% (12.00%)	**57.23% (13.33)%**
COREL-BN	39.50% [16]	46.65% (18.10%)	**48.74% (23.39%)**
COREL-BG	43.00% [16]	46.08% (7.16%)	**46.87%(9.00%)**
COREL-CN	42.50% [16]	54.13% (27.36%)	**55.59% (30.1%)**
COREL-CG	47.50% [16]	52.28% (10.06%)	**52.71% (10.97%)**
SCEF-I	**60.94%** [14]	59.99% (−1.55%)	60.35% (−0.96%)
SCEF-II	78.73% [14]	81.55% (3.58%)	**82.92% (5.32%)**
MSRC-I	**93.94%** [18]	86.60% (−7.81%)	88.82% (−5.45%)
MSRC-2	70.50% [19]	70.60% (0.14%)	**76.03% (7.84%)**
VOGEL	71.70% [17]	70.78% (−1.28%)	**72.54% (1.17%)**

The UVD-based method of Winn et al. outperformed the EBM by about 5.45% in the MSRC-1 collection, while the method due to Papadopulos et al. sightly outperformed EBM by 0.96%. The difference with Winn et al.'s method can be considered significative, nevertheless one should note that Winn et al. developed a method that learn features for the collection and task, resulting in a more computationally-expensive learning process and a more complicated implementation. In this respect, Shotton et al. have reported results on Winn et al.'s method over the MSRC-2 collection, the reported performance is of 67.6%; compared to our result in MSRC-2, the EBM outperformed Winn et al.'s method by more than 10%; thus giving evidence of the specificity of the UVD-based method for the MSRC-1 collection and its degraded performance when considering more classes.

We would like to emphasize that each of the methods considered for comparison have been proposed specifically for certain collections and the authors of such works have not tested their methods in other collections. On the other hand, the EBM method has been used similarly for all of the data sets we considered, with minor variations in the λ parameter, no further ad-hoc information or methodologies were adopted, thus proving the generality of our approach.

6.4. Annotation based image retrieval

In this section we report results on ABIR using the annotations provided by OVA-RF and OVA-RF+EBM for the SAIAPR TC-12 benchmark. The goal of these experiments is to show how the labels as generated by the EBM can be helpful for image retrieval by using keywords, regardless of the apparent low region-labeling performance reported in Section 6.2.

We evaluated the retrieval performance, under ABIR, when using labels generated with OVA-RF and OVA-RF+EBM by means of a 10-fold CV loop. Specifically, we repeated 10 trials (each using 90% of data for training and 10% for testing, where the data splits were different in each iteration) of the following process. The training subset was used for training an RF classifier; next the trained classifier was tested in the corresponding testing subset (OVA-RF); the output of the RF classifier was refined with the EBM (OVA-RF+EBM); then we retrieved images from the test set (as described below) using the labels generated with OVA-RF and OVA-RF+EBM; finally, we evaluated the retrieval performance; we report results averaged over the 10 testing sets.

The specific retrieval setting we adopted is as follows. For each of the 90 labels considered for the SAIAPR TC-12 collection, we used the label as query and retrieved those images (from the testing set) containing regions that were labeled (with either OVA-RF or OVA-RF+EBM) with such query-label. Then we evaluated the retrieval performance for that query by considering as relevant those retrieved images that, according to the ground-truth data, contain regions labeled with the corresponding label. We report precision, recall and the f_1 measure averaged over all of the queries (labels). Precision (P) is defined as the fraction of retrieved images that were indeed relevant; recall (R) is defined as the proportion of relevant retrieved images to the total of relevant images; the f_1 measure is defined as $f_1 = \frac{2 \times P \times R}{P+R}$. Table 8 reports averaged results of this experiment.

The best retrieval results were obtained when we used labels generated with the EBM; the difference was greater in recall, which means that more regions in images were labeled with the correct labels. Figs. 12 and 13 show the per-label average precision and recall, respectively, for images labeled with OVA-RF+EBM. We can see that for most labels both precision and recall are acceptable. As expected, precision and recall were greater for labels with many examples in the SAIAPR TC-12 collection [4].

Makadia et al. have reported ABIR results by using a KNN-based retrieval method, over 291 labels, on the IAPR TC-12 collection [46]. Despite such results are not directly comparable, they can serve to give us an idea of the meaning of the results shown in Table 8. The best average precision reported therein is of 0.28 while the best recall is 0.29, these results are much lower than ours; giving evidence on the usefulness of labels generated by OVA-RF+EBM for ABIR. One should note that the labels considered by Makadia et al. have been manually assigned, which alleviated for the authors the annotation uncertainty inherent in our OVA-RF+EBM method. Fig. 14 shows sample images retrieved for some selected labels; the three leftmost images show relevant retrieved images whereas the three rightmost images are non-relevant retrieved images. We can see, that even some (judged) irrelevant images are closely related to the relevant ones. We are currently exploring more elaborated techniques for performing ABIR.

Table 8
Results on ABIR for OVA-RF and OVA-RF+EBM.

Method	Precision	Recall	f_1
OVA − RF	0.2995	0.3640	0.3286
OVA − RF + EBM	**0.3049**	**0.3847**	**0.3402**

6.5. Discussion

We have presented experimental results that show the usefulness of OVA-RF and OVA-RF+EBM for RL; this subsection summarizes our main findings.

- We provided evidence suggesting that OVA, and in particular OVA-RF, offer a tremendous potential of improvement that can be exploited by techniques similar to ours for the task of RL. The random forest (RF) classifier was particularly well suited for OVA multiclass classification as it has shown satisfactory performance on highly imbalanced data sets. The PSMS method outperformed RF when the binary classifiers were compared one-to-one, although the different scales made difficult to apply EBM with PSMS-based classifiers.
- We showed that the proposed EBM is able to improve the initial labeling performance provided by OVA-RF; the improvements over 12 data sets were statistically significant. Our results suggest that the EBM does not depend on the segmentation method or the number of classes, what is more, the EBM was successfully applied to all databases, regardless of the diversity of

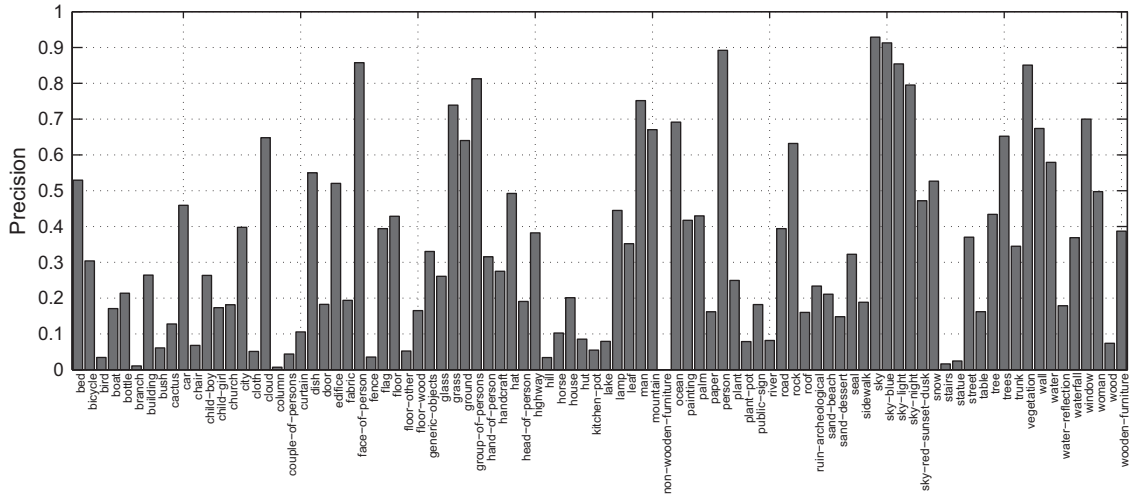

Fig. 12. Average per-label precision on ABIR for images labeled with OVA − RF + EBM.

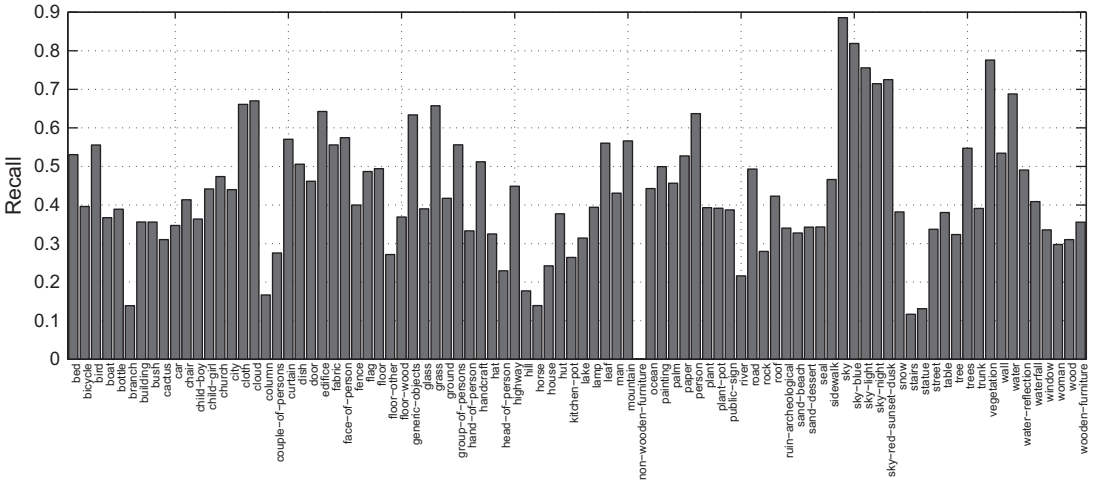

Fig. 13. Average per-label recall on ABIR for images labeled with OVA – RF + EBM.

Fig. 14. Sample relevant (columns 1–3) and non-relevant (columns 4–6) retrieved images for a few labels under ABIR with labels generated with OVA – RF + EBM.

content, resolution, features and number of regions per image, giving evidence of the generality of the method and its robustness to varied settings.

- We compared the proposed EBM with related techniques (some of which have been specifically designed for a particular data set), the EBM showed comparable and even superior performance in 9 out 11 databases (this difference was statistically significant). This result gives evidence that our method, despite being generic and requiring little information, can outperform more specialized strategies on region-level AIA.
- Finally, our results on ABIR showed that labels generated by the OVA-RF+EBM method can be useful as to support keyword-based image search. Given the low labeling performance on the SAIAPR TC-12 collection, this is a very interesting result.

Our results are superior to that reported by other researchers that have used manually labeled images for ABIR on the same collection.

7. Conclusions

We have introduced an energy-based model for generic region labeling. The proposed method refines the output of multiclass classification methods based on the one-vs-all formulation. Intuitively, the EBM maximizes the semantic cohesion among labels assigned to neighboring regions in images; taking into account statistics about the association between labels together with the predictions of the base classifier. Additionally, we presented an analysis on the suitability of OVA classification for the task of RL.

For our experiments we considered 12 data sets that show a considerable diversity in terms of number of regions, number of labels, labeling granularity, segmentation methods, resolution of images and domains. Our results revealed interesting facts about OVA and the EBM for RL. On the one hand, we provided evidence that shows that the OVA approach offers an important potential of improvement, in terms of labeling performance, that can be exploited by refinement techniques similar to ours. In this respect, a random forest classifier proved to be particularly well suited for RL under OVA. On the other hand, experimental results show that our EBM can effectively improve the labeling provided by the base classifier, the difference in performance is statistically significant. The EBM is highly efficient and it can be applied without modifications to different data sets. The heterogeneity of the considered databases show the generality of our approach and its robustness to different scenarios. Furthermore, results on image retrieval, show that the labels, as generated with our EBM, can be helpful for annotation-based image retrieval. We compared the performance of our methods to that of other techniques that have used the same collections. The results show that OVA is a very competitive method, although the best results were obtained with the EBM, which outperformed significantly the considered techniques.

Several future work directions arose through the development of our research. We are working on a new energy function that can take advantage of the PSMS based classifiers, as they are better than RF ones when compared one-to-one. We are developing a new strategy for retrieving images based on the labels assigned by our method. We are also working on a comparison of different visual features for the task of RL in support of image retrieval.

Acknowledgements

We are grateful with Nando de Freitas, A. López, E. Morales, F. Trinidad and L. Villaseñor for their comments on a draft of this paper. We also thank the comments made by reviewers that have helped us to improve this paper. We thank J. Vogel, S. Carsten, G. Papadopoulos, M. Grubinger, J. Winn, J. Shotton, K. Barnard, and P. Carbonetto, for making their data sets available. This project was supported by CONACyT under project grant 61335 and scholarship 205834. This work was done while Hugo Escalante was PhD student at INAOE.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D.A. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.
[2] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 1–60.
[3] H.J. Escalante, M. Montes, L.E. Sucar, Multimedia document indexing based on semantic cohesion, Information Retrieval, 2011, accepted for publication.
[4] H.J. Escalante, M. Grubinger, C.A. Hernández, J.A. González, A. López, M. Montes, E. Morales, L.E. Sucar, L. Villaseñor, The segmented and annotated IAPR TC-12 benchmark, Computer Vision and Image Understanding 114 (2010) 419–428.
[5] M. Inoue, On the need for annotation-based image retrieval, in: Proceedings of the ACM-SIGIR Workshop on Information Retrieval in Context, Sheffield, UK, 2004, pp. 44–46.
[6] J. Jeon, V. Lavrenko, R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models, in: SIGIR'03: Proceedings of the 26th International ACM-SIGIR Conference on Research and Development on Information Retrieval, Toronto, Canada, 2003, pp. 119–126.
[7] A. Hanbury, A survey of methods for image annotation, Journal of Visual Languages and Computing 19 (5) (2008) 617–627.
[8] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, J. Kaufhold, Evaluation of localized semantics: data, methodology, and experiments, International Journal of Computer Vision 77 (1–3) (2008) 199–217.
[9] H.J. Escalante, M. Montes, L.E. Sucar, Word co-occurrence and Markov random fields for improving automatic image annotation, in: Proceedings of the 18th British Machine Vision Conference, vol. 2, Warwick, UK, 2007, pp. 600–609.
[10] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, M. Strintzis, Combining global and local information for knowledge-assisted image analysis and classification, EURASIP Journal on Advances in Signal Processing 2007, 2007, p. 15 (Article ID 45842).
[11] C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 2008, pp. 1–8.
[12] C. Saathoff, M. Grzegorzek, S. Staab, Labeling image regions using wavelet features and spatial prototypes, Proceedings of the 3rd International Conference on Semantic and Digital Media Technologies: Semantic Multimedia, LNCS, vol. 5392, Springer, Koblenz, Germany, 2008, pp. 89–104.
[13] C. Saathoff, S. Staab, Exploiting spatial context in image region labelling using fuzzy constraint reasoning, in: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, IEEE, Klagenfurt, Austria, 2008, pp. 16–19.
[14] G. Papadopoulos, C. Saathoff, M. Grzegorzek, V. Mezaris, I. Kompatsiaris, S. Staab, M. Strintzis, Comparative evaluation of spatial context techniques for semantic image analysis, in: Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services, IEEE, London, UK, 2009, pp. 161–164.
[15] C. Hernandez, L.E. Sucar, Markov random fields and spatial information to improve automatic image annotation, Proceedings of the 2007 Pacific-Rim Symposium on Image and Video Technology, LNCS, vol. 4872, Springer, Santiago, Chile, 2007, pp. 879–892.
[16] H.J. Escalante, M. Montes, L.E. Sucar, Multi-class PSMS for automatic image annotation. Swarm Intelligence Journal, 2011.
[17] J. Vogel, B. Shiele, Semantic modeling of natural scenes for content-based image retrieval, International Journal on Computer Vision 72 (2) (2007) 133–157.
[18] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Proceedings of IEEE International Conference on Computer Vision, Beijing, China, 2005, pp. 1800–1807.
[19] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling appearance, shape and context, International Journal on Computer Vision 81 (1) (2008) 2–24.
[20] A. Llorente, R. Manmatha, S. Rüger, Image retrieval using Markov random fields and global image features, in: Proceedings of the ACM International Conference on Image and Video Retrieval, Xian, China, 2010, pp. 243–250.
[21] Y. Jiang, J. Wang, S. Chang, C. Ngo, Domain adaptive semantic diffusion for large scale context-based video annotation, in: Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 1420–1427.
[22] Y.J. Lee, K. Grauman, Object-graphs for context-aware category discovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 1–8.
[23] A. Llorente, S. Overell, H. Liu, R. Hu, A. Rae, J. Zhu, D. Song, S. Rüger, Exploiting term co-occurrence for enhancing automated image annotation, Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the CLEF, LNCS, vol. 5706, Springer, 2009, pp. 632–639.
[24] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human–object interaction activities, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 17–24.
[25] R. Rifkin, A. Klautau, In defense of one-vs-all classification, Journal of Machine Learning Research 5 (2004) 101–141.
[26] W. Li, M. Sun, Semi-supervised learning for image annotation based on conditional random fields, in: Proceedings of the International Conference on Image and Video Retrieval, LNCS, vol. 4071, Tempe, AZ, USA, 2006, pp. 463–472.
[27] J. Yuan, J. Li, B. Zhang, Exploiting spatial context constraints for automatic image region annotation, in: Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany, 2007, pp. 595–604.
[28] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001.
[29] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general context object recognition, in: Proceedings of the 8th European Conference on Computer Vision, LNCS, vol. 3021, Springer, Prague, Czech Republic, 2004, pp. 350–362.

[30] Y. LeCun, S. Chopra, R. Hadsell, M.A. Ranzato, F.J. Huang, Energy-based models, in: Predicting Structured Data, MIT Press, 2007, pp. 191–246 (Chapter 10).
[31] G. Winkler Image Analysis, Random Fields and Markov Chain Monte Carlo Methods, Number 27 in Applications of Mathematics, Springer, 2006.
[32] J. Besag, On the statistical analysis of dirty pictures, Journal of the Royal Statistical Society, Series B 48 (1986) 259–302.
[33] S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[34] Yuri Boykov, Olga Veksler, Ramin Zabih, Efficient approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (12) (2001) 1222–1239.
[35] M. Grubinger, Analysis and Evaluation of Visual Information Systems Performance, PhD Thesis. School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.
[36] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.
[37] V. Mezaris, I. Kompatsiaris, M. Strintzis, Still image segmentation tools for object-based multimedia applications, International Journal of Pattern Recognition and Artificial Intelligence 18 (2004) 701–726.
[38] A. Saffari, I. Guyon. Quickstart Guide for CLOP, Tech. Rep., Graz University of Technology and Clopinet, Graz, Austria, 2006.
[39] H.J. Escalante, M. Montes, L.E. Sucar, Particle swarm model selection, Journal of Machine Learning Research 10 (2009) 405–440.
[40] L. Breiman, Random forest, Machine Learning 45 (1) (2001) 5–32.
[41] C. Dahinden, Classification with tree-based ensembles applied to the WCCI 2006 performance prediction challenge datasets, in: Proceedings of the International Joint Conference on Neural Networks, Vancouver, BC, Canada, 2006, pp. 1669–1672.
[42] S. Dudoit, J. Fridlyand, Classification in Microarray Experiments, Chapter in Statistical Analysis of Gene Microarraty Data, CRC Press, 2002.
[43] I. Guyon, A. Saffari, G. Dror, Gavin Cawley, Analysis of the IJCNN 2007 competition agnostic learning vs. prior knowledge, Neural Networks 21 (2–3) (2008) 544–550.
[44] J. Demsar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
[45] G.Th. Papadopoulos, C. Saathoff, H.J. Escalante, V. Mezaris, I. Kompatsiaris, M.G. Strintzis. A comparative study of spatial context techniques for semantic image analysis, Computer Vision and Image Understanding, submitted for publication.
[46] A. Makadia, V. Pavlovi, S. Kumar, A new baseline for image annotation, in: Proceedings of the 10th European Conference on Computer Vision, LNCS, vol. 5304, Springer, Marseille, France, 2008, pp. 316–329.