Word Co-occurrence and Markov Random Fields for Improving Automatic Image Annotation

H. Jair Escalante, Manuel Montes and L. Enrique Sucar
Computer Science Department
National Astrophysics, Optics and Electronics Institute
Puebla, 72840, México,
hugojair@ccc.inaoep.mx, {mmontesg,esucar}@inaoep.mx

Abstract

In this paper a novel approach for improving automatic image annotation methods is proposed. The approach is based on the fact that accuracy of current image annotation methods is low if we look at the most confident label only. Instead, accuracy is improved if we look for the correct label within the set of the top$-k$ candidate labels. We take advantage of this fact and propose a Markov random field (*MRF*) based on word co-occurrence information for the improvement of annotation systems. Through the *MRF* structure we take into account spatial dependencies between connected regions. As a result, we are considering *semantic* relationships between labels. We performed experiments with iterated conditional modes and simulated annealing as optimization strategies in a subset of the Corel benchmark collection. Experimental results of the proposed method together with a $k-$nearest neighbors classifier as our annotation method show important error reductions.

1 Introduction

The task of assigning semantic labels (words) to images is known as image annotation. This is a very important step towards developing more precise image retrieval systems. For text-based image retrieval systems, annotations are indispensable features; while for content-based image retrieval methods, annotations can provide them with semantic information for improving their performance. Image annotation, however, is not an easy task; manual annotation is both infeasible for large collections and subjective. Therefore, there is an increasing interest in developing automatic methods for image labeling.

There are two ways of facing this problem, at image level and at region level. In the first case, labels are assigned to the entire image as an unit, not specifying which words are related to which objects within the image. In the second approach, which can be conceived as an object recognition task, the assignment of labels is at region level; providing a one-to-one correspondence between words and regions. The last approach can provide more semantic information for the retrieval task, although it is more challenging than the former. Within the region-level automatic image annotation (*AIA*) task, we can distinguish two approaches for assigning labels to regions, these are soft and hard annotation. Hard

BMVC 2007 doi:10.5244/C.21.60

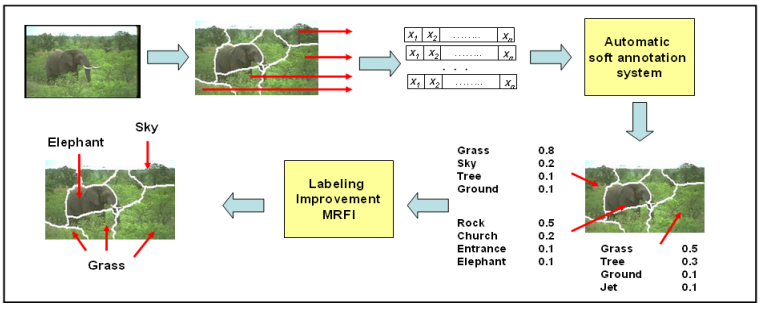

Figure 1: Graphical schema of our approach. We start from an image that is segmented into regions; attributes are obtained from each region; next these attributes are used with a soft-*AIA* method that returns a set of candidate labels, together with a relevance weight, for each region in the image. Then the method proposed in this paper is applied, and it returns an unique correct label for each image.

annotation consist of the task of assigning, with probability 1, an unique label to each region; soft annotation, on the other hand, ranks the labels according to their relevance to being the correct annotation for a given region. Accuracy of soft annotation systems is superior to that of hard systems, though assigning a set of labels to a single region is both confusing and impractical. On the other hand, accuracy of hard annotation systems is poor, though it is more understandable and practical assigning a unique label to each region.

In order to take advantage of the high precision of soft annotation methods as well as the clarity of hard approaches, we propose *MRFI*, a probabilistic model based on word co-occurrence information for improving image annotation systems. *MRFI* considers the top$-k$ candidate labels for each region within an image and, by using word co-occurrence information together with spatial context, it re-ranks each candidate label. Then we select the unique top label for each image, according to this ranking. In Figure 1 the proposed approach for improving *AIA* methods is graphically described. We used a $k-$nearest neighbor classifier as our *AIA* system and experiments on a subset of the benchmark Corel collection were performed. Experimental results show significant improvements by using *KNN+MRFI* over single *KNN*, furthermore *KNN+MRFI* outperforms several others state of the art annotation methods.

The rest of this document is organized as follows. In the next Section we review related work. In Section 3 some background information is described. Next in Section 4 the *MRFI* method is proposed. Then in Section 5 experimental results are presented. Finally, in Section 6 conclusions and future work directions are discussed.

2 Related work

A wide variety of methods for image labeling have been proposed since the late nineties. However, none of current methods have taken advantage of label's semantics for improving their performances. A very early attempt that used word co-occurrence information is the work by Mori et al [13], in which every word assigned to the entire image is inherited by each region; regions are visually clustered and probabilities of the clusters given

each word are calculated by counting the occurrence of common words within these clusters. A recent approach that attempts to take advantage of co-occurrence information is that proposed by Li et al [12]. They use a probabilistic support vector machine classifier for ranking candidate labels for each region within an image. Co-occurring words in the candidate labels for regions in the same image are weighted high; then candidate labels are re-ranked, top ranking labels are assigned as annotation for the entire image. Our approach is different to the previous methods because we obtained the co-occurrence information from an external corpus and considered spatial dependencies between connected regions. Instead of just considering co-occurrence of labels within the same image [12] or clusters of regions [13]. Moreover in such works co-occurrence information is used ad-hoc for their annotation method; while in this work we propose a method that can be used with other soft-annotation systems.

A work close in spirit to ours is due to Carbonetto et al [4]. In this work the authors introduce spatial information into a *MRF* for object recognition. This approach is different to the one we adopted; since Carbonetto et al define the potential function for discovering the unknown association between visual features extracted from each region and the considered labels; furthermore the *MRF* is entirely based on a single collection of annotated images. While in this work we use semantic information, obtained from an external source, for modeling word association between neighboring regions. Dealing with a different problem: that of selecting an unique label given a set (a subset of the vocabulary) of candidate ones; which can be seen as a re-ranking strategy. Conditional random fields (*CRF's*) have also been applied to pixel-level image labeling [9], and object recognition [14]. These works have obtained positive results in different scenarios, although their applicability is still limited to segmentation ([9]) and two-class object recognition ([14]). However using conditional random fields for *AIA* can be an immediate future work direction. The above described approaches take into account dependencies between connected regions [4, 9, 14]; although none of these have used semantic knowledge together with spatial context for improving performance of object recognition methods. *MRFI*, on the other hand, does not attempt to induce the *visual-features to word* relationship by considering spatial information. Instead *MRFI* takes advantage of semantic information and attempts to select the best configuration of labels for the regions contained in the same image. Semantic information is obtained off-line from a word co-occurrence matrix calculated from an external collection of manually annotated images.

3 Background

3.1 KNN as annotation system

The $k-$nearest neighbors (*KNN*) classifier is an instance based learning algorithm widely used in machine learning tasks. In this work we used this method as our annotation system due to the fact that it can outperform other state of the art methods (see Section 5); furthermore, *KNN* can be adapted to work in the hard and soft annotation schemas.

KNN starts from a training data set $\{X,Y\}$ consisting of N pairs of examples of the type $\{(x_1,y_1),\ldots,(x_N,y_N)\}$, with the $x_i's$ being $d-$dimensional feature vectors and the $y_i's$ being the labels of $x_i's$. In this work each x_i contains visual attributes extracted from a region. While each y_i is one of the $|V|$ labels we can assign to a region. The training phase of *KNN* consist of storing all available training instances. When a new instance, x_t,

needs to be classified *KNN* searches, in the training set, for $\{x_1^t, \ldots x_k^t\}$, the top $k-$objects more similar to x_t; then in a hard annotation schema it assigns to x_t the class of the most similar neighbor in the training set, we call this approach *1-NN*.

In order to apply *MRFI* with *KNN* as annotation method we need to turn *KNN* into a soft-annotation method. That is, candidate words for a given region should be ranked and weighted according to the relevance of the labels to being the correct annotation for such a region. We used the distance of the test instance to the top$-k$ nearest neighbors as relevance weight. In this way we can infer relevance weights directly related to the proximity of the neighbor to the test instance. Relevance weighting is obtained using Equation (1)

$$P^R(y_j^t) = \frac{d_j(x^t)}{\sum_i^k d_i(x^t)} \tag{1}$$

with $d_j(x^t)$ being the inverse of the Euclidean distance in the attribute space of instance x_j^t, within the $k-$nearest neighbors, to x^t, the test instance. As we can see, the sum of the priors for all the candidate labels is one, therefore this relevance weighting of *KNN* can be taken as the prior probability for the *MRFI* method. Note that this relevance weight is accumulative; that is, labels appearing more than once will accumulate their weights according to the times they appear in the top$-k$ labels. In this way we are implicitly accounting for repeated labels.

3.2 Obtaining co-occurrence information

Word co-occurrence is a form of word association that has been widely used by information retrieval models [1]. In the simpler schema, bags of words of documents and queries are compared (that is, word co-occurrences are calculated) for retrieving the documents whose bags of words are more *similar* to that of the query. This form of word association can be used with labels in the vocabulary for *AIA* tasks for taking into account semantic information between neighboring labels.

The co-occurrence information matrix (M_c) is a $|V|_X|V|$ square matrix in which each entry $M_c(w_i, w_j)$ indicates the number of documents (counted on an external corpus) in which words w_i and w_j appeared together. That is, we considered each pair of words $(w_i, w_j) \in V_X V$ and searched for occurrences, at document level, of (w_i, w_j). We did this for each of the $|V| * |V|$ pairs of words and for each document in our textual corpus. The collection of documents we considered for this work was the set of captions of a new image retrieval corpus: the *IAPR-TC12* [8] benchmark. This collection consists of around $20,000$ images that were manually annotated, at image level; therefore, if two words appear together in the captions of such collection, they are very likely to be visually related. Captions consist of a few text lines indicating visual and semantic content. From the entries of the M_c matrix we can estimate conditional and joint probabilities if we take: $P(w_i|w_j) = \frac{P(w_i,w_j)}{P(w_j)} \approx \frac{c(w_i,w_j)}{c(w_j)}$, and $P(w_i, w_j) \approx \frac{c(w_i,w_j)}{|D|}$, where $c(x,y)$ indicates the number of times x and y appear together in the corpus (that is, an entry of the M_c matrix); and $|D|$ is the number of documents in our textual corpus. If we repeat this process for each pair of words in the vocabulary we obtain a matrix of probabilities $((P_M))$, which may contain conditional or joint probabilities. Preliminary experiments showed that the use of conditional probabilities resulted in more significant improvements than those with joint probabilities; therefore, we used in this work conditional probabilities for (P_M).

A problem with the P_M matrix is the sparseness of data, that is, many entries of the matrix have zero values, which can affect the performance of our approach; this is a very common issue in natural language processing [6]. In order to alleviate this problem we applied a widely used smoothing technique known as interpolation smoothing [6], described on Equation (2)

$$P(w_i|w_j) \approx \Lambda * \frac{c(w_i,w_j)}{c(w_j)} + (1-\Lambda) * \frac{c(w_j)}{|W|} \tag{2}$$

where Λ is an interpolation parameter[1] and $|W|$ is the number of words in the collection. This formula is an interpolation between the empirical estimate ($\frac{c(w_i,w_j)}{c(w_j)}$) and the empirical distribution of the term w_j ($c(w_j)$). Therefore if two terms never co-occur in the co-occurrence matrix (M_c) we will not have a zero value in P_M.

4 *MRFI:* A Markov random field for improving *AIA*

A random field is a collection of random variables indexed by sites [11]. We consider a set of random variables $F = F_1, \ldots, F_M$ associated to each site in the site's system S. Each random variable takes a value f_i from a set of possible values L. A Markov random field (*MRF*) is a random field with the Markov property $P(f_i|f_{i-1}, f_{i-2}, \ldots f_1) = P(f_i|N(f_i))$, where $N(f_i)$ is the set of neighbors of f_i. A typical application of *MRF's* is to obtain the most probable configuration (F^*) for the *MRF*; given some restrictions represented by local probabilities, also known as potentials. We can express the joint probability of a *MRF*, $''F''$, given the observation, $''G''$, as the product of the potentials:

$$P_{F|G}(f) = \nu \prod_c P_c(X) \tag{3}$$

With ν constant, potentials ($P_c(X)$) can be thought of as restrictions that will favor or punish certain configurations of F. In this way, F^* can be considered as the configuration that have the highest compatibility with the local probabilities ($P_c(X)$). We can express the potentials as energy functions in exponential form, that is: $P_c(X) = e^{-U_c(X_c)}$, with $U_c(X_c)$ being an energy function. Then using Equation (3) we have an unique energy function $U_p(f) = \sum_c U_c(X_c)$. In consequence Equation (3) can be reformulated as:

$$P_{F|G}(f) = \frac{1}{Z} * \exp^{-U_p(f)} \tag{4}$$

with Z being a normalization constant. For a first order neighborhood, as the one we considered in this work, we have:

$$U_p(f) = \sum_c V_c(f) + \lambda \sum_o V_o(f) \tag{5}$$

Where V_c corresponds to P_F, the domain information given by the neighbors; and V_o corresponds to $P_{G|F}$, the information given by the observations; λ is a constant that weights the contribution of each term. In our case, we would like to select the best configuration of labels assigned to the regions in each image. Making a compromise between the visual

[1] Usually the value of Λ is chosen empirically. Intuitively a low value of Λ should be used with sparser data. After a few trial and error experiments we selected $\Lambda = 0.5$.

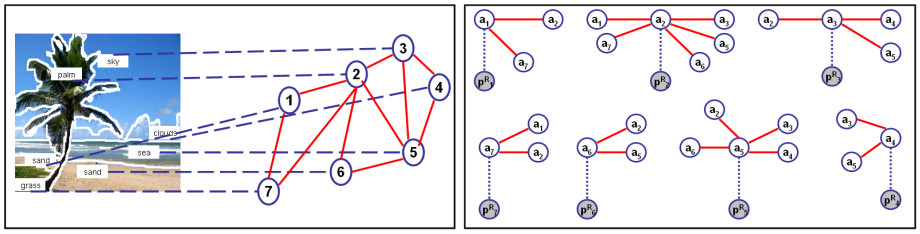

Figure 2: Left: graphical interpretation of *MRFI* for a given configuration of labels and regions. Right: spatial dependencies are shown for this configuration. The p_o^R's correspond to the relevance weight attached to each candidate label; the $a_i's$ represent the unknown association between connected regions.

properties of the region (V_o) and the semantics of its neighboring regions (V_c). Therefore, we used the above described framework for approaching this problem.

The observed variables in our task are the relevance weight attached to each label $p_1^R, \ldots p_{M_n}^R$, for each region R; and the top$-k$ candidate labels $w_1, \ldots w_K$, for each region. Observing this variables we define potential functions that exploit spatial dependencies between labels assigned to spatially connected regions within each image. The structure of *MRFI* and the dependencies it consider are shown in Figure 2. For this work we consider a region r_i is connected (spatially related) to another region r_j, if r_i is *next-to* r_j. Note that the next-to relation is symmetric and that *MRFI* depends on the segmentation. Moreover *MRFI* can not deal with problems like over-segmentation. However, as we will see in Section 5, if we have no available an accurate segmentation tool we can always divide an image into squared patches. Although poor, the use of this simple partition in *AIA* has outperformed methods based on sophisticated algorithms just has normalized cuts (see Section 5 and [4, 3]). Also we can make the square patches as small as we want; smaller patches will provide finer grain segmentations. Potentials for *MRFI* are defined in Equations (6) and (7) for the consideration of context and observation information, respectively.

$$V_c(f) = \sum_c (P(w_c|w_i))^n \tag{6}$$

$$V_o(f) = \left(\frac{1}{p_o^R(w_i)}\right)^n \tag{7}$$

Conditional probabilities in Equation (6) are obtained from the word co-occurrence matrix, as described in Section 3.2. While relevance weights p_o^R's, are obtained from the *AIA* system. The problem of selecting the correct annotation for each region within a given image reduces to the selection of the configuration that minimizes Equation (5). The selection of this *optimal* configuration is solved by standard optimization algorithms. In this work we performed experiments with two widely used algorithms: iterated conditional modes (*ICM* [2]) and simulated annealing with metropolis criteria (*SA* [10]). In Section 5 we report results of experiments with these two search strategies.

5 Experimental results

In order to evaluate the performance of *MRFI* several experiments on a subset of the Corel collection were performed. The data set we used is described in Table 1. It is a single

Data set	# Images	Words	Training blobs	Testing blobs
A-NCUTS	205	22	1280	728
A-P32	205	22	3288	1632

Table 1: Subset of the Corel image collection we used in the experimentation with *KNN-MRF*

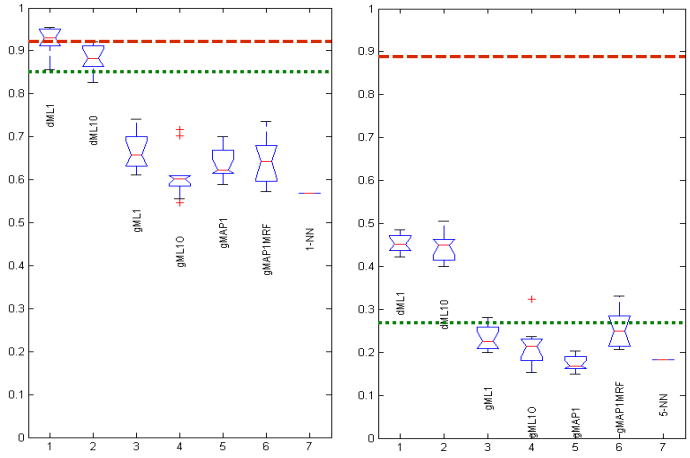

Figure 3: Comparison of *KNN* against other semi-supervised methods (*dML1* [7]; *dML1O*,*gML1*,*gMLO*, [3]; *gMAP1* [5]; *gMAP1MRF* [4]), using a Box-and-Whisker plot. The central box represents the values from the 25 to 75 percentile, outliers are shown as separate points. Left: accuracy at the first label. Right: accuracy at the top-5 labels. The upper dotted line represents a random bound, while the bottom dotted line represents a naïve method that always assigns the same label to all regions.

data set composed of 205 images segmented with normalized cuts [15] (*A-NCUTS*) and grid segmentation (*A-P32*). The attributes we considered for each region are the following: area, and color attributes. First we compared *KNN* against other semi-supervised object recognition methods [7, 4, 5, 3] (see caption of Figure 3), which are extensions and modifications to the reference work proposed by Duygulu et al [7]. In order to provide an objective comparison, we used the code provided by P. Carbonetto[2]. This code includes implementations of the above mentioned methods. In Figure 3 a comparison between *KNN* and the semi-supervised methods for the *A-NCUTS* data set is shown. In this plot, error is computed using the following equation:

$$e = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M_n}\left(1-\delta(a_{nu}^{-}=a_{nu}^{max})\right) \tag{8}$$

where M_n is the number of regions on image n, N is the number or images in the collection; and δ is an error function which is 1 if the predicted annotation a_{nu}^{max} is the same as the true label a_{nu}^{-}. Results with the test sets are averaged over 10 trials.The left plot in Figure 3 shows error at the first label (*hard annotation*). Error is high for all of the methods we considered, however *1-NN* outperforms in average all of the semi-supervised approaches.

[2] $http://www.cs.ubc.ca/\sim pcarbo/$

Method	**k**	**Its**	λ	**n**	**Context**	*Time*	**Improved**	**#−runs**	**AVG-I**
ICM-P32	20	100	0.1	1	Next-to	1.8	134	4500	41.3
ICM-NCUTS	20	100	5	0.5	Full	0.78	56	4500	-0.7
SA-P32	20	50	0.1	2	Next-to	1.5	144	2700	98.6
SA-NCUTS	20	25	10	0.5	Next-to	0.5	54	2700	27.5

Table 2: Parameters for the best configurations. k is the number of candidate labels in *KNN*; *Its* is for iterations; λ and n are parameters for Equation (5); context indicates the type of neighborhood considered; time is the average time in seconds required to analyze an image with *MRFI*. *Improved* is the number of annotations improved. *#−runs* is the number of experiments performed and *AVG-I* is the total of annotation improvements averaged by #−runs

gMl0 is the closest in accuracy to *1-NN*, though it obtains an average error which is above *1-NN* by 4.5%. In the right plot of Figure 3 we consider a label is correctly annotated if the true label is within the top−5 candidate labels, (*soft annotation*). As we can see, error for all methods is reduced, this clearly illustrates the fact that accuracy of annotation systems is high considering a set of candidate labels instead of the first one. In this case *gMAP* [5] outperforms *5-NN* by 0.9% in average. All other approaches obtain a higher average error than that of *5-NN*.

In the second experiment we compared the performance of *KNN+MRFI* to that of *KNN* alone as well as to the previous methods. Note that we have several parameters to fix for *MRFI*. These are: k, the number of candidate labels for each region; λ and n, parameters for Equation (5); the number of iterations is a parameter for the optimization algorithms; furthermore, we performed experiments with spatial context (see Figure 2) and with full spatial context, that is, assuming all regions in an image are connected to each other. Given that *MRFI* is an efficient method we could perform many experiments with both data sets in order to determine the average improvement of *MRFI+KNN* over single *KNN*. The parameters of the best configurations for each data set considering both optimization strategies are shown in Table 2. We also show the average of accuracy improvement and processing time. From Table 2 we can point out several interesting observations. First, as expected, the more candidate labels we consider, the more improvements we gain. We performed experiments with $k \in \{3,5,10,20\}$ and the best results were obtained with $k = 20$. *ICM* needs a higher number of iterations to converge than *SA*. A small value of λ works well for the *P32* data set, which means that a small weight is given to the co-occurrence information. While a high value of λ performs better for *NCUTS*, giving more importance to co-occurrence information. We can see that for *NCUTS* a value of $n = 0.5$ performs well, while this parameter do not significantly affected the performance of *MRFI*. The use of spatial information, through the *next-to* relation, results in larger improvements than if consider each region is connected to each other in the image. Improvements are consistent through the number of experiments performed. The lowest average improvement was obtained with *ICM-NCUTS*. While with the grid segmented data (*P32*) we obtained the largest improvement, 98 annotations per run in average; which is a very significant improvement. An important result showed in Table 2 is the processing time[3] required to process an entire image with *MRFI*. These results show the efficiency of *MRFI*.

[3] All experiments were carried out on a PC with 1 GB in RAM and a 2.7 GHz *pentium*^{R} processor

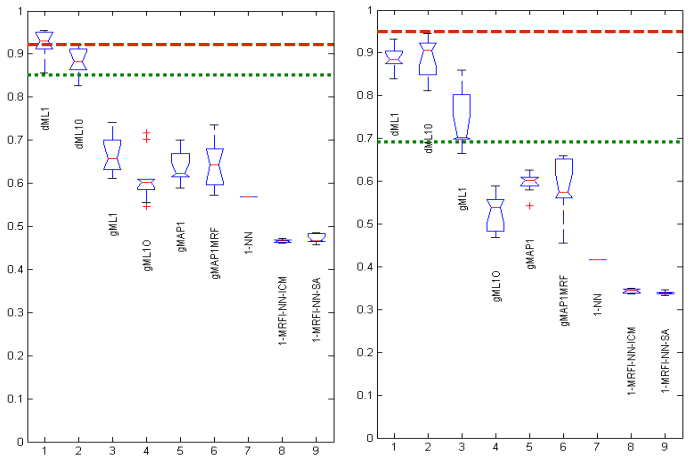

Figure 4: Comparison of *KNN* and *KNN+MRFI* against other semi-supervised methods (see caption of Figure 3) for images segmented with normalized cuts (left) [15] and with the grid approach (right); error is measured at the first label, see caption in Figure 3.

In all experiments performed using grid segmentation, which is faster than the other method, outperformed in accuracy segmentation with normalized cuts [15]. This result agrees with previous work [4, 5]. In *MRFI* this can be due to the fact that with grid segmentation (*P32*) the structure of the *MRF* is equal for all images. While for normalized cuts we have a different segmentation, according to the image's content, and therefore a different structure for the *MRF*. The use of *SA* instead of *ICM* does not result in significant improvements, *SA* outperformed *ICM* by 0.5%, which means that we have not many local minima. In Figure 4 we compare the best configurations of *MRFI* (Table 2) with the other methods. From Figure 4 we can clearly appreciate the improvement we can get by applying *MRFI+KNN*, instead of *1-NN* alone, for both data sets. The improvements of *MRFI+KNN* over *1-NN* are of 7.5% and 10.3% for the *P32* and *NCUTS* data sets, respectively. These percentages represent around 140 (for *P32*) and 46 (for *NCUTS*) annotations that were enhanced; this is a very significant improvement in accuracy. Furthermore, the difference in performance between *MRFI+KNN* and the other methods is dramatically increased. The semi-supervised method with closest average accuracy is *gML0*. *MRFI+KNN* improved *gML0* in average by 18.9% and 14.7% for the *P32* and *NCUTS* data sets, respectively. Results from this Section give evidence that *KNN+MRFI* is an effective image annotation method. Furthermore, *MRFI* can be applied with any other annotation system, though more experimentation should be performed in order to evaluate its impact with other methods.

6 Conclusions

We have presented *MRFI*, a method for the improvement of *AIA* systems. In *MRFI* spatial dependencies are considered through a *MRF* model. Semantic information between labels is incorporated using word co-occurrences. Co-occurrence information is calculated off-line from an external collection of captions, which is a novel approach. Experimen-

tal results of our method on a subset of the Corel collection, give evidence that the use of *KNN+MRFI* results in significant error reductions. Our method is efficient since the co-occurrence matrix is obtained off-line, and in most of the cases we just need a few iterations to obtain a good configuration (around 1.1 seconds per image). Furthermore, *MRFI* can be used with other *soft-annotation* systems.

The improvement of the co-occurrence matrix is an immediate step towards the enhancement of *MRFI*. Other future directions include the consideration of global image labels into *MRFI* and considering other models than *MRF's*, such as *CRF's* as well as experiments with probabilistic *AIA* methods.

Acknowledgements. We would like to thank K. Barnard, P. Carbonetto and M. Grubinger for making available their data and the reviewers by their useful commentaries that helped to improve this paper. This work was partially supported by CONACyT under grant 205834.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson E. L., 1999.

[2] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc. B*, 48:259–302, 1986.

[3] P. Carbonetto. Unsupervised statistical models for general object recognition. Master's thesis, C.S. Department, University of British Columbia, August 2003.

[4] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general context object recognition. In *Proc. of 8th ECCV*, pages 350–362, 2005.

[5] P. Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature eeighting for unsupervised learning. In *Proc. of the HLT-NAACL workshop on Learning word meaning from non-linguistic data*, pages 54–61, Morristown, NJ, USA, 2003.

[6] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. of the 34th meeting on Association for Computational Linguistics*, pages 310–318, Morristown, NJ, USA, 1996.

[7] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. 7th ECCV*, volume IV of *LNCS*, pages 97–112. Springer, 2002.

[8] M. Grubinger, P. Clough, and C. Leung. The iapr tc-12 benchmark -a new evaluation resource for visual information systems. In *Proc. of the International Workshop OntoImage'2006 Language Resources for CBIR*, 2006.

[9] X. He, R. Zemel, and M. Carreira. Multiscale conditional random fields for image labeling. In *Proc. of CVPR'04*, volume 2, pages 695–702. IEEE, 2004.

[10] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[11] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2nd edition, 2001.

[12] W. Li and M. Sun. Automatic image annotation based on wordnet and hierarchical ensembles. In *CICLING*, volume 3878 of *LNCS*, pages 417–428, Mexico, City, 2006.

[13] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *1st Int. Worksh. on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[14] A. Quattoni, M. Collins, and T. Darrel. Conditional random fields for object recognition. In *NIPS*, 2004.

[15] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI-IEEE*, 22(8):888–905, 2000.