Improving Automatic Image Annotation based on Word Co-occurrence

H. Jair Escalante, Manuel Montes, and L. Enrique Sucar

National Institute of Astrophysics, Optics and Electronics,
Computer Science Department
Puebla, 72840, México,
{hugojair,mmontesg,esucar}@ccc.inaoep.mx

Abstract. Accuracy of current automatic image labeling methods is under the requirements of annotation-based image retrieval systems. The performance of most of these labeling methods is poor if we just consider the most relevant label for a given region. However, if we look within the set of the top$-k$ candidate labels for a given region, accuracy of most of these systems is improved. In this paper we take advantage of this fact and propose a method (*NBI*) based on word co-occurrences that uses the naïve Bayes formulation for improving automatic image annotation methods. Our approach utilizes co-occurrence information of the candidate labels for a region with those candidate labels for the other surrounding regions, within the same image, for selecting the correct label. Co-occurrence information is obtained from an external collection of manually annotated images: the *IAPR-TC12* benchmark. Experimental results using a $k-$nearest neighbors method as our annotation system, give evidence of significant improvements after applying the *NBI* method. *NBI* is efficient since the co-occurrence information was obtained off-line. Furthermore, our method can be applied to any other annotation system that ranks labels by their relevance.

1 Introduction

Content based image retrieval (*CBIR*), the task of recovering images using visual features, has became an active research field since the early nineties [18, 24, 15, 9]. Typically, a query for a *CBIR* system consist of a visual example similar to the desired image and the task is to find, within the collection, images similar to such a visual example. However, this sort of querying is unnatural, since most of the time we would like to retrieve images by specifying queries in natural language (*images of a tiger, grass and water*), or even combining a sample image and natural language statements (*images of brown cows like in this photograph*). In consequence, it was recognized that in order to improve *CBIR* systems we would need to incorporate semantic information into the *CBIR* task. This semantic information generally consist of textual keywords (semantic descriptors, words, labels) indicating some semantic properties of the image.

Manually incorporating semantic information into images is both: expensive (in terms of human-hour costs) and subjective (due to the annotator criteria).

Therefore, recently there is an increasing interest on automatically assigning semantic information to images. The task of assigning semantic descriptors to images is known as image annotation or image labeling. There are two ways of approaching this problem: at image level or at region level. In the first case, often called weakly labeling, keywords 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 annotations is at region level within each image, providing a one-to-one correspondence between words and regions. This later approach is the one we considered in this work, since we believe that region-level semantics are more useful for discovering relationships between semantic concepts and visual objects within an image collection.

We say a region is correctly annotated if the more likely annotation, according to our annotation method, is the same as the true (manual) annotation. Most of the annotation methods that rank words according to their relevance to belong to a determined region fail in assigning the correct label just by taking the most confident word [17, 4, 5, 11, 16]. Accuracy improves if we search for the true label within the set of the top$-k$ ranked labels for a given region. However, assigning a set of $k-$labels to an unique region is confusing and unpractical. In this work we propose a method for improving automatic image annotation by taking advantage of word co-occurrence information. The problem we approach is the following: given a set of ranked candidate labels for a given image region, selecting the *(unique) correct label* for such region. The solution we propose takes advantage of word co-occurrence information of the candidate labels for the region we are analyzing and the corresponding candidate labels for other regions within the same image. We formulate this problem as a classification task using the naïve Bayes algorithm. Candidate labels are considered classes and measures of association between labels (based on word co-occurrences) are considered as attributes/features.

Our intuitive idea is that co-occurrence information between labels assigned to regions in the same image can help us to improve annotation accuracy; since annotations for regions appearing in the same image are very likely to be related. Given that word co-occurrences are obtained from the captions of a collection of manually-annotated images we can trust in that this information can give us an indicator of words association. Since words that tend to co-occur in the captions are very likely to be visually related, as they are used to describe the same image. We performed experiments on three subsets of the benchmark Corel data set ([18, 15, 6, 11, 14]), using a $k-$nearest neighbors (*knn*) [19] method as our annotation strategy. Experimental results show that *knn* combined with our approach can result in significant improvements over *knn* alone and over other *soft-annotation* methods as well. An advantage of the proposed approach is that it is simple and efficient as it is based on a naïve Bayesian classifier and the co-occurrence information is obtained off-line; also the approach can be applied to other annotation methods, provided they rank words for their relevance (soft-annotation).

The rest of this paper is organized as follows. In the next Section we briefly review related work on automatic annotation. In Section 3 we describe how the *knn* classifier was used for annotation. Next in Section 4, our proposal: *Naïve Bayesian Improver based on Co-occurrences* is described. In Section 5 we describe how we obtained the word co-occurrence matrix. Then in Section 6 we report experimental results on subsets of the Corel Collection. Finally, in Section 7 we present some conclusions and outline future work directions.

2 Automatic image annotation

A wide variety of methods for image labeling have been proposed since the late nineties. Maybe the first attempt was the work by Mori et al, in which each word assigned globally to the image is inherited by each (square) region within the segmented image [20]; regions are visually clustered and probabilities of the clusters given each word are calculated by measuring word co-occurrence among the clusters. In this work, however, word co-occurrence was measured among clusters of regions. A reference work for this task is the one proposed by Duygulu et al, on which image annotation is seen as a problem of machine translation [11]. The task consist of finding the one-to-one correspondence between vector-quantized regions and words (that is learning a lexicon) starting from weakly annotated images. This method received much attention and since its introduction several modifications and extensions have been proposed [22, 12, 4, 5]; furthermore, the data used by Duygulu et al have become a benchmark for comparing image annotation methods [18, 15, 14]. Several successful semi-supervised methods have been proposed[1] [6, 1, 2, 4, 11, 16, 5], some of which outperform the previous work [11]. The intuitive idea in most of these methods is to introduce latent variables for modeling the joint (or conditional) probability of words and regions. Then, when a new region needs to be labeled, these methods select the word that maximizes the joint probability between such a region and the words in the vocabulary; words are ranked according to the joint estimate. Hidden Markov models and Markov random fields have been introduced for consideration of dependencies between regions [4, 13]. From the supervised learning community some approaches have been proposed for image labeling [17, 6, 7]. A work close in spirit to ours is that due to Li et al [17], where a probabilistic support vector machine classifier is used for ranking labels for each region. Then, co-occurrences between candidate labels within the image are calculated in order to re-rank the possible labels. Our approach is different to the works by Mori et al and Li et al because we obtained the co-occurrence information from an external corpus, instead of considering co-occurrence of labels within the same image [17] or clusters of regions [20]. Furthermore, in such works co-occurrence information is used ad-hoc for their annotation method; while in this work we propose a

[1] Carneiro et al refer to such methods as unsupervised, though a more convenient term would be semi-supervised; since these methods start from weakly annotated images [6], therefore there exist some supervision

method that can be used with other annotation methods, provided they rank labels for its relevance, just as those proposed in [1, 2, 4, 11, 16, 5, 3, 17].

3 knn as annotation system

The *knn* classifier is an instance based learning algorithm widely used in machine learning tasks [19]. The zero training time of this algorithm makes it suitable for middle size data sets. Furthermore, it is adequate for domains in which new instances are continuously added to the data set. In this work we used this method as our automatic annotation system, due to the fact that it returns a set of ranked candidate labels for a given object; also, it does not require a training phase resulting in a fast method; and, further, this method outperforms other annotation systems, as we will show in Section 6.

knn needs a training data set $\{X, Y\}$ composed of N pairs 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 class of $x_i's$; for two class problems $y \in [0, 1]$. 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 it assigns to x_t a weighted combination of the labels belonging to $\{x_1^t, \ldots x_k^t\}$. We used the Euclidean distance as similarity function. For automatic image annotation at region level, we use *knn* in a multiclass learning setting, in which we have as many classes as words has the vocabulary. Instead of having two classes $[0, 1]$, we have $|V|$ classes $[1, \ldots, |V|]$, with $|V|$ being the number of different words in the collection. Since we would like to annotate regions, we need to extract features for each region, the features we considered for this work were color and shape statistics as described in Section 6. We decided to assign to a new instance the class of the most similar neighbor in our training set[2].

3.1 *knn* as a soft-annotation system

In order to apply the proposed approach 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 the relevance of the labels to being the correct annotation for such a region. A natural way of ranking labels is by using the ordering of labels that *knn* returns. However, within the set of the top$-k$ candidate labels, according to *knn*, these can be repeated; therefore, more confidence should be given to these repeated labels. Another problem with the *knn* ordering is that labels have not a relevance weight attached. This relevance weight, (which would be the equivalent of the posterior of the words given the region for probabilistic soft-annotation systems [1, 2, 4, 11, 16, 5, 3, 17]) should reflect the confidence we have on each candidate label. The relevance weight is an

[2] We do this when *knn* should return a single label for the region, this way of annotation is referred as *1-NN* through this document.

important component of our method since we take this weight (or posterior) as prior probabilities for the labels. Prior probabilities for the proposed method, should met the following: 1) they should reflect the confidence of the annotation method in the candidate labels and 2) the weight for the top$-k$ candidate labels should sum one, in order to be considered as prior probabilities.

We realized two intuitive ways of obtaining prior probabilities from the relevance ranking of *knn*. First we used the inverse of the distance of the test instance to the top$-k$ nearest neighbors; in this way we can infer prior probabilities directly related to the proximity of each neighbor to the test instance, as described in Equation (1).

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

where $d_j(x^t)$ is the inverse of the distance of instance x_j^t, within the $k-$nearest neighbors, to x^t, the test instance. This prior probability is accumulative, that is, labels appearing more than once will accumulate its priors according to the times they appear and their distance in each apparition. Note that we are implicitly counting for repetitions with this formulation.

The second intuitive way of obtaining prior probabilities is by considering the repetition of labels within the set of the top $k-$nearest neighbors of x^t, as described in Equation (2)

$$Pr_r(y_j^t) = \frac{rep(y_j^t)}{k} \tag{2}$$

with $rep(y)$ being a function that tells us the number of times label y is repeated within the $k-$nearest neighbors of x^t, note that this formulation is also normalized.

4 Naïve Bayesian improver based on co-occurrences

There are several automatic image annotation systems than rank labels in the vocabulary according to their relevance for a given region [17, 4, 5, 3, 11, 16, 1]. If we take the (top-one) most relevant label for a region it results on a poor performance of the annotation system. On the other hand, considering the top$-k$ possible labels for each region will result in an improvement on the system's accuracy. Unfortunately assigning a set of labels to a region is not straightforward; since this may cause confusion, adding uncertainty to the annotation and retrieval processes. However, if we measure the degree of association of each candidate label for a region with the candidate labels assigned to surrounding regions within the same image, we can determine which of the candidate labels is the most appropriate for the given region. We propose a naïve Bayes approach, abbreviated *NBI*, for the selection of the best candidate label by using co-occurrence information between candidate words of regions within the same image. We approach this problem as a learning task considering the candidate labels for a given region as classes and the (association with) candidate labels of surrounding regions as attributes.

A Bayesian classifier aims to estimate $P(H_{1,...,M}|E)$, that is the probability of each of the hypotheses (or classes) given some evidence (attributes). Then, according to decision theory [10], the H_i that maximizes $P(H_{1,...,M}|E)$ is selected as the most probable class. In our case we would like to select the label C^l, from a set of M candidate labels ($C^l_{1,...,M}$), that maximizes $P(C^l_{1,...,M}|A^k_{1,...,N})$. Taking as our evidence the top$-N$ candidate labels for the surrounding regions ($A^k_{1,...,N}$). Therefore, applying Bayes theorem for inverting the conditional probability, dropping the denominator and assuming conditional independence among attributes given the class, we have a naïve Bayesian classifier:

$$P(C^l_{1,...,M}|A^k_{1,...,N}) = P(C^l_{1,...,M}) * \prod_{i=1}^{N} P(A^k_i|C^l_{1,...,M}) \quad (3)$$

Where $P(C^l_{1,...,M})$ are the prior probabilities for each of the M candidate labels and $P(A^k_{1,...,N}|C^l_{1,...,M})$ are the conditionals of the candidate labels of other regions given the candidate labels for the region being analyzed. Therefore, we should calculate $P(C^l_{1,...,M}|A^k_{1,...,N})$ for each of the M candidate labels for a region and select the C^l that maximizes Equation (3) as the correct label for the region. The prior probability for each candidate label is the relevance ranking returned by the annotation method (see Equations (1) and (2)). The conditional probabilities $P(A^k_{1,...,N}|C^l_{1,...,M})$ are obtained from a co-occurrence matrix, as explained in the next Section. As we can see, in order to select a label for a given region, namely X, *NBI* considers the rank assigned to each candidate label of X by the annotation system; as well as the *semantic cohesion* between each candidate label of X and the set of candidate labels for regions surrounding X.

5 Obtaining co-occurrence information

The co-occurrence information matrix M_c consist of a $|V|_X|V|$ square matrix in which each entry $M_c(w_i, w_j)$ indicates the number of documents (on an external corpus) in which words w_i and w_j appeared together, where V is the set of words in the vocabulary[3]. That is, we considered each pair of words $(w_i, w_j) \in V_X V$ and searched for occurrences, at document level, of words (w_i, w_j). Then we count one co-occurrence if (w_i, w_j) appear together in a document. We did this for each of the $|V|.|V|$ pairs of words and for each document in our textual corpus. The documents we considered on this work were the captions of a new image retrieval collection: the *IAPR-TC12* [14] benchmark. This collection consists of about $20,000$ 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. Our matrix M_c then contains the co-occurrence information

[3] Co-occurrence matrix and code with implementations of the methods used in this paper can be obtained at *http://ccc.inaoep.mx/ hugojair/code/*.

for each pair of words within the vocabulary on such corpus. From the entries of the M_c matrix we can obtain conditional probabilities if we take: $P(w_i|w_j) = \frac{P(w_i,w_j)}{P(w_j)} \approx \frac{c(w_i,w_j)}{c(w_j)}$, where $c(x)$ indicates the number of times x appears in the corpus, which can be obtained from M_c. If we do this for each pair of words in the vocabulary we obtain our conditional probabilities matrix P_M.

A problem with this P_M matrix is the sparseness of data, that is, many entries of the matrix have zero values, this is a very common problem in natural language processing [8]. This problem is particularly damaging to naïve Bayes, because a zero value in Equation (3) will result on a zero confidence value for the class in consideration. In order to alleviate this problem we applied two widely used smoothing techniques: Laplace and Interpolation smoothing [8]. Laplace, also known as sum-one smoothing, is based on in Equation (4), while interpolation smoothing[4] is based on Equation (5).

$$P(w_i|w_j) \approx \frac{c(w_i, w_j) + 1}{c(w_j) + |V|} \tag{4}$$

$$P(w_i|w_j) \approx \lambda * \frac{c(w_i, w_j)}{c(w_j)} + (1 - \lambda) * c(w_j) \tag{5}$$

Where $|V|$ is the vocabulary size and λ is a parameter that weights the contribution of the original conditional probabilities and the counts of the conditioned term. These two smoothing techniques are the simplest ones [8], though more elaborated smoothing techniques could be applied to the M_c matrix. Laplace smoothing dramatically affects highly occurrence terms while increasing probabilities for low frequency terms. On the other hand, interpolation smoothing acts as a scaling of the original P_M matrix. In none of the smoothed matrices we have zero-valued entries now. As we will see on Section 6, the selection of smoothing technique affects the performance of our *NBI* approach. Therefore, an enhancement on the co-occurrence matrix will directly improve the performance of our method.

6 Experimental results

In order to evaluate our method we performed several experiments on three subsets of the benchmark Corel collection, which were also used in [11, 4]. First we evaluated the performance of *knn* as annotation method and we compared *knn* to other state of the art annotation methods [4, 5, 3, 11]. Then we evaluated how much accuracy improvement can we gain by applying *NBI* to *knn*. The subsets we used were made publicly available by Peter Carbonetto[5]. We used them due to the fact that the data sets are completely annotated at region level, which facilitates the evaluation of our method. The images of each subset were segmented with normalized cuts [23]. A sample segmented image is shown in Figure 1.

[4] We used $\lambda = 0.5$ for the experiments reported here.
[5] http://www.cs.ubc.ca/ pcarbo/

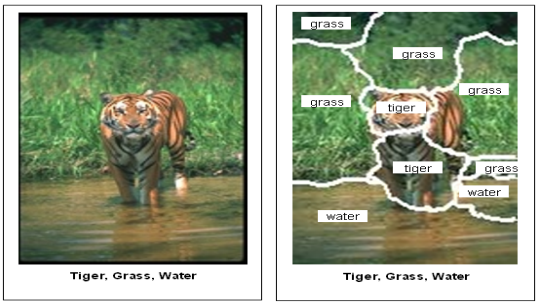

Fig. 1. Sample image with manual annotations from the Corel data set. Left: original image, right: image segmented with normalized cuts [23].

As features, we used color statistics and shape information from each region, resulting in 16 visual features. These features are described in detail in [4, 3]. The Corel subsets we used are described in Table 1. All of the results reported in this paper are obtained by applying our methods to the test sets of each data set.

Data set	**# Images**	**Words**	**Training blobs**	**Testing blobs**
A-NCUTS	205	22	1280	728
B-NCUTS	299	38	2070	998
C-NCUTS	504	55	3328	1748

Table 1. Subsets of the Corel image collection we used in the experimentation with *knn-NBI*

6.1 *knn* as automatic image annotation method

In the first experiment we evaluated the performance of *knn* as an automatic image annotation method. We measured accuracy as the percentage of correctly annotated regions, where a region is say to be correctly annotated if the label assigned to the region corresponds to its true label. Results of this experiment for different values of k, in the three Corel subsets we considered, are shown in Table 2. As we may expect, accuracy of the *knn* method increases as we consider more neighbors. Furthermore, *knn* is not much sensible to the number of classes being considered; for subset C with 55 classes, accuracy was higher than that of subset B with 38 classes.

We also compared the *knn* method against several semi-supervised annotation methods, proposed in [4, 5, 3, 11]. These methods are extensions and modifications to the work proposed in [11]. A description of these methods is over the scope of this paper, for a detailed description of the methods we encourage the

K	A-NCUTS	B-NCUTS	C-NCUTS	Average
1	36.8%	28.22%	29.11%	**34.99%**
5	61.5%	49.15%	54.57%	**59.29%**
10	72.8%	57.65%	63.95%	**68.42%**
15	77.1%	62.21%	69.73%	**73.05%**
20	81.5%	65.67%	72.76%	**76.27%**

Table 2. Percentage of correctly annotated regions, considering a region is correctly annotated if the true label is found within the top$-k$ neighbors

reader to follow the references. In order to provide an objective comparison, we used the code provided by Peter Carbonetto. This code includes implementations of the above mentioned methods and it is designed to work with the same Corel subsets we used. The error calculation and the plotting functions are also provided, which guarantee a fair comparison. In Figure 2 a comparison between the *knn* approach and the methods proposed in [11, 4, 5, 3], for the *A-NCUTS* data set, is presented[6]. In this plot, error is computed using the following equation:

$$e = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M_n}\big(1 - \delta(\bar{a}_{nu} = a_{nu}^{max})\big) \tag{6}$$

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 $\bar{a}_{nu}$. We ran 10 trials for each method, note that for *knn* a single trial could suffices since for every run the results are the same.

The left plot in Figure 2 shows error at the first label. Error is high for all of the methods we considered, however *knn* outperforms in average to all of the semi-supervised approaches; in the plot this is clear for most of the considered methods. The *gMl0* method [3] is the closer in accuracy to *knn*, although *gMl0* obtains a superior average error of 4.5%. In the right plot of Figure 2 we consider a label is correctly annotated if the true label is within the top-5 candidate labels. As we can see, error for all methods is reduced, this clearly illustrates the fact that accuracy of annotation methods is high considering a set of candidate labels instead of the first one. In this case the *gMAP* method [5] outperforms *5-NN-d* in average by 0.9% which is not a significant improvement. The other approaches obtain higher average error than that of *knn-d*. The ranking by distance (Equation (1)) is a better ranking strategy, this can be due to the fact that with this formula we are implicitly taking into account repetition information as well as the relevance based on distance.

Results from Figure 2 and those from Table 2 give evidence that the *knn* approach is an accurate method for image annotation, besides the simplicity of the

[6] We only used the *A-NCUTS* data set, since we later report results with *NBI* on this data set, and the labels of the *A-NCUTS* set are the only ones that are fully contained in our P_M matrix.

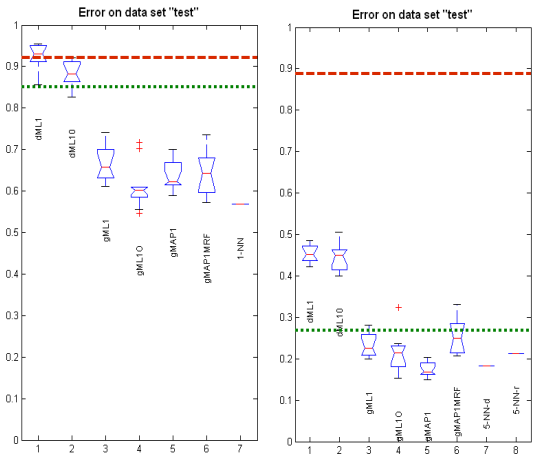

Fig. 2. Comparison of *knn* against semi-supervised methods proposed on [4, 5, 3, 11], using a Box-and-Whisker plot. The central box represents the values from the 25 to 75 percentile, outliers are shown as separate points. Left: accuracy at the first label. Right: accuracy considering the top−5 labels as candidate ones. Suffixes d and r stand for the way we ranked labels for *knn*. 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

method. However, we have to say that while the *knn* method needs of a training data set composed of pairs of feature regions and label, the semi-supervised methods start from a data set of weakly annotated image. It is surprising that, to best of our knowledge, *knn* has not been widely used as a method for image annotation given its simplicity and accuracy. Probably, the main reason is the need of a representative training data set. However, it can be possible to take advantage of semi-supervised learning algorithms and unlabeled data for obtaining good training data sets, as in [21].

6.2 *knn* + NBI: Improving annotation accuracy

We conducted several experiments with *knn* followed by *NBI* in order to measure the annotation improvement based on word co-occurrence. We have several parameters to consider when running *knn + NBI*: *1)* we varied the way we calculated the priors for *NBI* from *knn*, using the distance approach (Equation (1)) or using the repetition of labels (Equation (2)); *2)* the number of neighbors to consider for each region $k \in [3, \ldots, 7]$ in *knn*, *3)* the number of top candidate labels for surrounding images that we considered, $k_{img} \in [1, \ldots, 5]$; *4)* a binary valued variable, indicating if we should take the intersection of labels occurring on the surrounding regions, or if we should count for each label[7]. Due to the

[7] Where, if a label is repeated t-times we weight the conditional by a factor of t

efficiency of *knn* and *NBI* for the data sets we considered, we could ran experiments with all of the data subsets, though results in global accuracy are shown for the *A-NCUTS* only; since all of the words in this data set are present in our co-occurrence matrix P_M; while for the *B-NCUTS* and *C-NCUTS*, only a portion of the labels are present in P_M, therefore, accuracy improvement in these data sets should be measured differently, as we did in Figure 4.

Due to space limitations and the number of parameters considered, we report results of the best parameter configuration we obtained with *knn+NBI* over 100 runs, varying the parameters as described at the beginning of this section. The best configuration for each of the smoothing techniques is presented in Table 3.

Smoothing	**#-C's**	**#-A's**	**Intersection**	**Prior-type**
Interpolation	6	1	Yes	Distance
Laplace	6	1	Yes	Repetition

Table 3. Best configuration for each smoothing technique, we show the number of candidate labels for the region being analyzed (column 2), the number of candidate labels for the surrounding regions (column 3), if we used the intersection of candidate labels for surrounding images (column 4) and the way we obtained priors from *knn* to *NBI* (column 5).

We can appreciate consistency in the parameters for both configurations. The only difference is in the way that priors where calculated; with Laplace smoothing Equation (2) worked well, while for the Interpolation smoothing, Equation (1) worked best. Something not showed here is that in every run (different parameters) of the interpolation smoothing there was always an improvement over *knn* alone. Although with the Laplace smoothing, only half of the times (approximately) there was an improvement.

In Figure 3 we show the error as in Equation (6), this time we compared the methods proposed in $[4, 5, 3, 11]$ with *knn* and *knn+NBI* with the Laplace (*knn+NBI-Lap*) and interpolation (*knn+NBI-Inter*) smoothing, with the parameters described in Table 3. Accuracy of both smoothing techniques is very close (they differ by 0.002, which can not be appreciated in the plot), though interpolation smoothing performed much more better in average. The advantage of *knn+NBI* over *knn* is of about 6.5% for both smoothing techniques. Which is a significant improvement over *knn* alone; furthermore, the gain over the other annotation methods is clearly increased. The advantage of *knn+NBI* over the closer semi-supervised method (gML10) is of around 11% in average, which is an evident advantage.

You should note that *NBI* can improve an annotation method, provided the correct label is within the top$-k$ candidate labels returned by such method, we call this improvements the *gain*. We only consider that we have a *gain* when *1-NN* misclassified the region. On the other hand, *NBI* can decrease accuracy of annotation methods as well. This happens when *1-NN* selects the correct annotation, but *NBI* returns an incorrect label, we call this *lost*. In Figure 4 we

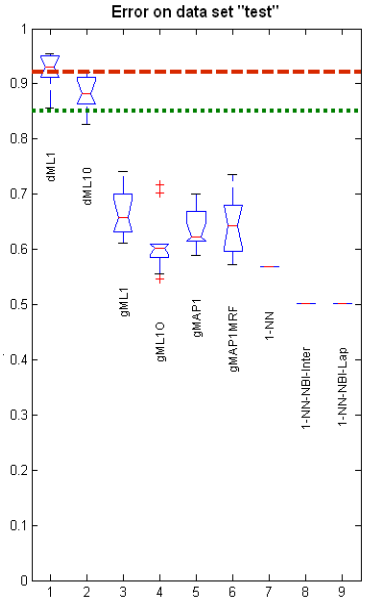

Fig. 3. Comparison of *knn* and *knn+NBI* against semi-supervised methods in [4, 5, 3, 11]; error is measured at the first label. We used the parameter configurations described in Table 3.

plot *gain* and *lost*, as defined above, in terms of percentages for the three data sets we considered.

From Figure 4 we can clearly appreciate that the gain is almost always superior to lost, note that each point in the plot is the result of averaging gain and lost over 100 trials, varying parameters for *knn+NBI*. From this plot, it is evident that the gain we obtain with *NBI* is significant over all data sets, and it is more evident on the *A-NCUTS* data set, independently of the smoothing technique used. The gain decreases for the B and C sets due to the fact that we can only apply *NBI* to a portion of the instances; because not all words are represented in the co-occurrence matrix. In some cases there is a loose in accuracy, however we consider that this could be significantly improved by having a more extensive corpus to estimate the co-occurrence matrix, and considering more robust models such as Markov random fields. Furthermore, a key contribution of this work is that it can be applied to other annotation methods that rank labels for their relevance.

7 Conclusions

In this paper we have presented a method for the improvement of automatic image annotation methods at region level. Our method, *NBI*, is based on the fact that accuracy of annotation methods at the first label is lower than that

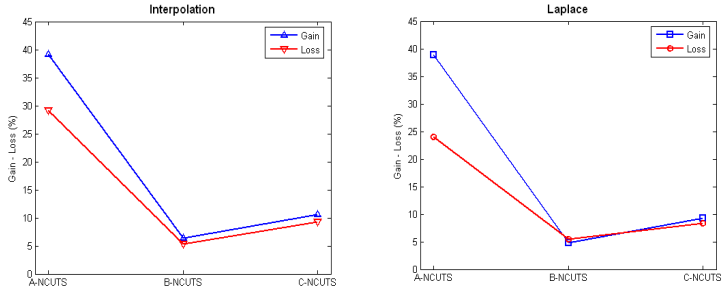

Fig. 4. Percentages of gain and lost of accuracy, for all the data sets. Results are averaged over 100 experiments, varying parameters. We plot results for both smoothing techniques we used: interpolation (Left) and Laplace (Right)

obtained if we consider the set of the top$-k$ candidate labels as annotations. *NBI* takes advantage of word co-occurrences among the candidate labels for a region and those of the other regions within the same image. Co-occurrence information is obtained off-line from an external collection of captions, which is a novel approach. Experimental results of our method on three subsets of the benchmark Corel collection, give evidence that the use of *NBI*, with *knn* as our annotation method, results in significant error reductions. We also show that the use of different smoothing techniques can affect the performance of *NBI*. Therefore, by building a more robust co-occurrence matrix and by considering more elaborated smoothing techniques we could obtain further improvements with *NBI*. Our method is efficient since we used a naïve Bayesian approach and the co-occurrence matrix is obtained off-line. Furthermore, *NBI* can be used with any other annotation method, provided it ranks labels for their relevance; even when the method does not ranks labels probabilistically, as we have shown here.

Future work includes the use of more robust probabilistic models, just like Markov random fields, for example. The improvement of the co-occurrence matrix is an immediate step towards the enhancement of *NBI*. Finally, we would like to test the *NBI* method with another annotation methods, and in other image collections, such as the *IAPR-TC12* benchmark [14].

Acknowledgements. We would like to thank Kobus Barnard and P. Carbonetto for making available their data and code on image annotation, and M. Grubinger for made available the $IAPR-TC12$ collection. Thanks too, to the members of the *INAOE-TIA-research group* by their comments and suggestions. This work was partially supported by CONACyT under grant 205834

References

1. K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. ICCV*, volume 2, pages 408–415. IEEE, 2001.

2. D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. of the 26th international ACM-SIGIR conf. on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.
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. G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on PAMI*, 29(3):394–410, 2007.
7. G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proc. of CVPR*, volume 2, pages 163–168, Washington, DC, USA, 2005. IEEE Computer Society.
8. 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.
9. R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval - approaches and trends of the new age. In *Proceedings ACM International Workshop on Multimedia Information Retrieval*, Singapore, 2005. ACM Multimedia.
10. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
11. 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.
12. G. Iyengar et al. Joint visual-text modeling for automatic retrieval of multimedia documents. In *Proc. the 13th MULTIMEDIA*, pages 21–30, New York, NY, USA, 2005. ACM Press.
13. A. Ghoshal, P. Ircing, and S. Khudanpur. Hmm's for automatic annotation and content-based retrieval of images and video. In *Proc. of the 28th int. conf. on Research and development in information retrieval*, pages 544–551, New York, NY, USA, 2005.
14. 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.
15. J. S. Hare, P. H. Lewis, P. G.B. Enser, and C. J. Sandom. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In Hanjalic A. Chang, E. Y. and Eds. Sebe, N., editors, *Proceedings of Multimedia Content Analysis, Management and Retrieval*, volume 6073, San Jose, California, USA, 2006. SPIE.
16. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS, 16*. MIT Press, Cambridge, MA, 2004.
17. 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.
18. Y. Liu, D. Zhang, G. Lu, and W. Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282, 2007.
19. T. Mitchell. *Machine Learning*. McGraw-Hill Education , October 1997.

20. 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.
21. K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.
22. J. Pan, H. Yang, P. Duygulu, and C. Faloutsos. Automatic image captioning. In *Proc. of the ICME*, 2004.
23. J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI-IEEE*, 22(8):888–905, 2000.
24. A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22(12):1349–1380, december 2000.