Can a probabilistic image annotation system be improved using a co-occurrence approach?

Ainhoa Llorente[1,2] and Stefan Rüger[1]

[1] Knowledge Media Institute, The Open University,
Walton Hall, Milton Keynes, MK7 6AA,
United Kingdom
[2] INFOTECH Unit, ROBOTIKER-TECNALIA,
Parque Tecnológico, Edificio 202, Zamudio, E-48170
Bizkaia, Spain
a.llorente@open.ac.uk, s.rueger@open.ac.uk

Abstract. *The research challenge that we address in this work is to examine whether a traditional automated annotation system can be improved by using external knowledge. Traditional means any machine learning approach together with image analysis techniques. We use as a baseline for our experiments the work done by Yavlinsky et al. [24] who deployed non-parametric density estimation. We observe that probabilistic image analysis by itself is not enough to describe the rich semantics of an image. Our hypothesis is that more accurate annotations can be produced by introducing additional knowledge in the form of statistical co-occurrence of terms. This is provided by the context of images that otherwise independent keyword generation would miss. We test our algorithm with two datasets: Corel 5k and ImageCLEF 2008. For the Corel dataset, we obtain statistically significant better results while our algorithm appears in the top quartile of all methods submitted in ImageCLEF 2008. Regarding future work, we intend to apply Semantic Web technologies.*

Key words: Automated Image Annotation, Statistical Analysis, Keyword Co-occurrence, Semantic Similarity

1 Introduction

Automated image annotation, also known as image auto-annotation, consists of a number of techniques that aim to find the correlation between low-level visual features and high-level semantics. It emerged as a solution to the time-consuming work of annotating large datasets. Most of the approaches use machine learning techniques to learn statistical models from a training set of pre-annotated images and apply them to generate annotations for unseen images using visual feature extracting technology. One limitation of these methods is the difficulty in distinguishing objects that are visually similar because the annotations are generated based on the correlation between keywords and image features such as colour, texture and shape. Without using additional information from the image context, a marble floor surface in a museum can be confused with a layer of ice in the arctic because both of them have got similar colour and texture. Another limitation of traditional systems is that each word is produced independently from the other annotated words, without considering that these words represent objects that co-occur in the same scene. Our work addresses this second limitation, the fact that words are generated independently in image annotation systems. Our intuition is that understanding how the human brain works in perceiving a scene will help to understand the process of assigning words to an image by a human annotator and consequently will help to model this process. In addition to that, having a basic understanding of the scene represented in an image, or at least a certain knowledge of other objects contained there, can actually help to recognize an object. In the previous example, if we had known that we were in a museum, we would have discarded the layer of ice in favour of the marble surface. On the other hand, the fact of knowing that there is a statue together with the unidentified object, it would have helped us to disambiguate the scene, and to think about a museum instead of the arctic. An attempt to identify the rules behind the human understanding of a scene was made by Biederman in [2]. In his work, the author shows that perception and comprehension of a scene requires not only the identification of all the objects comprising it, but also the specification of the relations among these entities. These relations mark the difference between a well-formed scene and an array of unrelated objects. In order to guarantee that all the keywords that annotate an image are coherent between each other we consider that, as they share the same context, the scene depicted in the image, they share a certain degree of semantic similarity. Among all the many uses of the concept semantic similarity we refer to the definition by Miller and Charles [18] who consider it as *"the degree of contextual interchangeability or the degree to which one word can be replaced by another in a certain context"*. Consequently, two words are similar if they refer to entities that are likely to co-occur together like "forest" and "tree", "sea" and "waves", "desert" and "dunes", etc. Semantic similarity is applied to our work in the form of statistical analysis techniques such as vector space models to correlate the annotation words with the context of the images.

The rest of this paper is organised as follows: Section 2 revises previous work on automated image annotation and Section 3 analyses some of their limitations. Section 4 describes how we have exploited the co-occurrence of keywords. Section 5 describes our algorithm while Section 6 shows the evaluation measures undertaken and provides results for the two datasets used. Finally, Section 7 contains our conclusions and plans for future work.

2 Related Work

Automated image annotation, also known as image auto-annotation, consists of a number of techniques that aim to find the correlation between low-level visual features and high-level semantics. It emerged as a solution to the time-consuming work of annotating large datasets.

Most of the approaches use machine learning techniques to learn statistical models from a training set of pre-annotated images and apply them to generate annotations for unseen images using visual feature extracting technology.

Automated image annotation can be divided with respect to the deployed machine learning method into co-occurrence models, machine translation models, classification approaches, graphic models, latent space approaches, maximum entropy models, hierarchical models and relevance language models. Another classification scheme makes reference to the way the feature extraction techniques treat the image either as a whole in which case it is called scene-orientated approach or as a set of regions, blobs or tiles which is called region-based or segmentation approach.

A very early attempt in using *co-occurrence* information was made by Mori et al. [21]. The process used by them starts by dividing each training image into equally rectangular parts ranging from 3x3 to 7x7. Features are extracted from all the parts. Each divided part inherits all the words from its original image and follows a clustering approach based on vector quantization. After that, conditional probability for each word and each cluster is estimated dividing the number of times a word i appears in a cluster j by the total number of words in that cluster j. The process of assigning words to an unseen image is similar to the carried out on the learning data. A new image is divided into parts, features are extracted, the nearest clusters are found for each divided part and an average of the conditional probability of the nearest clusters is calculated. Finally, words are selected based on the largest average value of conditional probability.

Duygulu et al. [7] improved the co-occurrence method using a *machine translation model* that is applied in order to translate words into image regions called blobs in the same way as words from French might be translated into English. The dataset used by them, Corel 5k dataset, has become a popular benchmark of annotation systems in the literature.

Monay and Gatica-Perez [19] introduced *latent variables* to link image features with words as a way to capture co-occurrence information. This is based on latent semantic analysis (LSA) which comes from natural language processing and analyses relationships between images and the terms that annotate them. The addition of a sounder probabilistic model to LSA resulted in the development of probabilistic latent semantic analysis (PLSA) [20].

Blei and Jordan [3] viewed the problem of modelling annotated data as the problem of modelling data of different types where one type describes the other. For instance, image and their captions, papers and their bibliographies, genes and their functions. In order to overcome the limitations of the generative probabilistic models and discriminative classification methods Blei and Jordan proposed a framework that is a combination of both of them. They culminated in *Latent Dirichlet Allocation*, [4] a model that follows the image segmentation approach and finds conditional distribution of the annotation given the primary type.

Jeon at al. [11] improved on the results of Duygulu et al. by recasting the problem as cross-lingual information retrieval and applying the *Cross-Media Relevance Model* (CMRM) to the annotation task. In addition to that, they showed that better ranked retrieval results could be obtained by using probabilistic annotation rather than hard annotation.

Lavrenko et al. [13] used the *Continuous-space Relevance Model* (CRM) to build continuous probability density functions to describe the process of generating blob features. The CRM model outperforms the CMRM model significantly.

Metzler and Manmatha [17] proposed an *Inference Network* approach to link regions and their annotations; unseen images can be annotated by propagating belief through the network to the nodes representing keywords.

Feng et al. [9] used a *Multiple Bernoulli Distribution* (MBRM), which outperforms CRM. MBRM differs from Continuous-space Relevance Model in the image segmentation and in the distribution of annotation words. CRM segments images into semantically-coherent regions while MBRM imposes a fixed-size rectangular grid (tiles) on each image. The advantage of this tile approach is that it reduces significantly the computational time. CRM models annotation words using a multinomial distribution opposed to MBRM which uses a multiple-Bernoulli distribution. This model focuses on the presence or absence of words in the annotation rather than in their prominence as it does the multinomial distribution. Image feature probabilities are estimated using a non-parametric kernel density estimation.

Other authors like Torralba and Oliva [23] focused on modelling *a global scene* rather than image regions. This scene-oriented approach can be viewed as a generalisation of the previous one where there is only one region or partition which coincides with the whole image. Torralba and Oliva supported the hypothesis that objects and their containing scenes are not independent. They learned global statistics of scenes in which objects appear and used them to predict presence or absence of objects in unseen images. Consequently, images can be described with basic keywords such as “street”, “buildings” or “highways”, using a selection of relevant low-level global filters.

Yavlinsky et al. [24] followed this approach using simple global features together with robust *non-parametric density estimation* and the technique of kernel smoothing. The results shown by Yavlinsky et al. are comparable with the inference network [17] and CRM [13]. Notably, Yavlinsky et al. showed that the Corel dataset proposed by Duygulu et al. [7] could be annotated remarkably well by just using global colour information.

3 Limitations of previous approaches

As a first step to understand what needs to be improved, we analysed different cases in which wrong keywords were assigned by a machine learning approach. The result of the study is the identification of two main categories of inaccuracies.

The first group corresponds to problems recognizing objects in a scene. This happens when a marble floor surface in a museum is confused with a layer of ice or when waves in the sea are taken for wave-like sand dunes in a desert. These problems are a direct consequence of the use of correlation between low-level features and keywords, as well as the *difficulty in distinguishing visually similar* *concepts.* One way to tackle these problems is to refine the image analysis parameters of the system, but this task is out of the scope of this work. Duygulu et al. also addressed these problems, suggesting that they are the result of working with vocabularies not suitable for research purposes. In their paper [7], they made the distinction between *concepts visually indistinguishable* such as “cat” and “tiger”, or “train” and “locomotive” in opposition to *concepts visually distinguishable in principle* like “eagle” and “jet”, which depend on the features selected.

In the second group of inaccuracies we find different levels of incoherence among tags, that range from the improbability to the impossibility of two objects being together in the real world. This problem is the result of *each annotated word being generated independently* without considering their context.

Other inaccuracies come from the *improper use of compound names* in some data collections. Compound names are usually handled as two independent words. For instance, in the Corel dataset, the concept “lionfish”, *a brightly striped fish of the tropical Pacific having elongated spiny fins*, is annotated with “lion” and “fish”. As these words never appear apart sufficiently often in the learning set, the system is unable to disentangle them. Methods for handling compound names can be found in the work done by Melamed [16].

Finally, it is important to mention the *over-annotation* problem. This situation happens when the ground-truth is made up of less words than the annotations. Over-annotation decreases the accuracy of the image retrieval as it introduces irrelevant words inside the annotations. This problem was also detected by Jin et al. [12] who proposed a system with flexible annotation length in order to avoid the over-annotation.

Our work attempts to overcome the limitations of words being generated independently by applying statistical analysis techniques. In order to go from low-level features to the high-level features of an image, semantic constraints should be considered, such as relations among entities and likelihood of each entity being present in a given scene. In this way, the accuracy of annotations will be improved when there is incoherence or improbability among the annotation words.

4 Exploiting keyword co-occurrence

The context of the images is computed using statistical co-occurrence of pair of words appearing together in the training set. This information is represented in the form of a co-occurrence matrix. The starting point for computing it is an image-word matrix A where each row represents an image of the training set and each column a word of the vocabulary. Each cell indicates the presence or absence of a word in the image. The co-occurrence matrix B is obtained after multiplying the image-word matrix A by its transpose A^T. The resulting co-occurrence matrix ($B = A^T.A$) is a symmetric matrix where each entry b_{jk} contains the number of times the word w_j co-occurs with the word w_k. The elements in the diagonal b_{jj} represent the number of images of the training set annotated by the word w_j. The dimension of the co-occurrence matrix corresponds to the number of words in the vocabulary. For example, in the Corel 5k dataset it is 374x374 while in ImageCLEF2008 is 17x17. Finally, the matrix is transformed in a conditional probability distribution after being normalised, dividing each element of a row by its Euclidean norm as suggested by Manning and Schütze in [15].

5 Description of the algorithm

The input for our algorithm are the top five keywords w_j and their associated probability $p(w_j|i_i)$ generated by the probabilistic framework [24]. Let *AnnoSet* be the annotations assigned to an image i_i ordered according to the decreasing probability:

$$\text{AnnoSet}(i_i)=\{(w_1,p(w_1|i_i));(w_2,p(w_2|i_i));(w_3,p(w_3|i_i));(w_4,p(w_4|i_i));(w_5,p(w_5|i_i))\}$$

Our algorithm only works with a selection of images from the test set for which the underlining system is *“confident enough”* i.e. at least one of the keywords has greater probability than a threshold α which is estimated empirically. Once an image i_i is selected the objective is to prune the keywords that are incoherent with the rest. The function $\textit{incoherence}(w_j, w_k)$ with $j, k = 1..5$ will detect whether a pair of keywords are semantically dissimilar or not. Two keywords w_j and w_k are semantically similar if their correlation value is greater than β. On the contrary, they are dissimilar if their correlation value is lower than γ. These parameters β and γ are estimated empirically and are dependent on the dataset used. If the system finds that the keywords w_j and w_k are incoherent, the function $\textit{lowerProbability}(w_k)$ will lower the probability of the keyword associated to the lowest probability w_k. Furthermore, the probability of each keyword w_l semantically similar to w_k is also lowered. This is done in order to ensure that all words incoherent with the context are removed from the annotation set. After modifying the probability values of these keywords, the function $\textit{generate}(\textit{AnnoSet}(i))$ sorts the keywords according to its probability and by selecting the five highest, new and more precise annotations are produced. A schema of the algorithm is the following:

```
For each image $i$ in testSet:
   if (max{$p(w_j|i)$ with $j = 1..5$} > threshold $\alpha$):
      for all pairs of keywords $(w_j, w_k)$ in AnnoSet($i$):
         if incoherence($w_j, w_k$):
            lowerProbability($w_k$)
            for each keyword $w_l$ in vocabulary $V$:
               if not incoherence($w_k, w_l$):
                  lowerProbability($w_l$)
Generate(AnnoSet($i$))
```

6 Results

We use as a baseline for our experiments the probabilistic framework developed by Yavlinsky et al. [24] because they used the Corel 5k dataset with the same experimental set-up than Duygulu et al. [7], which is considered a benchmark for automated image annotation systems. In addition to that, their evaluation measures showed state-of-the-art performance as evidenced in a review by Magalhães and Rüger [14]. We evaluate the performance of our algorithm (*Enhanced Method*) comparing it with the deployed by Yavlisnky et al. (*Trad. Method*) under different metrics.

6.1 Corel 5k dataset

The Corel 5k dataset is a collection of 5,000 images from 50 Corel Stock Photo CDs that comprises a training set of 4,500 images and a test set of 500 images. Images of the training set were annotated by human experts using a set of keywords ranging from three to five from a vocabulary of 374 terms. For evaluation purposes, we use two different metrics, the image annotation and the ranked retrieval. Under the image annotation metric, recall and precision of every word in the test set are computed. The number of words with non-zero recall *NZR*, provides an indication of how many words the system has effectively learned. Under the ranked retrieval metric, performance is evaluated with the mean average precision (MAP), which is the average precision, over all queries, at the ranks where recall changes where relevant items occur. The queries are 179 keywords that were selected based on their capacity for annotating more than one image from the test set. A comparison of the results using both methods is presented in Table 1.

Metric 1	**Trad. Method**	**Enhanced Method**
Words with NZR	86	91
Precision	0.1036	0.1101
Recall	0.1260	0.1318
Metric 2	**Trad. Method**	**Enhanced Method**
MAP	0.2861	0.2922

Table 1: Comparative results for the Corel dataset

The mean average precision (MAP) of our algorithm is 0.2922 which gives statistically significant better results than the value obtained by Yavlinsky et al., which were comparable to state-of-the-art automated image annotation. Interestingly, our algorithm is able to increase the number of words with non-zero recall from 86 to 91 as well as the precision and recall under Metric 1.

6.2 ImageCLEF 2008

Our algorithm was also tested with the collection of images provided by ImageCLEF 2008 for the Visual Concept Detection Task (VCDT) in [6]. This collection was made up of 1,800 training images and 1,000 test images, taken from locations around the world and comprising an assorted cross-section of still natural images. The results are presented under the evaluation metric followed by the ImageCLEF organisation which is based on ROC curves and under the image annotation metric. ROC curves [8] represent the fraction of true positives (TP) against the fraction of false positives (FP) in a binary classifier. The Equal Error Rate (EER) is the error rate at the threshold where FP=FN. The area under the ROC curve, AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

The results obtained are represented in Table 2.

Metric 1	**Trad. Method**	**Enhanced Method**
EER	0.288186	0.284425
AUC	0.776546	0.779423
Metric 2	**Trad. Method**	**Enhanced Method**
MAP	0.588489	0.589168

Table 2: Comparative results of ImageCLEF 2008

Such good results were obtained because the collection uses a vocabulary of 17 words which denotes concepts quite general, such as “indoor”, “outdoor”, “person”, “day”, “night”, “water”, “road or pathway”, “vegetation”,“tree”, “mountains”, “beach”, “buildings”, “sky”, “sunny”, “partly cloudy”, “overcast” and “animal”. In addition to that, our algorithm performed rather well appearing in the top quartile of all methods submitted in ImageCLEF 2008, however it failed to provide signicant improvement over the automated image annotation method. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm and in the nature of the vocabulary itself, where instead of incoherence we have mutually exclusive terms and almost no semantically similar terms.

7 Conclusions and Future work

The main goal of this work is to improve the accuracy of a traditional automated image annotation system based on a machine learning method. We have demonstrated that building a system that models the image context on top of another that is able to accomplish the initial identification of the objects increases significantly the mean average precision of an automated annotation system. Experiments has been carried out with two datasets, Corel 5k and ImageCLEF

2008. Our algorithm shows that modelling a scene using co-occurrence values between pair of words and using this information appropriately, helps to achieve better accuracy. However, it only obtained statistically better results than the baseline machine learning approach in the case of the Corel dataset where the vocabulary of terms were big enough. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm. This makes sense as a big vocabulary allows us to exploit properly all the knowledge contained in the image context. This is in tune with the opinion of most researches [10] as they believe that hundreds or thousands of concepts would be more appropriate for general image or video retrieval tasks.

Another important conclusion is the nature of the vocabulary, if it is quite general like in the case of the ImageCLEF 2008, the accuracy increases notably. On the other hand, the vocabulary used for annotating the Corel dataset is much more specific and consequently the algorithm decreases its accuracy as it needs to be precise enough to distinguish between animals belonging to the same family such as "polar bear", "grizzly" and "black bear".

Regarding future work, we want to improve the encouraging results shown in this paper by introducing Semantic Web technologies in order to further improve the algorithm. We plan to use ontologies to model generic knowledge (i.e. that can be used with different datasets) about images, and then exploiting them to additionally prune incoherent words and representing the relationships among objects contained in the scene.

Acknowledgements: This work was partially funded by the Pharos project sponsored by the Commission of the European Communities as part of the Information Society Technologies programme under grant number IST-FP6-45035.

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
2. I. Biederman. On the semantics of a glance at a scene. In *Perceptual organization.* Erlbaum, 1981.
3. D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
5. G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.
6. P. Clough, H. Müller, and M. Sanderson. The CLEF 2004 Cross-Language Image Retrieval Track. In *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum*, 2005.
7. P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112, 2002.

8. T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
9. S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.
10. A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 627–634, 2007.
11. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, 2003.
12. R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. In *Proceedings of the 12th International ACM Conferencia on Multimedia*, 2004.
13. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*, 2003.
14. J. Magalhães and S. Rüger. Information-theoretic semantic multimedia indexing. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, 2007.
15. C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.
16. I. D. Melamed. *Empirical methods for exploiting parallel texts*. PhD thesis, University of Pennsylvania, 1998.
17. D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, 2004.
18. G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Journal of Language and Cognitive Processes*, 6:1–28, 1991.
19. F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the 11th International ACM Conference on Multimedia*, pages 275–278, 2003.
20. F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: constraining the latent space. In *Proceedings of the 12th International ACM Conference on Multimedia*, 2004.
21. Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
22. A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.
23. A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, 2005.