

# A Study of Quality Issues for Image Auto-Annotation with the Corel Data-Set

Jiayu Tang, Paul H. Lewis

**Abstract**—The Corel Image set is widely used for image annotation performance evaluation although it has been claimed that Corel images are relatively easy to annotate. The aim of this paper is to demonstrate some of the disadvantages of data-sets like the Corel set for effective auto-annotation evaluation. We first compare the performance of several annotation algorithms using the Corel set and find that simple near neighbour propagation techniques perform fairly well. A Support Vector Machine (SVM) based annotation method achieves even better results, almost as good as the best found in the literature. We then build a new image collection using the Yahoo Image Search engine and query-by-single-word searches to create a more challenging annotated set automatically. Then, using three very different image annotation methods, we demonstrate some of the problems of annotation using the Corel set compared with the Yahoo based training set. In both cases the training sets are used to create a set of annotations for the Corel test set.

**Index Terms**—Corel Image set, Image Auto-Annotation, Support Vector Machine (SVM).

## I. INTRODUCTION

Image auto-annotation has been drawing more and more attention in recent years. So far, researchers have been focusing primarily on developing various auto-annotation algorithms [1], [2], [3], [4], [5], but very few have examined the effect of the data-set itself on the annotation result. Although good annotation algorithms are certainly what we really need to advance the process, the choice of appropriate data-sets for experiments is also important. An inappropriately designed data-set could give a biased measure of how well certain methods work. Westerveld and Vries [6] used the Corel images for evaluating their image retrieval technique and claimed that the Corel data-set is relatively easy. Tang and Lewis [7] examined this problem in the context of automatic image annotation.

This paper gives more extensive and detailed experiments and discussions on quality issues of image data-sets. We have developed and applied three auto-annotation methods to two image collections, one of which is built by capturing images from the web. Through the experiments and by comparisons of the results, we have examined several issues about image

data-sets, including problems when training sets and test sets contain many very similar images and data-sets with redundant information.

## II. TWO IMAGE COLLECTIONS

### A. The Corel Set

The first image set we consider is the widely used Corel Image set provided by Duygulu *et al.* [1] which is already separated into a training set with 4500 images and a test set with 500 images. Most of the images have 4 word annotations, while a few have 1, 2, 3 or 5. The vocabulary size of the whole set is 374 and that of the test set is 263. We note that in fact, the crucial vocabulary size is that of the training set since no other words are accessible for the auto annotation process. The vocabulary size of the Corel training set is 371. It will be shown in Section V that the simple colour structure descriptor (CSD)-based propagation method [7] achieves good results, compared with some state-of-the-art methods, for this image set. We argue that this is indeed because the Corel images are relatively easy to annotate and that this in turn is a result of the presence of groups containing many closely related images in the collection. This can actually be seen clearly from the following analysis. In the Corel training set there are 2705 images with 4 word annotations and these comprise a vocabulary of 342 different words. Among the 2705 images with 4 word annotations there are only 1833 different combinations of words. Assuming only random selections, the probability of getting such a low number of different combinations in a sample of 2705 is almost 0 ( $\sim 10^{-4796}$ ). Of course, some of the departure from randomness is due to the way in which some objects or image features appear frequently together in nature, for example trees and grass.

The fact that some of the images are close to each other in terms of both low-level features (such as color) and also the semantics and thus have the same combination of keywords for their annotations, is shown in Figure 1. A query image can be annotated correctly by a simple propagation method if there exists a training image that is very similar (both at the low-level and semantically) and if this is the one chosen for propagation.

### B. The Yahoo Set

In order to create a new image set avoiding, if possible, groups of similar images, a new image collection was created by querying the Yahoo Image Search engine<sup>1</sup> using each of the

Manuscript received July 7, 2006; revised September 29, 2006. This paper was recommended by Associate Editor E. Izquierdo.

The authors are with the Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom. (email:jt04r@ecs.soton.ac.uk; phl@ecs.soton.ac.uk)

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>

Digital Object Identifier 10.1109/TCSVT.2006.888941

<sup>1</sup><http://images.yahoo.com>



Fig. 1. Examples of similar Corel images, the number in the parenthesis being the file name of the image.

263 keywords from the Corel test set [1], such as ‘water’, ‘sky’ and ‘people’. For each keyword, the first 20 images returned by Yahoo are selected and annotated with the single query keyword used to retrieve it, resulting in a collection of 5260 images. All images are JPG color images, with a resolution of 120x80 on average. In some cases these annotations were not particularly appropriate because of the text based nature of the Yahoo image search. For example, image 2(a) is retrieved by using the word “water”, probably because there was an article about drinking water around the image and the word “water” appeared so many times that the Yahoo image search assumed it was a “water” image. In addition, the images are sparsely annotated because most images have more than one object. For example, images 2(c) and 2(d) contain multiple objects, but are only annotated with a single word. It is a more challenging set because, unlike the Corel set, the collection is less likely to contain groups of images with very similar content. The implication is that effective training with the Yahoo set<sup>2</sup> will be more difficult than with Corel.

In order to illustrate the self-similarity problem of the Corel set, we computed on each data set (Corel and Yahoo) the Euclidean distance between each image and its nearest neighbour (NN) in the CSD feature space. As shown in Fig. 3, the X axis represents the value of distance, while the Y axis represents the number of images that have a NN at this distance. The average distance for the Corel set and Yahoo set are 210 and 241 respectively. Statistically, 23.5% of the Yahoo images have a NN with a distance less than the average value of the Corel images. In contrast, up to 73.4% of the Corel images have a NN with a distance less than the average value of the Yahoo image.

<sup>2</sup>Available at: <http://www.ecs.soton.ac.uk/~phl/YahooSet.tar.gz>



Fig. 2. Examples of Yahoo images. The top images are inappropriately annotated, and for the bottom images only one object is annotated.

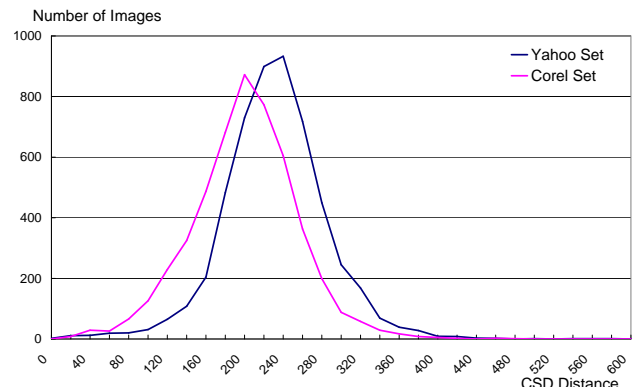


Fig. 3. The curves show, on the Corel and Yahoo set respectively, the CSD Euclidean distance between each image and its nearest neighbour (NN).

### III. THREE AUTO-ANNOTATION METHODS

We have implemented and used three very different approaches to image auto-annotation for the main comparison of methods. These methods are designated as CSD-Prop, SvdCos and CSD-SVM. We have introduced CSD-Prop [7] as a propagation method based on a global feature vector, the MPEG-7 Colour Structure Descriptor (CSD) [8]. SvdCos is a more complex region based method using correlation statistics, based on the work of [5]. Finally CSD-SVM is a multi-class and multi-label image classification method. Due to limited space, the details of CSD-Prop and SvdCos are not described here. Interested readers are referred to our previous work [7].

Image auto-annotation can be handled as a multi-class and multi-label classification problem. Multi-class means there are more than two classes, each of which is represented by a keyword, while multi-label means each image belongs to multiple classes. For example, images of 1(a) belong to

the classes “Birds”, “Sea”, “Sun” and “Waves”. Multi-label classification is more difficult than single-label classification. In the following, we propose to turn the multi-label problem into a single-label problem, using the CSD-SVM method, and describe how optimal parameters can be found.

CSD-SVM is a method based on Support Vector Machines (SVM), which is a very popular technique for classification. Two common ways of multi-class classification by SVM are “one-vs-all” and “one-vs-one”. As for multi-class classification, we choose the “one-vs-one” method, which uses a voting scheme. A binary classifier is trained for each possible combination of two classes. The class with the highest votes for the test document wins. In order to predict multiple annotations for each test image, rather than using only the class with the highest vote, we use the top  $n$  classes that have the highest votes,  $n$  being the number of annotations to be predicted. Another issue is that each training image has multiple labels, which makes it a multi-label training problem. We turn it into a general single-label training problem by duplicating each training image based on the number of words it has, and assign one and only one of the words to each copy of the image. Therefore, when applying the “one-vs-one” method, there are cases in which some training documents belong to both classes.

We used LIBSVM [9] to classify images that are represented by the Colour Structure Descriptor (CSD) as used for the CSD-Prop method. The radial basis function (RBF) [10], is used as the kernel function. Two parameters need to be optimized in LIBSVM, namely the penalty parameter  $C$  and the kernel parameter  $\gamma$  [9]. As recommended in [10], a grid-search method is applied on  $C$  and  $\gamma$  to find the optimal values, using  $v$ -fold cross-validation. The pair  $(C, \gamma)$  achieving the highest cross-validation accuracy is finally used to classify the test documents. However, calculating the cross-validation accuracy is not as straightforward as that on general single-label training problems. Instead, for each training image being used for testing, we have to compare the predicted label with all the words attached to this image before the image is duplicated. If the predicted label is one of them, the image is regarded as being correctly classified.

#### IV. EVALUATION METRICS

The *Mean Per-word Precision and Recall* and *Keyword Number with Recall*  $> 0$ , as used by previous researchers [1], [3], [4], [11], are adopted for evaluating annotation effectiveness. Per-word precision is defined as the number of images correctly annotated with a given word, divided by the total number of images annotated with this word. Per-word recall is defined as the number of images correctly annotated with a given word, divided by the total number of images having this word in its ground-truth or manual annotations. Per-word precision and recall values are averaged over the set of test words to generate the mean per-word precision and recall. A keyword has recall  $> 0$  if it is predicted correctly once or more, otherwise not.

We also introduce *Mean Per-image Precision and Recall* and *Cumulative Correct Annotations* for evaluation. Per-image

precision is the number of correctly predicted words for a given image divided by the total number of words predicted for that image, and per-image recall is the number of correctly predicted words divided by the number of manual annotations for that image. Per-image precision and recall are then averaged over all the test images to get the mean per-image precision and recall. Cumulative Correct Annotations is the total number of correct annotations.

## V. RESULTS AND DISCUSSION

### A. Comparison with state-of-the-art methods

We applied the previously described annotation algorithms to the Corel set and predict 5 words for each test image. Table I compares the CSD-Prop, SvdCos and CSD-SVM methods with the results of some state-of-the-art methods taken from the literature when the Corel training set is trained to annotate the Corel test set. These methods are the Translation model [1], the CRM model [3], the MBRM model [4], and the Mix-Hier model [11].

It is interesting to note in Table I that our simple CSD-Prop method achieves results almost as good as the best results from the more advanced methods. We argue that this is due to the training set and test set containing very similar images as illustrated in Figure 1. For example, in our experiment, the CSD-Prop method successfully predicts the word “Kauai”, which is even unlikely for a human being to learn, for the test image 1(c). It results from the training set containing the image 1(d) which is the closest in terms of Euclidean distance in the CSD feature space.

The best performance for the CSD-SVM algorithm is found to be at  $(C = 2^5, \gamma = 2^{-1})$  using grid-search and cross-validation. From the results, our CSD-SVM method performs reasonably well in the experiments when compared with the other methods considered. Although it gets a slightly lower number of words with recall  $> 0$  than the CSD-Prop method, overall the CSD-SVM achieves better results than CSD-Prop in view of the higher precision and recall measures. Again we argue that the reason why CSD-Prop method achieves a higher number of words with recall  $> 0$  is because it benefits from the fact that the Corel training set and test set have many very globally similar images in common. Difficult words, such as the place name “Kauai”, are easily learned by just comparing image similarity. The CSD-SVM method is also comparable in performance with the Mix-Hier method, which achieves the best results from the state of the art literature based methods in terms of mean per-word precision and recall.

### B. An Examination of word combinations

In this section, annotation effectiveness is analysed further in terms of word combinations. We consider the combination of four words that are correctly predicted, since most of the Corel images have 4 ground-truth labels, and for a single test image the auto-annotation methods predicted a maximum of four correct words. The SvdCos method is excluded from the analysis as it only managed to get 4 words correct 5 times, which is much less than that of CSD-Prop (45 times) and CSD-SVM (53 times).

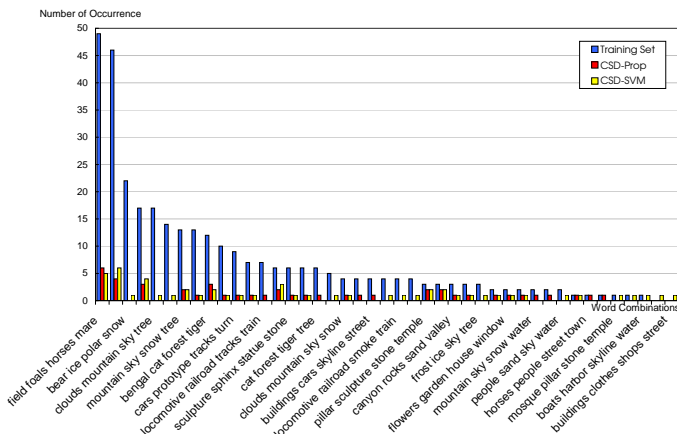


Fig. 4. Four Word combinations that are correctly predicted by CSD-Prop and CSD-SVM, being ordered by the number of occurrence of each combination in the Corel training set.

The analysis was conducted as follows on the predicted annotations both for CSD-Prop and CSD-SVM. Firstly, we find all different kinds of 4 word combinations from the predicted annotations on the test set, under the condition that each word is correct. Then, for each combination found, we search the training set to see if such a combination of annotations exists and if it does, how many times it occurs. Considering the propagation nature of CSD-Prop, it is not surprising that for CSD-Prop, all predicted 4 word combinations are found in the training set. For CSD-SVM, 51 out of 53 are found. The fact that almost all of the correct 4 word predictions exist in the training set, implies that this CSD based method may only be learning the relations between the whole image and the corresponding word sets or object sets from the training set which is certainly the situation for the CSD-Prop method. For correctly predicted four word combinations, Figure 4 shows their number of occurrences in the training set and in the predicted annotations by CSD-Prop and CSD-SVM. Note that for the last two combinations, the number of occurrence in the training set equals zero. This means that CSD-SVM managed to predict two combinations that do not exist in the training set, as shown in Figure 5. In the training set, the word combinations “clouds, sun, water, tree” and “buildings, clothes, shops, street” do not exist, but CSD-SVM managed to predict them correctly. Moreover, it can be seen that the other words predicted by CSD-SVM predict, “palm” and “people” for images 1061 and 119088 respectively, are actually reasonable annotations though not in the ground truth. The words predicted by CSD-Prop are included for comparison. All in all, if image auto-annotation is recognised as a problem of object recognition, the relations between objects and words, rather than the whole image and words, are really what need to be discovered. A good annotation method should be able to predict object combinations in the test images, no matter how these objects occur in the training set, either together in single images or separately in different ones. The use of global descriptors in annotation algorithms severely limits the possibility of achieving this.



		
	1061	119088
Ground-truth	clouds, sun, water, tree	buildings, clothes, shops, street
CSD-Prop	fountain, palace, light, reflection, palm	costume, <b>street</b> , village, <b>buildings</b> , people
CSD-SVM	<b>sun</b> , <b>water</b> , palm, <b>tree</b> , <b>clouds</b>	<b>street</b> , people, <b>shops</b> , <b>clothes</b> , <b>buildings</b>

Fig. 5. Two combinations predicted by CSD-SVM that do not exist in the training set, words in bold being correct.

C. Comparison between our three methods when different training sets are used.

For each of the three methods, we used the Corel training set and the Yahoo set for training respectively, to annotate the Corel test set, each image being predicted by 5 words. However, for fair comparison, only one random word out of the complete set of captions (normally 4) is used for each Corel training image, since each Yahoo image has only one caption. For example, for Fig. 1(a), we randomly choose one of the words “Birds”, “Sea”, “Sun” and “Waves” as the only annotation for this image and discard the others. Note that the whole set of labels of the test images are kept for evaluation. Table II compares the three methods using the two different image sets for training.

It can be seen that the CSD-Prop method performs better than the SvdCos method when it is trained on the Corel training set, but worse than the SvdCos method when trained on the Yahoo set. In other words, the CSD-Prop method degrades more rapidly when it moves from an easy training set (Corel) to a more difficult set (Yahoo). The CSD-SVM method maintains relatively good performance in both cases.

It can be seen that, even though only about 25% of the annotations of the Corel training set are used, the results of these methods did not decrease as much when compared with those referred to in Table I, where 100% of the annotations are used. The results of the CSD-SVM method are even comparable with that of the CRM [3] method, which uses 3 times more annotations. This implies redundant information exists in the Corel training set. For example, both image 1(a) and 1(b) belong to the training set. Since they are so similar to each other in both low-level features and semantics, there is no need for an annotation method to learn on both, especially for computationally expensive methods or when the training set is extremely large. Potentially, reduction techniques [12] can be used to condense the training set, reducing the size while retaining important training information.

We conclude that it is relatively easy to annotate the Corel test set using the Corel training set, and that the CSD-Prop method does not transfer as well as the SvdCos and CSD-SVM methods to the more challenging Yahoo dataset. In addition, it can be argued that using a challenging data set, good auto-annotation approaches should perform substantially better than, propagation-based approaches. Finally we conclude that simple sets like the Corel set should be used with caution for

TABLE I  
COMPARISON BETWEEN CSD-PROP, SVD-COS, CSD-SVD AND SOME OTHER STATE-OF-THE-ART METHODS USING THE COREL IMAGES

Models	Translation	CRM	MBRM	Mix-Hier	CSD-Prop	SvdCos	CSD-SVM
words with recall>0	49	107	122	137	130	102	127
Results on 49 best words							
Mean Per-word Recall	0.34	0.70	0.78	–	0.80	0.59	0.84
Mean Per-word Precision	0.20	0.59	0.74	–	0.58	0.51	0.74
Results on all 263 words							
Mean Per-word Recall	0.04	0.19	0.25	0.29	0.27	0.15	0.28
Mean Per-word Precision	0.06	0.16	0.24	0.23	0.20	0.15	0.25

TABLE II  
COMPARISON BETWEEN THE THREE METHODS ON DIFFERENT TRAINING SETS

Training Set	Corel(4500)			Yahoo(5260)		
Test Set	Corel(500)					
Models	CSD-Prop	SvdCos	CSD-SVM	CSD-Prop	SvdCos	CSD-SVM
Words with recall>0	107	100	94	46	58	59
Results on all 263 words						
Mean Per-word Recall	0.19	0.15	0.187	0.053	0.057	0.067
Mean Per-word Precision	0.14	0.11	0.153	0.038	0.040	0.053
Results on all 500 test images						
Cumulative Correct Annotations	577	349	767	102	123	118
Mean Per-image Recall	0.327	0.196	0.434	0.058	0.069	0.066
Mean Per-image Precision	0.231	0.140	0.306	0.040	0.049	0.047

effective evaluation of annotation methods.

## VI. CONCLUSIONS AND FUTURE WORK

The three image auto-annotation methods, CSD-Prop, Svd-Cos and CSD-SVM, have been used to annotate the Corel test set, by training on two different training sets, the Corel training set and the Yahoo training set. The Yahoo training set was constructed by obtaining images from the Yahoo Image Search Engine for 263 query words.

Through the experiments described in this paper, we have demonstrated and discussed some issues about the data-sets for image annotation. Firstly, we show how the simple propagation method CSD-Prop achieves fairly good results on the Corel set. It is argued that the Corel test set is a relatively easy set to be annotated when training on the Corel training set, because the Corel training and test sets contain many very globally similar images. As the Corel set has been popular for experiments on image auto-annotation, it is recommend that researchers be aware of the disadvantages of data sets like the Corel for effective annotation evaluation. Secondly, we have shown that the Corel test images can still be annotated well even when only 25% of the training information is used and it is argued that the training set contains redundant information. Choosing the best sub-set for training is worth exploring if only for computational efficiency. Training set reduction techniques [12] are of potential use for reducing the size of training sets and simultaneously filtering out the noise. We are planning to explore its application in image auto-annotation.

## REFERENCES

- [1] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *The Seventh European Conference on Computer Vision*, Copenhagen, Denmark, 2002, pp. IV:97–112.
- [2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.
- [3] J. Jeon, V. Lavrenko, and R. Manmatha., "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03 Conference*, 2003, pp. 119–126.
- [4] S. L. Feng, R. Manmatha, and V. Lavrenko., "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of the International Conference on Pattern Recognition (CVPR 2004)*, vol. 2, 2004, pp. 1002–1009.
- [5] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME 2004)*, 2004, pp. 1987–1990.
- [6] T. Westerveld and A. P. de Vries, "Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data," in *Proceedings of SIGIR Multimedia Information Retrieval Workshop 2003*, Aug 2003, pp. 135–142.
- [7] J. Tang and P. H. Lewis, "Image auto-annotation using 'easy' and 'more challenging' training sets," in *Proceedings of 7th International Workshop on Image Analysis for Multimedia Interactive Services*, 2006, pp. 121–124.
- [8] J. M. Martinez, "Mpeg-7 overview," N6828 ISO/IEC JTC1/SC29/WG11, Tech. Rep., October 2004. [Online]. Available: [www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm](http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm)
- [9] C. Chang and C. Lin, "Libsvm: a library for support vector machines," 2001. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [10] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification." [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
- [11] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *CVPR (2)*, 2005, pp. 163–168.
- [12] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257–286, 2000.