

# Automatic Image Annotation of News Images with Large Vocabularies and Low Quality Training Data

J. Jeon and R. Manmatha  
Center for Intelligent Information Retrieval  
Computer Science Department  
University of Massachusetts  
Amherst, MA 01003  
[jeon,manmatha]@cs.umass.edu

## ABSTRACT

A traditional approach to retrieving images is to manually annotate the image with textual keywords and then retrieve images using these keywords. Manual annotation is expensive and recently a few approaches have been proposed for automatically annotating images. These techniques usually learn a statistical model using a training set of images annotated with keywords and use this model to automatically annotate test images. While promising, these techniques have generally been tested on a few thousand images, with vocabularies of a few hundred words or less and using relatively high quality training data where the keywords are categories/objects and are directly correlated with the visual data.

Here, we investigate the problem of automatically annotating a large dataset of news photographs using low quality training data and a large vocabulary. We use 56,117 images and captions from Yahoo News Photos for our training and test data. The captions in the training portion of this data often contain a great deal of text most of which does not directly describe the image and as labels are, therefore noisy. We use the Normalized Continuous Relevance Models for our annotation and discuss how to speed up the model (by a factor of 10) using a voting technique. An improved distance measure also improves precision. To handle noisy text data and the large vocabulary of 4073 words, we investigate using different kinds of words for training and show that words which describe the content of the picture are significantly more useful for annotating images. Previous work on annotating images has largely dealt with high quality keywords.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval Models; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—*Object Recognition*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

## General Terms

Algorithms, Measurement, Experimentation

## Keywords

Image annotation, image retrieval, relevance models

## 1. INTRODUCTION

With the advent of digital imagery, the number of digital images has been growing rapidly and there is a need for effectively searching such image retrieval systems. Systems using non-textual (image) queries have been proposed but many users find it hard to represent their information needs using abstract image features. Most users prefer textual queries and this has been usually achieved by manually providing keywords or captions and searching over these captions using a text query. Manual annotation is an expensive and tedious procedure and most images are never likely to be captioned in this way. Recently, a number of researchers[2, 4, 8, 11, 12, 13, 16, 14, 20] have proposed automatic ways of annotating images using statistical models by training over a set of annotated images. Most of this work has used the Corel Stock Photos since all the images are well organized under semantic concepts and the manually attached labels usually describe visual objects in the image. The Corel dataset is very convenient to train a system for given labels or concepts. To provide practical image retrieval systems, it is necessary to increase the size of the annotation vocabulary. But it is unrealistic to get this kind of high quality training data for large number of words or concepts since such manual annotations are expensive.

One source of more training images is news images like Yahoo News Photos. A large number of news photographs with captions are available. The images are used to illustrate a story and the captions are usually obtained from the story. Captions are very different from the keywords typically used to annotate say Corel images. While some of the words are related to the visual content of the image, many of the words have nothing to do with the visual content of the image. For example in Figure 1, by looking at the image alone we can not infer that it is located on “Capitol Hill”. Nor can we infer the date. Except for the word “Greenspan” all the words in the second sentence have little do with the visual content of the picture. Many words have no visual meaning or are only tenuously related to the visual content of the image. Figure 2 shows some of the training examples related to the



Figure 2: Training examples for the word “PEACE” showing little in common

word “peace”, that is training images which have the word “peace” in the caption. As can be seen, these have nothing in common and thus “peace” is a poor word to use for annotating images. This is fairly typical of the news photographs on this site and probably many other sites. However, since such collections abound in practice, it would be useful to use such collections for training and testing. In this paper,

speed up the model we propose a voting approach which improves speed by an order of magnitude without sacrificing performance. To improve accuracy, we propose a new distance measure which takes into account geometric constraints. Experimental results show that the new distance measure doubles the recall and precision of retrieval.

Our main contribution is the investigation of techniques to perform annotation inspite of the poor quality of the training data. We, therefore, investigated a number of different ways of using the training data. The first set, *EntireCaption*, is generated by removing stop words and low frequency words and stemming words from the captions. The size of the vocabulary is 4,073. The second set utilizes the observation that the first sentence of the photograph is most closely related to the visual content of the image and therefore, the *FirstSentence* set, is acquired by following exactly the same process with the *EntireCaption* set but only from the first sentences of the captions. There are a total of 2,563 words in the *FirstSentence* set. The third set, is based on the observation that news photographs often contain pictures of people and hence the *NamedEntity* set, uses only the **PERSON** type named entities extracted from the first sentence of the captions. The number of unique named entities are 1,200. We perform both annotation and retrieval experiments to evaluate the results. For the annotation experiments on the *EntireCaption* set, we take each of the 4073 words as a query and evaluate its retrieval performance. The mean average precision for all 4,073 one word queries is 0.04. This result is a little misleading since a large number of queries in this set consist of non-visual words such as *abroad*, *abuse*, *add*, *mind*, *peace*. If one looks at the top 5 images of the top 500 queries, the average precision is much higher at 0.4. Surprisingly the results from the *FirstSentence* set do not change much from those for the *EntireCaption* set. However, using the names of people alone (note that we do not do any face detection) improves the mean average precision to 0.07.

Finally, we also divide up the words in the *FirstSentence* set into four categories - Non-Visual words(e.g. benefit), words which may have a Mixture of visual and non-visual meanings depending on context (e.g. base or bank), People’s names and Visual words (which are often objects, e.g. basket, bird or other words like black, sky). We show that performance on Visual words is substantially better than the performance on other kinds of words.

Previous systems have been tested using small vocabularies with the photographs taken from a single source or limited domains. Our dataset is much more diverse and closer to real-life dataset. In practical applications, speed is also important. It takes only a couple of seconds to automatically annotate one image using our system. Our work points to the challenges in doing image annotation and retrieval on




Corel Dataset	VIDEOTREC Dataset
	
PEOPLE, POOL, SWIMMERS, WATER	Face, female_face, outdoors, tree, news_subject
Yahoo Photo News Dataset	
	Federal Reserve Chairman Alan Greenspan , left, confers with Joint Economic Committee Chairman Sen. Robert Bennett , R-Utah, right, and Vice Chairman Jim Saxton, R-N.J., center, on Capitol Hill Wednesday, April 21, 2004. Greenspan told the committee America’s economic recovery has good momentum and that low, short-term interest rates will have to rise as some point.

Figure 1: Examples of different training datasets. The Corel dataset and VIDEO TREC dataset have accurate annotations that describe important visual objects in the photos. The captions in the Yahoo Photo News dataset contain many noisy words that are not related to the visual items.

we investigate the problem of automatically annotating and retrieving photographs from Yahoo News Photos. As mentioned above one of the main challenges in annotating this collection is that the captions provide poor quality training data. The vocabulary of 4,073 words is also substantially larger (and so even if the training data was not of poor quality) this would be a challenging task. Finally, the large number of images (56,117 images with roughly 25000 training and 25000 test images) also requires algorithms which can rapidly annotate the pictures.

Our approach to this problem uses the Normalized Continuous Relevance Model for annotation. Previous work [16] has shown that this model outperform a number of other models both on the Corel Data collection and on a subset of the Trec Video dataset for the image annotation task. To

such datasets. The surprising result is that the text data in the form of captions makes a big difference to the performance of the systems and a great deal of care is required in choosing the appropriate words. This is of course a non-trivial task. The poorer performance on larger vocabularies also indicates that much larger training datasets may be required. Finally, if high possible using high quality training data makes a big difference.

This paper is organized as follows. We discuss related work in section 2. This is followed by a discussion of the relevance models and our modifications to improve them in section 3. Section 4 discusses experimental details and results. Finally the last section is the conclusion of the paper.

## 2. RELATED WORK

In image annotation one seeks to annotate an image with its contents. Unlike more traditional object recognition techniques [1, 9, 22, 23] we are not interested in specifying the exact position of each object in the image. Thus, in image annotation, one would attach the label “car” to the image without explicitly specifying its location in the picture. For most retrieval tasks, it is sufficient to do annotation. Object detection systems usually seek to find a specific foreground object, for example, a car or a face. This is usually done by making separate training and test runs for each object. During training positive and negative examples of the particular object in question are presented. However, in the annotation scheme here background objects are also important and we have to handle a few thousand different object types and visual events at the same time. The model presented here learns all the annotation words at the same time. Object recognition and image annotation are both very challenging tasks.

Recently, a number of models have been proposed for image annotation [2, 4, 8, 13, 14]. Duygulu *et al.* [8] described images using a vocabulary of blobs. First, regions are created using a segmentation algorithm like normalized cuts. For each region, features are computed and then blobs are generated by clustering the image features for these regions across images. Each image is generated by using a certain number of these blobs. Their *Translation Model* applies one of the classical statistical machine translation models to translate from the set of blobs forming an image to the set of keywords of an image.

*Correlation LDA* proposed by Blei and Jordan [4] extends the Latent Dirichlet Allocation (LDA) Model to words and images. This model assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. This mixture of latent factors is then used to generate words and regions. Expectation-Maximization is used to estimate this model. [8] used 5,000 Corel photos and a vocabulary of 371 words for annotation and [4] used a similarly sized dataset. Barnard *et al.* [2] proposed and tested various statistical models to learn the joint probabilities of image regions and words. They used 16,000 Corel photos with 155 words and automatically annotated 10,000 test images.

Li and Wang [12] proposed a two-dimensional multiresolution hidden Markov models to relate images and concepts. They used 60,000 Corel photos with 600 concepts. While they claim to have used all 60,000 images (the training and test split is not clear), they only present evaluations for 4,060 test images.

In [13] the authors assumed that image annotation could be viewed as analogous to the cross-lingual retrieval problem and proposed the cross-media relevance model for this problem. [13] used the same discrete features as [8] and showed that a considerable improvement in performance was attained. A continuous relevance model was proposed in [14] to use continuous features with significant improvement in performance. Given the problems with variable length annotations, [20] proposed a Bernoulli model to improve annotation performance while in [16], the authors proposed the Normalized Continuous Relevance Model to pad the annotations to fixed length and still use a multinomial to achieve the same effect (see next section for more details). They showed that the performance of the model on the same dataset was considerably better than the models proposed by Duygulu *et al.* [8] and Mori *et al.* [18]. They used 5,000 Corel images with 371 words and 5,200 key frames from NIST’s TREC Video dataset with 137 words.

In this paper, we use Yahoo News Photos with 4,073 words and annotate 25,000 test images. The size of the vocabulary is much bigger than other experiments and the number of test images are also bigger than any other experiments.

## 3. CONTINUOUS RELEVANCE MODELS

The *Continuous Relevance Models* [14] is a statistical model that calculates the joint probabilities of a set of words and image features. The model learns the joint probabilities from annotated training samples. Each training image may be partitioned into regions using a general purpose segmentation algorithm or by simply using a partition of the image into rectangles. Previous work [20] has shown that a regular grid is significantly better than using existing segmentation algorithms and hence we use a 5x5 regular grid for all the images, giving 25 rectangular regions for each image. After partitioning, a feature vector is extracted from each region. The feature vector contains 23 features that represent the color, texture and shape information of the region.

As a result, each training image is represented by a set of feature vectors  $\mathbf{r} = \{r_1, r_2, \dots, r_{25}\}$  along with a set of annotation words  $\mathbf{w} = w_1, \dots, w_m$ .

The joint probability  $P(\mathbf{w}, I)$  of a test image  $I = \mathbf{r}$  being associated with words  $\mathbf{w}$  is computed as an expectation over the training samples.

$$P(\mathbf{w}, \mathbf{r}) = \sum_{J \in T} P(J)P(\mathbf{w}, \mathbf{r}|J) \quad (1)$$

where  $T$  is the training dataset,  $J$  is a training sample in  $T$  and

$$P(\mathbf{w}, \mathbf{r}|J) = \prod_{w \in \mathbf{w}} P(w|J) \prod_{r \in \mathbf{r}} P(r|J) \quad (2)$$

The model assumes the words and regions are all conditionally independent given  $J$ . For our purposes  $P(J)$  is assumed constant. The annotation component  $P(w|J)$  is modelled using a smoothed multinomial distribution.

$$P(w|J) = \lambda \frac{N_{w,J}}{N_J} + (1 - \lambda) \frac{N_w}{C} \quad (3)$$

where  $N_{w,J}$  is the number of times  $w$  occurs in the annotation of  $J$ ,  $N_J$  is the length of the annotation,  $N_w$  is the total number of times  $w$  occurs in the training set and  $C$  is the number of all annotations for all the training samples. This estimation reflects the *prominence* of the word  $w$  in

$$P \left( \begin{array}{c|c} \text{img1} & \text{img2} \\ \hline \text{img1} & \text{img2} \end{array} \right) = P \left( \begin{array}{c|c} \text{img1} & \text{img2} \\ \hline \text{img1} & \text{img2} \end{array} \right)$$

**Figure 3: Problems with the original Continuous Relevance Model.**  $P(\text{photo1}|\text{photo2})$  denotes the probability of generating photograph1 from photograph2. The left images are visually equivalent and the right images are not, but we get the same probabilities for both pairs of images because all the images have exactly the same regions regardless of region positions.

the annotation. For example, If  $\lambda = 1$  and some image  $J_1$  is annotated with a single word “face”, then  $P(\text{“face”}|J_1) = 1$ . If some other image  $J_2$  is annotated with 10 different words including “face”, then  $P(\text{“face”}|J_2) = \frac{1}{10}$ . In some cases, this property of capturing the prominence is desirable. However, this can also lead to the undesirable property that if two images have faces, the probability of the face in one image is much higher than that of the other merely because of the length of the annotation. In particular this causes problems when we want to rank across images. The *Normalized Continuous Relevance Models* [16] (see also [20] for a different take on the problem) addresses this problem. The intuition is that annotations for the training samples can be made of equal length by padding  $(N^* - N_J)$  instances of a special “null” word to the annotation of  $J$ . More explicitly this implies that we use  $N^* = \max_J N_J$  instead of  $N_J$  in equation (3). In this paper we use *Normalized Continuous Relevance Models* because of their improved performance [16]. For the feature component  $P(r|J)$ , the model uses a non-parametric kernel-based density estimate. Let  $\mathbf{r}_J = \{r_1 \dots r_n\}$  be the set of regions of image  $J$ .

$$P(r|J) = \frac{1}{|J|} \sum_{i=1}^n K(r, r_i) \quad (4)$$

where  $|J|$  is the number of regions in  $J$  and  $K$  is a Gaussian kernel that measures the closeness of two feature vectors.

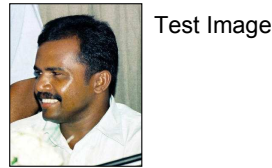
$$K(r|r_i) = \frac{\exp\{-\frac{1}{2}(r - r_i)^T \Sigma^{-1}(r - r_i)\}}{\sqrt{2^d \pi^d |\Sigma|}} \quad (5)$$

The model assumes that the covariance matrix  $\Sigma = \beta \cdot D$ , where  $D$  is the identity matrix and  $\beta$  is a scalar value that denotes the bandwidth of the kernel. The values of  $\lambda$  and  $\beta$  are determined from the held-out portion of the training data.

This joint probability distribution allows us to find the most likely annotations for a new unlabelled image  $\mathbf{r}$  by searching for words  $\mathbf{w}$  that maximize the joint probability  $P(\mathbf{w}, \mathbf{r})$ .

### 3.1 Modified Distance Measure

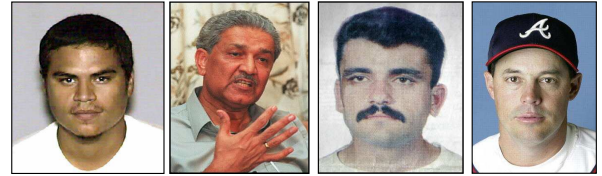
One of the problems in the original model is that the model does not take account the positions of the feature vectors. Figure 3 shows an example. If two images contain the same set of feature vectors, the two images are identical in the model regardless of the positions of the feature vectors. The original model is adapted from the Cross-Lingual Relevance Models[15] which uses a “bag of words” (bigrams have been shown to have little advantage in information retrieval) In images, however, the positions of the regions are important and so we modify the original model to incor-



Original CRM – top 5 closest training images



Modified CRM – top 5 closest training images



**Figure 4: The effect of using position constraints.** Top row: test image. Middle row: 4 closest training images using the old probability estimate. Bottom row: 4 closest images using the new estimate.

porate this information. We replace equation (4) with the following equation.

$$P(r|J) = \frac{1}{|J|} \sum_{r_i \in J} K(r, r_i) W(r, r_i) \quad (6)$$

$$W(r, r_i) = 1 / (L2(r_{position}, r_{i,position}) + 1) \quad (7)$$

where  $W$  is a weight function and  $L2$  is a function that calculates the Euclidean distance. If the positions of two feature vectors are close, the weight is big otherwise the weight is small.

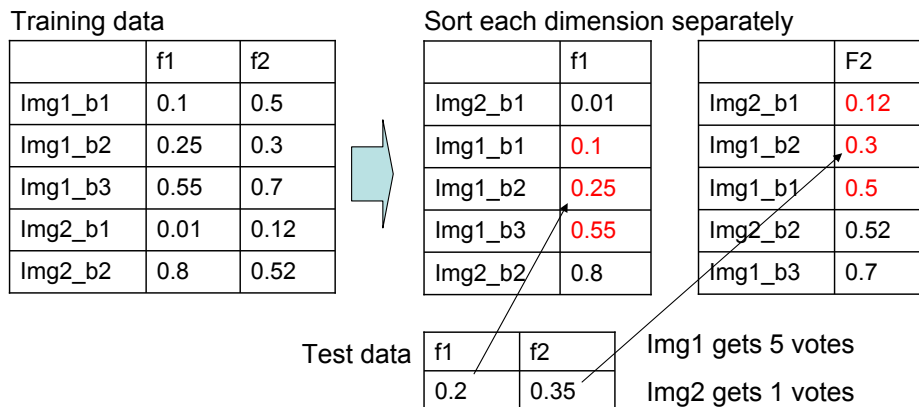
Using this simple modification, we can double the recall and precision of the retrieval and annotation performance. Figure 4 and 6 show the effect of new distance measure.

### 3.2 Speed-Up Using a Voting Scheme

The original Continuous Relevance Models are computationally expensive since every training sample is used to calculate the expectation value of the joint probability. This is too computationally expensive given that there are 25,000x25 feature vectors in the training dataset and each feature vector is 23-dimensional. Using a Pentium III 550Mhz machine, it takes 108 seconds to automatically annotate one image with 25,000 training samples. It requires about 750 hours to annotate 25,000 test images. Our approach to speeding up the model, uses the observation that most of the training images are very far in feature space and hence contribute very little to the final probability.

The speedup technique, therefore, uses only a subset of the training samples for a given query. First, we sort each dimension separately. Given a feature vector from a test image, for each dimension, we find the closest item from the sorted lists using the binary search algorithm. Training images that are close to the selected items get votes. By collecting the votes for all the feature vectors in the test





**Figure 5: Voting scheme.** This simple example assumes each feature vector is 2 dimensional and the window size is 3. Image 1 is closer to the test image since it gets more votes.

image, we can rank the training images by the number of votes. We use only the top 300 training images obtained in this manner instead of using all 25,000 images. Figure 5 shows a simple example of the voting scheme.

We tested this voting scheme using the Continuous Relevance Model on a Corel dataset which has 5,000 images. Table 1 shows the results of the scheme. With a very small loss in the quality of annotations we get a 7 times speed up. For larger datasets, the speedup is even more. In the experiments with the Yahoo News Photos, we get almost 10 times speed up. We successfully annotated 25,000 images in 18 hours using 4 Pentium III 550Mhz processors.

We also experimented with R-Trees and AV-files but in our case, none of these techniques even came close in terms of performance.

	Recall	Precision	time
No Voting	0.12	0.11	3h 30min
Voting, Window 100	0.09	0.09	27min
Voting, Window 200	0.10	0.10	28min
Voting, Window 300	0.11	0.11	30min

**Table 1: The results of the voting scheme using the Continuous Relevance Model on 5,000 Corel photos. We get 7 times speed up with a small loss in annotation accuracy. By adjusting the size of the window, we can trade-off the accuracy with the speed.**

## 4. EXPERIMENTS

### 4.1 Data Collection and Image Processing

Our data was obtained using a robot crawler that automatically downloaded Yahoo New Photos and corresponding captions everyday. We collected 56,117 samples over four weeks. 25,000 images from the first two weeks was used to build the training dataset and 25,000 images from the last two weeks was used to construct the test dataset. Photographs from the last day of the second week were used as the development dataset to fine-tune the system parameters. Note that about 6,117 images were omitted to leave a gap in time between the two sets. We partitioned each

image into a 5x5 rectangular grid and extracted feature vectors from each region (rectangle). The following is a list of the 23 features used.

Feature	Dimension
Average of RGB components	3
Average of LAB components	3
Texture, Garbor filter, 4 direction, 3 scale	12
Oriented edge energy, 4 direction	4
Ratio of edge to non-edge	1

### 4.2 Word Filtering

As described previously, the captions are noisy with many words that are not related to the visual objects in the photographs. These ‘noisy’ words need to be filtered out. Three different set of words are used. The first (baseline) *EntireCaption* set is obtained by simply removing all the stop words, stemming the remaining words and removing low frequency words (less than 50 occurrences) - resulting in 4,073 unique words. This set does includes many non-visual terms. The second *FirstSentence* set uses only the first sentences of the captions. The filtering process is the same as for the *EntireCaption* set. The second set contains 2,563 words. For the third *NamedEntity* set, ‘PERSON’ type named entities are extracted from the first sentences. For example, for the image Figure 1, ‘Alan Greenspan’, ‘Robert Bennett’ and ‘Jim Saxton’ are extracted from the caption and all other words are eliminated. After removing low frequency (less than 10 occurrences) named entities, 1200 named entities are left. While not perfect, in many cases the names of people who are in the photographs are extracted using this third method.

Sentence boundaries are determined using **MXTERMINATOR** program developed by Adwait Ratnapakkhi while **NER** named entity extractor developed by Wei Li and Andrew McCallum is used to mine people’s names.

### 4.3 Evaluation Results

Each word in the vocabulary is used as a query and the test images ranked according to the probability of that word (in other words by retrieving one word queries). Evaluation is done by using the captions obtained from Yahoo. Using all the words in the vocabulary as queries is not really desirable since many of these words have little to do with the images and are poor queries.

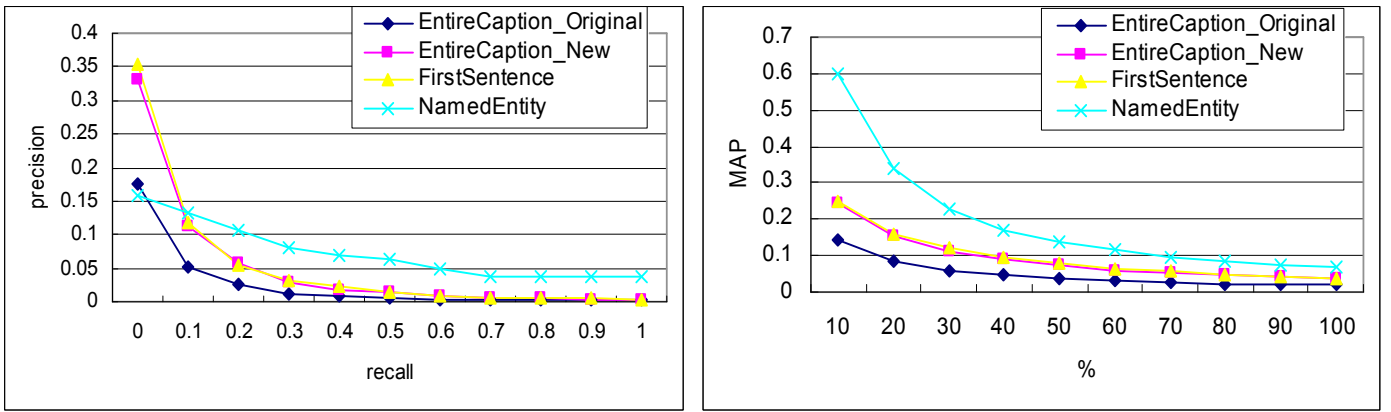


Figure 6: The left figure is a 11 point Recall Precision graph for automatic annotation. *EntireCaption\_New* performs better than the original model *EntireCaption\_Original*. Both use the same queries. Named entities perform much better except at very low recall. The right figure shows mean average precisions for the top 10% good queries, top 20% good queries and so on and shows that for the *NamedEntity* experiment for some queries we can do very good retrieval (e.g. at 10% MAP = 0.6)



Figure 7: Retrieval results for different text queries. Top five results shown for each query. Top row - query "kerry". Second and third row show the retrieval results for "Saddam". The results from *NamedEntity* is better because the vocabulary includes fewer noisy words.

The left chart in Figure 6 shows a 11 point recall/precision graph for all possible one word queries. The *EntireCaption\_Original* experiment uses the original Normalized Continuous Relevance Models and all the other experiments use the modified model. The *EntireCaption\_Original* experiment and the *EntireCaption\_New* experiment use the *EntireCaption* vocabulary set. The *FirstSentence* experiment uses the *FirstSentence* set and the *NamedEntity* experiment uses the *NamedEntity* set. *EntireCaption\_New* has much better results than *EntireCaption\_Old* because of the successful modification of the original model. *FirstSentence* has slightly better results than *EntireCaption\_New* but the difference is small. Therefore the hypothesis that the first sentences are more useful (contain more visual-words than other sentences) is not proved in our experiments. *NamedEntity* has much better results than the *EntireCaption\_New* and the results show that peo-



Figure 8: Example of poor image retrieval and poor quality training data. First row - top 5 images for query "car". Second row shows some sample training images that have "car" as an annotation. There are no cars in these images.

ple names are more closely related to the visual content of the photos than other random words. This is inspite of the fact that explicit face detection is not performed on the images. These results coincide with our intuition. Overall performance for all queries is not that good - for *FirstSentence*, the mean average precision over all queries is 0.04. Performance for the top ranks is much better. The poor performance is caused by the many words in the vocabulary which do not have anything to do with the image. We now look at more specific words. The right chart in Figure 6 shows mean average precision for the top 10% good queries, top 20% good queries and so on. The graphs show that for some queries, we can provide very good retrieval. The reason is many queries (words) are non-visual words. We manually analyze which queries (words) are good or bad - in terms of their visual relation to the image) in Table 4.3 and Figure 9. The non-visual category consists of words which have no visual correlates while the mixture category includes words which can be visual or non-visual depending on the context. The visual category includes words which usually denote objects or other visual features in the image. The analysis shows that queries that consist of visual words

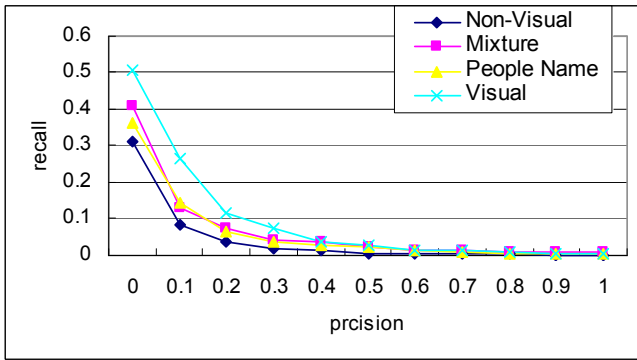


Figure 9: 11 point recall/precision graphs for each category of the queries in *FirstSentence* experiment.

have higher average precisions than non-visual queries. As the figure shows at low ranks, the performance is even better. As shown in Figure 2, it is almost impossible to related non-visual words and photographs since there are no common visual features for these words. Users usually use only visual words for their queries including people names when they use image search engines. The experimental results need to be improved but shows the potential of our approach. Some retrieval examples are shown in Figure 7 and 8.

Category	Nums	MAP	Example
Non-Visual	1390	0.0287	benefit, best, born
Mixture	449	0.0491	bank, base, bill
Person’s Name	456	0.0472	Baron, Bernard, Blair
Visual	207	0.0734	basket, bird, black

Table 2: Analysis of the queries. We manually classify all the queries for the *FirstSentence* experiment into 4 categories. The “Mixture” category contains words that have both visual and non-visual meanings. The mean average precision for visual words are much better than the MAP of the non-visual words. The analysis shows our model’s ability for relating words and images.

#### 4.4 Analysis of Some Examples

Figure 10 shows examples of automatic annotations. We used the top 30 words to annotate each image. Both images in the figure get the correct main keywords, *cricket* and *golf*, as their automatic annotations. The right images in the figure show the closest training images. The training images are not exactly the same as the test images but visually very similar, so the model borrows the annotations from the training images with high weights. Usually, it is not uncommon to find visually similar images from old news photo database for new news photos since news stories often (not always) use typical images to illustrate a point. Figure 11 is an interesting example. The test image and the closest training image are globally similar but locally different. While the model correctly uses the annotation word *meet*, many of the specific words in the training image cause problems. Without face detection/recognition, it is difficult to distinguish individual people (especially if they don’t occur very often). Clearly, there needs to be a way of

Test image	Annotation	Closest Train
	Original → new zealand s chri harris watch south africa s run past have hit bowl final international <b>cricket</b> park napier new zealand tuesday march south africa made inn fotopress ross Automatic → s,match,second, <b>cricket</b> ,international,feb,february,new,bowl,south,tuesday,stadium,africa,christchurch,zealand,ross,fotopress,finish,inn,wicket,iraq,celebrate,soldier,kill,al,jacque	
	Original → craig australia play shot th green third ford championship <b>golf</b> resort miami florida getty image file scott Automatic → three,th,s,image,february,john,la,lead,california,getty,hole,hit,course,dale,stroke,round,rd,pine,invitation, <b>golf</b> ,donald,buick,feb,annual,show,month,age,half,match,media	

Figure 10: Some good results. The images on the left are test images and the right images are the corresponding closest training images. The training and test images are not exactly the same but sufficiently similar, so the model borrows the annotations of the training images with high probability and successfully annotates “cricket” and “golf”.

Test image	Annotation	Closest Train
	Original → pakistan delegate <b>hussain r mohammad</b> share view unite nation food agriculture authority emergency regional meet avian influenza control animal asia <b>saeed khan</b> Automatic → s,minister,february,meet,group,bank,iraq,g,canada,central,seven,mark,reserve,ration, <b>greenspan</b> ,finance,federal,chairman,bocca, <b>alan</b> , <b>carney</b> ,world,u,bilateral,global,resort,feb,r,florida,	

Figure 11: A bad result. The example fails to correctly annotate the person names in the test image. One of the reasons is that we do not have enough training data for “Hussain R Mohammad” and “Saeed Khan”. Another reason is our image features are not designed to distinguish details of human faces. The closest training image has very similar visual settings with the test images, so the model annotates the test image with “Alan” and “Greenspan”. The automatic annotation “meet” correctly describes the content of the test image.

realizing that many of the words in the training image are specific to the particular image and should be eliminated and future work is needed to determine this. Figure 8 shows the retrieval results for the query *car*. The figure shows some training samples that have *car* as their annotation. These samples do not have cars in them but are related in some manner to cars. This example show the problem of the low quality training (caption) data.

## 5. CONCLUSION

In this paper, we explore the problem of automatically annotating a large number of images given poor quality training data. The Continuous Relevance Model was modified to improve both speed and performance. Our results show that captions are noisy sources of training data and some words are much visually meaningful and useful for annotation than other words. Our work shows that there is a need to incorporate more powerful text processing techniques to distinguish between “visual” and “non-visual” words.

## 6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation under grant NSF IIS-9909073. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

## 7. REFERENCES

- [1] S. Agarwal and D. Roth. Learning a Sparse Representation for Object Detection, IN *Proc. ECCV*, pages 113-130, 2002.
- [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*, 3:1107–1135, 2003.
- [3] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, Vol.2, pages 408-415, 2001.
- [4] D. Blei, Michael, and M. I. Jordan. Modeling annotated data. To appear in the *Proceedings of the 26th annual international ACM SIGIR conference*
- [5] D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [6] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, 19(2):263-311, 1993.
- [7] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*, Lecture Notes in Computer Science, 1614, pages 509-516, 1999.
- [8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, pages 97-112, 2002.
- [9] R. Fergus, P. Perona and A Zisserman Object Class Recognition by Unsupervised Scale-Invariant Learning. IN Proc. CVPR'03, vol II pages 264-271, 2003.
- [10] H. Muller, S. Maillat and T. Pun. The Truth about Corel - Evaluation in Image Retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, London, UK, 2002.
- [11] J. Jeon and R. Manmatha. Using Maximum Entropy for Automatic Image Annotation To appear In *Proceedings of the 5th International Conference on Image and Video Retrieval*, 2004
- [12] Jia Li and James Ze Wang. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9): 1075-1088 (2003)
- [13] J. Jeon, V. Lavrenko and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 119–126, 2003
- [14] V. Lavrenko, R. Manmatha and J. Jeon. A Model for Learning the Semantics of Pictures. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, NIPS'03, 2004.
- [15] V. Lavrenko, M. Choquette, and W. Croft. Cross-lingual relevance models. *Proceedings of the 25th annual international ACM SIGIR conference*, pages 175–182, 2002.
- [16] V. Lavrenko, S.L. Feng and R. Manmatha. Statistical Models for Automatic Video Annotation and Retrieval. In *the International Conference on Acoustics, Speech and Signal Processing*, ICASSP, Montreal, QC, Canada, 2004.
- [17] V. Lavrenko and W. Croft. Relevance-based language models. *Proceedings of the 24th annual international ACM SIGIR conference*, pages 120-127, 2001.
- [18] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
- [19] J. M. Ponte, and W. B. Croft, A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR Conference*, pages 275–281, 1998.
- [20] S.L. Feng, R. Manmatha and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proceedings of CVPR 2004*, Dublin, Ireland
- [21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [22] H. Schneiderman, T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. Proc. IEEE CVPR 2000: 1746-1759
- [23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR'01*, pages 511-518, 2001.