XML PATH BASED RELEVANCE MODEL FOR AUTOMATIC IMAGE ANNOTATION

Manjeet Rege, Ming Dong and Farshad Fotouhi

Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
{rege, mdong, fotouhi}@wayne.edu

ABSTRACT

This is the first paper that proposes automatic image annotation using the semantics of XML. In this paper, we propose XPRM - XML Path based Relevance Model for automatic image annotation. Our experimental results show that the proposed model has considerable advantage over single word annotations in performing automatic semantic annotation.

1. INTRODUCTION

Conventional CBIR systems require the user to retrieve images based on low-level image attributes such as color, texture, etc. As users are unfamiliar with the content of the images in the image database, this does not result in an efficient image retrieval. Also, it puts upon the user the added responsibility of performing image retrieval based on low-level image information. Ideally, users would prefer querying an image database by performing semantic querying without having a need to know the contents of the images in the database. To capture the semantics of images, manual image annotation has been practiced by image repository librarians. As this method is unscalable, automatic image annotation has received extensive attention recently. Though researchers have attempted linking images and words in various ways, to the best of our knowledge there has been no effort in annotating images using the semantic structure of XML [1].

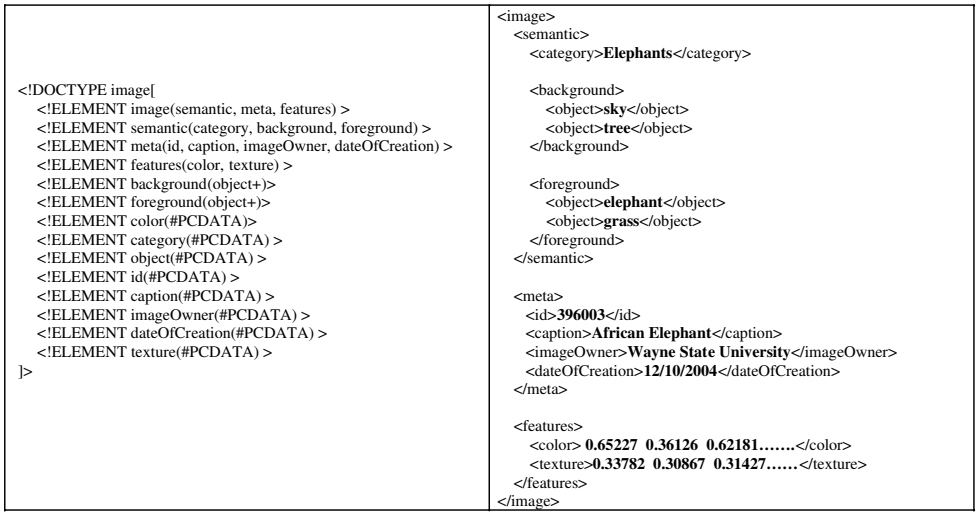
This is the first paper that proposes to annotate images using XML annotation paths. We propose the XML Path based Relevance Model (XPRM) that performs automatic image annotation using the semantics of XML. We partition each image into a set of rectangular regions and low-level features are extracted from each region. The model performs semantic annotation of new test images based on a XML representation of training images that store semantic, low-level data and other meta information. XPRM models the joint probability distribution of XML annotation paths and low-level image features. The key idea behind XPRM is that the XML annotation paths capture the semantic content of the image in a much efficient way instead of the traditional single word annotations. Our experimental results on an image database consisting of Corel images supports this fact.

The rest of the paper is organized as follows. Section 2 briefly discusses the related work. In Section 3, we first establish the advantage of XML path annotation over single word annotation in 3.1 and then present the proposed annotation model in 3.2. In section 4, we present our experimental results. Finally, we conclude in section 5.

2. RELATED WORK

There has been prior work done on schema or ontology based image annotation [2, 3]. Hyvonen et al [2] propose ontology based image retrieval and annotation of graduation ceremony images by creating hierarchical annotation. They used Protege as the ontology editor for defining the ontology and annotating images. In [3], Schreiber et al perform ontology based annotation of ape photographs. As in [2], they too annotate ape images using the same ontology defining and annotation tool and use RDF Schema as the output language. MPEG-7 [4] proposes to store low-level features, annotations and other meta information in one XML file. Rege et al [5] propose to annotate human brain images using MPEG-7. However, the major drawback of the above approaches is that annotation of images needs to be performed manually. There is an extra effort needed from the user's side in creating the ontology and performing the detailed manual annotation. Recently, statistical automatic image annotation has also been performed using single words [6, 7, 8]. As we point out in section 3.1, single word annotations do not have enough semantic meaning associated with it. So far there has been no work done on automatic image annotation using the semantics of XML which enables efficient image annotation to represent domain knowledge. The proposed XML Path based Relevance Model is the first work in this direction.

0-7803-9332-5/05/$20.00 ©2005 IEEE



```
<!DOCTYPE image[
   <!ELEMENT image(semantic, meta, features) >
   <!ELEMENT semantic(category, background, foreground) >
   <!ELEMENT meta(id, caption, imageOwner, dateOfCreation) >
   <!ELEMENT features(color, texture) >
   <!ELEMENT background(object+)>
   <!ELEMENT foreground(object+)>
   <!ELEMENT color(#PCDATA)>
   <!ELEMENT category(#PCDATA) >
   <!ELEMENT object(#PCDATA) >
   <!ELEMENT id(#PCDATA) >
   <!ELEMENT caption(#PCDATA) >
   <!ELEMENT imageOwner(#PCDATA) >
   <!ELEMENT dateOfCreation(#PCDATA) >
   <!ELEMENT texture(#PCDATA) >
]>
```

```
<image>
   <semantic>
      <category>Elephants</category>

      <background>
         <object>sky</object>
         <object>tree</object>
      </background>

      <foreground>
         <object>elephant</object>
         <object>grass</object>
      </foreground>
   </semantic>

   <meta>
      <id>396003</id>
      <caption>African Elephant</caption>
      <imageOwner>Wayne State University</imageOwner>
      <dateOfCreation>12/10/2004</dateOfCreation>
   </meta>

   <features>
      <color> 0.65227  0.36126  0.62181…….</color>
      <texture>0.33782  0.30867  0.31427……</texture>
   </features>
</image>
```

Fig. 1. An example of an XML schema and the corresponding XML representation of an image

3. PROPOSED FRAMEWORK

3.1. XML Path based image annotation

We propose to perform automatic image annotation using the semantic meaning associated with XML paths. Suppose we have an image of an *elephant* with single words annotation of "elephant, grass, tree, sky". It is obvious that this kind of annotation does not have enough semantic meaning associated with it. On the other hand, consider now that the same image is represented in an XML format. For the sake of illustration, consider the XML schema and the corresponding XML representation of the image shown in Fig 1. This XML schema stores foreground and background object information along with other meta information with keywords along various XML file paths. If we now consider the XML path from the root node of the XML file to the keyword as an annotation, then it has a more semantic meaning associated with it. In the case of the *elephant* image, semantically meaningful XML annotations would be "image/semantic/foreground/object=elephant, image/semantic /foreground/object=grass, image/semantic/ background/ object=sky, image/semantic/background/object=tree".

Harnessing the semantic structure of XML to represent image annotation, gives us an efficient way to represent and annotate images. Note that one can represent and annotate images to represent domain knowledge by conforming to any XML schema. We simply use the foreground and background object information as a running example in this paper to demonstrate the advantage of the proposed model over other works.

3.2. XPRM:Image Annotation Model

Let the set of XML annotation paths be presented by X, T denote the training images in XML format and let t be an image belonging to T. Let x_t be a subset of X containing the annotation paths for t. As image segmentation is a computationally intensive and also an erroneous activity, we represent each image using n rectangular regions of equal size. We extract low-level features from each rectangular region and construct a feature vector.

Consider an image q not in the training set. Let $f_q = \{f_{q1}, f_{q2}, ..., f_{qn}\}$ denote the feature vector for q. In order to perform automatic annotation of q, we model the joint probability of f_q and any annotation path subset x of X.

$$P(x, f_q) = P(x, f_{q1}, f_{q2}, ..., f_{qn}) \tag{1}$$

We use the training set T of annotated images to estimate the joint probability of observing x and $\{f_{q1}, f_{q2}, ..., f_{qn}\}$ by computing the expectation over all the images in the training set.

$$P(x, f_{q1}, f_{q2}, ..., f_{qn}) = \sum_{t \epsilon T} P(t) P(x, f_{q1}, f_{q2}, ..., f_{qn}|t) \tag{2}$$

We assume that the events of observing x and $f_{q1}, f_{q2}, ..., f_{qn}$ are mutually independent to each other and express the joint probability in terms of P_A, P_B and P_C as follows,

$$P(x, f_{q1}, f_{q2}, ..., f_{qn}) = \sum_{t \epsilon T} \{P_A(t) \prod_a P_B(f_a|t) \prod_{path \epsilon x} P_C(path|t) \prod_{path \notin x} (1 - P_C(path|t))\} \quad (3)$$

where P_A is the prior probability of selecting each training image, P_B is the density function responsible for modelling the feature vectors and the XML annotation paths are modelled using a multiple-Bernoulli distribution P_C.

In the absence of any prior knowledge of the training set, we assume that P_A follows a uniform prior and can be expressed as

$$P_A = \frac{1}{||T||} \quad (4)$$

where $||T||$ is the size of the training set.

For the distribution P_B, we use a kernel-based density estimate.

$$P_B(f|t) = \frac{1}{n} \sum_i \frac{exp\{-(f - f_i)^T \Sigma^{-1} (f - f_i)\}}{\sqrt{2^k \Pi^k |\Sigma|}} \quad (5)$$

where f_i belongs to $\{f_1, f_2, ..., f_n\}$ the set of all low-level features computed for each rectangular region of image t. Σ is the diagonal covariance matrix which is constructed empirically for best annotation performance.

In the XML representation of images, note that every annotation path can either occur or might not occur at all for an image. Moreover, as we annotate images based on object presence and not on prominence in an image, an annotation path if it occurs can occur only once in the XML representation of the image. As a result, we assume that the density function P_C follows a multiple Bernoulli distribution and is given by,

$$P_C(path|t) = \frac{(\gamma \alpha_{path,t} + N_{path})}{(\gamma + ||T||)} \quad (6)$$

where γ is a smoothing paramater, $\alpha_{path,t} = 1$ if the path occurs in the annotation of image t, else it is zero. N_{path} is the total number of training images that contain this $path$ in their annotation.

4. EXPERIMENTAL RESULTS

Since the proposed work models the joint probability of low-level features and XML annotation paths, we needed an image database that represented images in XML format where each XML file contained annotation, low-level features and other meta data information stored along different paths. In the absence of such a publicly available data, we had to manually create an XML representation for each image. Currently, our image database contains 1500 Corel images comprising of 15 image categories with 100 images in each category. Each image in the database has been represented in an XML format conforming to the schema shown in Fig 1. In order to obtain preliminary results to demonstrate the proof of concept for our proposed work, we performed our experiments on 5 randomly selected image categories - "Air Force", "Fabulous Fruit", "Elephants", "Beach" and "Buses". We randomly selected 70 percent of this dataset for training while the remaining were used for testing the performance of the model. As the focus of this paper is on models and not on features, we use some of the features standardized by MPEG-7 [4]. Since the proposed work is the first one in its kind to automatically annotate images using XML paths, we were unable to make a direct comparison with any other model. However, our annotation and retrieval results are comparable to ones obtained by [6, 7, 8].

4.1. Automatic Annotation Results

In order to evaluate automatic image annotation using XPRM, given a test image we calculate the joint probability of the low-level features extracted from this image and the XML annotation paths in the vocabulary. We select the top 4 paths with the highest joint probability as the automatic annotation for this image. Other approaches [6, 7, 8] simply try to emulate the ground truth annotation. In other words, they try to come up with the exact words appearing in the ground truth annotation and perform no semantically relevant automatic annotation. Figure 2 clearly demonstrates the contribution of XPRM in this regard. We can see that the XPRM annotation has more semantic meaning than even the original Corel annotation. Note that "image/semantic/foreground /object=trunk" does not appear in the top XPRM annotation, our model has annotated this image with "image/semantic/ background/object=tree", although the ground truth does not contain the "tree" annotation. This we believe is because XPRM learns the object context. As we have *tree* present in the other *elephant* images, the model learnt the context between the *elephant* and *tree* and hence provided the necessary additional semantic annotation.

We also evaluate the automatic image annotation using recall and precision as in [6, 7, 8]. We calculate recall and precision for every annotation path in the test set defined as follows: $recall = \frac{q}{r}$, $precision = \frac{q}{s}$, where q is the number of images correctly annotated by an annotation path, r is the number of images having that annotation path in the ground-truth annotation and s is the number of images automatically annotated by that annotation path. We report the results for all the 148 paths in the test set as well as the 23 best paths as in [6, 8]. Table 1 shows the annotation results.

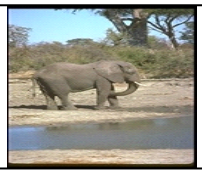
Image			
Original Annotation	plane, jet, wheels, sky	lime, close-up, food, fruit	elephant water trunk sky
XPRM Annotation	image/semantic/background/object=sky, image/semantic/foreground/object=plane, image/semantic/foreground/object=jet image/semantic/foreground/object=wheels	image/semantic/forgeround/object=food, image/semantic/forgeround/object=fruit, image/semantic/forgeround/object=lime, image/semantic/foreground/object=close-up,	image/semantic/background/object=sky, image/semantic/background/object=tree, image/semantic/forgeround/object=elephant, image/semantic/forgeround/object=water,

Fig. 2. Examples of top XPRM annotation in comparison with original Corel annotation

Table 1. XPRM Annotation Results

No of paths with recall > 0 is 50		
Annotation Results	all 148 paths	top 23 paths
Mean per path recall	0.22	0.83
Mean per path precision	0.21	0.73

4.2. Retrieval Results

Suppose a user wanted to find all images containing an *airplane*. To do so, we need to rank images according to the probability of annotation with the top most images having the highest probability of having an *airplane*. Hence, ranked retrieval is important in such a scenario.

With XPRM, the user can perform a semantic query and hence achieve better results. For example, the user might be interested in retrieving images that have an *airplane* in the background only. Single word queries would retrieve all images that have an *airplane* somewhere in the image. Usually, a user is not likely to view more than the first 10 to 20 images for a query retrieval performed on a large image database. As a result, even though images with *airplane* in the background might be retrieved, it is unlikely that the user might see them. The user query as a result has not been answered satisfactorily in this scenario. However with XPRM, such a retrieval can be easily performed by a path based query like "image/semantic/background/object=plane". Since the images in the database are represented and annotated in XML format too, we retrieve images that have *plane* as the background object. This is unachievable with single word queries and hence is a major drawback that XPRM successfully overcomes. Moreover, as the images in the database are stored in an XML format, we can perform retrieval by using XML query tools such as XPath and XQuery. In Table 2, we also report the mean average precision obtained for ranked retrieval as in [8].

Table 2. XPRM Mean Average Precision

All 148 paths	Paths with recall > 0
0.34	0.38

5. CONCLUSIONS

We propose the XML Path based Relevance Model that performs automatic image annotation using the semantics of XML. The proposed framework clearly has greater advantage over other conventional annotation approaches due to the semantic annotation of images using XML paths.

6. REFERENCES

[1] http://www.w3.org/XML, "World Wide Web Consortium, eXtensible Markup Language (XML)" .

[2] E.Hyvonen, A.Styrman, and S.Saarela, "Ontology-based image retrieval," in *Proc. of XML Finland Conf.*, 2002.

[3] A.T.Schreiber, B.Dubbeldam, J.Wielemaker, and B.Wielinga, "Ontology based photo annotation," *IEEE Intelligent Systems*, vol. 16, no. 3, pp. 66 – 74, 2001.

[4] B.S.Manjunath, *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, John and Sons, Incorporated, 2002.

[5] M.Rege, M.Dong, F.Fotouhi, M.Siadat, and L.Zamorano, "Using Mpeg-7 to build a human brain image database for image-guided neurosurgery," in *Proc. of SPIE International Symposium on Medical Imaging*, 2005.

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. of ECCV*, 2002, pp. 97–112.

[7] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Proc. of NIPS 03*, 2004.

[8] S.L.Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. of IEEE CVPR*, 2004.