Tag Recommendation for Photos

Gowtham Kumar Ramani, Rahul Batra, Tripti Assudani

December 10, 2009

Abstract. We present a real-time recommendation system for photo annotation that can be used in Flickr. We start with a database of 25000 images and 24 carefully chosen tags. We first extract global features that represent the shape of the image. We then train one classifier per tag to detect if the tag is relevant. Each classifier makes use of a fisher kernel followed by a Support Vector Machine. To obtain the fisher kernel, we train a Gaussian Mixture Model (GMM) on the data. For computational feasibility, we approximate the fisher information matrix by considering only the diagonal elements. Finally, for each test image, we sort the soft scores of the classifiers to rank the tags by relevancy. We compute performance metrics and compare them with previously published results. Finally we suggest possible future directions.

Key words. Photo annotation, Gaussian Mixture Model, Fisher Kernel, Support Vector Machine

1 Introduction

In recent years, tagging - the act of adding keywords (tags) to objects - has become a popular means to annotate various web resources, such as web page bookmarks, academic publications, and multimedia objects. The tags provide meaningful descriptors of the objects, and allow the user to organise and index her content. This becomes even more important, when dealing with multimedia objects that provide little or no textual context, such as bookmarks, photos and videos.

Tags may be recommended for images based either on related tags that have been previously applied to the image or on features that are extracted from an image. The former language-based approach can be applied only to images that have been tagged previously. One has to rank the recommendations based on how closely they relate to the applied tags [1]. The latter approach, though error prone, can be applied even when the image is not tagged before.

In [2], the authors describe GIST in which they represent global features using a very low-dimensional representation (termed spatial envelope) of the image.

In [3], the authors present a method to extract the signal present in collaborative communities in order to extract knowledge about tag usage. It describes methods to categorize sets of tags as places, landmarks, and visual descriptors. They use clustering of visual features followed by mutual information measurement to predict tags that can be learnt from visual features.

We built a recommendation system that identifies relevant tags based on visual features extracted by GIST. Fig. 1 shows an example of our Tag recommendation system in action.

For our work, we used the MIR FLICKR data set[4]. This image collection consists of 25000 images that were downloaded from the social photography site `Flickr.com` through its public API.

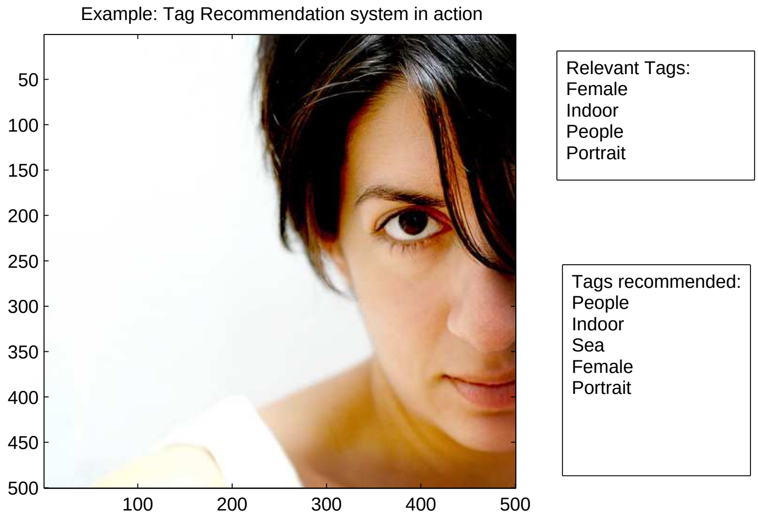

Figure 1:

The color images are representative of a generic domain and are of high quality. This is guaranteed by the high "interestingness" (`http://www.flickr.com/explore/interesting/`) of the images: this image score represents an evolving measure of quality determined by factors such as where clickthroughs on the image are coming from, who comments on it and when, or who marks the image as a favorite.

The rest of the report is structured as follows: In section 2, we describe our dataset and the feature extraction algorithm. In section 3, we describe implementation of the classifier. In section 4, we discuss results and compare our algorthim with prior work. In section 5, we conclude and suggest future directions.

2 Feature Extraction and Tag Selection

The MIR FLICKR data set contains a set of 24 most commonly occuring tags. Surprisingly, only 3% of the 25000 images cannot be described by any tag with this limited vocabulary.

We extract global features of the image using GIST [2]. It has been proven to perform well in scene classification with a overall accuracy of 83.7%.

We computed the mutual information between each tag and the visual features extracted by GIST as described in [3]. The lowest mutual information was 0.01 bits for `dog` and the highest was 0.16 bits for `clouds`. Due to high mutual information, it seemed reasonable that all 24 tags can be used for classification.

3 Our algorithm

Fig. 3 shows the block diagram of our algorithm.

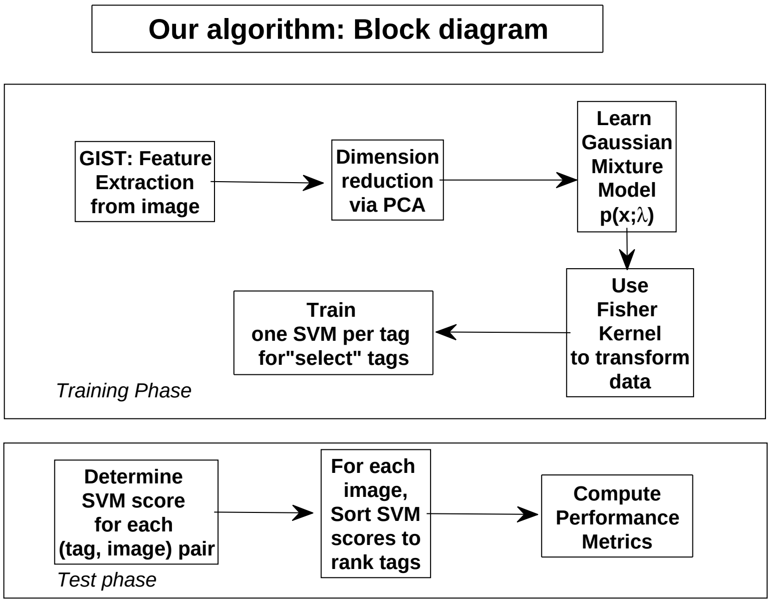
Our algorithm: Block diagram
GIST: Feature Extraction from image
Dimension reduction via PCA
Learn Gaussian Mixture Model p(x;λ)
Use Fisher Kernel to transform data
Train one SVM per tag for"select" tags
Training Phase
Determine SVM score for each (tag, image) pair
For each image, Sort SVM scores to rank tags
Compute Performance Metrics
Test phase

Figure 2:

3.1 Training a GMM

We represent each image by its $512-$dimensional feature vector extracted via GIST. As usual, we split the data in to 70% training and 30% test set.

We begin by performing a PCA to reduce the dimensionality of the data to 45. We find that more than 95% of the variance is in the first 50 dimensions.

Next, we fit a GMM to the visual features using $k = 18$ Gaussians. We follow [5] when we restrict the covariance of the distribution to be diagonal.

3.2 Fisher Kernel

The fisher kernel function is given by

$$g(x) = F^{-1/2} \nabla_\lambda \log p(x; \lambda)$$

where $F = E_X \left(\nabla_\lambda \log p(X; \lambda) \right) \left(\nabla_\lambda \log p(X; \lambda) \right)^T$ is the fisher information matrix. Again, as suggested in [5], we approximate F by a diagnoal matrix.

The fisher kernel approach combines a generative model (fitting a GMM) with a discriminative model (SVM classifier).

3.3 Final SVM classifier

Once the kernel transformation is complete, we build one SVM per tag to detect if the tag is relevant. Finally we sort the soft scores produced by the SVM to rank the recommended tags by relevancy.

4 Performance

We first define some performance metrics for a tag recommendation system:

- **Mean Reciprocal Rank (MRR):** Average of the reciprocal rank of the most relevant tag in the ranked list.
- **Success@K:** Probability that atleast one among the tags ranked K or better is relevant.
- **Precision@K:** Probability that a tag ranked K or better is a potential tag.

We report MRR, Success@1, Success@5 and Precision@5 for our vision-based algorithm and compare it with the reference algorthim presented in [1].

—	MRR	Success@1	Success@5	Precision@5
Our Algorithm	0.67	0.53	0.86	0.38
Reference Algorithm	0.79	0.70	0.95	0.52

Note that [1] presents a language-based algorithm which is generally known to outperform any vision-based approach. However, our algorithm seems to compare reasonably well with the language-based algorithm.

5 Future work

Though GIST extracts global visual features amazingly well, it does not identify local features which can often determine tags. For example, tags like `cloud` and `sea` are easy to recommend while tags like `animals` and `dogs` are far difficult. However, learning local features requires image segmentation which cannot be accomplished easily in real-time. One suggestion is to divide the image in to 4 quadrants and perform tag recommendation on each quadrant independently. Then the soft scores in each quadrant can be added and the tags can be ranked accordingly.

A second approach is to use the output of the vision-based approach as an input to a language-based approach and determine/recommend co-occuring tags (e.g. Jackard Coefficient). This is different from conventional approaches where the language-based approach runs in parallel to the vision-based approach. It helps to improve the tag vocabulary while retaining a low vocabulary that is crucial for a vision-based approach to succeed.

References

[1] Brkur Sigurbjrnsson and Roelof van Zwol, “Flickr Tag Recommendation based on Collective Knowledge”, WWW 2008 / Refereed Track: Rich Media.

[2] Aude Oliva and Antonio Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope”, International Journal of Computer Vision 42(3), 145175, 2001.

[3] E. Moxley, J. Kleban, J. Xu, and B. S. Manjunath, “Not All Tags Are Created Equal: Learning Flickr Tag Semantics for Global Annotation.”, IEEE ICME, New York, Jun. 2009.

[4] Mark J. Huiskes and Michael S. Lew, “The MIR Flickr Retrieval Evaluation”, ACM International Conference on Multimedia Information Retrieval (MIR’08), Vancouver, Canada.

[5] Perronnin, F., Dance, C., "Fisher kernels on visual vocabularies for image categorization.", Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.