A bipartite graph model for associating images and text

S H Srinivasan
Technology Research Group
Yahoo, Bangalore, India.
shs@yahoo-inc.com

Malcolm Slaney
Yahoo Research Lab
Yahoo, Sunnyvale, USA.
malcolm@ieee.org

Abstract

The joint modeling of image and textual content is even more important now because of the the availability of large databases of image-rich web pages and the tagging phenomenon. Much of the current work focused on one-way association (image to text or tags). The association is often captured by building a model with hidden variables. In this paper, we propose a simple model based on random walks on bipartite graphs for joint modeling of image and textual content. We show its effectiveness for several tasks — automatic image annotation, tag association, tag localization, and spurious tag detection. Such random walk models are useful for other tasks such as web search.

1 Introduction

The availability of image data with associated text is now very common. Images on the Internet occur in the context of web pages. The text surrounding an image is usually a good description of the image content. In addition, users of sites such as `flickr.com` contribute tags to describe each photo. All such data can be used to learn the joint statistics of visual and linguistic content and can be used for tasks such as *automatic image annotation*, *tag localization*, *tag clustering*, etc.

In this paper, we describe a new model for text–image content associations. We use the "bag of visual words" paradigm to describe visual content and bipartite graphs to model associations between linguistic and visual content. Using this model it is possible to

1. provide descriptions for new images (automatic image annotation)
2. detect tag associations
3. locate image regions with descriptions
4. detect spurious tags

This paper is organized as follows. Section 2 provides an overview of the related work. Section 3 discusses the bipartite graph model. Section 4 presents the experimental results on two databases. We conclude the paper with a critical discussion on the proposed technique.

2 Related Work

Joint keyword–image modeling has a relatively short but rich history. Two important issues in joint modeling are: image representation and the statistical modeling. Images are represented as either collections of blobs [Barnard *et al.*, 2003] or as collections of salient points [Bosch *et al.*, 2006]. Each blob is described by features – color and texture vectors. There are several techniques for detection of interest points [Schmid *et al.*, 2000]. Interest points are usually represented by Scale Invariant Feature Transform or SIFT [Lowe, 2004]. The feature vectors are often vector quantized for representational simplicity. The vector quantized features are a form of "visual word" and then the joint modeling problem is a machine translation problem [Duygulu *et al.*, 2002].

Once the representational issues are taken care of, there are several choices for the statistical modeling itself. Some of the statistical models described in the literature are Probabilistic Latent Semantic Indexing [Hofmann, 1999], Latent Dirichlet Allocation [Blei *et al.*, 2002], Correspondence LDA [Blei and Jordan, 2003], Bernoulli model [Feng *et al.*, 2004], and 2D HMMs [Li and Wang, 2003].

To make things concrete, we describe one such model – Gaussian-Multinomial LDA. Let r_n be the region descriptors and w_m the caption words. Let θ be the Dirichlet random variable generating the latent factors or "topics". The latent factors themselves are represented by z_n and v_m. Note that there are two sets of hidden variables — one for region descriptors and one for caption words. Then the joint distribution of image regions, caption words, and the latent variables is given by [Blei and Jordan, 2003]

$$\begin{aligned} p(\boldsymbol{r}, \boldsymbol{w}, \theta, \boldsymbol{z}, \boldsymbol{v}) &= p(\theta|\alpha)(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n, \mu, \sigma)) \\ &\quad \times(\prod_{m=1}^{M} p(v_m|\theta)p(w_m|v_m, \beta)). \end{aligned}$$

where N is the number of regions, M the number of words, μ and σ the means and variances of the Gaussians used to model image features, β the parameter for the multinomial distribution used for generating words. Estimation of the latent variables is usually complex. The next section introduces a simple parameterless model for joint modeling.

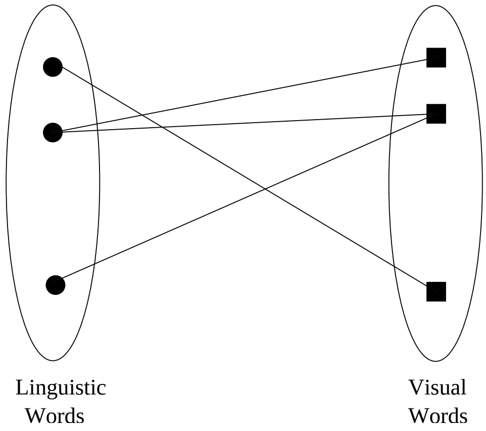

Figure 1: Bipartite graph model. The two partitions correspond to *linguistic words* and *visual words*. The two partitions are denoted by superscripts L and V in the main text. The sizes of the two partitions are M and N respectively. If π is an $(M+N) \times 1$ probability vector over the nodes of the graph, we write $\pi = [\pi^L \;\; \pi^V]$ where π^L is an $M \times 1$ and π^V is an $N \times 1$ probability vector corresponding to nodes in the partitions L and V respectively.

3 Bipartite graph model

We use the "bag of visual words" approach for representing image content. We start with a database of images and associated keywords. We perform salient-point detection on each image, describe the salient points through SIFT descriptors, and represent each descriptor by VQ codebook index. We thus have a set of linguistic and visual words for each image.

We construct a bipartite graph with linguistic words as one partition and visual words as the other. (See Figure 1.) There is an edge between a linguistic word w and a visual word v if w is used to describe an image in which v occurs. Each edge is labeled with a probability proportional to the number of times the visual word v occurs in the image. The edge probabilities are normalized in the usual manner. (The sum of edge probabilities at any node is 1.) We perform random walk on this bipartite graph and the stationary distribution is represented as π_0. If P_A is the connection diagram or transition matrix of the bipartite graph, then the stationary probability π_0 is given by

$$\pi_0 = P_A \pi_0.$$

π_0 has two components: π_0^L (stationary distribution over linguistic words) and π_0^V (stationary distribution over visual words). This corresponds to classical notion of stationary probability.

To find visual words associated with a given linguistic word w, we perform random walk starting at the node corresponding to w and with *restart* probability $\lambda > 0$. The notion of random walk with restarts for bipartite graphs was introduced by Sun *et al.* [Sun *et al.*, 2005]. The stationary distribution π_0 is independent of the initial probability for ergodic Markov chains.

Now assume that we want to restart at a linguistic word w with probability λ. Let q_w be a vector with 1 at the position corresponding to w and 0 at other positions. The stationary probability of the random walk with restarts is given by [Sun *et al.*, 2005]

$$\pi_w = (1-\lambda) P_A \pi_w + \lambda q_w.$$

Note that the stationary probability with restarts depends on the initial state w. We can write π_w as $[\pi_w^L \;\; \pi_w^V]$ corresponding to the two partitions. We use this to find several associations – linguistic word to visual word associations and associations between linguistic words themselves.

A note on the general strategy for finding associations follows. Assume that we are interested in associations between linguistic words and visual words. We expect visual words which have high values in π_w^V to be associated with w. This is true for large sparse graphs. The graphs constructed in the experiments reported in this paper are small (less than 2000 nodes) and dense. So π_w^V is "biased" by the dense connectivity of the graph despite restarts at w. Assuming that π_0^V captures the "bias", we use $\pi_w^V - \pi_0^V$ to measure association of visual words with w.

Word to visual word association: To find the visual words associated with w, we threshold the values of $\pi_w^V - \pi_0^V$. This gives the visual words most strongly associated with w. We call $\pi_w^V - \pi_0^V$ the *association strength* or *association score* of the visual words with the linguistic word w.

Word associations: $\pi_w^L - \pi_0^L$ is a measure of associations of linguistic words with w. Using this, we can capture the scene similarity between linguistic words.

Word to visual word associations can be used for region labeling and detecting spurious labels. The following section contains the details.

4 Experiments

We have used two different databases in our experiments. A brief description of these follows. We used Harris corner detector for salient point detection. The 128-dimensional SIFT descriptor was used to describe the salient points. The SIFT descriptor captures texture information around the keypoints. We did not use any color descriptor.

FP database: The 13-category database is used in [Fei-Fei and Perona, 2005]. This consists of images belonging to 13 categories: bedroom, suburb, kitchen, livingroom, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, and office. The database has a total of 3859 images. The images are in gray scale with an average size 270 × 246. The images in the database fall into two higher-level categories: natural (coast, forest, mountain, opencountry) and human-made (others). We used 1300 images (100 per category) for training and the remaining 2559 for testing.

The training images altogether contained approximately $425,000$ salient points and these descriptors were vector quantized to 1024 visual code words (centroids) using the LBG technique [Linde *et al.*, 1980].

UW database: The second database is from University of Washington `http://www.cs.washington.edu/`

`research/imagedatabase`. This database consists of approximately 1500 images. The images belong to 21 classes: arborgreens, australia, barcelona, cambridge, campusinfall, cannonbeach, cherries, columbiagorge, football, geneva, greenlake, greenland, indonesia, iran, italy, japan, leaflesstrees, sanjuans, springflowers, swissmountains, and yellowstone. The images are annotated using keywords such as trees, sky, etc. Unlike FP database, each image is annotated with multiple keywords. Annotations exist for 1109 images. We used 917 images for training and 94 for testing.[1] The average image size is 778×554.

For this database too, we used the Harris corner detector and SIFT descriptor. Since the images are large in size, the training images contained approximately 2 million salient points. Out of these, 200,000 descriptors were randomly chosen as training vectors for obtaining the vector quantization codebook using the LBG technique. This was done to speed up the VQ calculations. The number of visual words were 2048.

We performed four different experiments: predicting similar keywords based on associated common visual words, automatically predicting the category of an image based on visual words, labeling regions of an image, and detecting spurious tags. These are described in the following subsections.

4.1 Keyword association

Keywords used to describe similar images are related. We demonstrate this over the 13-category database. Each image is described by one keyword and the keywords are disjoint. So any relation between keywords is inferred through image content similarity. Figure 2 shows the association of keywords as given by the bipartite graph model. We display association score ($\pi_w - \pi_0$) for $w =$"bedroom", "forest", and "mountain". For the keyword "bedroom", the keywords with high association scores are: "kitchen", "livingroom", and "office". For "forest", these are "opencountry" and "mountain". For "mountain", the associated keywords are "coast", and "suburb".

4.2 Automatic annotation

In the previous subsection, we associated the keywords with each other. We can also associate keywords and visual words. We first find the association scores of visual words with a given keyword. Figure 3 shows the association scores for "bedroom". We choose words which are higher than a threshold. Let W be the keyword and let A_W be the set of visual words which are associated with W. For a test image I, let V_I be the set of visual words that occur in I. $|V_I \cap A_W|$ is a measure of how W describes I.

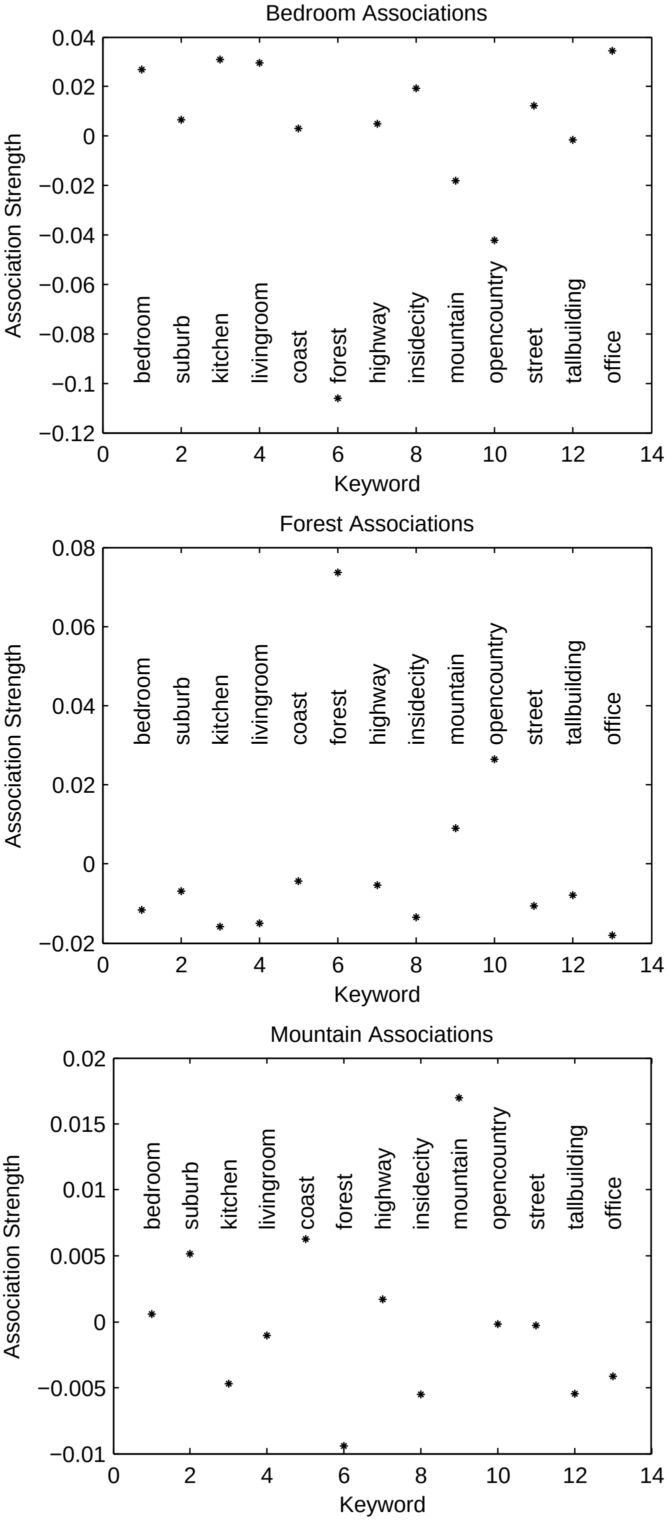

Figure 2: Association strengths for keywords: bedroom, forest, and mountain. Other keywords with high associaiton with the chosen keyword have high association stregths.

[1] All the experiments reported in the paper used Harris corner detector. Some experiments, not reported here, used the Harris-Affine detector. For sake of comparison, out of 1109 images which have annotations, we eliminated images that had no keypoints detected by the Harris-Affine detector. We used Linux binaries from `http://www.robots.ox.ac.uk/~vgg/research/affine/` for both Harris corner detection and Harris-Affine salient point detection.

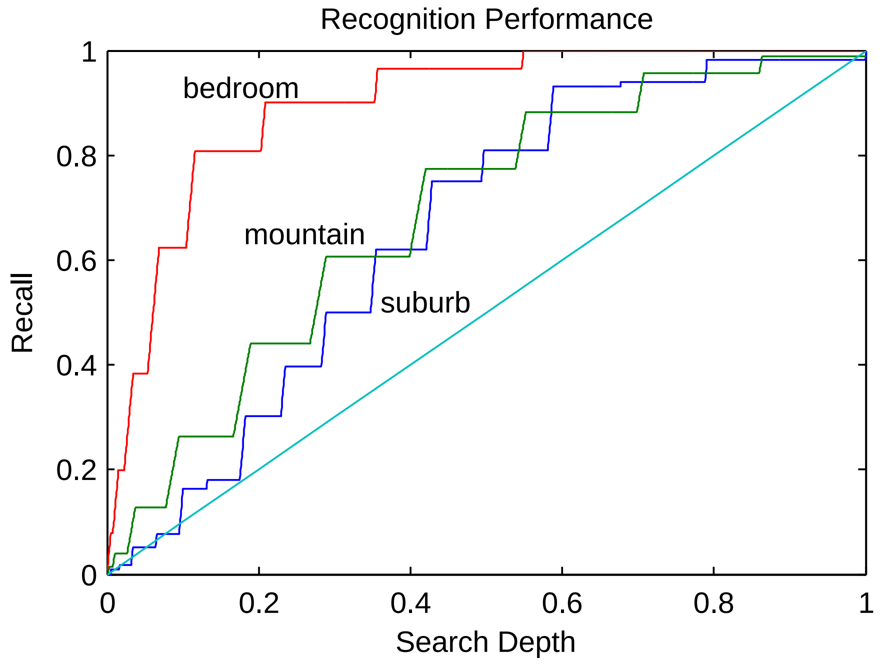

Figure 4: ROC for automatic annotation. *Depth* denotes the fraction of the images top ranking results and *recall* the fraction of the images of a given category present in it. If the images of a given category are randomly distributed in the results, the performance corresponds to the $45°$ line.

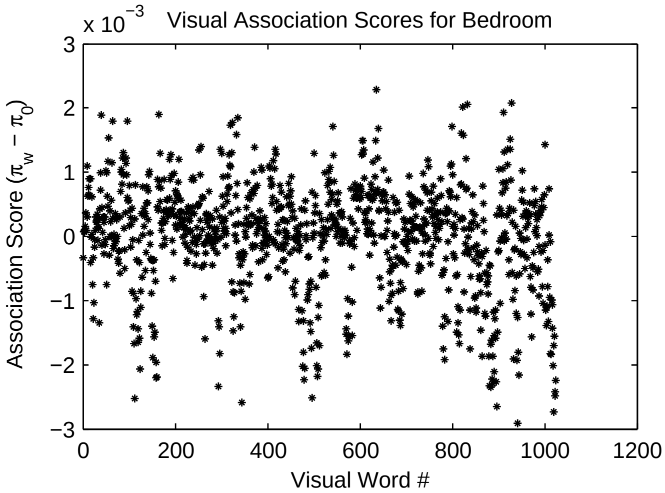

Figure 3: Association scores of visual words with “bedroom”.

The test set contains 2559 images from the FP database. For each of the categories “bedroom”, “mountain”, and “suburb” we rank ordered the the test set based on the above score. Based on this ranking, the *depth vs recall* is shown in Figure 4.

4.3 Region annotation

We use the UW database for a region annotation task since each image has multiple annotations and there is no explicit association of the keywords with image regions. The task is to associate each keyword with image regions in an *unsupervised* manner.

We used 917 training images for constructing the bipartite graph. The stationary distribution of this graph is shown in Figure 5. For the linguistic word “tree”, the associated probabilities are shown in Figure 6 and the difference of the two in Figure 7. Visual words which have high scores are selected from this data. These visual words are the visual correlates of “tree”.

The visual words corresponding to “tree” on a test image are highlighted in Figure 8. The same figure contains a example for “building”. Figure 9 shows the occurrences of “building” visual words on the tree image and “tree” visual words on the building image. The occurrences are few in number. The positive examples (Figure 8) show salient points on an image corresponding to the correct keyword label, while negative examples (Figure 9) show visual words that are essentially false positives.

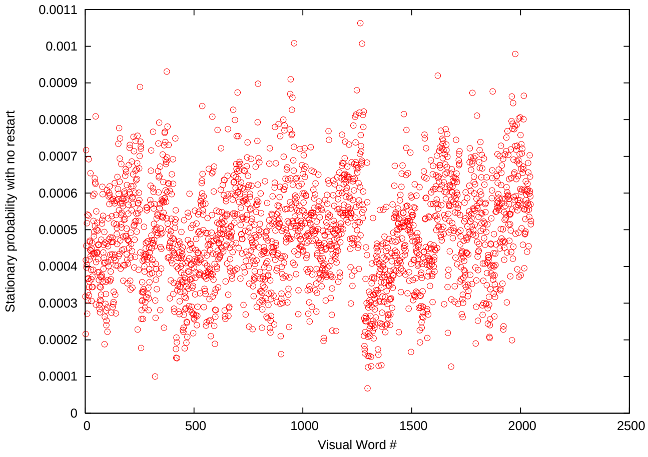

Figure 5: Stationary probability of the underlying Markov chain for the University of Washington dataset. Only the probabilities correspond to visual words (π_0^V) are shown.

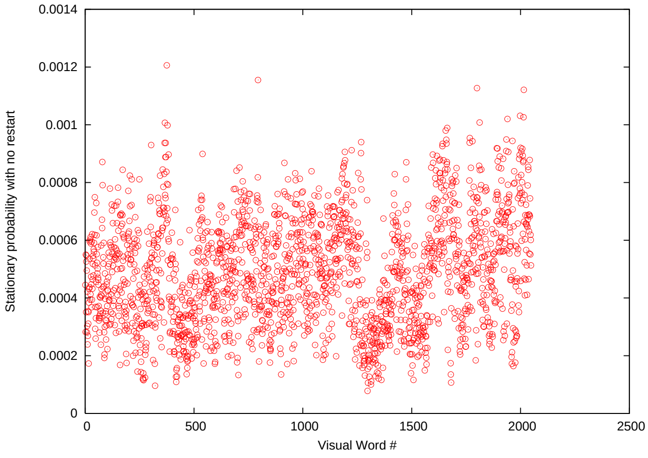

Figure 6: Stationary probability for the word "trees" (π_{trees}^V). (University of Washington dataset)

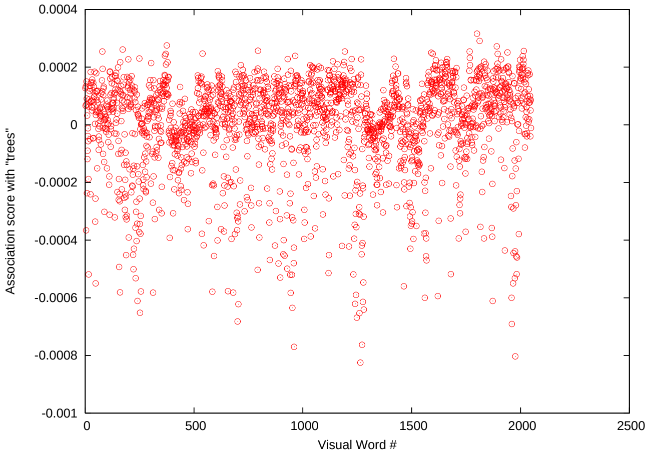

Figure 7: Association scores ($\pi_{trees}^V - \pi_0^V$) for the linguistic word "trees". (University of Washington dataset)

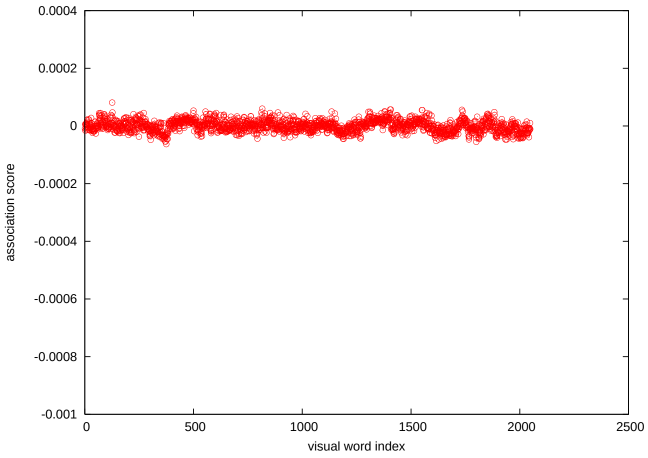

Figure 10: Spurious tag detection. The association scores of visual words with a tag that is used to describe *all* the images. The association scores are small. See Figure 7 for comparison.

4.4 Spurious tag detection

Often, tags used to describe an image are not about objects *in* the image but *about* the image itself. Tags like "Tokyo" do not describe anything in the image; they are about the image. We provide a framework for detecting such tags. We added a spurious tag to *all* the images in the University of Washington dataset. Figure 10 shows the association score for this tag. For ease of comparison, the axes scales are the same as that of Figure 7. The association scores are small for the spurious tag because the tag is not consistent with any one category.

5 Discussion

In this paper, we have proposed a model for image-keyword associations based on bipartite graphs. The results show that the model performs well on several tasks. Since random walks on graphs are the backbone of modern Internet search [Brin and Page, 1998], it is likely that the technique scales well to very large datasets. Before large scale experimentation, a systematic study on the effect of various parameters like — sparse versus dense description, use of color, effect of number of visual words — is required.

References

[Barnard *et al.*, 2003] K Barnard, P Duygulu, N de Freitas, D Forsyth, D Blei, and M I Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 2003.

[Blei and Jordan, 2003] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR conference*, 2003.

[Blei *et al.*, 2002] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In *NIPS*, 2002.

[Bosch *et al.*, 2006] A Bosch, A Zisserman, and X Munoz. Scene classification via pLSA. In *European Conference on Computer Vision*, 2006.

[Brin and Page, 1998] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 1998.

[Duygulu *et al.*, 2002] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, 2002.

[Fei-Fei and Perona, 2005] L Fei-Fei and P Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[Feng *et al.*, 2004] S L Feng, R Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, pages 1002–1009, 2004.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, 1999.

[Li and Wang, 2003] Jia Li and James Ze Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell*, 25(9):1075–1088, 2003.

[Linde *et al.*, 1980] Y Linde, A Buzo, and R M Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 1980.

[Lowe, 2004] D G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[Schmid *et al.*, 2000] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 2000.

[Sun *et al.*, 2005] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, 2005.

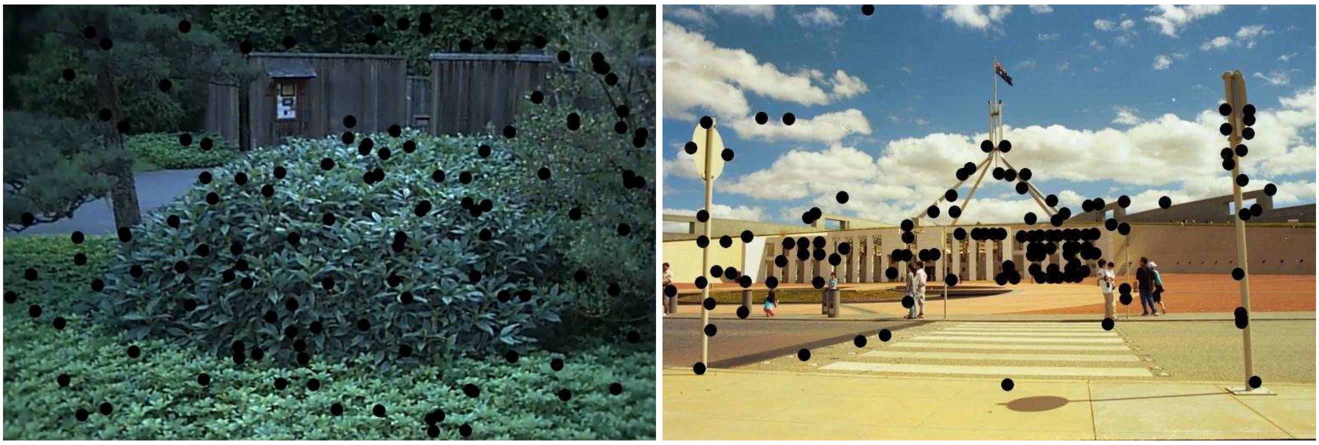

Figure 8: Annotation regions for “tree” (Left) and “building” (Right). (Left) The visual words in a test image corresponding to linguistic word “tree” are by black circles. (Right) The visual words for “building” highlighted on a test image.

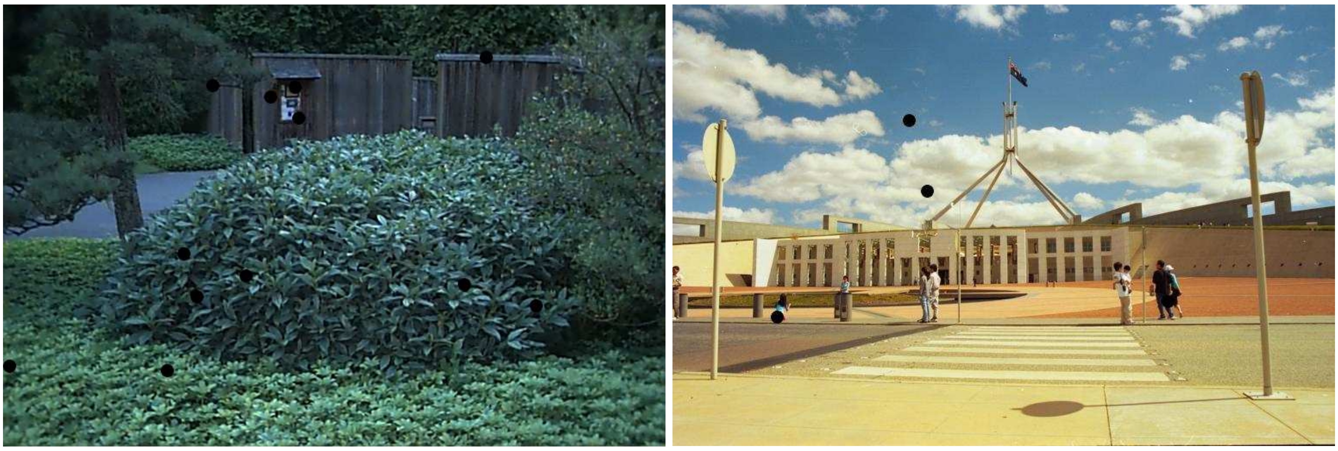

Figure 9: ”Cross words”. (Left) “Building” visual words on tree image. (Right) “Tree” visual words on building image. There are 12 occurrences of “building” visual words on the tree image and 4 occurrences of “tree” keywords on the building image.