Geometric LDA: A Generative Model for Particular Object Discovery

James Philbin[1] Josef Sivic[2] Andrew Zisserman[1]

[1] Visual Geometry Group, Department of Engineering Science, University of Oxford
[2]INRIA, WILLOW Project-Team, Ecole Normale Superieure, Paris, France

Abstract

Automatically organizing collections of images presents serious challenges to the current state-of-the art methods in image data mining. Often, what is required is that images taken in the same place, of the same thing, or of the same person be conceptually grouped together.

To achieve this, we introduce the Geometric Latent Dirichlet Allocation (*g*LDA) model for unsupervised particular object discovery in unordered image collections. This explicitly represents documents as mixtures of particular objects or facades, and builds rich latent topic models which incorporate the identity and locations of visual words specific to the topic in a geometrically consistent way. Applying standard inference techniques to this model enables images likely to contain the same object to be probabilistically grouped and ranked.

We demonstrate the model on a publicly available dataset of Oxford images, and show examples of spatially consistent groupings.

1 Introduction

Our goal is to automatically "discover" significant objects and scenes in photo collections.

In the statistical text community, latent topic models such as probabilistic Latent Semantic Analysis (pLSA) [13] and Latent Dirichlet Allocation (LDA) [2] have had significant impact as methods for "semantic" clustering. Given a collection of documents such as scientific abstracts, with each document represented by a bag-of-words vector, the models are able to learn common topics such as "biology" or "astronomy". The models are able to associate relevant documents, even though the documents themselves may have few words in common.

Given the success of these models, several vision papers [9, 19, 20, 21] have applied them to the visual domain, replacing text words with visual words [7, 23]. The discovered *topics* then correspond to discovered visual *categories*, such as cars or bikes in the image collection. However, in the visual domain there are strong geometric relations between images, which simply do not exist in the text domain. There has been only a limited exploration of these relations in visual latent models: for incorporating segmentation [4, 20, 27, 28]; or for a grid-based layout of images and objects [3, 10, 14, 22].

In this paper we develop a generative latent model with geometric relations at its core. It is an extension of LDA, with a geometric relation (an affine homography) built into the generative process. We term the model *g*LDA for "*Geometric* Latent Dirichlet Allocation". The latent topics represent objects as a distribution over visual words *and their positions* on a planar facet, like a "pin-board". The visual words in an image (including

BMVC 2008 doi:10.5244/C.22.39

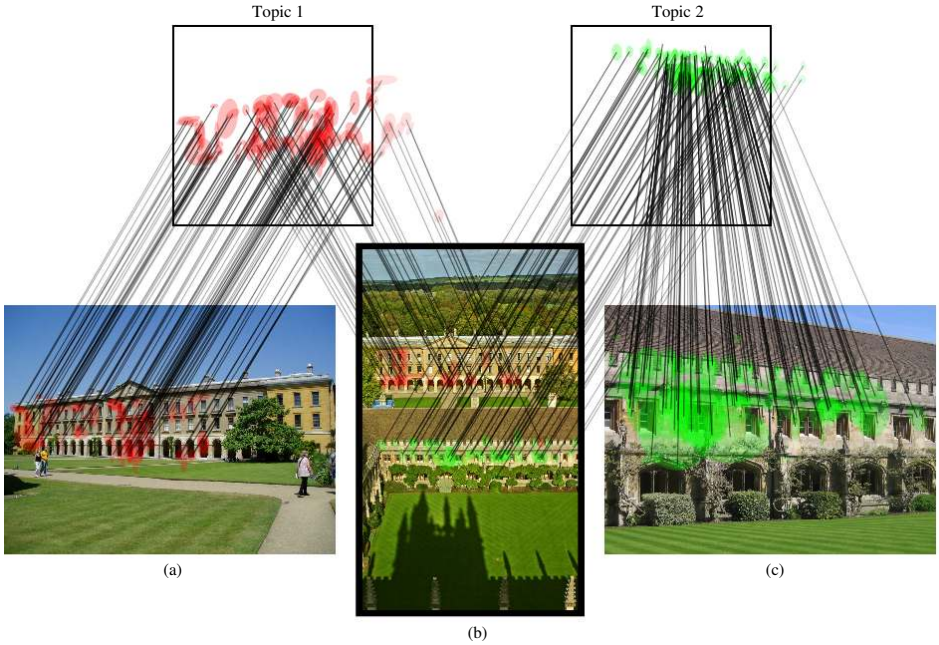

Figure 1: **The generative model**. The two topic models (above) generate the visual words and their layout in the three images (below). Each topic model can be thought of as a virtual pinboard, with the words pinned at their mapped location. Image (a) is generated only from topic 1 with a single affine transformation, and image (c) from topic 2, again with a single transformation. Image (b) is a composite of topic 1 under one homography (for the rear building) and topic 2 under a different homography (for the front building). This is a small subset of the images and topics learnt from the set of images shown in figure 4. The lines show the inliers to each topic model. The *g*LDA model correctly identified the Georgian facade (topic 1) and cloisters (topic 2) as being separate objects, despite the linking image (b), and has correctly localized these two objects in all three images.

location and shape) are then generated by an affine geometric transformation from this pinboard topic model. The generative process is illustrated in figure 1. We show that this model can be learnt in an unsupervised manner by a modification of the standard LDA learning procedure which proposes homography hypotheses using a RANSAC-like procedure. The results demonstrate that this model is able to cluster significant objects in an image collection despite large changes in scale, viewpoint, lighting and occlusions. Additionally, by representing images as a mixture over topics, the method effortlessly handles the presence of multiple distinct objects.

The model can be thought of as an automated version of the Total Recall [6] retrieval system – in Total Recall, a human selects a region of an image to issue as a search query for an object; the system then returns images containing the object, and in turn uses these images to issue new queries (spatial consistency is used for the relevance feedback), in the process building up a latent model of the object. In *g*LDA every image of the corpus provides a query, and the topic models learnt are those queries which have significant returns.

To demonstrate the model we target the discovery of significant buildings or facades in a collection of flickr images. We use the 5K image Oxford Building dataset from [1]. This contains images of many Oxford landmarks as well as many "distractor" images. It

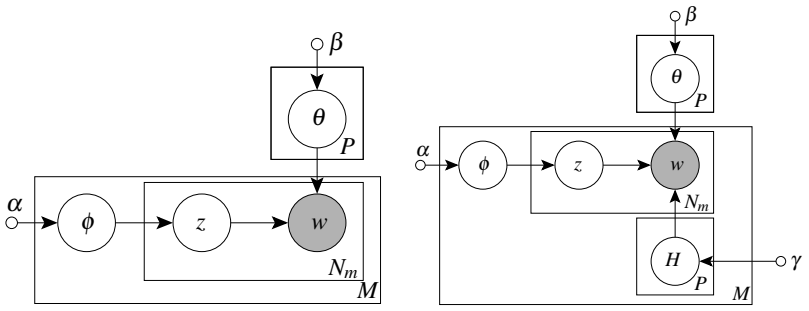

Figure 2: (a) The standard LDA model. (b) The *g*LDA model. M, P and N_m are the number of documents, topics and words (in document m) respectively.

provides a ground truth annotation for certain landmarks that we use for measuring the performance of *g*LDA.

There is limited prior work on clustering particular objects under changes in viewpoint and scale from image collections. The two most convincing recent examples are clustering particular objects (such as people, scenes) in video [18, 24]. For object and scene categories, Sudderth *et al.* [25] introduced spatial transformations in a generative framework describing locations of multiple objects and object parts within the scene, but their focus is on modelling intra-class variations rather than scale and viewpoint variations for particular objects.

2 Object Discovery

In this section, we review the standard LDA model and then describe how geometric information is incorporated in the *g*LDA model.

2.1 Latent Dirichlet Allocation (LDA)

We will describe the LDA model with the original terms 'documents' and 'words' as used in the text literature. Our visual application of these (as images and visual words) is given in the following sections. Suppose we have a corpus of M documents, $\{\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_M\}$, containing words from a vocabulary of V terms, where $\mathbf{w}_i$ is the frequency histogram of word ids for document i. A document is generated in the LDA model by picking a distribution over topics and then picking words from a topic dependent word distribution.

Figure 2(a) shows the various components of this model. The document specific topic distribution ϕ is sampled from a Dirichlet prior with parameters α. Similarly the topic specific word distribution θ is sampled from a Dirichlet prior with parameters β. The z variable is a topic indicator variable, one for each observed word, w. The aim is to find the topic distributions which best describe the data by evaluating the posterior distribution $P(\mathbf{z}|\mathbf{w}, \alpha, \beta) \propto P(\mathbf{z}|\alpha)P(\mathbf{w}|\mathbf{z}, \beta)$. These last two terms can be found by integrating out θ and ϕ respectively. Inference can be performed over this model by using a Gibbs sampler [12] with the following update formula:

$$P(z_{ij} = k|\mathbf{z}_{-ij}, \mathbf{w}) = \frac{n_{i\cdot k} + \alpha}{n_{i\cdot\cdot} + P\alpha} \times \frac{n_{\cdot jk} + \beta}{n_{\cdot\cdot k} + V\beta}. \quad (1)$$

In this equation, z_{ij} is the topic assigned to the jth word in the ith document, n_{ijk} is the number of words from document i, word id j assigned to topic k. A $\cdot$ denotes a summation over that parameter. P and V denote the number of topics and words respectively. $\mathbf{z}_{-ij}$ denotes the current topic assignments for all words except the ijth. Note that in equation (1), the first term assigns higher probability to topics occurring more frequently in the particular document, and the second term assigns higher probability to words more frequently occurring in the particular topic.

2.2 Geometric LDA

In gLDA, the topics of the LDA model are augmented with the spatial position of the visual words. Given a set of such latent topics, which may be thought of as pin-boards (with the visual words pinned at their positions), an image is generated by picking a distribution over the pinboards, and for each pinboard sampling an affine homography; the image is then formed by the composition of the visual words from each topic mapped under the corresponding homography – i.e. for each document, we pick a distribution over topics, and then for each word pick a topic from this distribution, together with a transformation from the latent spatial model to the document, and then pick a word plus spatial location and shape. The gLDA model is shown in figure 2(b).

Note, an image will not contain all the words belonging to a topic. This is necessary in the visual domain because not all visual words will be detected – there are errors due to feature detection (such as drop out, or occlusions), feature description and quantization. Others have handled this situation by learning a sensor model [8].

gLDA adds extra spatial transformation terms, H, to the LDA model and augments the word terms, $\mathbf{w}$, to contain both the visual identity and spatial location of the word in the image. These image specific transformations, H, describe how the words for a particular topic occurring in an image can be projected into the "pin-board" model for that topic. H is assumed to be an affine transformation, so that the model can account for moderate changes in viewpoint between the topic and the image.

For a particular word in an image, the joint probability of the gLDA model factors as follows

$$P(w,z,\theta,\phi,H|\alpha,\beta,\gamma) = P(w|z,\theta,H)P(z|\phi)P(\theta|\beta)P(\phi|\alpha)P(H|\gamma). \tag{2}$$

The generative distributions could be further specified and inference on the model carried out in a similar manner to [25]. However, to avoid the expense of generatively sampling the transformation hypotheses, we instead approximate the joint as described next.

2.3 Approximate inference

To perform approximate inference in the gLDA model we make the following simplifying assumptions: (i) Rather than modelling the full generative likelihood of the pinboard model, $P(w|z,\theta,H)$, we ignore the location of individual visual words and assume that the likelihood is independent of transformation H, i.e. $P(w|z,H,\theta) \approx P(w|z,\theta)$, where $P(w|z,\theta)$ is the standard LDA multinomial likelihood; (ii) To take into account the degree of spatial match between the pinboard model and an image we estimate the probability $P(H|\gamma)$ of transformation hypothesis H using a RANSAC matching procedure, commonly used in multi-view geometry. Here we take $P(H|\gamma)$ to be proportional to the number of inliers between the particular pinboard and the image.

Figure 3: Some of the Oxford landmarks.

The two approximations above allow us to integrate out θ and ϕ using Dirichlet priors, $P(\theta|\beta)$ and $P(\phi|\alpha)$, as in standard LDA to obtain the following joint probability over all images in the corpus $P(\mathbf{z},\mathbf{H}|\mathbf{w},\alpha,\beta,\gamma) \propto P(\mathbf{z}|\alpha)P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{H}|\gamma)$. The goal is to obtain samples from this joint. The pinboard assignments z_{ij} are resampled using a modified Gibbs sampler (1) with the following update formula

$$P(z_{ij}=k|\mathbf{z}_{-ij},\mathbf{w}) = \frac{n_{i\cdot k}+\alpha}{n_{i\cdot\cdot}+P\alpha} \times \frac{n_{\cdot jk}+\beta}{n_{\cdot\cdot k}+V\beta} \times \frac{q_{ik}+\gamma}{q_{i\cdot}+P\gamma} \tag{3}$$

where q_{ik} is the number of inlier visual words between the pinboard k and image i, and γ is a smoothing constant preventing assigning zero probability to topics with no inliers. Note how inlier counts q_{ik} influence re-sampling of pinboard indicators z_{ij} by assigning higher probability to pinboards with higher number of inliers. The transformation hypotheses are estimated using RANSAC (for details see section 3.2). For each document we alternate between resampling indicators $\mathbf{z}_{i\cdot}$ given the current transformation hypotheses $\mathbf{H}$ and resampling the transformation hypothesis $\mathbf{H}_{i\cdot}$ given the current indicators $\mathbf{z}$.

Note that the interleaved sampling of pinboard assignments $\mathbf{z}$ using (3) and transformation hypothesis $\mathbf{H}$ using RANSAC can be viewed as data driven Markov Chain Monte Carlo in the spirit of [26].

3 Dataset and Spatial Consistency

3.1 Dataset

For evaluating the performance of the gLDA model both quantitatively and qualitatively, we use the Oxford dataset [1]. This consists of 5,062 high resolution (1024×768) images automatically retrieved from Flickr, together with groundtruth occurrences for 11 different landmarks in Oxford. A sample of 5 landmark images is shown in figure 3. Note that the dataset also contains many images of other buildings and non-buildings.

To generate visual features for this dataset, we detect Hessian interest points and fit affine covariant ellipses [16]. For each of these affine regions, we compute a 128-dimensional SIFT descriptor [15]. A large discriminative vocabulary of 500K words is generated using an approximate k-means clustering method [17]. Each descriptor is assigned to a single cluster centre to give one visual word. On average, there are $\sim$3,300 regions detected per image. Once processed, each image in the dataset is represented as a set of visual words which include spatial location and the affine feature shape.

3.2 Spatial scoring using RANSAC

Our discriminative gLDA model relies on being able to score the spatial consistency between two spatially distributed sets of visual words (e.g. between the pinboard model and an image) and return an approximate transformation between the two sets of visual words as well as a matching score. The score is based on how well the feature locations are predicted by the estimated transformation.

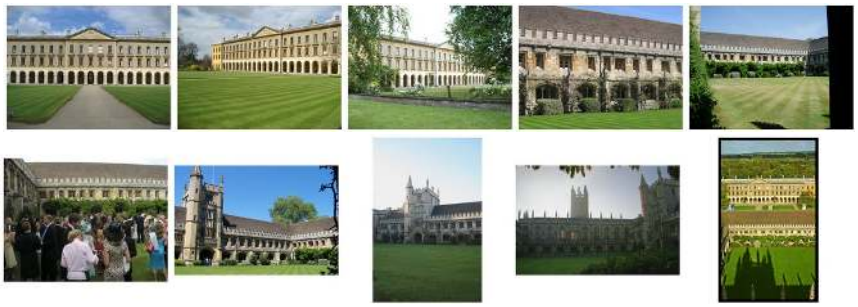

Figure 4: A sample of 10 images from a connected component associated with Magdalen college. The component contains two separate buildings: A Georgian building and a college cloisters, linked by the aerial photo shown (bottom right). Within the cloisters there are two distinct "facades", one of the wall, the other of a tower. Our method is able to extract all three "objects" (cloisters, tower and building) completely automatically. The total size of the component is 42 images.

We use a deterministic variant of RANSAC [11]. It involves generating hypotheses of an approximate transformation [17] and then iteratively re-evaluating promising hypotheses using the full transformation. By selecting a restricted class of transformations for the hypothesis generation stage and exploiting shape information in the affine-invariant image regions, we are able to generate hypotheses with only a *single* feature correspondence. We enumerate all such hypotheses, resulting in a deterministic procedure. The inliers for a given transformation are the set of words which approximately agree with that transformation. The size of this inlier set for the best transformation is directly used as a measure of matching quality as outlined in section 2.3.

4 Object Discovery in Large Image Collections

There are issues of scalability in applying the *g*LDA model directly to large image collections. Every word in the dataset must be present in a topic model and RANSAC is repeatedly run between these large topic models and each document. This becomes prohibitively slow when the amount of data becomes large.

4.1 Pre-clustering

To apply our method to large datasets we run a much simpler pre-clustering step on the data, which splits it into a number of disjoint image sets. The aim is to pull all images that might possibly contain the same object into the same cluster while excluding all images which definitely do not contain the object. To do this we build a pairwise matching graph between every pair of images in the dataset using fast methods from particular object retrieval [17]. This is done by building an inverted index for all the identities of the visual words in the dataset, then querying for each image in turn using this index and performing spatial verification only for the top 500 returns. The worst case complexity of building the graph is $O(N^2)$ in the number of images, N, but, in practice, the time taken is close to linear in N.

We compute connected components over this graph thresholding at a particular similarity level. This similarity is specified by the number of spatially consistent inliers between each image pair. In general, the connected components now contain images linked together by some chain of similarities within the cluster, but will not necessarily be of the

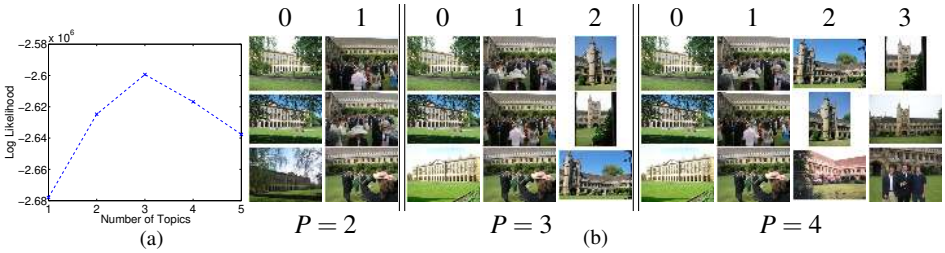

Figure 5: Automatically choosing the number of topics. We run the *g*LDA model over the connected component shown in figure 4 for different numbers of topics plotting the log likelihood in (a). (b) shows the top three documents (ranked by $P(z|d)$ in columns) for each topic for different numbers of topics, P. In this case three topics are chosen which separate the building, cloisters and tower.

same object. For example, "linking" images containing more than one object will join other images of these objects into a single cluster (see figure 4).

Applying the graph pre-clustering method to the 5K Oxford dataset, linking all images with at least 25 spatially consistent inliers between them results in 323 separate components containing 2 or more images. The size of the largest component is 396 images. The *g*LDA model can now be applied independently in each of these connected components. The scale of the problem within each run of the model is much reduced and completely irrelevant images are not considered.

4.2 *g*LDA implementation details

Topic initialization. For each connected component of the matching graph the topics are initialized by first obtaining P separate clusters (using agglomerative clustering with the average linkage as the similarity score). For each cluster, we project each document's words to a normalized size in the pinboard models: a transformation is found that projects each image to a fixed size square in the topic model and these are used to initialize the locations and shapes of the visual words in the model. Although this isn't strictly necessary for the *g*LDA model, it greatly improves convergence speed and generally leads to improved results.

Prior parameters. The *g*LDA model (section 2.3) includes priors for the per document topic distribution, α, the per topic word distribution, β, and the hypothesis likelihood, γ. Empirically we find that using $\alpha = 200.0$, $\beta = 1.0$ and $\gamma = 20.0$ gives reasonable results and we use these parameter values for all subsequent experiments.

Choosing the number of topics. To select the number of topics within each connected component, we run 100 iterations of the Gibbs sampler described in section 2.3 changing the number of topics from 1 to 8, then choose the Markov chain with the highest likelihood (see figure 5) [12]. We note here that it is better to choose too many topics than too few as the model explicitly allows for documents to be a mixture of topics. In general, the optimal number of topics found will vary with the choice of hyper-parameters, α and β.

Running the model. After the number of topics has been selected, we run the model for a further 100 iterations. We find that with the geometric information, the *g*LDA model

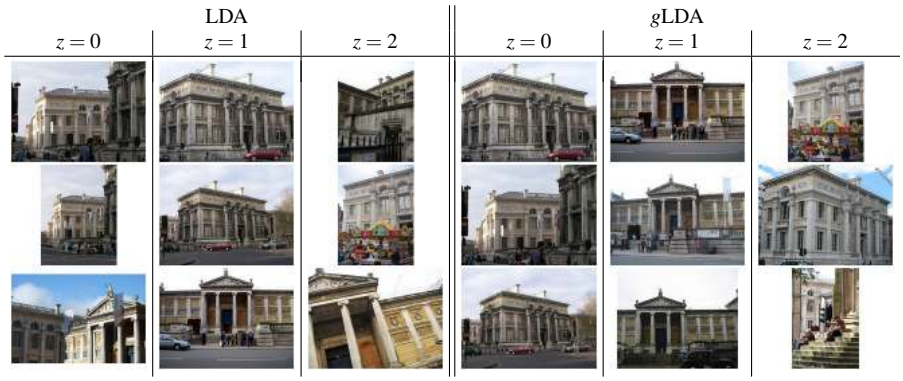

Figure 6: Comparing *g*LDA to standard LDA for a connected component containing images of the Ashmolean, for $P = 3$. The top three images are shown for each topic, ranked by $P(z|d)$ in columns. LDA has failed to differentiate the facade of the Ashmolean museum from a separate building.

tends to converge to a mode extremely quickly and running it longer brings little appreciable benefit.

Scalability. The time taken to run the *g*LDA model varies from a fraction of a second per iteration for a component of less than 5 images up to about 55s per iteration for the largest component of 396 images on a 2GHz machine.

5 Results

In this section we examine the performance of the *g*LDA both qualitatively and quantitatively. For the quantitative evaluation we determine if the discovered topics coincide with any of the groundtruth labelled Oxford landmarks.

Evaluation on the Oxford dataset. Within each connected component, we use the document specific mixing weights $P(z|d)$ to produce a ranked list of documents for each discovered topic. We then score this ranked list against the groundtruth landmarks from the Oxford dataset using the average precision measure from information retrieval. For each groundtruth landmark, we find the topic which gives the highest average precision – the results are listed in table 1.

The topic model often effectively picks out the particular landmarks from the Oxford dataset despite knowing nothing *a priori* about the objects contained in the groundtruth. Most of the gaps in performance are explained by the topic model including neighbouring facades to the landmark object which frequently co-occur with the object in question. The model knows nothing about the extents of the landmarks required and will include neighbouring objects when it is probabilistically beneficial to do so. We also note that sometimes the connected components don't contain all the images of the landmark – this is mainly due to failures in the initial feature matching.

Robustness to imaging conditions. Due to the richness of the pinboard models, the *g*LDA method is able to group images of a specific object despite large imaging variations (see figure 7). Standard LDA often struggles to cluster challenging images due to the absence of the extra spatial information.

0.896 0.856 0.834 0.651 0.371

Figure 7: Due to the richness of the topic pinboards, *g*LDA is able to group these images (which are all of the same landmark – the Sheldonian theatre) despite large changes in scale, viewpoint, lighting and occlusions. $P(z|d)$ is shown underneath each image.

Groundtruth Landmark	LDA max AP	*g*LDA max AP	Component recall
all_souls	0.90	**0.95**	0.96
ashmolean	0.49	**0.59**	0.60
balliol	**0.23**	**0.23**	0.33
bodleian	0.51	**0.64**	0.96
christ_church	0.45	**0.60**	0.71
cornmarket	**0.41**	**0.41**	0.67
hertford	0.64	**0.65**	0.65
keble	**0.57**	**0.57**	0.57
magdalen	**0.20**	**0.20**	0.20
pitt_rivers	**1.00**	**1.00**	1.00
radcliffe_camera	0.82	**0.91**	0.98

Table 1: The performance of *g*LDA on the Oxford dataset compared to LDA. The scores list the average precision (AP) of the best performing topic for each groundtruth landmark. *g*LDA always outperforms or does as well as standard LDA for object mining. The last column shows the recall for the component containing the best performing topic – the highest AP score either method could have returned. Figure 6 examines the differences in results for the Ashmolean landmark.

Comparison with standard LDA. In figure 6 we compare *g*LDA to standard LDA. The parameters were kept exactly the same between the two methods (except for the spatial term). LDA was initialized by uniformly sampling the topic for each word and run for 500 iterations to account for its slower Gibbs convergence. From the figure we can see that the LDA method has been unable to properly split the Ashmolean facade from an adjacent building. Table 1 compares the methods quantitatively. In all cases gLDA is superior (or at least equal) to LDA.

As well as being able to better discover different objects in the data, the *g*LDA method can localize the occurrence of particular topics in each image instead of just describing the mixture. This can be seen in figure 1 which displays three images from the Magdalen cluster with correspondences to two automatically discovered topics.

6 Conclusion and Future Work

We have introduced a new generative latent topic model for unsupervised discovery of particular objects and building facades in unordered image collections. In contrast to previous approaches, the model incorporates strong geometric constraints in the form of affine maps between images and latent aspects. This allows the model to cluster images of particular objects despite significant changes in scale and camera viewpoint. We have shown that the *g*LDA model outperforms the standard LDA model for discovering particular objects in image datasets.

The model can be generalized in several directions – for example using a fundamental

matrix (epipolar geometry) as its spatial relation instead of an affine homography; or adding a background topic model in the manner of [5]. There is also room for improving the computational efficiency in order to apply the model to larger datasets.

Acknowledgements. We are grateful for financial support from EPSRC, Microsoft and the Royal Academy of Engineering.

References

[1] http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/.

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

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE PAMI*, 30(4), 2008.

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proc. ICCV*, 2007.

[5] C. Chemuduguntu, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, 2007.

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

[7] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[8] M. Cummins and P. Newman. Probabilistic appearance based navigation and loop closing. In *Proc. IEEE International Conference on Robotics and Automation (ICRA'07)*, 2007.

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

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proc. ICCV*, 2005.

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.

[12] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[13] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001.

[14] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. 2007.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[18] T. Quack, V. Ferrari, and L. Van Gool. Video mining with frequent itemset configurations. In *Proc. CIVR*, 2006.

[19] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV*, pages 883–890, 2005.

[20] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. CVPR*, 2006.

[21] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005.

[22] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proc. CVPR*, 2008.

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, volume 2, pages 1470–1477, Oct 2003.

[24] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Proc. CVPR*, Jun 2004.

[25] E. Sudderth, A. Torralba, W. T. Freeman, and A. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1–3), 2008.

[26] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IEEE PAMI*, 62(2):113–140, 2005.

[27] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007.

[28] J. Winn and N. Joijic. Locus: Learning object classes with unsupervised segmentation. In *Proc. ICCV*, 2005.