Int J Comput Vis
DOI 10.1007/s11263-010-0363-5

Geometric Latent Dirichlet Allocation on a Matching Graph for Large-scale Image Datasets

James Philbin · Josef Sivic · Andrew Zisserman

Received: 23 October 2009 / Accepted: 27 May 2010

Abstract Given a large-scale collection of images our aim is to efficiently associate images which contain the same entity, for example a building or object, and to discover the significant entities. To achieve this, we introduce the *Geometric Latent Dirichlet Allocation* (gLDA) model for unsupervised discovery of particular objects in unordered image collections. This explicitly represents images 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.

Additionally, to reduce the computational cost of applying the gLDA model to large datasets, we propose a scalable method that first computes a *matching graph* over all the images in a dataset. This matching graph connects images that contain the same object, and rough image groups can be mined from this graph using standard clustering techniques. The gLDA model can then be applied to generate a more nuanced representation of the data. We also discuss how "hub images" (images representative of an object or landmark) can easily be extracted from our matching graph representation.

We evaluate our techniques on the publicly available Oxford buildings dataset (5K images) and show examples of automatically mined objects. The methods are evaluated quantitatively on this dataset using a ground truth labeling for a number of Oxford landmarks. To demonstrate the scalability of the matching graph method, we show qualitative results on two larger datasets of images taken of the Statue of Liberty (37K images) and Rome (1M+ images).

Keywords Object discovery · Large-scale retrieval · Topic/generative models

1 Introduction

In image collections, and especially in collections of tourist photographs collected from sites such as Flickr, certain scenes and objects tend to be photographed much more frequently than others. Our objective in this work is to obtain an association based not on the entire image, but on the *objects* contained in the images – we want to associate a set of images containing the same objects, even if a particular pair of images is quite dissimilar. The objects may vary significantly in scale, viewpoint, illumination or even be partially occluded. The extreme variation in imaging conditions presents serious challenges to the current state of the art in image-based data mining.

The ability to associate images based on common objects has many potential applications: the frequently occurring objects in a large collection can quickly be perused to form a visual summary; the clusters can provide an access mechanism to the collection; image-based particular object

J. Philbin (✉) · A. Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK
e-mail: james@robots.ox.ac.uk

A. Zisserman
e-mail: az@robots.ox.ac.uk

J. Sivic
INRIA – Willow Project, Laboratoire d'Informatique de l'Ecole Normale Supérieure (CNRS/ENS/INRIA UMR 8548), Paris, France
e-mail: josef@di.ens.fr

Published online: 15 June 2010

retrieval could use such methods as a filter to reduce data requirements and so reduce search complexity at query time; and techniques such as automatic 3D reconstruction which take as an input multiple views of the same object can then be applied to these image collections (Agarwal et al. 2009; Schaffalitzky and Zisserman 2002; Snavely et al. 2006), and can discover canonical views (Simon et al. 2007).

This work presents two contributions towards this objective: firstly, we introduce a geometrically consistent latent topic model, that can discover significant objects over an image corpus; secondly we propose methods for efficiently computing a *matching graph* over the images, where the images are the nodes and the edge strength is given by the overlap in visual content between the images. Using this matching graph together with inexpensive graph-based clustering techniques allows us to partition the corpus into smaller sets of images where our more expensive geometric latent topic model can then be learned. This makes the entire process scalable to large datasets.

Latent topic models such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann 2001) and Latent Dirichlet Allocation (LDA) (Blei et al. 2002) have had significant impact as methods for "semantic" clustering in the statistical text community. 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 can then be used to associate relevant documents, even though the documents themselves may have few words in common.

Given the success of these models, several vision papers (Fei-Fei and Perona 2005; Quelhas et al. 2005; Russell et al. 2006; Sivic et al. 2005) have applied them to the visual domain, replacing text words with visual words (Csurka et al. 2004; Sivic and Zisserman 2003). 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 do not exist in the text domain. There has been only a limited exploration of these relations in visual latent models: for incorporating segmentation (Cao and Fei-Fei 2007; Russell et al. 2006; Wang and Grimson 2007; Winn and Joijic 2005); or for a grid-based layout of images and objects (Bosch et al. 2008; Fergus et al. 2005; Fritz and Schiele 2008; Li et al. 2007; Sivic et al. 2008).

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 word identities *and their positions* on a planar facet, like a pinboard or bulletin board (we will use the term "pinboard" from now on). The visual words in an image (including location and shape) are then generated by an affine geometric transformation which projects words from the pinboard topic models. The generative process is illustrated in Fig. 1. We show that this model can be learned 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, the method effortlessly handles the presence of multiple distinct objects. It is similar in spirit to Simon and Seitz's (2008) use of pLSA for inferring object segmentations from large image collections, though we do not require the full 3D scene reconstruction of their method, which is found by performing an expensive bundle adjustment.

Our second contribution is a method to efficiently generate a sparse matching graph over a large image corpus. Each image is a node in the graph, and the graph edges represent the spatial consistency between sub-areas of the pairs of images linked by the edge – if the images contain a common object then the edge strength will reflect this. The graph is used to reduce the computational complexity of learning the *g*LDA model on large datasets. We can generate this graph using efficient text-based query mechanisms (Nister and Stewenius 2006; Philbin et al. 2007; Sivic and Zisserman 2003) coupled with accurate spatial verification, using each image in turn as a query. Given this graph, standard clustering methods can be applied to find images containing the same object. We are then able to efficiently learn a *g*LDA model using only subsets of images which are known to share a common object.

Until recently the two most convincing examples for data-mining employing some spatial consistency were (Quack et al. 2006; Sivic and Zisserman 2004) where the methods were applied in video to cluster particular objects (such as people or scenes). However, since 2008, four papers (Chum and Matas 2008; Crandall et al. 2009; Li et al. 2008; Quack et al. 2008) have appeared with differing approaches to the large scale mining problem, all using Flickr image collections.

Chum and Matas (2008) explore random sampling for clusters on a 100K corpus using the min-hash method of Chum et al. (2007a). This is a very efficient first step, and avoids the more costly building of a complete matching graph employed here. However, as the number of visual words in common between images decreases, the chance of discovering a cluster "seed" in Chum and Matas (2008) decreases, so that potential clusters mined in the complete graph can be missed.

Quack et al. (2008) mine a large Flickr corpus of 200K photos, but as a first step use geo-tagging information to decimate the corpus into sets no larger than 4K. The set is

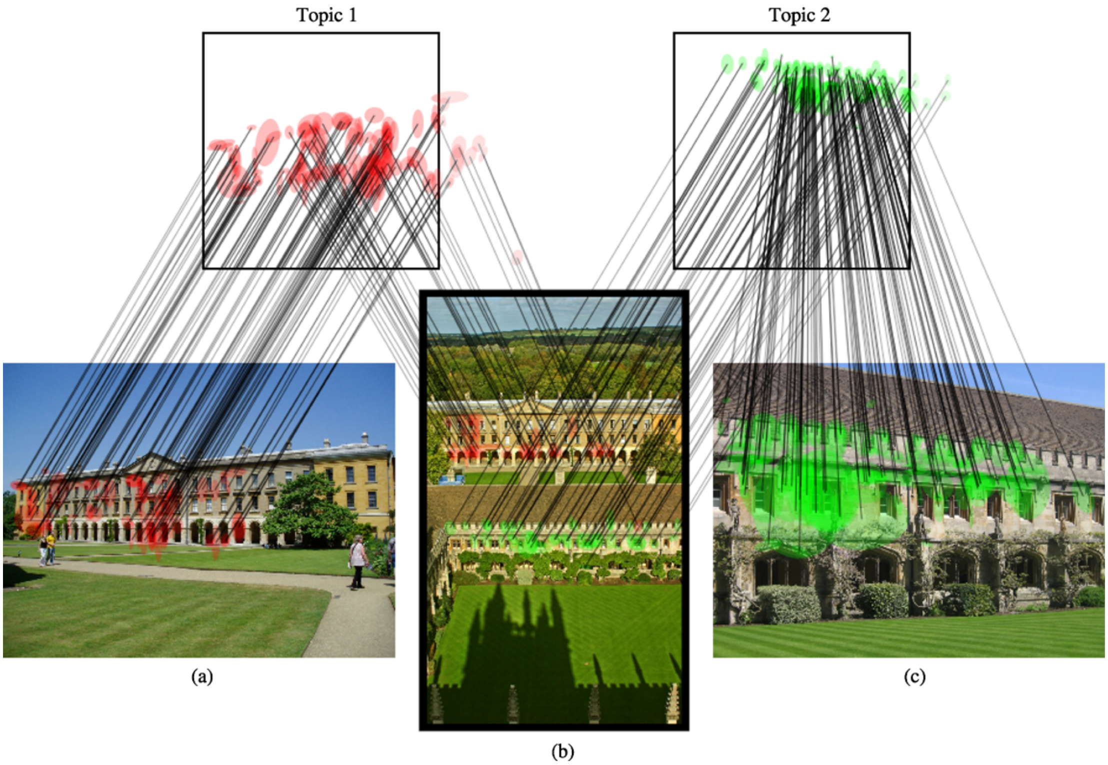


Fig. 1 The *g*LDA generative model. The two topic models (*above*) generate the visual words and their layout in the three images (*below*). Each topic model can be thought of as a virtual pinboard, with the words pinned at their mapped location. Image (**a**) is generated only from topic 1 with a single 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 learned from the set of images shown in Fig. 9. 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 (topics), despite the linking image (**b**), and has correctly localised these two objects in all three images

then partitioned into clusters using a combination of spatial consistency (as here) and textual similarity. Crandall et al. (2009) use an extremely large collection (33M) of Flickr images, but as a first step partition the data using mean shift clustering on the GPS location, similarly to (Quack et al. 2008). They then define a "matching graph" of images within a cluster using text and visual features, but *not* spatial consistency between images, and extract canonical or representative images for particular landmarks by spectral clustering.

Li et al. (2008) mine a 45K Statue of Liberty Flickr photo collection (the corpus differs from the one used here). Their approach is to first cluster the images using the GIST descriptor. Again, this decimates the problem, and spatially consistent clustering can then proceed efficiently within a cluster. As in Chum and Matas (2008) this first step avoids the expense of building a complete matching graph, but because *images* are matched, rather than objects, the risk is that images with more extreme changes in viewpoint will not be assigned to the same cluster, and will not be associated in subsequent cluster merging. There is clearly an interesting comparison to be made on the measures of speed vs what is missed, between the method presented here and the methods of Chum and Matas (2008), Li et al. (2008).

The remainder of the paper is arranged as follows: Sect. 2 describes the three datasets used for evaluation; Sect. 3 describes our procedure for building a complete matching graph of an image dataset including a brief review of the image retrieval methods used; Sect. 4 describes the *g*LDA model and the inference procedure; finally, Sect. 5 demonstrates the methods on the three datasets.

The work presented in this paper was originally published in Philbin et al. (2008), Philbin and Zisserman (2008). It has been expanded here to include a full description of the *g*LDA model and its implementation together with additional examples.

2 Datasets

We use three datasets of varying sizes, all collected automatically from Flickr by searching for images with particular text tags. However, many of the images retrieved bear no relation to the tag initially searched for, as the manual anno-

Table 1 Statistics for each image collection

Dataset	# images	# descriptors
Oxford	5,062	16,334,770
Statue of Liberty	37,034	44,385,173
Rome	1,021,986	1,702,818,841

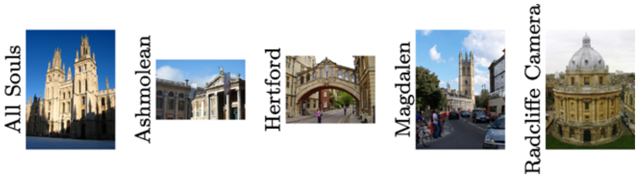


Fig. 2 Some of the Oxford landmarks. The Oxford dataset includes 11 "landmarks" – common buildings/views of Oxford for which a manually generated groundtruth is available

tation on Flickr tends to be extremely noisy. Some statistics for the datasets used are given in Table 1.

Oxford buildings dataset (5K images) For groundtruth evaluation, we use the *Oxford Buildings* dataset available from http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/. This consists of 5,062 high resolution (1024 × 768) images automatically retrieved from Flickr by searching on particular Oxford landmark tags, such as "Ashmolean". The dataset also provides groundtruth for the occurrences of 11 different Oxford landmarks. A sample of 5 landmark images is shown in Fig. 2. Note that the dataset contains many images of other buildings and non-buildings (a random sample is shown in Fig. 3).

Statue of Liberty dataset (37K images) This is a larger dataset of 37,034 images downloaded from Flickr containing a tag for the "Statue of Liberty". Although all of these images were tagged with the Statue of Liberty, the annotations are extremely noisy and the dataset contains a large number of other, unrelated scenes. The images were provided by Simon et al. (2007).

Rome dataset (1M images) This is a much larger dataset of 1,021,986 images collected from Flickr tagged with "Rome". The dataset contains a large number of tourist and other sites generally taken in Rome, including sites such as the Sistine Chapel and the Colosseum. Again, the images were provided by Simon et al. (2007).

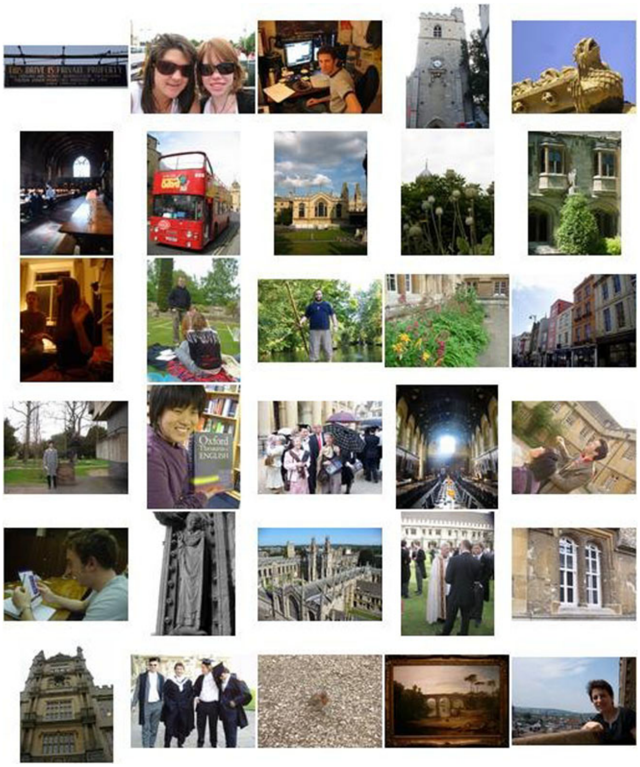


Fig. 3 A random sample of images from the Oxford dataset

3 Building a Matching Graph

In this section we explore using a cheap clustering step, which partitions the dataset into a number of disjoint sets of images. The aim is to associate all images that might possibly contain the same object into the same cluster whilst discarding images which definitely have no object in common. We achieve this using a 'matching graph' – a graph of the entire dataset with a node for each image and an edge connecting nodes i and j when images i and j share some common, spatially verified sub-region. Once this cheap clustering step has completed, we can go onto apply more expensive models to each subset in turn.

The process of building the graph relies for its efficiency on a visual words representation and inverted index, as reviewed in Sect. 3.1. In overview, the graph is built in the following way: Initially the graph is empty. For each image of the dataset in turn, we query using the whole image over the entire corpus. The top 400 results from the inverted index are spatially verified as described in Sect. 3.2. Images retrieved with more than a threshold number of verified inliers (we use 20 inliers in the following) to the query image contribute a new edge to the graph linking the query image to the retrieved image. This is repeated for each image in the corpus. The weights on the edges are given by $\frac{N_{I_m}}{(N_q+N_r)/2}$, where N_{I_m} is the number of spatially verified inliers and N_q, N_r are the number of visual words in the query and result respectively. This normalises for the effect of variation in the number of detected visual words in each image.

The graph generated is generally very sparse – for example, the matching graph for the 5K Oxford set contains

24,561 edges (a thousand times less than if every image matched to every other).

3.1 Particular Object Retrieval

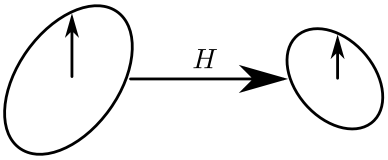
The search engine uses the vector-space model (Baeza-Yates and Ribeiro-Neto 1999) common in information retrieval. The query and each document (image) in the corpus is represented as a sparse vector of term (visual word) occurrences and search proceeds by calculating the similarity between the query vector and each document vector, using an L_2 distance. The document vectors are weighted using the simple tf-idf weighting scheme used in text retrieval. This downplays the contribution from commonly occurring, and therefore uninformative, words.

For computational speed, the word occurrences are stored in an inverted index which maps individual visual words (i.e. from 1 to K) to a list of the documents in which they occur. Only words which occur in the query need to be considered and generally this is a small percentage of the total (words not in common do not contribute to the distance). In the worst case, the computational complexity of querying the index is linear in the corpus size, but in practise it is close to linear in the number of documents which match a given query, which can provide a substantial saving. Note also that this method is trivially scalable as the corpus can be distributed to many computing nodes where each node can query in parallel and the result vectors concatenated.

To generate visual features we detect Hessian interest points and fit affine covariant ellipses (Mikolajczyk and Schmid 2004). For each of these affine regions, we compute a 128-dimensional SIFT descriptor (Lowe 2004). For the Oxford and Statue of Liberty datasets, a large discriminative vocabulary of 500K words is generated using an approximate k-means clustering method (Philbin et al. 2007) on all the descriptors of all the images in the corpus. For the Rome dataset, a random subsample of 50M descriptors is used for clustering to 1M cluster centres. Each descriptor is assigned to a single cluster centre to give one visual word. On average, there are ∼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 Verification

We use a deterministic variant of RANSAC (Fischler and Bolles 1981), which involves generating hypotheses of a restricted (affine) transformation (Philbin et al. 2007) and then iteratively re-evaluating promising hypotheses using the full transformation, similar in spirit to Chum et al. (2003). By selecting a restricted class of transformations for the hypothesis generation stage and by exploiting shape information in the affine-invariant image regions, we are able to generate hypotheses from only a *single* feature correspondence. Each feature in the image is represented as an ellipse – therefore, each pair of matched features can define a 5 degree-of-freedom transformation between the two images. By including an "upness" constraint (that images are taken upright) we can define a restricted affine transformation (see Fig. 4). 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. Note that although the initial homography does not allow in-plane rotation (due to the "upness" constraint), by iteratively computing the full transform the system can handle significant rotation between images.



$$H = \begin{bmatrix} a & 0 & t_x \\ b & c & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

Fig. 4 Spatial verification: a restricted affine homography of the form shown is computed for every elliptical match between two images

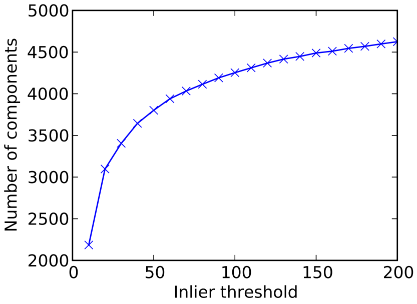


Fig. 5 Examining the number of connected components found as a function of inlier threshold for the Oxford dataset

The size of this inlier set for the best transformation is used in the matching graph to determine the edge strength, and in the gLDA model to score the support for a latent topic in an image.

3.3 Connected Components

One of the simplest operations for splitting the data is to find the connected components on the matching graph. This greatly reduces the complexity of any subsequent clustering step, as now much smaller groupings of images need to be considered. Finding the connected components of a graph can be computed in linear time in the number of graph nodes using depth-first search (Cormen et al. 1990). An example

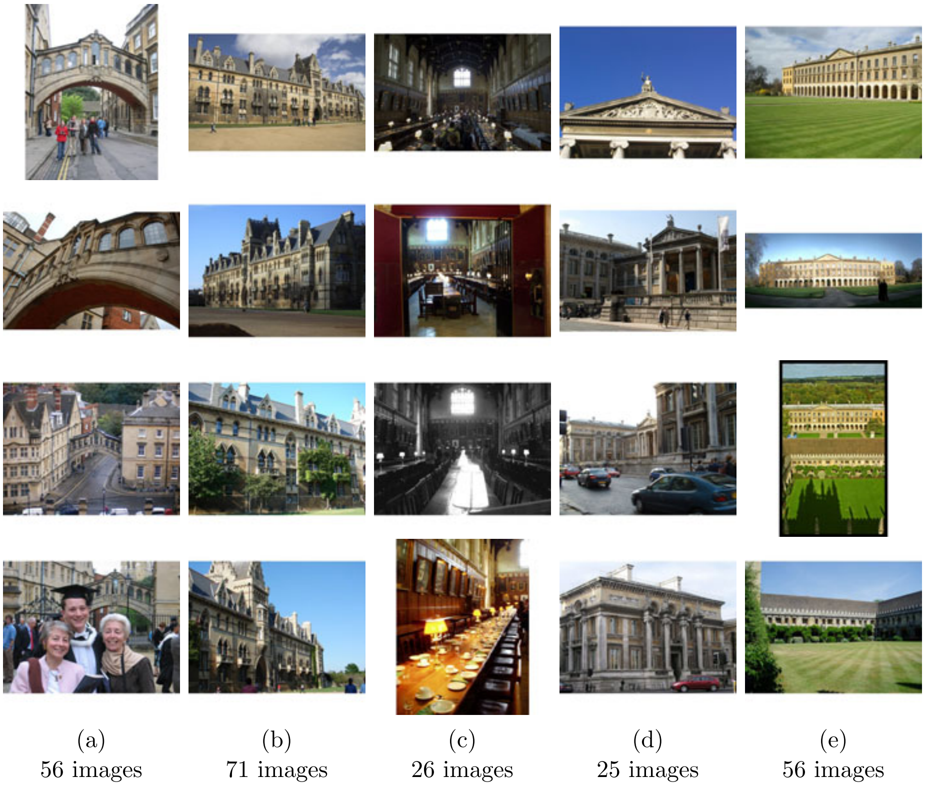


Fig. 6 Examples of the connected components automatically found on the 5K Oxford dataset. Some components are already extremely accurate in isolating individual buildings/landmarks (see (**a**)–(**c**)). (**d**) and (**e**) show examples of components linking disjoint objects via connecting views. The number of images in each component is shown beneath the label. Note the significant variation of scale and viewpoint within each component

of the sub-graph automatically discovered for a connected component is shown in Fig. 7.

Even though the method is crude, it can be surprisingly effective at pulling out commonly occurring objects in photo datasets. It naturally achieves a “transitive association” over the views of an object: views A and C may have no matches, even though they are of the same object. However, provided A links to B, and B links to C, then A and C will be transitively associated. This lack of matches (e.g. between A and C) may arise from detector drop out, SIFT descriptor instability, partial occlusion etc, and was the subject of the “Total Recall” method of Chum et al. (2007b) where missed matches were corrected at run time by the additional overhead of a form of query expansion. More recently, Turcot and Lowe (2009) have used a matching graph to address this missed matches problem by off line processing, thus avoiding the run time cost.

This transitive advantage is also a problem though, in that it joins too much together – a “connecting image” (one that contains multiple disjoint objects) pulls all images of these objects into a single connected component. Figure 6 shows some examples of connected components found on the Oxford dataset. Building the matching graph involves setting a threshold on the number of inliers which defines a pair of images as being connected, and this governs the number of connected components obtained. Setting this threshold too low links all images into a single component; too high and no image connects to any other (see Fig. 5). Figures 6(d) and 6(e) show examples of connected components joining disjoint objects via connecting images. We examine the scalability of the graph matching procedure in Sect. 5.1.

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 same object. For example, “linking” images containing more than one object will join other images of these objects into a single cluster (see Fig. 9).

3.4 Hub Images

Although not directly used in this work, we note here that one can rank the images within each connected component to pull out canonical or “hub” images. A simple but effective method is to rank images according to the number of spatially verified connections they make to other images within the component. This corresponds to the *degree* of each node within the graph. Figure 8 shows the three highest and three lowest ranked images according to *degree*. The images showing more common or canonical views of the object are ranked highly – those showing strong differences in imaging conditions are ranked lowly. Though not done here, simple extensions to this method might include using spec-

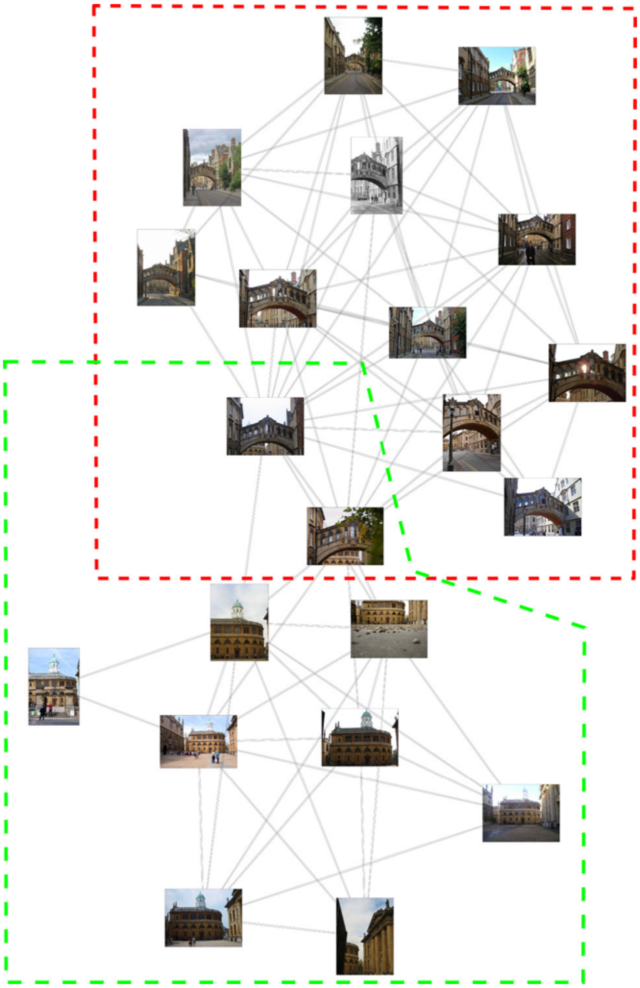


Fig. 7 A portion of the full Oxford matching graph for a single connected component. The images in the *top* (*red*) and *bottom* (*green*) *regions* contain the "Bridge of Sighs" and the "Ashmolean Theatre" respectively. Note the two connecting images which contain both objects

tral clustering or computing pagerank (Crandall et al. 2009; Kim and Torralba 2009; Quack et al. 2008).

4 Object Discovery

In this section, we review the standard LDA model (Blei et al. 2002), and then describe its extension to the *g*LDA model which incorporates geometric information.

4.1 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, $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$, 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.

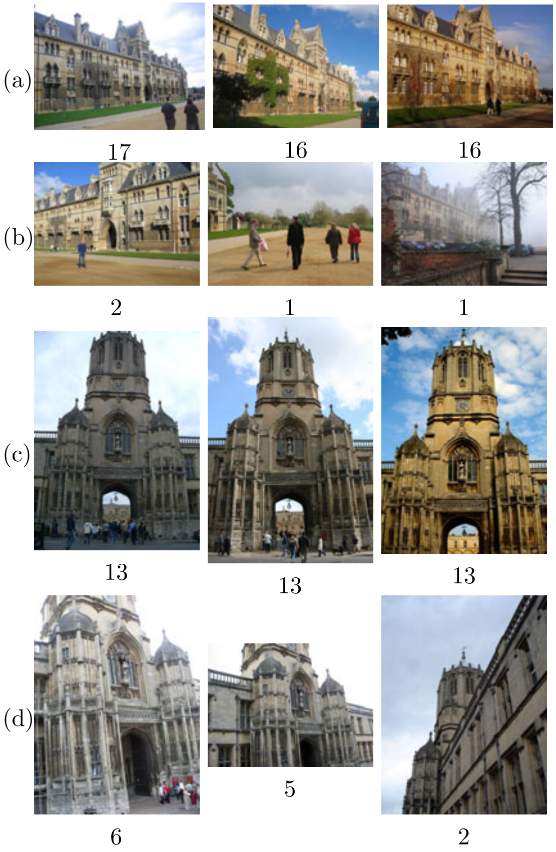


Fig. 8 Examples of hub images. The three highest (**a**) and lowest (**b**) images ranked by degree for a connected component of Christ Church College. The three highest (**c**) and lowest (**d**) images ranked by degree for the Thom Tower, Oxford. The degree is listed beneath each image

Figure 10(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

$$P(\mathbf{z}|\mathbf{w}, \alpha, \beta) \propto P(\mathbf{z}|\alpha) P(\mathbf{w}|\mathbf{z}, \beta) \quad (1)$$

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 (Griffiths and Steyvers 2004) with the following update formula:

$$P(z_{ij} = k|\mathbf{z}_{-ij}, \mathbf{w}) \propto \frac{n_{i\cdot k} + \alpha}{n_{i\cdot\cdot} + T\alpha} \times \frac{n_{\cdot jk} + \beta}{n_{\cdot\cdot k} + V\beta} \quad (2)$$

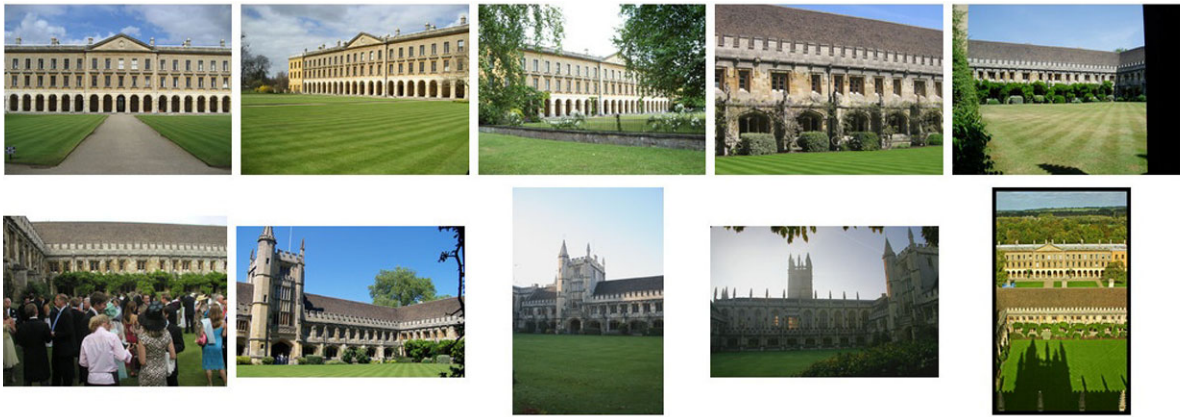


Fig. 9 A sample of 10 images from a connected component associated with Magdalen college. The component contains two separate buildings: A Georgian building and a college cloisters, linked by the aerial photo shown (*bottom right*). Within the cloisters there are two distinct "facades", one of the wall, the other of a tower. Our full *g*LDA method is able to extract all three "objects" (cloisters, tower and building) completely automatically. The total size of the component is 42 images

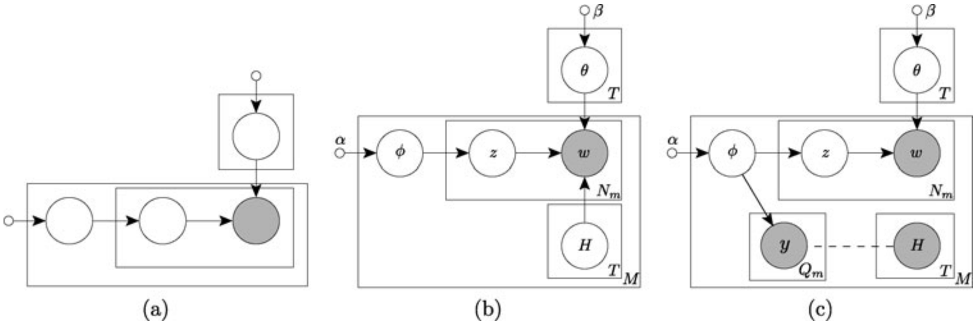


Fig. 10 (**a**) The standard LDA model. (**b**) The *g*LDA model. M, T and N_m are the number of documents, topics and words (in document m) respectively. (**c**) The simplified *g*LDA model used for one step of the approximate inference. Q_m is the number of inlier visual words in document m. See text for details

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. T 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 (2), 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.

4.2 Geometric LDA

In *g*LDA, the topics of the LDA model are augmented with the spatial position and shape of the visual words, and a geometric transformation between topics and documents is introduced. 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 first picking a distribution over the pinboards (topics) and sampling an affine homography, $\mathbf{H}$, for each pinboard; and then forming the image as the composition of the visual words from each topic mapped under the corresponding homography.

Note, as in LDA, 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 quantisation. Others have handled this situation by learning a sensor model (Cummins and Newman 2007).

The *g*LDA model is shown in Fig. 10(b). *g*LDA adds extra spatial transformation terms, **H**, to the LDA model and the word terms, **w**, contain both the identity and spatial position and shape of the visual word in the image. These image specific transformations, **H**, describe how the words for a particular topic occurring in an image are projected from the "pin-board" model for that topic. **H** are assumed to be affine transformations, so that the model can account for moderate changes in viewpoint between the topic and the image.

The joint probability of the *g*LDA model factors as follows

$$P(\mathbf{w}, \mathbf{z}, \mathbf{H}, \theta, \phi|\alpha, \beta)$$
$$= P(\mathbf{w}|\mathbf{z}, \mathbf{H}, \theta) P(\mathbf{z}|\phi) P(\theta|\beta) P(\phi|\alpha) P(\mathbf{H}) \tag{3}$$

The generative distributions could be further specified and inference on the model carried out in a similar manner to Sudderth et al. (2008). However, to avoid the expense of generatively sampling the transformation hypotheses, we instead approximate the joint as described next.

4.3 Approximate Inference

The goal of the inference in the *g*LDA model is to estimate topic assignments **z** together with transformation hypotheses **H** given observed visual words **w**. For approximate inference, we use an iterative procedure, which alternates between (i) estimating **H** given the current estimate of **z**, and (ii) estimating **z** given the current estimate of **H**. In step (i), the transformations **H** between the image and the topic-pinboards are estimated from the words assigned to each topic **z** directly using efficient tools from multiview geometry. The outcome of this step are a set of *inlier* words for each homography **H**. These inliers, together with the associated homography are *observed*. In step (ii) the *number* of inliers for each topic-pinboard for each image influences (positively) the assignment of words to topic-pinboards **z**, as can be seen by glancing ahead to (9).

For these steps we need to book keep which words are inliers to which transformation/topic-pinboard. For this, we introduce indicator variables **y**, where $y_{ij} = k$ specifies that in image i word j is assigned to topic-pinboard k. **y** is only defined for the inlier words and is not defined for those words that are not an inlier to any transformation. Note that, unlike for **z** where a word is only assigned to one topic, an inlier word can be assigned to multiple transformation/topic-pinboards. As will be seen, it is only the total number of inliers between an image and a topic-pinboard that is used by (9), and we denote this count as $q_{i \cdot k}$, in analogy with $n_{i \cdot k}$, for the number of inliers to topic-pinboard k in image i, where $q_{i \cdot k} = |\forall j : y_{ij} = k|$.

Given the now observed **y** and **H**, the *g*LDA model of Fig. 10(b) is approximated for step (ii) as in the graphical model of Fig. 10(c). In the approximation, the observed words no longer depend directly on **H** (in essence this ignores the shape and position of the visual words generated by the transformation from the pinboard-topic). Instead, it is assumed that the inlier indicators **y** depend on the topic proportions in each image, and these inliers are determined from **H** (as indicated by the dotted line, with the actual computation described in Sect. 4.4).

We now work through the derivation of (9), starting from the graphical model of Fig. 10(c). The joint probability of the approximate model factors as

$$\begin{aligned} &P(\mathbf{w}, \mathbf{z}, \mathbf{y}, \theta, \phi | \alpha, \beta) \\ &\quad = P(\mathbf{w}|\mathbf{z}, \theta) P(\theta|\beta) P(\mathbf{z}, \mathbf{y}|\phi) P(\phi|\alpha) \\ &\quad = P(\mathbf{w}|\mathbf{z}, \theta) P(\theta|\beta) P(\mathbf{z}|\phi) P(\mathbf{y}|\phi) P(\phi|\alpha). \end{aligned} \tag{4}$$

Note, we assume that **y** and **z** are conditionally independent given ϕ. However, when ϕ is not observed (as here), inlier indicators **y** influence topic assignments **z** through ϕ.

As in standard LDA, parameters ϕ and θ can be integrated out (Griffiths and Steyvers 2004):

$$\begin{aligned} &P(\mathbf{w}, \mathbf{y}, \mathbf{z} | \alpha, \beta) \\ &\quad = \int P(\mathbf{w}|\mathbf{z}, \theta) P(\theta|\beta)\, d\theta \int P(\mathbf{z}|\phi) P(\mathbf{y}|\phi) P(\phi|\alpha)\, d\phi \\ &\quad = P(\mathbf{w}|\mathbf{z}, \beta) P(\mathbf{z}, \mathbf{y}|\alpha) \end{aligned} \tag{5}$$

The two integrals can be performed analytically. The first integration gives:

$$P(\mathbf{w}|\mathbf{z}, \beta) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^T \prod_{k=1}^{T} \frac{\prod_j \Gamma(n_{\cdot jk} + \beta)}{\Gamma(n_{\cdot\cdot k} + V\beta)}, \tag{6}$$

where $\Gamma(\cdot)$ is the standard Gamma function and $n_{\cdot jk}$ is the number of visual words with id j assigned to pinboard-topic k over all images. Note that (6) is the same as in standard LDA (Griffiths and Steyvers 2004). The second integration results in

$$P(\mathbf{z}, \mathbf{y}|\alpha) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \right)^M \prod_{i=1}^{M} \frac{\prod_k \Gamma(n_{i \cdot k} + q_{i \cdot k} + \alpha)}{\Gamma(n_{i \cdot\cdot} + q_{i \cdot\cdot} + T\alpha)} \tag{7}$$

where $n_{i \cdot k}$ is the number of words in document i assigned to pinboard-topic k, and $q_{i \cdot k}$ is the number of *inlier, i.e. spatially verified* words in document i assigned to pinboard-topic k. Note that the observed inlier counts $q_{i \cdot k}$ can be viewed as a document specific prior (virtual word counts) biasing the probability towards topics with a higher number of inliers.

Similar to the standard LDA, evaluating the posterior distribution

$$P(\mathbf{z}|\mathbf{y}, \mathbf{w}, \alpha, \beta) \propto P(\mathbf{z}, \mathbf{y}|\alpha) P(\mathbf{w}|\mathbf{z}, \beta) \tag{8}$$

is intractable. However, similar to Griffiths and Steyvers (2004), we can sample from high probability regions of the **z** space using a Gibbs sampler with the following update formula:

$$P(z_{ij} = k | \mathbf{z}_{-ij}, \mathbf{w}, \mathbf{y}) = \frac{n_{i \cdot k} + q_{i \cdot k} + \alpha}{n_{i \cdot\cdot} + q_{i \cdot\cdot} + T\alpha} \times \frac{n_{\cdot jk} + \beta}{n_{\cdot\cdot k} + V\beta} \tag{9}$$

This defines a multinomial over topic assignments for a single word, z_{ij} which is sampled to give the new topic assignment. By comparing this update formula to that of standard LDA given in (2), it is evident how the aggregate inlier counts $q_{i \cdot k}$ influence re-sampling of the pinboard-topic indicators z_{ij} by assigning a higher probability to topic-pinboards with a higher number of inliers.

In summary, the approximate inference proceeds in the following iterative two stage procedure. Firstly, the pinboard assignments, **z**, are resampled with the Gibbs sampler (9). This is a very simple change to the Gibbs update formula from LDA, but it makes the model much easier to learn than if the full coupling between **w** and **H** had been modelled (Fig. 10(b)). Secondly, given the current assignment of topic-pinboards, **z**, the transformation hypotheses **H** together with inlier indicators **y** are estimated using RANSAC (for details see Sect. 4.4). The pinboard-topic assignments **z** depend in-directly on **H** through the (observed) inlier indicators **y**. Conversely, changing **z** by re-assigning a particular visual word to a different pinboard influences transformations **H** and inlier indicators **y** during the RANSAC procedure.

Note that the interleaved sampling of pinboard assignments **z** using (9) and inliers **y** with transformation hypothesis **H** using RANSAC can be viewed as data driven Markov Chain Monte Carlo in the spirit of Tu et al. (2005).

4.4 Spatial Scoring Using RANSAC

The discriminative *g*LDA 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 and is given by the number of inliers. The transformation is estimated using the deterministic variant of RANSAC described in Sect. 3.2.

The pinboards are updated as follows – every word in the corpus with $z_{ij} = k$ is contained in the pinboard model for topic k projected from the original document i using the current transformation hypothesis, $\mathbf{H}_{ik}$. Terms projected into the pinboard need not be inliers under the current transformation but may become inliers in a further step of alternating Gibbs sampling. This is observed in practise.

4.5 *g*LDA Implementation Details

Topic initialisation For each connected component of the matching graph the topics are initialised by first obtaining T separate clusters (using agglomerative clustering with average linkage as the similarity score). For each cluster, we project each document's words to a normalised 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 initialise the locations and shapes of the visual words in the model. Although this is not strictly necessary for the *g*LDA model, it greatly improves convergence speed and generally leads to improved results.

Prior parameters The *g*LDA model (Sect. 4.3) includes priors for the per document topic distribution, α, and the per topic word distribution, β. Empirically we find that using $\alpha = 200.0$, $\beta = 1.0$ 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 Sect. 4.3 changing the number of topics from 1 to 8, then choose the Markov chain with the highest likelihood (see Fig. 11) (Griffiths and Steyvers 2004). 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 hyperparameters, 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 tends to converge to a mode extremely quickly (<50 iterations) and running it longer brings little appreciable benefit.

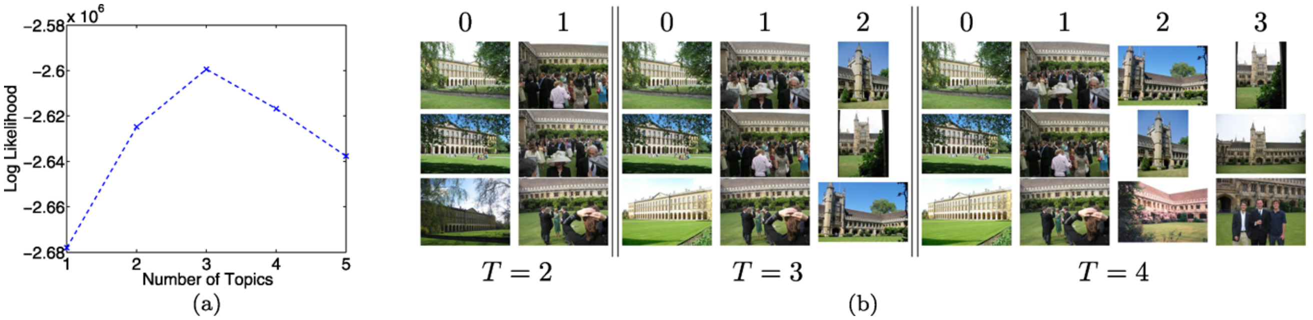


Fig. 11 Automatically choosing the number of topics. (**a**) The log likelihood of the *g*LDA model fitted to the connected component shown in Fig. 9 for different numbers of topics. (**b**) The *top three documents* (ranked by $P(z|d)$ in *columns*) for each topic for different numbers of topics, T. In this case three topics are automatically chosen which separate the building, cloisters and tower

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

5 Results

5.1 Matching Graph

In this section we show results of building a matching graph over each of the three datasets.

For the Oxford dataset, clustering using connected components found 323 separate components (clusters of more than one image) using an inlier threshold of 20. The size of the largest component is 396 images (of the Radcliffe Camera, a popular Oxford tourist attraction).

Scalability To demonstrate the scalability of our matching graph method, Figs. 12 and 13 show samples from automatically discovered object clusters for the Statue of Liberty (37K images) and Rome (1M+ images) datasets, respectively. Searching for every image in the 37K Statue of Liberty dataset takes around 2 hours on a single 3 GHz machine. The Rome data (1M+ images) was much more challenging – it took 1 day on 30 machines to generate the matching graph on this corpus. Though expensive, this demonstrates the ability of our methods to scale across multiple machines.

5.2 *g*LDA

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 2. The component recall column refers to the maximum recall of the object over all connected components and so gives an upper bound on the possible improvement (as LDA and *g*LDA look within components).

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.

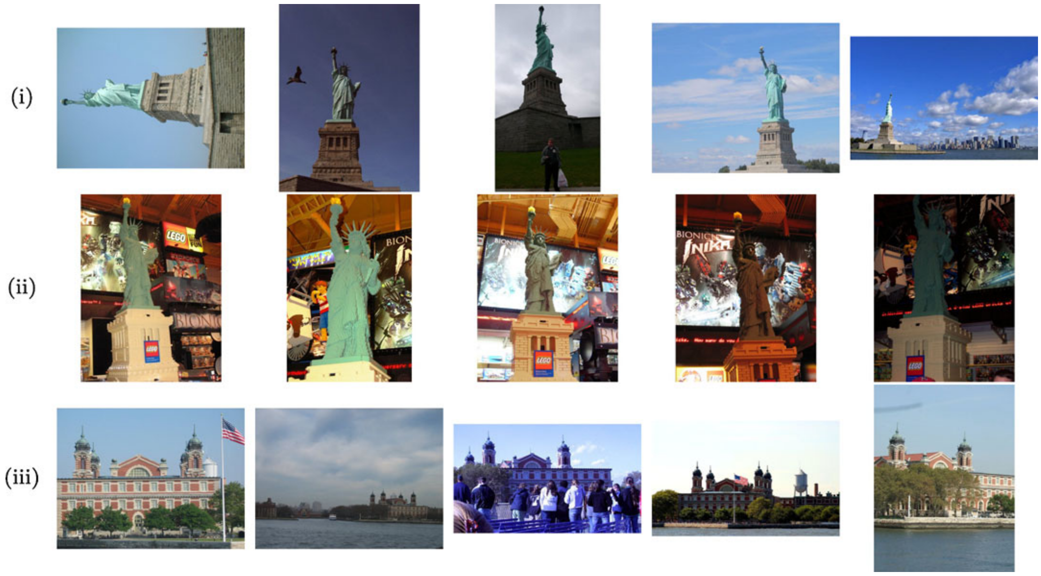


Fig. 12 Random samples of the three largest clusters automatically found from the Statue of Liberty dataset as connected components on the matching graph. Note the extreme variety of imaging conditions (changes in scale, viewpoint, lighting and occlusion) **(i)** the Statue of Liberty (11170 images). **(ii)** A lego Statue of Liberty (59 images). **(iii)** An Ellis Island building (52 images)

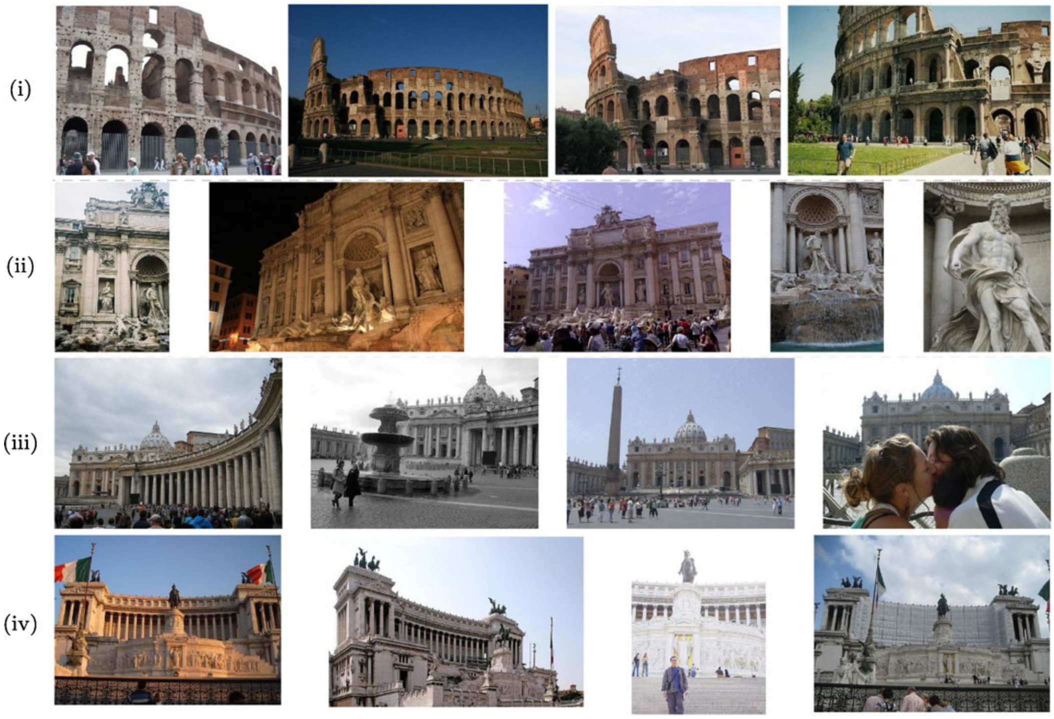


Fig. 13 Random samples of the four largest clusters automatically found from the 1M+ image Rome dataset as connected components on the matching graph. (**i**) Coliseum (18676 images). (**ii**) Trevi Fountain (15818 images). (**iii**) St Peter's Square, Vatican (9632 images). (**iv**) "Il Vittoriano" (4869 images)

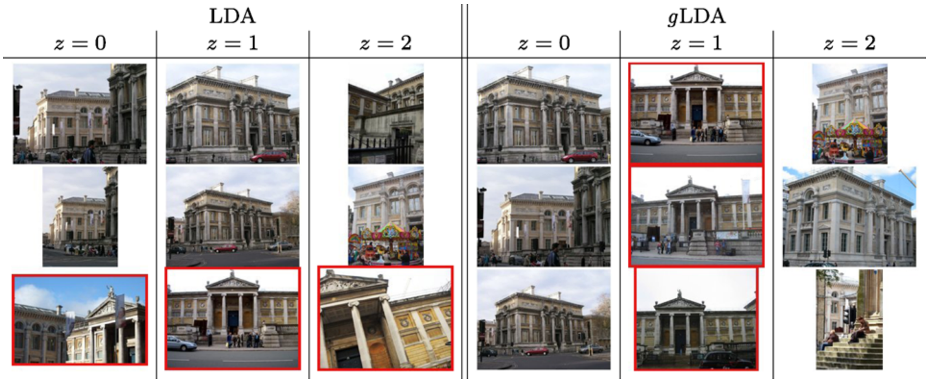


Fig. 14 Comparing gLDA to standard LDA for a connected component containing images of the Ashmolean, for $T = 3$. The *top three images* are shown for each topic, ranked by $P(z|d)$ in columns. Notice that LDA has confused the Ashmolean facade (outlined in *red*) between the three topics whereas gLDA has used the stronger spatial constraints to correctly isolate the building facade

Figure 18 shows a visualisation of two topics discovered by gLDA. It is easy to see that gLDA has correctly found and localised these particular objects in the dataset images. Figure 17 shows three topics automatically discovered by gLDA.

Robustness to imaging conditions Due to the richness of the pinboard models, the gLDA method is able to group images of a specific object despite large imaging variations (see Fig. 15). Standard LDA often struggles to cluster challenging images due to the absence of the extra spatial information.

In Fig. 16, we show the results of running the gLDA method on a 200 image sub-sample from one of the connected components of the Rome dataset, corresponding to the Trevi Fountain. We see that, by forcing a larger number of topics, the gLDA method can also pick out different views of a single object or facade. In this case the model has discovered a night-time view, and two daytime views of the fountain differing in viewpoint.

Comparison with standard LDA In Fig. 14 we compare gLDA to standard LDA. The parameters were kept exactly the same between the two methods (except for the spatial term). LDA was initialised 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 prop-

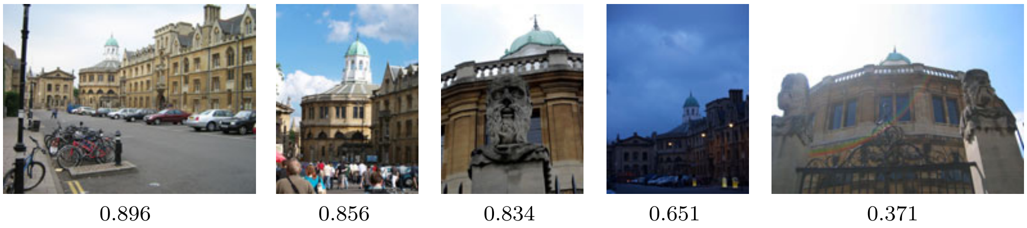


Fig. 15 Due to the richness of the topic pinboards, gLDA is able to group these images (which are all of the same landmark – the Sheldonian theatre) despite large changes in scale, viewpoint, lighting and occlusions. $P(z|d)$ is shown underneath each image

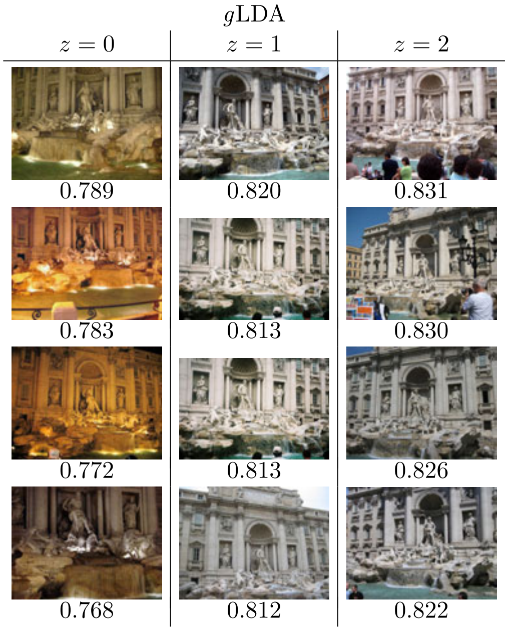


Fig. 16 Results of running gLDA on a 200 image sub-sample of one of the connected components (corresponding to the Trevi fountain) for the Rome dataset. Here, the model predicted $T = 1$ using the likelihood method, but we forced $T = 3$. When this is done, the gLDA model tends to discover views differing in lighting or viewpoint. $P(z|d)$ is listed beneath each image

erly split the Ashmolean facade from an adjacent building.

For a quantitative comparison we use the landmarks from the Oxford dataset. This is an indirect test of performance, because it requires that the landmarks correspond to a discovered topic (and is not split between connected components). For each landmark the component that has highest average precision (AP) is selected. The AP is computed as the area under the precision-recall curve for each landmark. The gLDA and LDA scores are then given for the best performing topic. Note, the AP for the component is an upper bound on the AP for the topics within that component. The results are given in Table 2. 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 gLDA method can localise the occurrence of particular topics in each image instead of just describing the mixture. This can be seen in Fig. 1 which displays three images from the Magdalen cluster with correspondences to two automatically discovered topics.

Table 2 The performance of gLDA on the Oxford dataset compared to LDA. The scores list the average precision (AP) of the best performing topic for each groundtruth landmark. gLDA 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 14 examines the differences in results for the Ashmolean landmark

Groundtruth landmark	LDA max AP	gLDA 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

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 gLDA model outperforms the standard LDA model for discovering particular objects in image datasets.

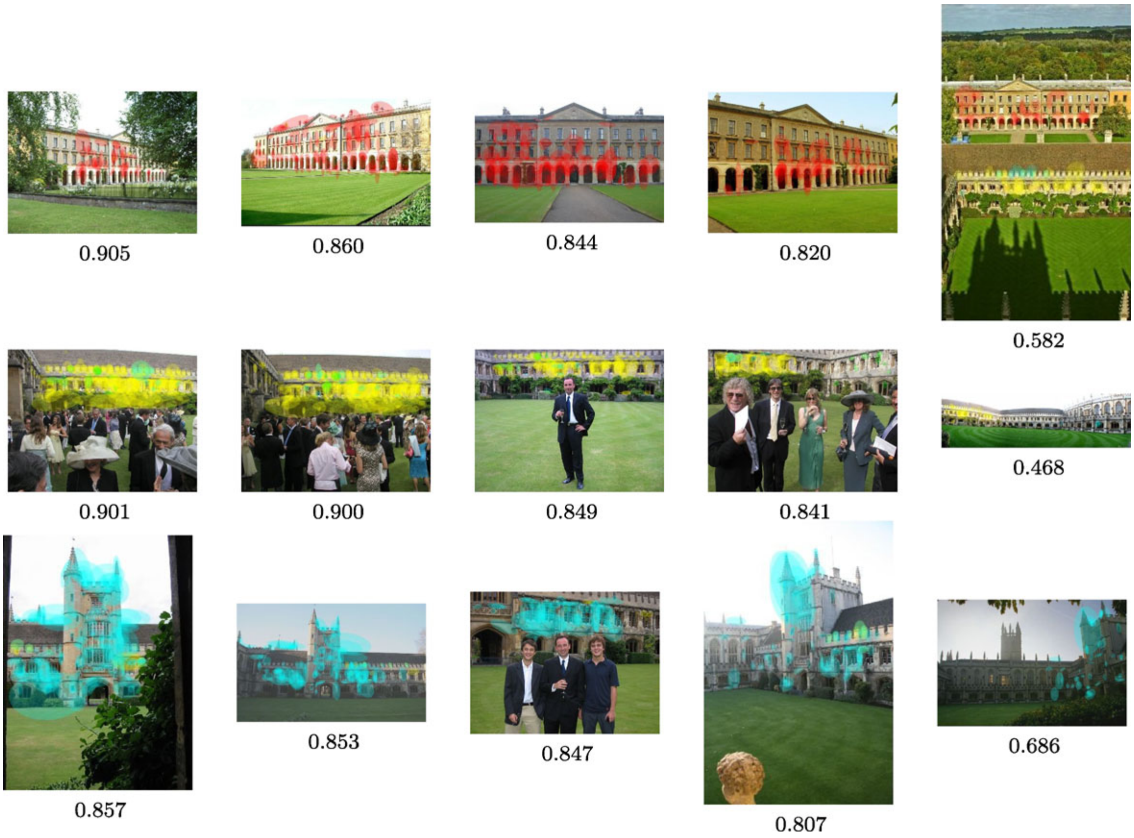


Fig. 17 (Color online) Example images from three topics (one per row) automatically discovered by *g*LDA from a component of Hertford college, Oxford. The visual words are coloured according to the topic they belong to: 0 – *red*, 1 – *yellow*, 2 – *blue*, 3 – *green* (not shown). $P(z|d)$ is listed beneath each image

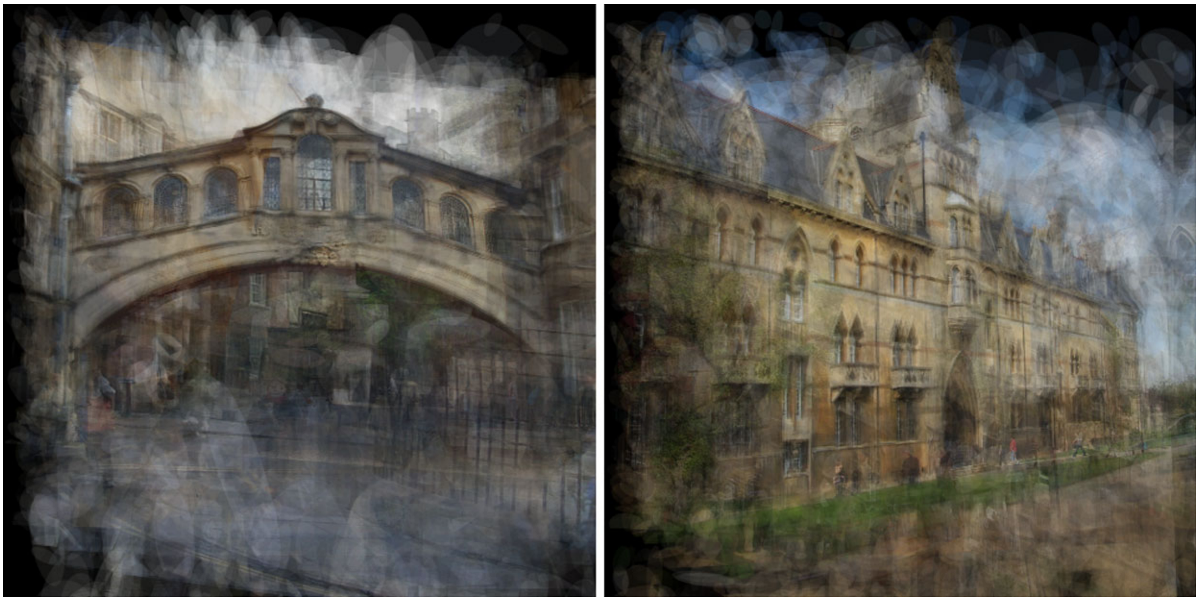


Fig. 18 Visualising the topics discovered by *g*LDA. The image data underlying each word in the topic has been projected into the canonical frame for visualisation. Here, two discovered topics are shown for different connected components in the Oxford matching graph. This topic visualisations have been generated from all the images in the respective connective components (56 images and 71 images)

To make the model tractable we also introduced a matching graph clustered using connected component clustering, that can be used to quickly organise very large image collections, and demonstrated this on image collections of over 1M images.

The *g*LDA model can be generalised 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 Chemuduguntu et al. (2007). There is also room for improving the computational efficiency in order to apply the model to larger datasets.

Acknowledgements We are grateful for discussions with Michael Berry and Michael Isard, and for financial support from EPSRC, Microsoft, the Royal Academy of Engineering, ERC grant VisRec No. 228180 and ANR grant ANR-07-BLAN-0331-01.

References

Agarwal, S., Snavely, N., Simon, I., Seitz, S., & Szeliski, R. (2009). Building Rome in a day. In *Proc. ICCV*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: ACM Press.

Blei, D., Ng, A., & Jordan, M. (2002). Latent Dirichlet allocation. In *NIPS*.

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

Cao, L., & Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proc. ICCV*.

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

Chum, O., & Matas, J. (2008). Web scale image clustering: large scale discovery of spatially related images. Technical Report CTU-CMP-2008-15, Czech Technical University in Prague.

Chum, O., Matas, J., & Kittler, J. (2003). Locally optimized RANSAC. In *DAGM* (pp. 236–243).

Chum, O., Philbin, J., Isard, M., & Zisserman, A. (2007a). Scalable near identical image and shot detection. In *Proc. CIVR*.

Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007b). Total recall: automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*.

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (1990). *Introduction to algorithms*. New York: McGraw-Hill.

Crandall, D., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. In *Proc. WWW*.

Csurka, G., Bray, C., Dance, C., & Fan, L. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (pp. 1–22).

Cummins, M., & Newman, P. (2007). Probabilistic appearance based navigation and loop closing. In *Proc. IEEE international conference on robotics and automation (ICRA'07)*.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, Jun 2005.

Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2005). Learning object categories from Google's image search. In *Proc. ICCV*.

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

Fritz, M., & Schiele, B. (2008). Decomposition, discovery and detection of visual categories using topic models. In *Proc. CVPR*.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci.*, *101*, 5228–5235.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, *43*, 177–196.

Kim, G., & Torralba, A. (2009). Unsupervised detection of regions of interest using iterative link analysis. In *NIPS*.

Li, L.-J., Wang, G., & Fei-Fei, L. (2007). Optimol: automatic online picture collection via incremental model learning. In *Proc. CVPR*.

Li, X., Wu, C., Zach, C., Lazebnik, S., & Frahm, J.-M. (2008). Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. ECCV*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, *60*(2), 91–110.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *Int. J. Comput. Vis.*, *1*(60), 63–86.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proc. CVPR*.

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

Philbin, J., Sivic, J., & Zisserman, A. (2008). Geometric LDA: a generative model for particular object discovery. In *Proceedings of the British machine vision conference*.

Philbin, J., & Zisserman, A. (2008). Object mining using a matching graph on very large image collections. In *Proceedings of the Indian conference on computer vision, graphics and image processing*.

Quack, T., Ferrari, V., & Van Gool, L. (2006). Video mining with frequent itemset configurations. In *Proc. CIVR*.

Quack, T., Leibe, B., & Van Gool, L. (2008). World-scale mining of objects and events from community photo collections. In *Proc. CIVR*.

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

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

Schaffalitzky, F., & Zisserman, A. (2002). Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. ECCV* (Vol. 1, pp. 414–431). Berlin: Springer-Verlag.

Simon, I., & Seitz, S. M. (2008). Scene segmentation using the wisdom of crowds. In *Proc. ECCV*.

Simon, I., Snavely, N., & Seitz, S. M. (2007). Scene summarization for online image collections. In *Proc. ICCV*.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., & Freeman, W. T. (2005). Discovering object categories in image collections. In *Proc. ICCV*.

Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T., & Efros, A. A. (2008). Unsupervised discovery of visual object class hierarchies. In *Proc. CVPR*.

Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV*.

Sivic, J., & Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proc. CVPR*, Jun 2004.

Snavely, N., Seitz, S., & Szeliski, R. (2006). Photo tourism: exploring photo collections in 3D. In *Proc. ACM SIGGRAPH* (pp. 835–846).

Sudderth, E., Torralba, A., Freeman, W. T., & Willsky, A. (2008). Describing visual scenes using transformed objects and parts. *Int. J. Comput. Vis.*, *77*(1–3).

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

Turcot, P., & Lowe, D. (2009). Better matching with fewer features: the selection of useful features in large database recognition problems. In *ICCV workshop on emergent issues in large amounts of visual data (WS-LAVD)*.

Wang, X., & Grimson, E. (2007). Spatial latent Dirichlet allocation. In *NIPS*.

Winn, J., & Joijic, N. (2005). Locus: learning object classes with unsupervised segmentation. In *Proc. ICCV*.