Scene Segmentation Using the Wisdom of Crowds

Ian Simon and Steven M. Seitz

University of Washington
{iansimon,seitz}@cs.washington.edu

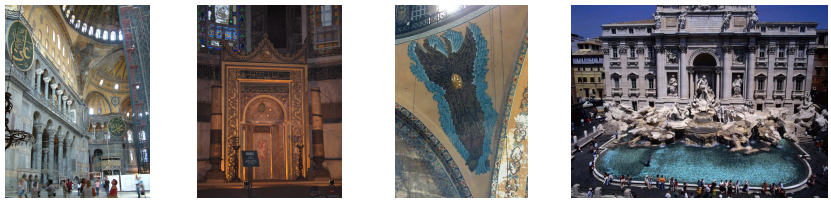
Abstract. Given a collection of images of a static scene taken by many different people, we identify and segment interesting objects. To solve this problem, we use the distribution of images in the collection along with a new *field-of-view* cue, which leverages the observation that people tend to take photos that frame an object of interest within the field of view. Hence, image features that appear together in many images are likely to be part of the same object. We evaluate the effectiveness of this cue by comparing the segmentations computed by our method against hand-labeled ones for several different models. We also show how the results of our segmentations can be used to highlight important objects in the scene and label them using noisy user-specified textual tag data. These methods are demonstrated on photos of several popular tourist sites downloaded from the Internet.

1 Introduction

With billions of photos now online, community photo collections found on Flickr [1] and other photo sharing sites offer tremendous opportunities for computer vision research. Recently, much work in the field has been devoted to making use of such photo collections for visualization [2], hole filling [3], learning object category models [4], dense 3D scene reconstruction [5], and geolocation [6]. While these recent efforts have capitalized on the *quantity* and *variety* of online images to enable various applications, a second important source of information is the *distribution* of photos — i.e., which photos do people choose to take?

The distribution of photos in a large collection holds valuable semantic information about the content of the scene. In this paper, we seek to leverage this information to automatically identify and segment interesting objects in a scene. While extremely challenging to solve purely by analyzing pixels in a single image, this problem is much more tractable with a large image collection. For example, a robust interest operator is obtained by simply finding features (e.g. SIFT [7]) that appear in numerous photos. By identifying oft-photographed features, this operator tends to highlight the parts of the scene that people find most interesting. The fact that this works robustly is a powerful demonstration of *the wisdom of crowds* [8], where a community of people combine to provide information that is otherwise difficult to obtain.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 541–553, 2008.



(a) (b) (c) (d)

Fig. 1. (a) A "bad" image of the Hagia Sophia. (b,c) Two "good" images of objects in the Hagia Sophia. Good images provide useful segmentation cues. (d) An image of the Trevi Fountain showing why photo collections of some scenes may not contain many good images. In this case, there are interesting statues in the façade, but it is difficult for photographers to get close enough to photograph them individually.

While detecting interesting features is straightforward via simple counting methods, identifying interesting *objects* is more challenging, as it necessitates segmentation — another difficult, ill-posed computer vision problem. To address this segmentation problem, we propose a new *field-of-view* cue for inferring perceptual grouping information from large photo collections. The key idea is very simple—it's based on the observation that people tend to take photos of "objects" as opposed to arbitrary regions of the scene, and these objects are usually framed within the field of view of the photo. Consequently, if two points are on the same object, those two points will likely appear together in many photos. We may therefore use the co-occurence of features in many images as a cue for grouping them together.

The main contribution of this paper is the introduction of this field-of-view cue and techniques for leveraging it for identifying and segmenting objects in images and point clouds. We also demonstrate the use of this analysis to display the relative importances of objects, automatically compute textual labels for these objects from noisy user-contributed Flickr tags (which are attached to images, not image regions), and browse a scene using an interactive map.

1.1 Related Work

The problem of decomposing a set of images into recurring objects in an unsupervised manner has been the subject of much recent work in computer vision, such as Fergus et al. [9], Sivic et al. [10] and Sudderth et al. [11]. However, to our knowledge, ours is the first paper to address this problem for static scenes, where objects always have the same 3D context and variation among images arises from differing camera positions and viewing directions. A closely related problem is addressed by Campbell et al. [12], who use a camera fixation cue that is similar to our field-of-view cue to segment a single 3D object from a set of images.

In a similar vein to our paper, Russell et al. [13] use grouping cues in an unsupervised approach to object category segmentation. They compute multiple

segmentations for each image and then apply a latent topic model where the *segments* serve as documents, relying on the fact that each object will appear against several different backgrounds, so "bad" segments will contain multiple latent topics. In our case, however, as the scene is static, any image region will almost always appear in the same context. We instead rely on the field-of-view of the photographs to provide our grouping cues.

Some of our applications draw on other work in related areas. Simon et al. [14] use the distribution of photos in an online collection to find canonical images of a scene. In their work, these images are intended to serve as a summary of the scene, and they do not attempt to extract individual objects, nor do they segment the scene or any of the images. A similar problem is addressed by Epshtein et al. [15], who use the distribution of viewing frusta in a photo collection to organize the photos into a hierarchy. Barnard et al. [16] learn the joint distribution of image regions and text labels and use this distribution to predict labels for regions in new images. Instead of attempting to predict noisy tags, we find strong associations between clusters and tags and use them to select good labels from the set of tags submitted by Flickr users in Section 4.2.

1.2 Our Approach

As discussed in the previous section, much existing work deals with identifying and segmenting object categories from visual scenes, where each object may appear against different backgrounds in different images. We address a different question: how can one identify and segment interesting objects from a static scene? The fact that the background changes was important for previous work, and enabled a solution using latent topic models. While handling static scenes may seem like a simpler problem, the fact that the background is not changing is a challenge — different cues are needed.

We instead use information provided by the distribution of photos taken of the scenes. To segment the scenes, we use what could be called a *field-of-view* or *incidence* constraint. For scenes that contain individual objects of interest, we hypothesize that multiple photographers are going to take pictures "of" each

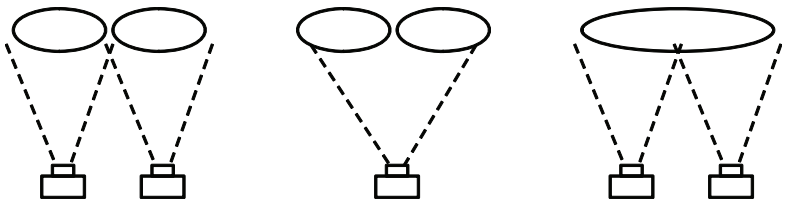


Fig. 2. (a) A case in which field-of-view cues and spatial cues agree, indicating a pair of objects. (b) A case in which field-of-view cues indicate a single object but spatial cues indicate a pair of objects. (c) A case in which field-of-view cues indicate a pair of objects but spatial cues indicate a single object.

object: pictures in which the object is prominent and takes up most of the frame. We call such images "good", and other images "bad". Good images, for our purposes, are ones which provide useful segmentation cues. Figure 1 shows a few example images of both types. When enough images are good, we find that field-of-view constraints can be used to accurately segment the scene into interesting objects. In the remainder of the paper, we give a simple scene model which takes advantage of field-of-view constraints and evaluate the effectiveness of this model on several different scenes. We also demonstrate the three applications of scene segmentation mentioned above: an "interestingness" viewer, an automatic object labeler that uses user-submitted Flickr tags, and an overhead map browser.

2 Algorithm

Given a set of images of a scene, we first follow the strategy of Snavely et al. [2]:

1. Use the SIFT keypoint detector [7] to extract feature regions from all images. These regions are represented using the SIFT descriptor.
2. For each pair of images, perform feature matching on the descriptors. Prune this set of matches by using RANSAC [17] to estimate a fundamental matrix, removing all inconsistent matches.
3. Organize the matches into tracks, or connected components of features, removing tracks that contain multiple features in the same image.
4. Perform structure from motion on the set of feature tracks, which returns a set of 3D point locations for all valid tracks, as well as camera parameters for each image.

We now operate on two structures, the (sparse) point-image incidence matrix, indicating which points appear in which images, and the set of 3D point locations. Let V be the set of images and X be the set of points, with $M = |V|$ and $N = |X|$. We use $x_j \in V_i$ if point j is visible in image i. Our goal is to compute a clustering C over the points X, where two points belong to the same cluster if they are part of the same object.

Our approach is to use both field-of-view cues and spatial cues to segment the scene. Figure 2 illustrates how these cues work at a basic level. Field-of-view cues encourage points seen in the same view to be part of the same object, while spatial cues encourage objects to be spatially localizable. We use image incidence to enforce field-of-view cues and a single 3D Gaussian distribution per object to enforce spatial cues. To construct a probabilistic model that takes advantage of both types of cues, we combine a mixture of 3D Gaussians (Figure 3a), which uses spatial cues only, and pLSA [18] (Figure 3b), which uses incidence cues only. Our combined model (Figure 3c) uses both types of cues.

We briefly review the probability distributions specified by a mixture of Gaussians and pLSA, and combine them into a single model. A mixture of Gaussians corresponds to the following distribution:

$$P(C, X|\pi, \mu, \Sigma) = \prod_j P(c_j|\pi) P(x_j|c_j, \mu, \Sigma) \tag{1}$$

$$P(c_j|\pi) \sim \text{Mult}(\pi)$$
$$P(x_j|c_j, \mu, \Sigma) \sim \mathcal{N}(\mu_{c_j}, \Sigma_{c_j})$$

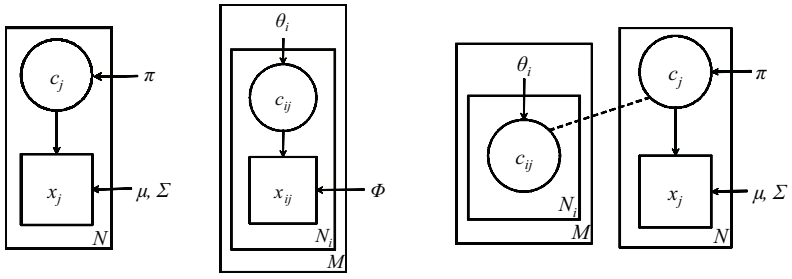
In this model, there is a class variable c_j associated with each point x_j. The class variable is drawn from a multinomial distribution with parameter π, and the point locations are drawn from 3D Gaussians with parameters μ_{c_j} and Σ_{c_j}, where the point class c_j specifies which Gaussian to use. The pLSA model corresponds to the following distribution:

$$P(C, X|\theta, \Phi) = \prod_i \prod_{j|x_j \in V_i} P(c_{ij}|\theta_i) P(x_{ij}|c_{ij}, \Phi) \tag{2}$$
$$P(c_{ij}|\theta_i) \sim \text{Mult}(\theta_i)$$
$$P(x_{ij}|c_{ij}, \Phi) \sim \text{Mult}(\Phi_{c_{ij}})$$

In ordinary pLSA, x_{ij} would be a tally of the number of times word j appears in document i. However, in our case the words are 3D points, none of which can appear more than once in a single image. In the pLSA model, there is a class variable c_{ij} for each point-image *incidence*. In other words, a point can belong to different objects in different images. This is not really desirable in our case, so we restrict our model to use a single class variable per point. In addition, we introduce a spatial term. Our combined model corresponds to the following distribution:

$$P(C, X|\theta, \pi, \mu, \Sigma) = \left(\prod_i \prod_{j|x_j \in V_i} P(c_{ij}|\theta_i) \right) \times \left(\prod_j P(c_j|\pi) P(x_j|c_j, \mu, \Sigma) \right) \tag{3}$$
$$P(c_{ij}|\theta_i) \sim \text{Mult}(\theta_i)$$
$$P(c_j|\pi) \sim \text{Mult}(\pi)$$
$$P(x_j|c_j, \mu, \Sigma) \sim \mathcal{N}(\mu_{c_j}, \Sigma_{c_j})$$
$$c_{ij} = c_j$$

Instead of directly replacing the multinomial topic distributions from pLSA with 3D Gaussian distributions, we add a single class variable c_j for each 3D point x_j and tie the values of the incidence class variables c_{ij} with c_j. The resulting model is not strictly a valid Bayesian network, but could easily be converted to one in which the c_{ij} variables are eliminated and c_j is directly conditioned on all of the images in which the point appears. We find it is easier to think about tied variables, and the above joint density can easily be maximized (locally) using the EM algorithm [19].



(a) (b) (c)

Fig. 3. Graphical models for (a) a mixture of Gaussians on 3D point locations, (b) the pLSA model, and (c) the combined model that uses both point location and image incidence information

3 Evaluation

We tested our model on six scenes, each of which contains between 300 and 3000 images downloaded from Flickr: Trafalgar Square, the Pantheon in Rome, Hagia Sophia, Trevi Fountain, Old Town Square in Prague, and Piazza Navona. These scenes all contain component objects which have names and can be identified visually. In order to test our segmentations, we created ground-truth clusterings for each scene manually, as follows. For each scene, we reconstructed a set of 3D points using the system of Snavely et al. [2]. We then assigned 3D points to clusters by manually selecting regions in images and grouping all points in the selected region with a particular cluster. Though any hand-labeling of this nature is somewhat arbitrary, we attempted to label objects as uncontroversially as possible. We also used Wikipedia [20] text and images to decide which objects should be included when there was some uncertainty. In other cases, there is a natural segmentation implied by the scene. For example, in the Trafalgar Square scene we labeled each building and statue as a separate object. Since the reconstructed scenes contain hundreds of thousands of 3D points, many of which don't belong to an easily nameable object, we only hand-label a small fraction of the points, and evaluate our algorithm on these points only. Note that the aforementioned manual steps were used only to create the ground truth used for evaluation purposes. The segmentation algorithms themselves are fully automatic.

For all scenes, we tested three different models: a mixture of 3D Gaussians, the pLSA model, and our combined model that uses both spatial and incidence cues. Each model was tested using multiple different values of k, the number of clusters. For each value of k, we used 5 different random initializations and ran 100 iterations of the EM algorithm for each, then kept the single result with highest joint probability for each value of k. We created hard cluster assignments by assigning each point to its most likely cluster (or, for pLSA, the cluster under which the point has highest probability). We evaluate clusterings using

Table 1. Median values of the VI clustering metric $VI(C, C^*)$ for each algorithm and scene, over multiple runs of EM. Lower values indicate the computed clustering is closer to the hand-labeled clustering.

	Trafalgar	Pantheon	Hagia Sophia	Trevi	Prague	Navona
mixture of Gaussians	1.15	1.36	0.63	**0.81**	0.35	0.68
pLSA	2.07	1.70	0.64	3.12	1.13	1.46
combined model	**0.69**	**0.38**	**0.53**	2.07	**0.20**	**0.45**

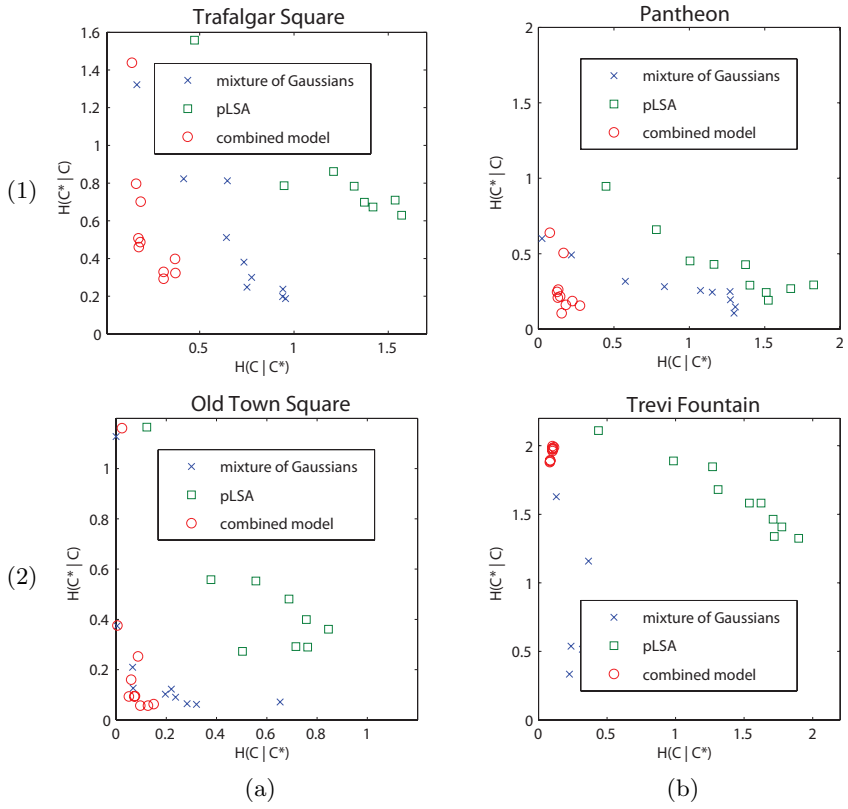
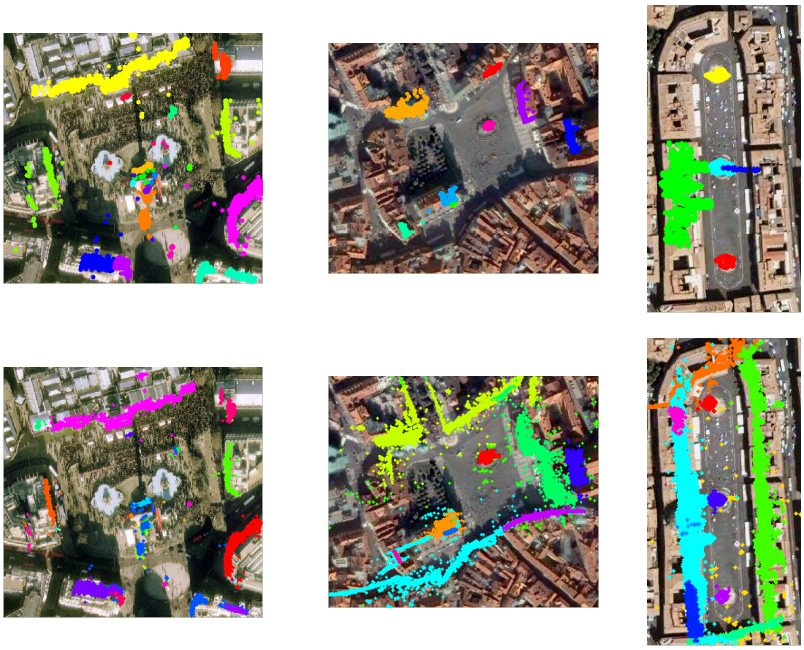


Fig. 4. Evaluation of clustering C against ground truth clustering C^* for the (1a) Trafalgar Square, (1b) Pantheon, (2a) Old Town Square, and (2b) Trevi Fountain datasets. The horizontal axis $H(C|C^*)$ is a measure of over-segmentation, and the vertical axis $H(C^*|C)$ is a measure of under-segmentation. Note that for the Trevi Fountain, both pLSA and the combined model are prone to undersegmentation, as most images of the Trevi Fountain are of the entire façade. This causes field-of-view cues to prefer larger objects, even when there are interesting details within the façade.

Meila's Variation of Information metric [21]. Given a ground truth clustering C^* and computed clustering C, the VI metric $VI(C, C^*) = H(C|C^*) + H(C^*|C)$ measures the amount of information lost and gained between the two clusterings.



(a) (b) (c)

Fig. 5. Satellite view of ground truth (top row) and computed segmentations (bottom row) for (a) Trafalgar Square, (b) Old Town Square, and (c) Piazza Navona. Each color corresponds to a different cluster. For the Trafalgar Square scene, not all reconstructed points are shown in the computed segmentation to avoid clutter. In all figures in this paper, we use the same color for each cluster in each scene. As there is no explicit correspondence between ground-truth clusters and computed clusters, the ground-truth clusterings do not use the same color scheme.

For two identical clusterings, this value is zero. Also, by looking at the two conditional entropy terms separately, we can get a sense of how over- and under-segmented C is.

Table 1 contains median VI distances for each of the three algorithms and six scenes. The combined model performs best on all scenes except the Trevi Fountain, which suffers from undersegmentation as photographers cannot get close enough to the façade to take closeup images of the interesting objects. Figure 4 shows our over- and under-segmentation results on four of the scenes. In general, when the collection of photographs contains many images of interesting objects, our combined model does well. Since what we are testing is whether or not field-of-view cues provide additional information that is useful for segmentation, our results demonstrate that this is true for many scenes. Visualizations of the clusterings themselves are shown in Figures 5 and 6.

(a) (b) (c) (d)

Fig. 6. Image views of computed segmentations for (a,b) Trafalgar Square, (c) Old Town Square, and (d) Piazza Navona

4 Applications

4.1 Importance Viewer

Once we have computed a 3D point segmentation for a scene, we can use this segmentation to compute and display additional information about the scene. In this section, we describe two such applications: highlighting interesting objects in images and labeling image regions using noisy user-submitted Flickr tags.

Given an image in the collection, we want to identify regions which belong to objects that are important or interesting. We define an object as interesting if there are many photos of it in the collection. As this tends to overly reward large background objects, we also penalize objects for size. Our importance score is:

$$\text{imp}(c) = \alpha \frac{1}{|\Sigma_c|} \sum_i \theta_i(c) \tag{4}$$

Here, α is a scene normalization coefficient that enforces a fixed total importance, $\frac{1}{|\Sigma_c|}$ penalizes clusters proportional to the determinant of their covariance matrices, and $\sum_i \theta_i(c)$ rewards clusters for appearing in many images. To visualize importance in an image, we assign each pixel to its nearest feature point in the image and highlight the pixel with intensity proportional to the importance of the cluster to which this point is assigned, falling off as distance to this point increases. Some resulting importance images can be seen in Figure 7.

(a) (b) (c)

Fig. 7. Importance images computed for (a) Piazza Navona, (b) Prague, and (c) the Pantheon (top) and Trafalgar Square (bottom). Importance is indicated by color saturation, with different hues for different clusters.

4.2 Region Labeling

Flickr and most other photo sharing sites give users the ability to attach textual tags to entire images. (Flickr also provides functionality for leaving rectangular notes on image regions, but this feature is much less utilized.) In general, these tags are noisy and the majority of them do not correspond to actual objects in the scene. For many objects, however, we are able to compute accurate tags by examining tag-cluster co-occurrence statistics. To find good object tags, we first apply pLSA to the tags with fixed image-topic distributions θ computed from the point clustering. This gives us a distribution over tags for each cluster $P(t|c)$, which also gives us the joint distribution $P(c,t)$. For a particular cluster c and tag t, we compute the following score:

$$\text{score}(c,t) = P(c,t)\,(\log P(c,t) - \log P(c)P(t)) \tag{5}$$

This gives high scores to cluster-tag pairs that appear much more frequently than would be expected given just the marginals, which indicates that the cluster and tag are probably related. We then assign the tag with highest score to each cluster if the score exceeds a specified threshold, otherwise we assign no tag to the cluster. Region tagging results can be seen in Figure 8.

4.3 Interactive Map Viewer

Many systems have been created to support interactive browsing of the visual content of a scene [2,14,15,22]. Our segmentations allow for the possibility of

Fig. 8. Region tags automatically computed from our segmentation, and used to label two images of (a) Trafalgar Square, (b) Old Town Square, and (c) Piazza Navona. We compute these labels using noisy user-submitted tags downloaded from Flickr, automatically associating a single tag (or no tags) with each cluster. In these images, we manually moved the labels to make them more readable.

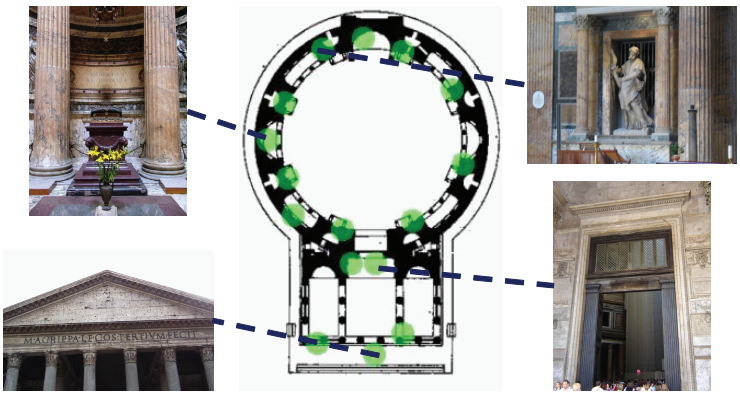


Fig. 9. An interactive floor plan viewer showing the Pantheon (center). By moving the mouse over one of the highlighted circles, the user can see an image of the object at that location (sample images shown on left and right).

object-centric interactive scene viewers. We have created a simple interactive viewer based on an overhead map or floor plan, shown in Figure 9.

To create the interactive floor plan, we first segment the scene using the algorithm from Section 2. We then manually align the scene points with an

overhead view. Since we only want to include segments which can be localized at a reasonably-sized spot on the map, we remove all segments larger than a size threshold. To choose the representative image for each segment, we compute the Kullback-Leibler divergence between the distribution of scene points in each image and each cluster.

5 Conclusion

In this paper we have proposed a new field-of-view cue that can be used to extract objects from static 3D scenes, along with a probabilistic model that takes advantage of this cue and several applications of our method. We stress that the probabilistic model is intended to evaluate the usefulness of field-of-view cues, and not to provide a complete solution to the scene segmentation problem. Note in particular that we are not incorporating any of the more standard image segmentation cues such as intensity, color, contour, or region information, except through feature matching for estimating point correspondences. Including such additional terms and using more sophisticated models, like Latent Dirichlet Allocation [23] or Dirichlet process mixtures would likely further improve upon our results. Still, our model that takes into account only field-of-view and spatial proximity is able to achieve segmentations that are good enough to enable a variety of applications.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0743635, the Office of Naval Research, the University of Washington Animation Research Labs, Adobe, Microsoft, Google, and an endowment by Rob Short and Emer Dooley.

References

1. Flickr, `http://www.flickr.com`
2. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3D. In: SIGGRAPH Conference Proceedings, pp. 835–846 (2006)
3. Hays, J., Efros, A.A.: Scene completion using millions of photographs. In: SIGGRAPH Conference Proceedings, vol. 26(3) (2007)
4. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Proceedings of the 10th IEEE International Conference on Computer Vision, vol. 2, pp. 1816–1823 (2005)
5. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: Proceedings of the 11th IEEE International Conference on Computer Vision, pp. 1–8 (2007)
6. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)

7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Surowiecki, J.: The Wisdom of Crowds. Random House, New York (2004)
9. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271 (2003)
10. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering objects and their localization in images. In: Proceedings of the 10th IEEE International Conference on Computer Vision, vol. 1, pp. 370–377 (2005)
11. Sudderth, E., Torralba, A., Freeman, W., Willsky, A.: Describing visual scenes using transformed objects and parts. International Journal of Computer Vision 77(1-3), 291–330 (2007)
12. Campbell, N., Vogiatzis, G., Hernandez, C., Cipolla, R.: Automatic 3D object segmentation in multiple views using volumetric graph-cuts. In: Proceedings of the British Machine Vision Conference, vol. 1, pp. 530–539 (2007)
13. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1605–1614 (2006)
14. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: Proceedings of the 11th IEEE International Conference on Computer Vision, pp. 1–8 (2007)
15. Epshtein, B., Ofek, E., Wexler, Y., Zhang, P.: Hierarchical photo organization using geo-relevance. In: Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, pp. 1–7. ACM, New York (2007)
16. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research 3(6), 1107–1135 (2003)
17. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981)
18. Hofmann, T.: Probabilistic latent semantic analysis. Proceedings of Uncertainty in Artificial Intelligence, 289–296 (1999)
19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38
20. Wikipedia, `http://www.wikipedia.org`
21. Meilă, M.: Comparing clusterings: an information based distance. Journal of Multivariate Analysis 98(5), 873–895 (2007)
22. Ahern, S., Naaman, M., Nair, R., Yang, J.H.: World Explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: Proceedings of the 2007 Conference on Digital Libraries, pp. 1–10 (2007)
23. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003)