Synergistic Clustering of Image and Segment Descriptors for Unsupervised Scene Understanding

Daniel M. Steinberg, Oscar Pizarro, Stefan B. Williams
Australian Centre for Field Robotics, The University of Sydney
Sydney, NSW, 2006
{d.steinberg, o.pizarro, stefanw}@acfr.usyd.edu.au

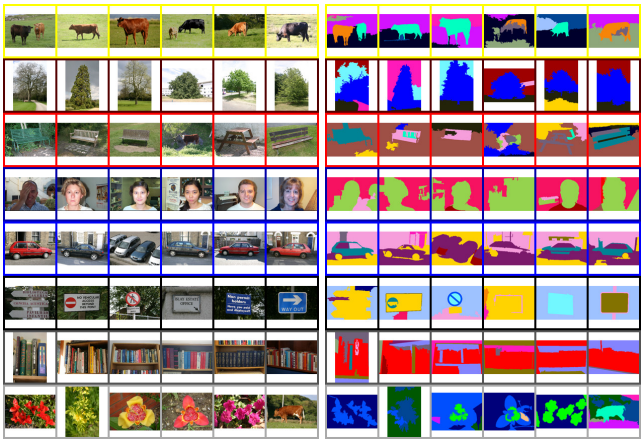
1. Overview

In many applications, the quantity and rate at which visual data is collected can far outpace a humans ability to label or annotate even a small percentage of it. One example of this is the collection of scientific visual data by autonomous agents such as planetary rovers, unmanned air vehicles (UAVs), or autonomous underwater vehicles (AUVs). Unsupervised "scene understanding" algorithms could summarise this data in the absence of any annotations. A human expert would then only need to view these summaries before directing their attention to relevant subsets of the data for subsequent analysis.

In [7], we present a Bayesian graphical model specialised for truly unsupervised scene understanding applications. We refer to it as the *multiple-source clustering model* (MCM). It is able to model multiple albums of images at both scene and object levels without human supervision. Rather than relying on human-generated scene labels, it infers scene-types by clustering images. It uses a whole-image descriptor *as well as* a latent distribution of "object" types to represent images. These object-types are formed by simultaneously clustering image and segment descriptors – hence multiple-source. See Figure 1 for an example of the MCM's output.

The structure of our model is such that scene-types can influence the objects found in an image (we would likely find trees in a forest). This is conceptually similar to the work in [9], which is inspired by research on the human visual cortex [6]. Global visual features are used to understand the context of a scene without explicitly registering the individual objects that compose the image. This scene recognition provides context that aids the recognition of objects, which otherwise may be difficult to recognise in isolation. Also, in the MCM the co-occurrence and distribution of objects within an image can influence the type of scene it belongs to (cows and grass likely make a rural scene).

We present a fast, deterministic variational inference algorithm for the MCM in [7]. We also introduce a simple cluster search heuristic used in conjunction with the variational inference algorithm to automatically determine the number of scene and object clusters.



(a) Image clusters. (b) Segment clusters.

Figure 1. (a) A random selection of images from 8 of the 15 image clusters found by our proposed model on the MSRC dataset and (b) some of the (28) corresponding segment clusters. The image clusters have a normalised mutual information (NMI) score of 0.731, the segment clusters have an NMI of 0.580. No training or annotation data is used.

Sparse code spatial pyramid match (ScSPM) [11] descriptors that have been compressed using randomized PCA are used to represent whole images. These descriptors preserve aspects of an image's spatial structure and dominant features. We use a fast mean-shift algorithm to over-segment images, and extract features from these segments using a dense independent component analysis transform, which preserves colour and texture information. We find these representations complimentary and greatly enhance the MCM's performance.

Table 1. Image clustering/classification performance for the UIUC sport dataset. The algorithms above the mid rule are unsupervised, the algorithms below are weakly or fully supervised. The MCM achieves the best NMI score, and the second best derived accuracy score.

Algorithm	**NMI** (std.)	**Acc.** (% (std.), #0)
MCM	**0.641** (0.018)	74.1 (1.5), 1
VDP+ScSPM [2]	0.557	63.4, 2
SC+ScSPM [12]	0.429 (0.02)	58.9 (2.4), 1.1
Du *et. al.* [1] no LSBP	0.389	60.5
Du *et. al.* [1] LSBP	0.418	63.5
Li *et. al.* [4]	0.276	54
sLDA [10] (annots.)	0.438	66
sLDA [10]	0.446	65
Li *et. al.* [3]	0.466	69.11
DiscLDA+GC [5]	0.506	70
SVM+ScSPM [11]	0.549	72.9
CA-TM [5]	0.592	**78**

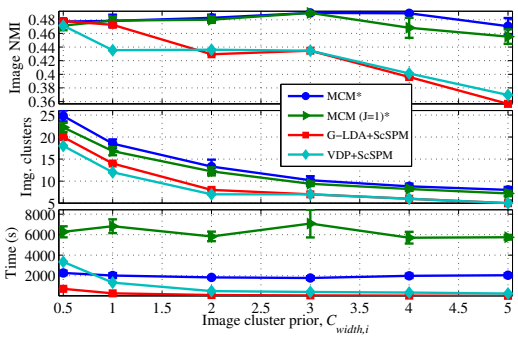


Figure 2. AUV dataset image clustering performance. Here the MCM and G-LDA can model the 12 AUV deployments as separate "corpora". The VDP and MCM (J=1) only model a single corpus. * Denotes 8 Xeon 2.2 GHz cores were used.

2. Key Results

The primary result of [7] is in showing that the MCM can synergistically cluster both image and segment descriptors and that it outperforms unsupervised models that only consider one source of information. It is also competitive with weakly-supervised and supervised models for scene understanding, see for example Table 1. We are able to compare unsupervised and supervised techniques using standard measures derived from confusion matrices and contingency tables, i.e., mean accuracy and normalised mutual information (NMI) [8]. Finally, we demonstrate our model operating on a dataset of 100,647 images collected from multiple deployments of an autonomous underwater vehicle. Interestingly, we find that modelling these deployments as separate "corpora" in the MCM produced significant run time reductions as opposed to modelling all of the deployments as a single corpus, see Figure 2.

3. Conclusion

In [7] we demonstrate that fully unsupervised, annotation-less algorithms for scene understanding can be competitive with supervised and weakly-supervised algorithms. The proposed MCM can use contextual information from scene-types to improve object discovery and is able to use object co-occurrence and proportion information to greatly improve scene discovery. We also demonstrate that the MCM is able to run on large datasets gathered by autonomous robots, enabling fully automated data gathering and interpretation pipelines. Like many weakly- and supervised scene understanding models, the MCM is effective at discovering scene-types, but not as effective at object discovery – which is a much harder problem. Focusing on the unsupervised object discovery and recognition aspects of such models will be a useful area of future research.

References

[1] L. Du, L. Ren, D. Dunson, and L. Carin. A Bayesian model for simultaneous image clustering, annotation and object segmentation. *Advances in Neural Information Processing Systems*, 22:486–494, 2009. 2

[2] K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational Dirichlet process mixtures. *Advances in Neural Information Processing Systems*, 19:761, 2007. 2

[3] L. Li, M. Zhou, G. Sapiro, and L. Carin. On the integration of topic modeling and dictionary learning. *International Conference on Machine Learning, ICML*, 2011. 2

[4] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, CVPR 2009*, pages 2036–2043. IEEE, 2009. 2

[5] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *Computer Vision and Pattern Recognition, CVPR*, pages 2743 –2750. IEEE, 2012. 2

[6] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23, 2006. 1

[7] D. M. Steinberg, O. Pizarro, and S. B. Williams. Synergistic clustering of image and segment descriptors for unsupervised scene understanding. In *International Conference on Computer Vision (ICCV)*, Darling Harbour, Sydney, 2013. IEEE. 1, 2

[8] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2003. 2

[9] A. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun. ACM*, 53(3):107–114, 2010. 1

[10] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, CVPR*, pages 1903–1910. IEEE, 2009. 2

[11] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, CVPR*, pages 1794–1801. IEEE, 2009. 1, 2

[12] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, volume 17, pages 1601–1608, 2004. 2