Synergistic clustering of image and segment descriptors for unsupervised scene understanding

Daniel M. Steinberg,
Oscar Pizarro,
Stefan B. Williams
Australian Centre of Field Robotics
The University of Sydney, NSW

In many applications the quantity and rate at which visual data is collected can far outpace a human's ability to label or annotate even a small percentage of it. For example, the collection of scientific visual data by autonomous agents such as planetary rovers 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.

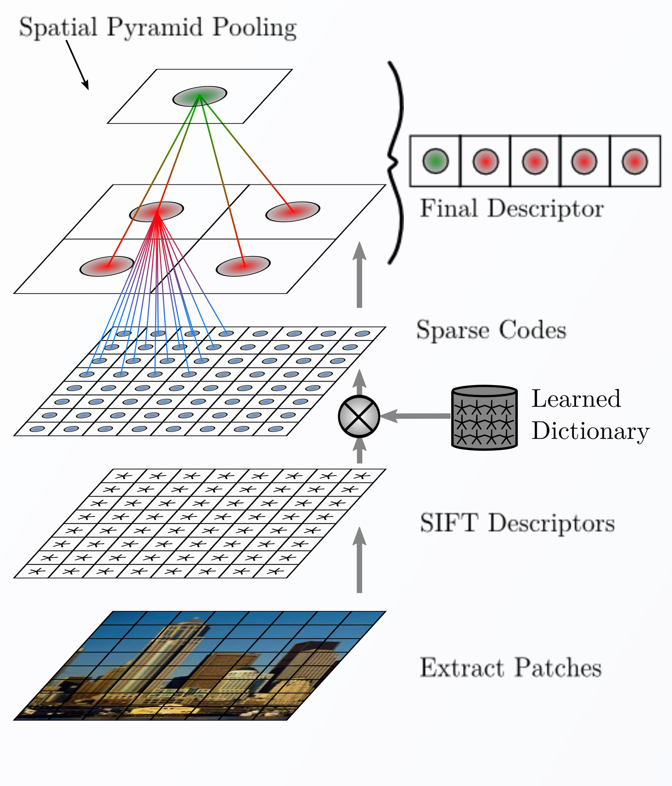
Introduction & Aim

- **Scene understanding**: frameworks that incorporate and model multiple sources of visual, annotation or other information to improve some joint visual inference task (i.e. scene recognition and object detection with image annotations).
- Scene understanding is an active research area, and many algorithms exist for **weakly or fully supervised** applications.
- A few of these algorithms can be used in situations where only visual data is available, though they may operate in a reduced capacity [4] or have not been benchmarked thoroughly in this setting [7, 12].
- We present a Bayesian graphical model specialised for **truly unsupervised** (visual data only) scene understanding applications.

Image Representation

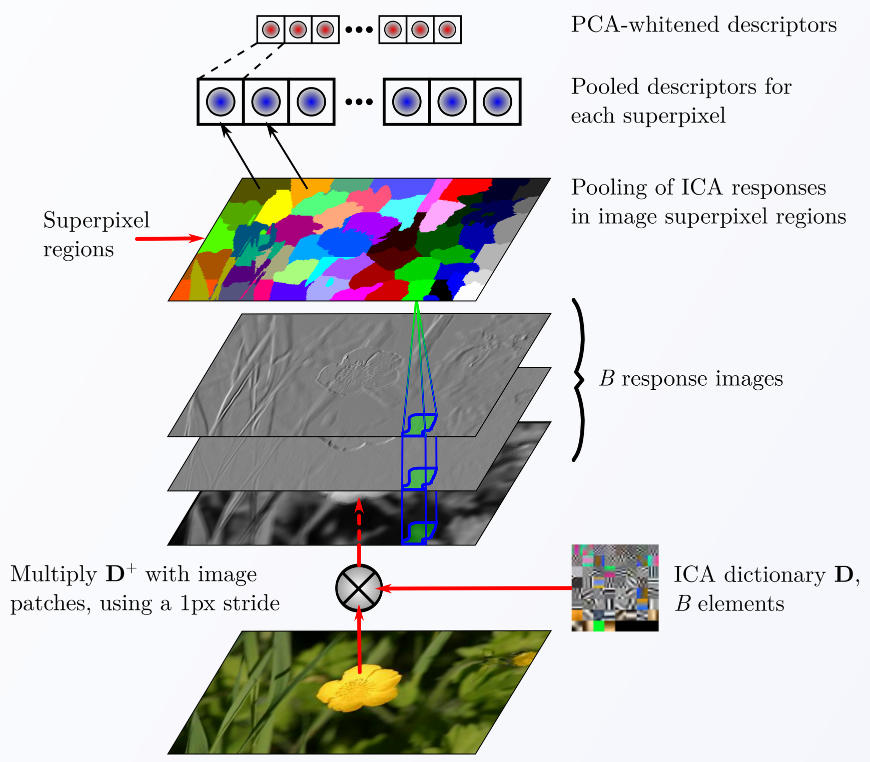
We use a whole image descriptor as well as a latent distribution of "object" types to represent images. These object-types are formed by simultaneously clustering image and segment features.

Image features:



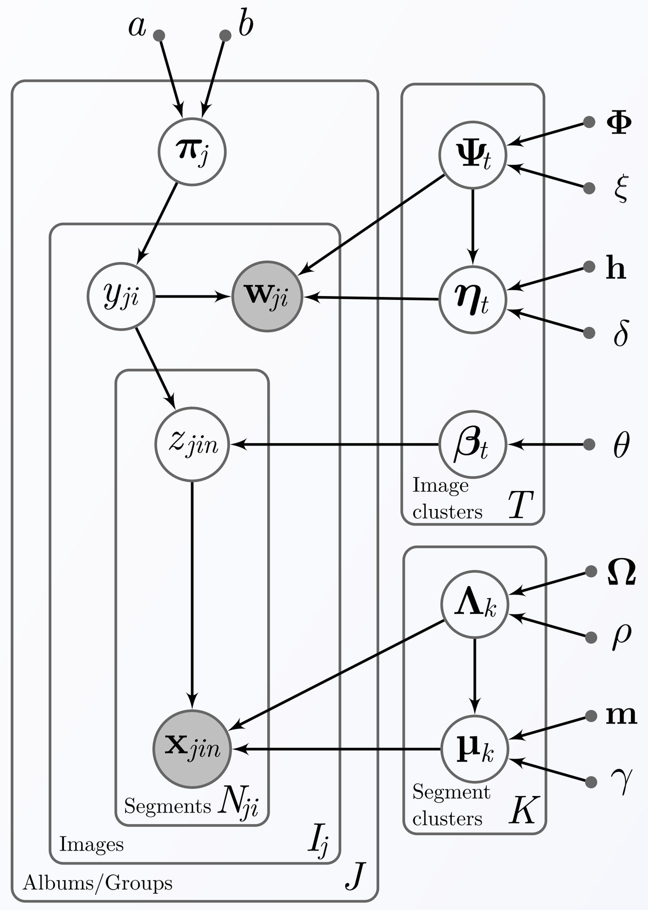
- Based on sparse coding spatial pyramid matching (ScSPM) [27]
  - Encodes spatial layout of the image

Segment features:



- Made especially for this task
- Based on pooling independent component analysis (ICA) responses in segments
- Segmentation is done by the mean shift algorithm

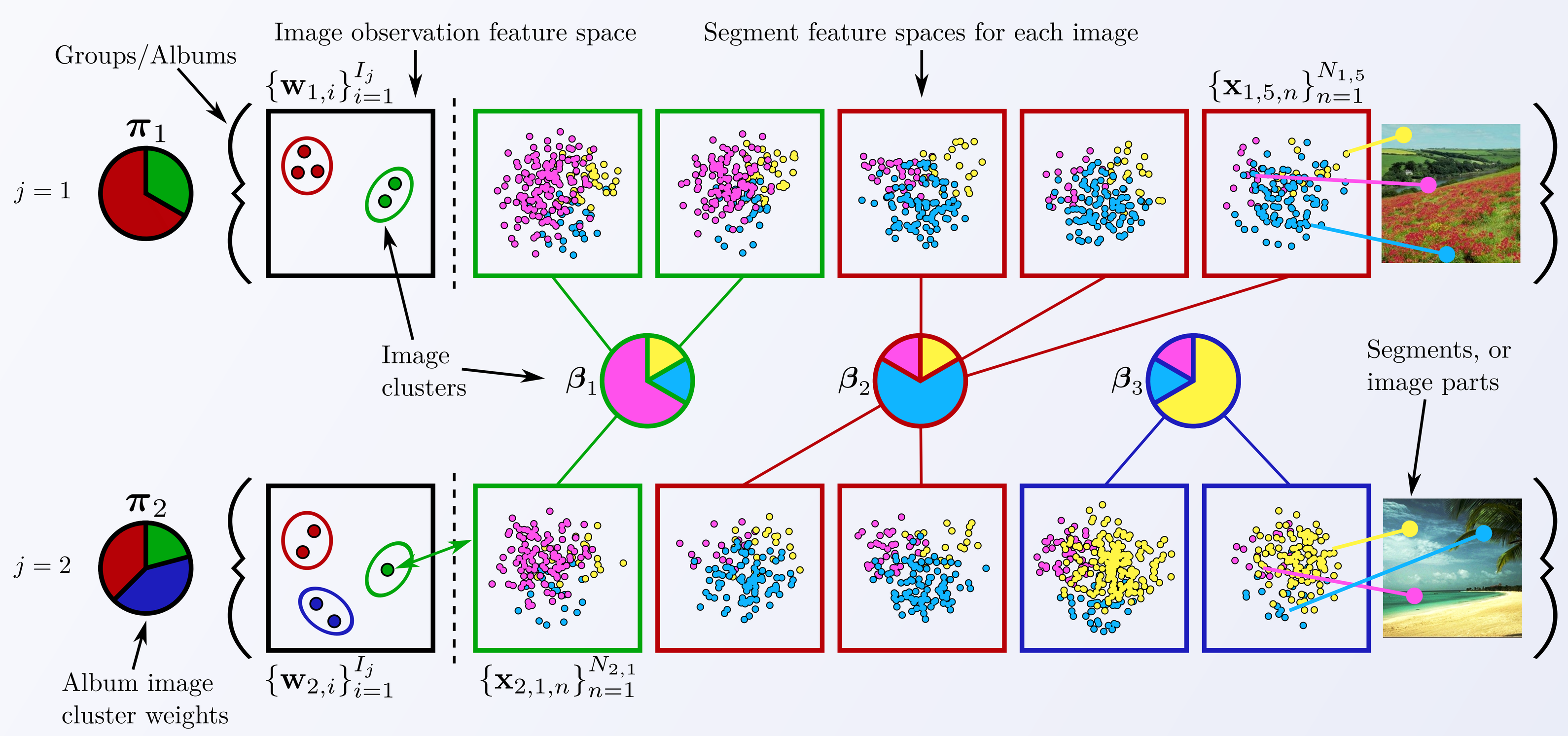
Multiple-source Clustering Model



MCM's generative story:

1. Draw T image cluster parameters $\boldsymbol{\beta}_t$, $\boldsymbol{\eta}_t$ and $\boldsymbol{\Psi}_t$ from $\mathrm{GDir}(a, b)$, $\mathcal{N}\big(\mathbf{h}, (\delta\boldsymbol{\Psi})^{-1}\big)$ and $\mathcal{W}(\boldsymbol{\Phi}, \xi)$ respectively.
2. Draw K segment cluster parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ from $\mathcal{N}\big(\mathbf{m}, (\gamma\boldsymbol{\Lambda}_k)^{-1}\big)$ and $\mathcal{W}(\boldsymbol{\Omega}, \rho)$ respectively.
3. For each group or album, $j \in \{1, \ldots, J\}$:
   (a) Draw mixture weights $\boldsymbol{\pi}_j \sim \mathrm{GDir}(a, b)$.
   (b) For each image, $i \in \{1, \ldots, I_j\}$:
      i. Choose an image cluster $y_{ji} \sim \mathrm{Categ}(\boldsymbol{\pi}_j)$.
      ii. Draw an image observation from the chosen image cluster $\mathbf{w}_{ji} | (y_{ji} = t) \sim \mathcal{N}(\boldsymbol{\eta}_t, \boldsymbol{\Psi}_t)$.
      iii. For each image segment $n \in \{1, \ldots, N_{ji}\}$:
         A. Choose a segment cluster $z_{jin} | (y_{ji} = t) \sim \mathrm{Categ}(\boldsymbol{\beta}_t)$.
         B. Draw a segment observation from the segment cluster $\mathbf{x}_{jin} | (z_{jin} = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$.

- Global visual features, $\mathbf{w}_{ji}$, are used to understand the **context of a scene.** This scene recognition provides **context that aids the recognition of objects**.
- That is, discovered scene-types or image clusters, *T*, can influence the objects or segment clusters, *K*, found in an image (e.g. we would likely find trees in a forest).
- Also, the **co-occurrence and distribution of objects** within an image, $\boldsymbol{\beta}_t$, can **influence the type of scene** it belongs to (e.g. cows and grass likely make a rural scene).
- The hyperparameters and **number of clusters** are learned using **variational Bayes**.

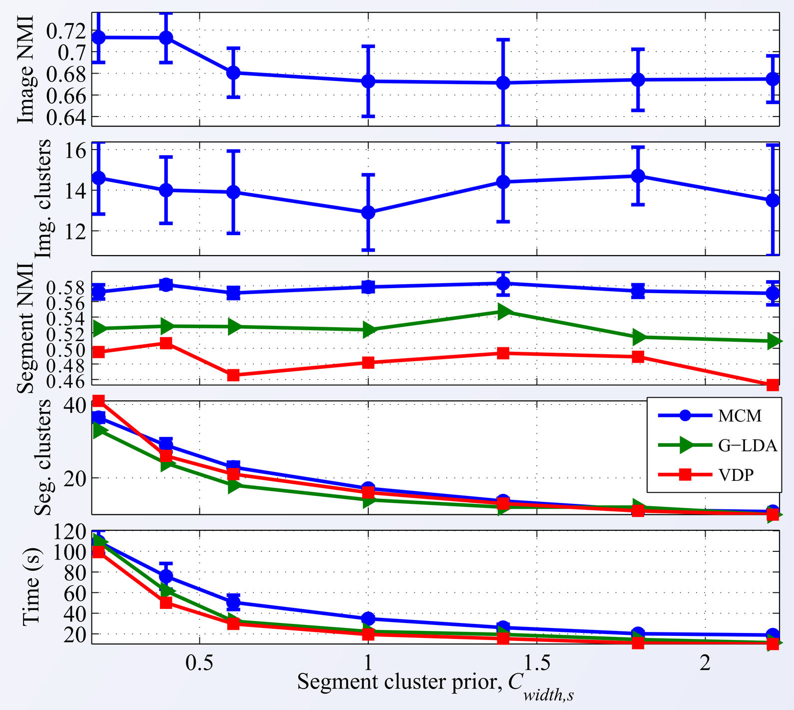


Some Results

- The MCM was compared against other unsupervised, weakly supervised and fully supervised algorithms on four datasets:
  - MSRC,
  - LabelMe,
  - UIUC Sports,
  - 100K underwater images from and AUV.
- **NMI** - *normalised mutual information* is a clustering metric. A value of 1.0 is when the labels and clusters perfectly agree.

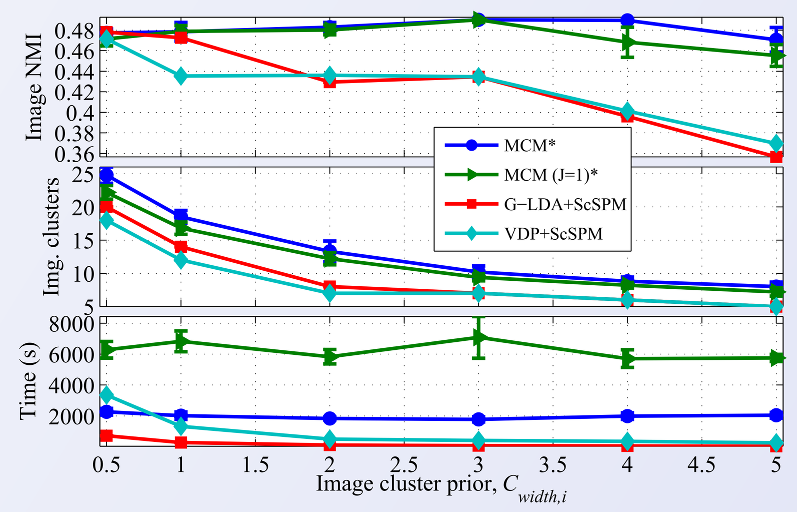
UIUC sports dataset scene recognition

Algorithm	**NMI** (std.)	**Acc.** (% (std.), #0)
MCM	**0.641** (0.018)	74.1 (1.5), 1
VDP+ScSPM [10]	0.557	63.4, 2
SC+ScSPM [28]	0.429 (0.02)	58.9 (2.4), 1.1
Du *et. al.* [7] no LSBP	0.389	60.5
Du *et. al.* [7] LSBP	0.418	63.5
Li *et. al.* [13]	0.276	54
sLDA [25] (annots.)	0.438	66
sLDA [25]	0.446	65
Li *et. al.* [12]	0.466	69.11
DiscLDA+GC [15]	0.506	70
SVM+ScSPM [27]	0.549	72.9
CA-TM [15]	0.592	**78**



MSRC scene and object discovery

$C_{width,s}$ is a prior tuning parameter that influences how many segment clusters the unsupervised algorithms find.

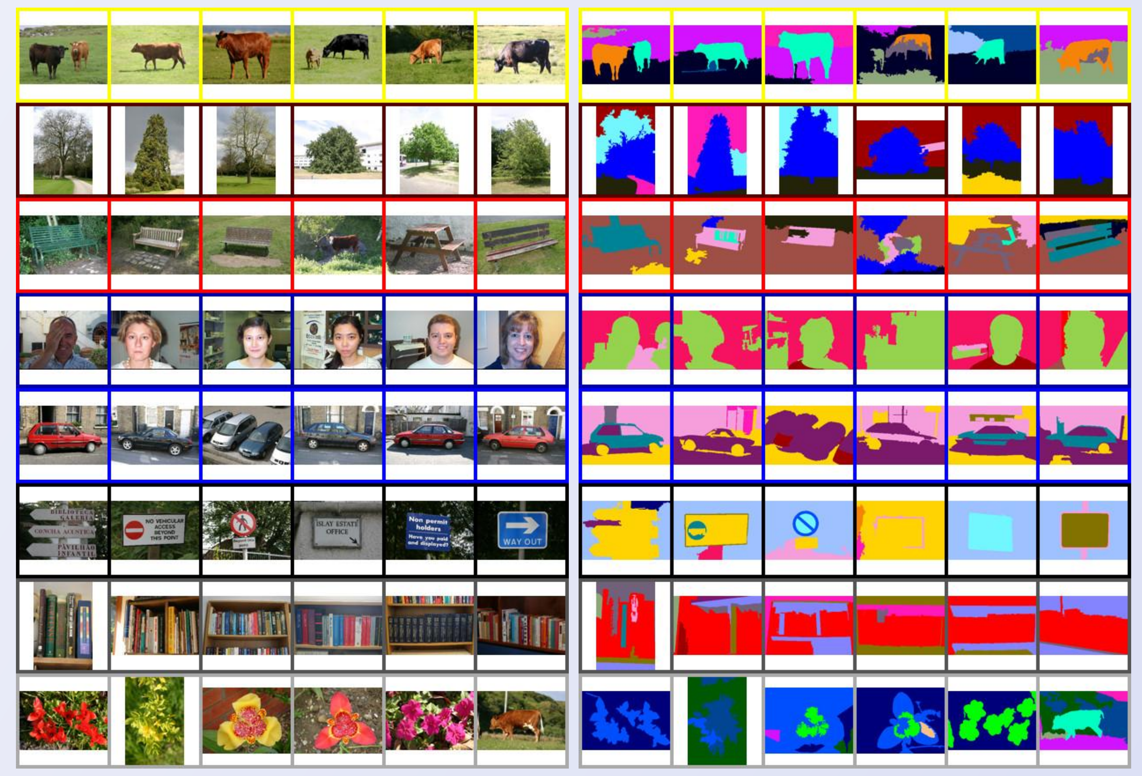


Robotic dataset scene discovery

This dataset is comprised of 100,000 images of the seafloor captured by an autonomous underwater vehicle (AUV). $C_{width,i}$ is a prior tuning parameter that influences how many image clusters the unsupervised algorithms find. *These algorithms were run using 8 cores as opposed to one.

Sample MSRC result

This is a single result of the MCM clustering the MSRC dataset. Random samples of the scene clusters are indicated by the row-wise coloured squares, and the segment clusters have been shown in the figure on the right. Here the image NMI was 0.731, and segment NMI was 0.58.



Conclusion

- This paper has demonstrated 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 performance.
- We have also demonstrated that the MCM is able to run on **large datasets** gathered by autonomous robots, enabling **fully automated** data gathering and interpretation pipelines.

References

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In International Conference Computer Vision, ICCV, pages 1–8. IEEE, 2007.

[7] 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.

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

[12] 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.

[13] 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.

[15] 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

[25] 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.

[27] 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.

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

ACFR AUSTRALIAN CENTRE FOR FIELD ROBOTICS

The paper, supplementary material and code can be found at: www.daniel-steinberg.info/publications.html