Latent Dirichlet Allocation for Image Segmentation and Source Finding in Radio Astronomy Images

Anna Friedlander
School of Engineering and Computer Science
Victoria University of Wellington
P.O. Box 600
Wellington 6140, New Zealand
anna.friedlander@ecs.vuw.ac.nz

Marcus Frean
School of Engineering and Computer Science
Victoria University of Wellington
P.O. Box 600
Wellington 6140, New Zealand
marcus.frean@ecs.vuw.ac.nz

Melanie Johnston-Hollitt
School of Chemical and Physical Sciences
Victoria University of Wellington
P.O. Box 600
Wellington 6140, New Zealand
melanie.johnston-hollitt@vuw.ac.nz

Christopher Hollitt
School of Engineering and Computer Science
Victoria University of Wellington
P.O. Box 600
Wellington 6140, New Zealand
christopher.hollitt@ecs.vuw.ac.nz

ABSTRACT

We present exploratory work into the application of the topic modelling algorithm latent Dirichlet allocation (LDA) to image segmentation in greyscale images, and in particular, source detection in radio astronomy images.

LDA performed similarly to the standard source-detection software on a representative sample of radio astronomy images. Our use of LDA underperforms on fainter and diffuse sources, but yields superior results on a representative image polluted with artefacts — the type of image in which the standard source-detection software requires manual intervention by an astronomer for adequate results.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Segmentation; I.5.1 [**Pattern Recognition**]: Models—*Statistical*; J.2 [**Physical Sciences and Engineering**]: Astronomy

General Terms

Algorithms

Keywords

Source detection, image segmentation, latent Dirichlet allocation, radio astronomy, pixel classification, flood-filling

IVCNZ '12 November 26 - 28 2012, Dunedin, New Zealand

1. INTRODUCTION

1.1 Source Detection in Radio Astronomy

The sheer volume of data to be produced by the next generation of radio telescopes makes detection of astronomical objects (sources) by manual processing impracticable [12].

The majority of automated source detection algorithms can be described as flood-filling or region-growing driven by (possibly transformed) pixel intensities [11], but these do not find all objects of interest. Spatially extended sources, particularly those that are faint, are poorly handled by existing automated approaches, as are sources in the presence of artefacts, and sources in images in which the signal-to-noise ratio varies across the image [8, 12, 14].

Radio astronomy images can be thought of as primarily background with an unknown number of spatially extended sources. Identifying the sources requires distinguishing them from background, a task made difficult by the diversity within and between sources. The variability of background is lower than that of sources and in that sense it is easier to identify. Additionally, there are relatively few source pixels compared to the number of background pixels (in contrast to non-astronomical images with relatively many foreground/object pixels; see Fig. 2).

We approach the problem of source detection as one of identifying and excluding regions of background and merging what remains into a modest number of sources. This requires that we specify a method for labelling a region as likely or unlikely to be background.

One plausible approach is to assume background is equivalent to "non-signal" with some noise, and this is the basis for many existing algorithms [11]. However, this assumption can fail. Background pixels may not be restricted to a single narrow range of pixel intensities, or to just one such band, and may lie in source intensity ranges. For example, see the

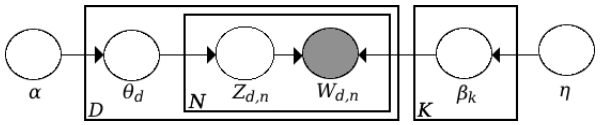

Figure 1: Graphical model for LDA. Nodes are random variables (shaded are observed; unshaded are latent), directed edges show dependence. Boxes denote replication. Image adapted from [2].

large artefacts in Fig. 4.

A second possible approach is to use a human domain expert to identify "valid" background regions, and use this to build a probabilistic model for segmentation. However, such manual intervention is problematic given the volumes of data to be produced by next generation telescopes [12].

We propose a method for producing a model of background without manual intervention, extracted from image data containing both background and sources. This task is nontrivial. Individual pixels in an astronomical image are not spatially independent (source pixels are more likely to be found with other source pixels, and similarly, background pixels are more likely to be found together than with source pixels), but regions of the image may contain only background pixels, only source pixels, or an unknown mixture. This motivates our use of the "mixed-membership model" latent Dirichlet allocation (LDA).

We propose source detection in radio astronomy images via flood-filling based on a probabilistic model of pixel intensities inferred by LDA. We also present an additional application of this technique in segmenting greyscale images.

1.2 Latent Dirichlet Allocation

Given the variation in intensity of background pixels in radio astronomy images, which contain an unknown mixture of source and background pixels, we use the topic model latent Dirichlet allocation to learn distributions of pixel intensity ranges in radio astronomy images.

A "topic model" is a generative model for documents[1] based on latent topics, where topics are modelled as distributions over a vocabulary and documents are modelled as mixtures of topics [21].

Topics are discovered by fitting the generative model to data and finding the best set of latent variables to explain the observed data; for example, the best mixture of topics in a document and distributions of words in a topic [21].

LDA [3] is one such generative probabilistic model for sets of discrete data such as collections of documents, where a document is a mixture of topics, and topics are distributions over the vocabulary of words in the collection.

Each document in a collection of documents is represented as a "bag of words" in LDA. Under the generative model described by LDA, a document is generated by first drawing topic proportions for that document. Given the document's topic proportions, a topic is drawn for each word that will be in the document. The actual word is then generated by drawing it from the distribution corresponding with its assigned topic [1].

[1]A description of a probabilistic procedure for generating documents which is used to form a conditional probability density function and infer the latent topics (rather than actually generate documents).

Figure 2: Image segmentation: each pixel is assigned to the topic it was most likely generated by (for illustrative purposes, each topic is assigned a greyscale value, and each region is coloured according to its topic). Left to right, top to bottom: the original image, followed by the segmented image with two, three, four, five, and six topics. Increasing the number of topics may increase the level of detail revealed by segmentation. Image adapted from [10].

More formally, and with reference to the graphical model in Fig. 1, the generative model assumes there are K topics, each of which is a multinomial distribution over the words in the vocabulary of the document collection. The topic distributions are drawn from a Dirichlet distribution with parameter vector η (a Dirichlet distribution can be informally thought of as a distribution of multinomial distributions). There are D documents in the collection, each with topic proportions θ_d drawn from a Dirichlet distribution with parameter vector α (where $\theta_{d,k}$ is the topic proportion for the k^{th} topic in the d^{th} document). The n^{th} word in the d^{th} document is assigned topic $z_{d,n}$ (drawn from θ_d), with the *observed* word $w_{d,n}$ drawn from the multinomial topic distribution for topic $\beta_{z_{d,n}} \in \{\beta_1 \,..\, \beta_K\}$ [3].

The per-word topic assignments z_i can be inferred via Gibbs sampling[2] [1].

The distribution in Eq. (1) can be iteratively sampled from to infer each latent topic assignment z_i given the observed words w_i in each document d_i in the collection, and all other topic assignments z_{-i}.

$$p(z_i = j | z_{-i}, w_i, d_i) \propto \frac{C^{WT}_{w_{ij}} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \frac{C^{DT}_{d_{ij}} + \alpha}{\sum_{k=1}^{K} C^{DT}_{d_i k} + K\alpha} \quad (1)$$

To perform Gibbs sampling, each word in each document in the collection is randomly assigned a topic, and two count matrices are created: C^{WT} of topic assignments to each word in the vocabulary (with C^{WT}_{wj} the number of times topic j is assigned to word w in the collection), and C^{DT} of topic assignments per document (with $C^{DT}_{d_i k}$ the number of times topic k is assigned in document d_i) [21].

One iteration of Gibbs sampling involves decrementing the matrices at the entry corresponding to each word in the collection in turn, allocating that word a topic from the distribution in Eq. (1), and incrementing the matrices accordingly [21]. Sampling is run until equilibrium is reached.

The first term in Eq. (1) describes the probability of word w_i under topic j (the number of times word w_i is assigned

[2]An algorithm for sampling from a difficult to sample multivariate probability distribution by iteratively generating an instance of each variable conditioned on all others.

Table 1: Astronomical images and their sources.

Image	Description	Source
A1	ATLSB[3]survey region A at 50" resolution	[18, 22]
A2	ATLSB survey region A at 6" resolution	[18, 22]
B1	ATLSB survey region B at 50" resolution	[18, 22]
B2	ATLSB survey region B at 6" resolution	[18, 22]
C	ATLAS CDFS[4]	[13]

topic j as a proportion of the number of times any word is assigned topic j); the second term describes the probability of topic j under the current topic distribution in document d_i (the number of times topic j is found in document d_i as a proportion of all topic assignments in document d_i). The distributions of words per topic, β, and topics per document, θ, can be calculated using the first the second terms respectively. The α and η hyperparameters can be inferred or may be set empirically [21].

In essence, LDA uncovers latent topics in a document collection, where words that are likely to co-occur in documents in the collection are found together with high probability within a particular topic or topics (weighted by their overall representation in the document collection) [21].

As an example, a collection of documents might have a vocabulary of words ["ball", "game", "win", "film", "actor", "scene"]. The first three words in the vocabulary might be found to occur together in documents with high frequency, but rarely with the last three words (and vice versa). Two topics might be extracted accordingly, a "sports" topic (under which the first three words are highly likely and the latter three unlikely) and, similarly, a "movie" topic. A document might be primarily made up of words from one topic, or a mixture of both (for example, a review of a sports movie).

1.3 Application of LDA to images

We make the following analogy between document collections and images: a single greyscale image is a document collection, which comprises d non-overlapping subimages (the documents). The image "vocabulary" is constructed by taking a histogram of pixel intensities in the entire image, where each of w bins (pixel intensity intervals) is a word in the vocabulary. The number of occurrences of a word w_i for document d_j is the count of pixels in subimage d_j that fall into bin w_i of the overall image histogram. Topics are normalised distributions over bins.

Using this model, Gibbs sampling (as described in section 1.2) can be run on greyscale images to uncover latent "topics": distributions of pixel intensities that commonly co-occur in the image, for example a "background topic".

These topics can then be used to segment the image on a pixel-by-pixel or region-by-region basis. This can be done by assigning a pixel/region a topic based on the most likely topic to have generated the pixel/region. This can be calculated using the probability mass function of the multinomial distribution (Eq. (2), where x_i is the count of pixels in the i^{th} bin, $\sum_{i=1}^{k} x_i = n$, and p_i is the probability of the i^{th} bin under a particular topic).

$$\Pr(X_1 = x_1, .., X_k = x_k) = \frac{n!}{x_1!..x_k!} p_1^{x_1}..p_k^{x_k} \qquad (2)$$

[3]Australia Telescope Low Surface Brightness.

[4]Australia Telescope Large Area Survey Chandra Deep Field-South.

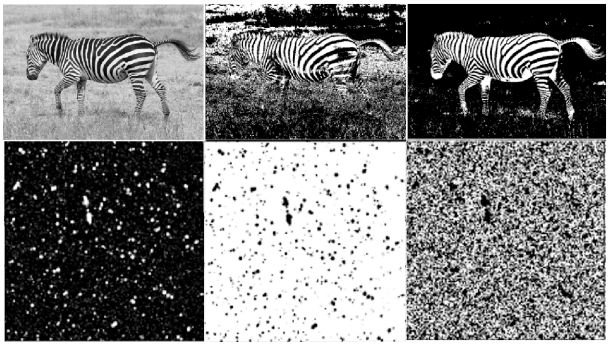

Figure 3: The number of bins in the pixel intensity histogram can affect results. Top row from left to right: a greyscale JPEG image, the segmented image using 10 bins, and 100 bins. Bottom row from left to right: an astronomical image (with contrast adjusted to see sources), the segmented image using 100 bins, and 1000 bins. Image sources: [9, 18, 22].

For source detection in radio astronomy images, flood-filling[5] can then be performed on the segmented image to identify the location and size of sources in the image.

1.3.1 Related work in image segmentation

This application of LDA to source-detection in images differs from the approach taken in [4, 5, 6, 16, 17, 19, 20, 23, 24], in which derivations of LDA and other topic models are applied to image segmentation and object and scene classification tasks. Our approach relies only on pixel intensity and location, whereas previous approaches employ techniques to extract image interest points and pre-segment images before applying LDA.

Our use of LDA-derived probabilities as a precursor to flood-filling is more powerful than many thresholding algorithms (such as those discussed by Gonzalez and Woods in [7]); in LDA, commonly co-occurring bins needn't be adjacent intensity ranges. Consider an image in which the background is made up of medium-intensity pixels while foreground objects comprise both dark and bright pixels. LDA allows the topics to reflect this, with medium-intensity bins found in one topic, and bright and dark bins in the other.

2. METHODS

LDA was performed for segmentation and source detection in several radio astronomy images (Table 1). Non-astronomical greyscale images were segmented as a demonstration of this application of LDA (see Figs. 2 and 3).

Astronomical images were in FITS format [25]; non-astronomical images were greyscale JPEG images.

For each image a histogram of pixel intensities was generated. For astronomical images 100 or 1000 bins were used; for JPEG images, 10 or 100 bins. Each bin (pixel intensity range) is a "word" in the "vocabulary".

The image was decomposed into subimages ("documents"), and counts of pixels in bins were calculated for each. A range of subimage sizes was trialled for each image.

Gibbs sampling was run to infer per-word topic assignments z_i, on the distribution in Eq. (1), as described in

[5]Labelling contiguous regions of pixels that have same topic label as a single region.

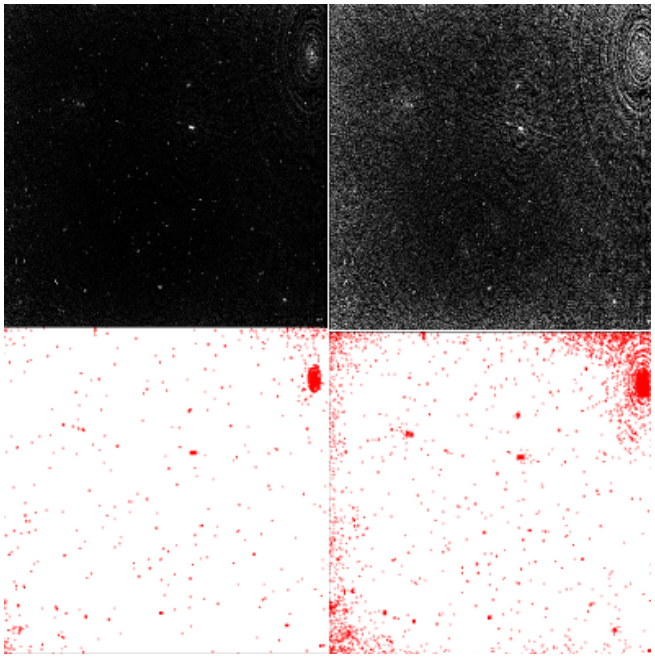

Figure 4: An astronomical image (top left) with contrast adjusted to see artefacts (top right), seen as concentric circles and radial spikes. LDA (bottom left) falsely identified fewer artefact pixels as sources than *Duchamp* (bottom right). Image source: [13].

section 1.2. The α and η vectors were set empirically ($\alpha_i = \alpha_j = 0.1\ \forall i, j$; $\eta_i = \eta_j = 0.01\ \forall i, j$) [21].

Gibbs sampling was run for 100 iterations. The distributions of words for each topic (β_k for $k \in \{\beta_1\ ..\ \beta_K\}$) were calculated from the hundredth sample. An average over samples was not taken as it was found that the sampler converged quickly after which the topic distributions changed very little if at all: sample 1000 was virtually identical to sample 500 and sample 100.

To segment the images using the inferred topic distribution, each pixel in the image was assigned the topic that it was most likely generated by using Eq. (2). When considering a single pixel, this equation simplifies to just p_i for a given bin i. That is, if the pixel being considered falls into bin i in the overall pixel intensity histogram, the topic with the greatest probability for bin i is assigned. This may be weighted by the topic's overall proportion in the collection, however this was not done for the current paper.

The performance of LDA was compared with source catalogues generated via both manual inspection by an astronomer (a "ground-truth" reference) [18, 22] and via the thresholding and region growing astronomical source detection software *Duchamp* [26]. Precision (the proportion of true sources found of all reported sources: $precision = \frac{tp}{tp+fp}$) and recall (the proportion of sources found out of all sources in the image: $recall = \frac{tp}{tp+fn}$) were calculated. (Where tp = "true positive", fp = "false positive", fn = "false negative" [15]).

3. RESULTS

The results of both LDA and *Duchamp* were compared to a source catalogue generated by manual inspection by an astronomer. Although LDA and *Duchamp* perform roughly equivalently with respect to spatially extended, multi-component sources, LDA had more false positives false negatives than *Duchamp* (see Fig. 5 for an example).

Table 2: Performance of LDA vs *Duchamp*

Image	LDA		*Duchamp*	
	Precision	Recall	Precision	Recall
A1	0.83	0.93	0.98	1.0
A2	0.98	0.99	1.0	1.0
B1	1.0	0.99	1.0	0.96
B2	1.0	0.89	1.0	1.0

Table 2 shows the performance of LDA and *Duchamp* as compared to source catalogues generated via manual inspection by an astronomer on sources for which the total flux (intensity) is less than 1.63 mJy[6]. LDA performed similarly to *Duchamp*.

LDA sometimes reported a single source where *Duchamp* correctly separated several; however this is due to the post-processing flood-filling, rather than the LDA algorithm itself. In other cases LDA correctly identified sources that *Duchamp* mistakenly merged.

Bright peak pixels seem key to detection by LDA. For example, in image A2 LDA detects several sources below 1.63 mJy, all of which have peak pixels at least 6σ above the rms noise[7]; in contrast LDA's false negatives above 1.63 mJy all have peak pixels less than 6σ above rms noise.

LDA identified fewer artefact pixels as sources than *Duchamp* in the artefact polluted Image C (Fig. 4). This is a clear demonstration of the strength of using a probabilistic model of background. To avoid the effects of such artefacts using *Duchamp* or similar software, an astronomer would need to manually decompose the image into a number of smaller regions and manually adjust region thresholds. Our implementation of LDA avoids such manual interventions.

4. DISCUSSION

LDA performed similarly to the standard source-detection software *Duchamp* [26] on a representative sample of radio astronomy images, particularly for sources with integrated source flux 2.5σ above the rms noise.

The two algorithms performed similarly with respect to extended, multi-component sources, but LDA had more false positive detections and non-detections than *Duchamp*.

Bright peak pixels seem to be essential for a source to be detected by the current implementation of LDA. The current implementation of LDA is therefore unlikely to detect any diffuse sources (spatially extended sources with low brightness overall and no bright peak pixels).

LDA outperformed *Duchamp* on the image polluted with artefacts — an image that would require labourious manual intervention by an astronomer to detect sources using software such as *Duchamp*. This is a clear demonstration of the utility of the probabilistic model employed.

4.1 Parameter and computational issues

In document collections there is a natural segregation of words and documents; LDA was developed for such discrete data [3]. However, there is no natural segregation of pixels

[6] $1Jy \equiv 1 \times 10^{-26} W/Hz/M^2$

[7] Noise was manually determined by an astronomer.

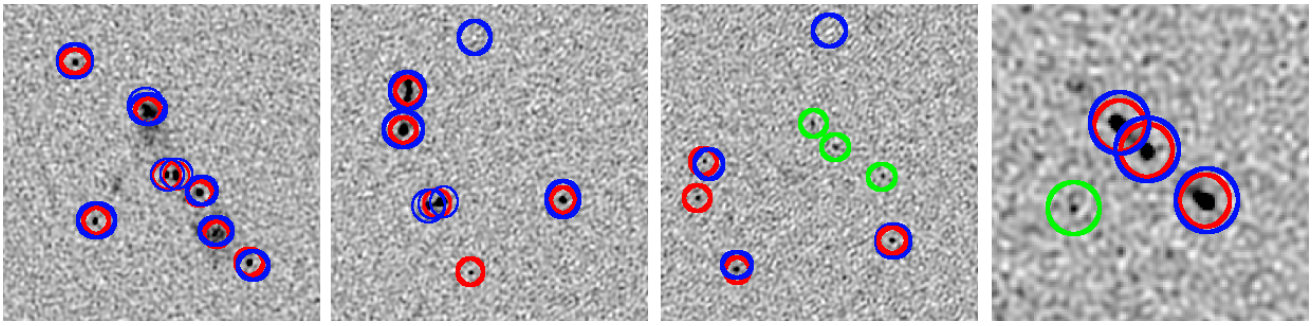

Figure 5: Performance of LDA (blue) and *Duchamp* [26] (red) on four representative regions from Image A2. From left: 1. A region containing a radio galaxy with two large jets (and no detected core) seen in projection with other point sources. Both algorithms identify multiple components. 2. A false positive and a false negative for LDA. 3. A false positive and a false negative for LDA; *Duchamp* detects three sources less than 2.5σ above the rms, missed by LDA (green). 4. A radio galaxy with three components, and a point source. Both algorithms split the radio galaxy into three components. Only *Duchamp* detects the point source less than 2.5σ above the rms (green). Image source: [18, 22].

into intensity ranges or images into regions. In the current approach we used histogram binning and decomposition of the image into subimages. This introduces new parameters to be set. In practice, it was found that the final topics extracted did not vary over a wide range of subimage sizes chosen, however, results did vary based on the number of bins in the histogram (see Fig. 3). The approach taken to histogram binning in the current paper is likely to be responsible for our implementation of LDA's poor performance on faint sources in radio astronomy images; this should be addressed in future implementations.

The number of topics must be set manually. For source detection two may be sufficient; however for image segmentation in general a different number of topics may give different results. Fig. 2 shows a grayscale image segmented with two to six topics. With two topics, the object in the image is clearly segmented from the background; increasing the number of topics reveals more details of the image; in general terms increasing the number of topics might be expected to increase the level of detail shown, but may introduce irrelevant detail, for example the segmentation of the sky in Fig. 2.

LDA is a "bag of words" model and so ignores the natural ordering of pixel intensities. However this may be a benefit, rather than a drawback, as this allows objects made up of non-neighbouring pixel intensity ranges to be correctly segmented from images.

Gibbs sampling to infer the latent topics in LDA is expensive both in terms of computation and time. One Gibbs sample involves iterating through each pixel in the image, allocating each a topic based on the current distributions of words to topics and topics to documents, and so is linear in the number of pixels multiplied by the number of topics. As it is not unusual for astronomical images to be 8000×8000 pixels, this can be computationally difficult. Additionally, at least one large three dimensional array (indexing words by documents by topic) must be kept in memory. However, LDA need not necessarily be run for every image, nor on the whole image. LDA could be run on a small representative section of one image in a collection of similar images in order to extract topics for the whole collection. This would reduce the computational expense of the approach.

4.2 Future work

The approach described in this paper shows how LDA can be used for image segmentation and source detection.

The use of the final topic distributions — segmentation by assigning each pixel a hard topic label and source detection by flood-filling on the segmented image — is crude. A more nuanced approach would eliminate this hard assignment and take a more probabilistic approach to region labelling.

Given the reliance on bright peak pixels for source detection by LDA, more work needs to be done to improve LDA's performance on faint sources. In the particular cases analysed, the addition of more bins in the low range of pixel intensities would likely improve performance; in the general case, the optimal use of bins should be investigated.

Performance of LDA in segmenting non-astronomical greyscale images could be assessed by comparing the obtained segmentation with human segmentation of the same images, using a large public database of images [10]. This would also allow comparison to the results obtained by other algorithms.

5. CONCLUSIONS

The current paper presents a preliminary investigation into use of the topic model latent Dirichlet allocation for image segmentation in greyscale images and source detection in astronomical images. Our method builds a probabilistic model of "non-source" pixel distributions.

LDA performed similarly to the standard source-detection software *Duchamp* [26] on a representative sample of radio astronomy images, however, for fainter sources and in particular diffuse sources, there is still some work to be done to explore if the LDA method will be an improvement over existing algorithms.

A particular success of the approach is the superior result obtained in Image C, which is polluted with artefacts, as compared to the relatively poor performance of *Duchamp*.

The algorithm could be refined to take a more probabilistic approach to region labelling rather than the hard assignment described in the current paper, along with further exploration of the optimal pixel binning strategy.

6. ACKNOWLEDGMENTS

We gratefully acknowledge the financial assistance provided by a KAREN Capability Build Fund grant in support of radio astronomy.

7. REFERENCES

[1] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] D. Blei and J. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text mining: classification, clustering, and applications*, volume 10, pages 71–94. Chapman & Hall/CRC, 2009.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[5] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pages 524–531. IEEE, 2005.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pages 1816–1823. IEEE, 2005.

[7] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, New Jersey, second edition, 2002.

[8] C. Hollitt and M. Johnston-Hollitt. Feature detection in radio astronomy using the circle Hough transform. *Publications of the Astronomical Society of Australia*, 29(3):309–317, 2012.

[9] M. M. Karim. Zebra_running_ngorongoro.jpg. http://en.wikipedia.org/wiki/File:Zebra_running_Ngorongoro.jpg.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, pages 416–423. IEEE, 2001.

[11] M. Masias, J. Freixenet, X. Lladó, and M. Peracaula. A review of source detection approaches in astronomical images. *Monthly Notices of the Royal Astronomical Society*, 2012.

[12] R. Norris, A. Hopkins, J. Afonso, S. Brown, J. Condon, et al. EMU: Evolutionary map of the universe. *Publications of the Astronomical Society of Australia*, 28(3):215–248, 2011.

[13] R. P. Norris, J. Afonso, and P. N. Appleton. Deep ATLAS radio observations of the Chandra Deep Field-South/Spitzer Wide-area Infrared Extragalactic Field. *The Astronomical Journal*, 132(6):2409–2423, 2006.

[14] R. P. Norris, J. Afonso, D. Bacon, R. Beck, M. Bell, et al. Radio continuum surveys with square kilometre array pathfinders. *(In Press)*, 2012.

[15] D. Olson and D. Delen. *Advanced data mining techniques*. Springer Verlag, 2008.

[16] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pages 883–890. IEEE, 2005.

[17] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pages 1605–1614. IEEE, 2006.

[18] L. Saripalli, R. Subrahmanyan, K. Thorat, R. Ekers, R. Hunstead, H. Johnston, and E. Sadler. ATLBS extended source sample: The evolution in radio source morphology with flux density. *The Astrophysical Journal Supplement Series*, 199(27), 2012.

[19] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. Technical Report AIM-2005-005, Massachusetts Institute of Technology, 2005.

[20] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pages 370–377. IEEE, 2005.

[21] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of latent semantic analysis*, pages 424–440. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA, 2007.

[22] R. Subrahmanyan, R. Ekers, L. Saripalli, and E. Sadler. ATLBS: the Australia telescope low-brightness survey. *Monthly Notices of the Royal Astronomical Society*, 402(4):2792–2806, 2010.

[23] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pages 1331–1338. IEEE, 2005.

[24] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 1577–1584. MIT Press, Cambridge, MA, 2008.

[25] D. Wells, E. Greisen, and R. Harten. FITS — a flexible image transport system. *Astronomy and Astrophysics Supplement Series*, 44:363, 1981.

[26] M. Whiting. Duchamp: a 3D source finder for spectral-line data. *Monthly Notices of the Royal Astronomical Society*, 421(4):3242–3256, 2012.