METRIC LEARNING FOR HYPERSPECTRAL IMAGE SEGMENTATION

Brian D. Bue[1], *David R. Thompson*[2], *Martha S. Gilmore*[3], *Rebecca Castaño*[2]

[1] Rice University, Electrical and Computer Engineering, 6100 Main St., Houston, TX, 77006

[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109

[3] Wesleyan University, Earth and Environmental Sciences, Weslayan Station, Middletown, CT 06459

ABSTRACT

We present a metric learning approach to improve the performance of unsupervised hyperspectral image segmentation. Unsupervised spatial segmentation can assist both user visualization and automatic recognition of surface features. Analysts can use spatially-continuous segments to decrease noise levels and/or localize feature boundaries. However, existing segmentation methods use task-agnostic measures of similarity. Here we learn task-specific similarity measures from training data, improving segment fidelity to classes of interest. Multiclass Linear Discriminant Analysis produces a linear transform that optimally separates a labeled set of training classes. This defines a distance metric that generalizes to new scenes, enabling graph-based segmentations that emphasizes key spectral features. We describe tests based on data from the Compact Reconnaissance Imaging Spectrometer (CRISM) in which learned metrics improve segment homogeneity with respect to mineralogical classes.

Index Terms— Segmentation, Metric Learning, CRISM

1. HYPERSPECTRAL IMAGE SEGMENTATION

Unsupervised hyperspectral image segmentations can reveal spatial trends that show the physical structure of the scene to an analyst. They highlight borders and reveal areas of homogeneity and change. Segmentations are independently helpful for object recognition, and assist with automated production of symbolic maps. Additionally, a good segmentation can dramatically reduce the number of effective spectra in an image, enabling analyses that would otherwise be computationally prohibitive. In particular, using an oversegmentation of the image instead of individual pixels can reduce noise and potentially improve the results of statistical post-analysis.

Recent work in hyperspectral image segmentation include the watershed transform [1], Markov Random Fields [2], and the Felzenszwalb graph segmentation algorithm [3]. Generally speaking, these techniques cluster pixels based on spatial proximity and a measure of spectral similarity. Existing hyperspectral segmentation approaches generally use task-agnostic distance measures that treat all channels equally or weight them based on global statistical properties of the dataset. Such metrics are often confused by noise, instrument artifacts, or spectral variations that are irrelevant to semantic categories of interest. Learning a task-specific similarity metric from labeled data can ameliorate this problem. Methods to learn such metrics include Information Theoretic Metric Learning (ITML) [4], Neighbourhood Components Analysis (NCA) [5], and variants of Generalized Relevance Learning Vector Quantization (GRLVQ) [6].

This work aims to improve unsupervised segmentations by learning a task-relevant measure of spectral similarity from expert-labeled training data. We employ a multiclass Linear Discriminant Analysis (LDA) based approach to learn this measure. Learned measures produce segmentations that are not only are more visually cohesive, but also quantitatively more accurate in separating known materials into disjoint segments, in comparison to segmentations produced using unweighted metrics. We evaluate this technique by comparing a set of expert-labeled mineral class maps to the segmentation maps produced by learned metrics, and provide a results on a case study focusing on several well-analyzed CRISM images [7]

2. METRIC LEARNING FOR HYPERSPECTRAL IMAGE SEGMENTATION

We use the Felzenszwalb segmentation algorithm for its simplicity and computational efficiency [3, 8]. This is an agglomerative clustering approach that joins pixels into groups based on a pairwise distance $\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)$ between adjacent pixels $\mathbf{x}_i$ and $\mathbf{x}_j$. The segmentation algorithm represents the image as an 8-connected grid of nodes; each node corresponds to a single pixel. All pixels are initially treated as separate segments and iteratively joined into larger groups. We weight edges between the nodes according to $\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)$; previous studies have used spectral angle distance and Euclidean (Euc) distance. The maximum internal edge weight of a segment S, $\mathrm{Int}(S)$, is defined as the largest edge weight in its minimum spanning tree, $\mathrm{MST}(S)$.

$$\mathrm{Int}(S) = \max_{\mathbf{x}_i, \mathbf{x}_j} \mathrm{d}(\mathbf{x}_i, \mathbf{x}_j) \; \forall \, \mathbf{x}_i \in S, \; \mathbf{x}_j \in S, \; (\mathbf{x}_i, \mathbf{x}_j) \in \mathrm{MST}(S)$$

The smallest edge weight that joins two neighboring segments S_a and S_b (i.e. the most similar pixel pair on their border) defines the cross-segment distance:

$$\mathrm{Dif}(S_a, S_b) = \min_{\mathbf{x}_i, \mathbf{x}_j} \mathrm{d}(\mathbf{x}_i, \mathbf{x}_j) \; \forall \, \mathbf{x}_i \in S_a, \mathbf{x}_j \in S_b, \; (\mathbf{x}_i, \mathbf{x}_j) \in E$$

Two adjacent segments are merged when the cross-segment distance is larger than the minimum of both internal weights, weighted by a constant k and inversely proportional to a segment's area $|S|$.

$$\mathrm{MInt}(S_a, S_b) = \min \left(\mathrm{Int}(S_a) + \frac{k}{|S_a|}, \mathrm{Int}(S_b) + \frac{k}{|S_b|} \right) \qquad (1)$$

Larger k values cause a preference for larger segments, but is not a minimum segment size – smaller segments are allowed when there is a sufficiently large difference between spatially neighboring segments. However, in some cases, a minimum segment size is desirable, so as a final step, we merge small segments below a user-defined threshold ≥ 1 with their spectrally-closest neighbors.

We augment the segmentation algorithm with a task-specific Mahalanobis distance metric learned from training data. The (squared) Mahalanobis distance between samples $\{\mathbf{x}_i, \mathbf{x}_j\} \in \mathbb{R}^D$ is: $\mathrm{d}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{M} = \mathbf{A}^{\mathbf{T}} \mathbf{A}$ is a

$D \times D$ linear transformation matrix. We seek to learn the matrix $\mathbf{A}$ which best separates a set of samples belonging to C classes. Our approach employs multiclass linear discriminant analysis (LDA) to maximize the ratio of between-class vs. within-class separation S:

$$S = (\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_b \boldsymbol{\alpha})(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_w \boldsymbol{\alpha})^{-1} \quad (2)$$

Here, $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_w$ are the between and within class scatter matrices, respectively. By selecting the top $C-1$ eigenvectors of ${\boldsymbol{\Sigma}_w}^{-1}\boldsymbol{\Sigma}_b$, we define a projection into a $C-1$ dimensional subspace that captures variability between features with respect to training data [9]. To prevent Equation 2 from becoming ill-posed due to an insufficient number of training samples, we regularize $\boldsymbol{\Sigma}_w$ by a parameter $\gamma_{\text{LDA}} \in [0, 1]$ (selected via cross-validation) according to: $\boldsymbol{\Sigma}_w = (1-\gamma)\boldsymbol{\Sigma}_w + \gamma \mathrm{I}$, where $\mathbf{I}$ is the identity matrix.

We also learn a Mahalanobis distance using Information Theoretic Metric Learning [4]. ITML calculates the matrix $\mathbf{M}$ by maximizing the relative entropy between a multivariate Gaussian parametrized by a set of training samples, and another multivariate Gaussian belonging to a known, well-behaved Mahalanobis distance function. This maximization is constrained such that similar classes remain nearby one another and dissimilar classes remain far apart in the space defined by the learned metric. The ITML algorithm takes a parameter γ_{ITML} which controls the tradeoff between satisfying similarity/dissimilarity constraints and maximizing the divergence between the Gaussians. We use the code provided by the authors [10] to learn the metric and select γ_{ITML}.

3. EVALUATING SEGMENTATION RESULTS WITH RESPECT TO CLASS KNOWLEDGE

We attempt a *superpixel* segmentation in which the image is conservatively oversegmented; that is, we accept that single surface features may be split into multiple segments, but try to ensure that each individual segment - or superpixel - has homogeneous mineralogy [8]. We compare superpixels produced using each metric to a set of expert-labeled classes defined by a planetary geologist. The geologist identified the primary constituents in each of the images we study, along with the image pixels containing the purest examples of each mineral, and defined class maps for the materials using the ENVI spectral angle mapper (SAM) function [11]. As a final step, the geologist examines the spectral angles for each class and define thresholds to filter out ambiguous or mixed materials. We exclude these pixels from the following performance evaluation.

Because we seek an *oversegmentation* of an image, each expert-labeled class will likely be split into multiple segments. However, when we use a learned metric to segment each image, we expect the resulting segments to be better separated with respect to the training classes – i.e., pixels in each segment are more likely to belong to a single training class, rather than multiple classes – in comparison to metrics which do not account for class relationships. We define two measures to quantify the degree to which the resulting segments partition distinct mineralogical classes. The first measure is the conditional entropy of the class map given the segmentation map, H(class|seg). H(class|seg) quantifies the remaining uncertainty for a random variable – in our case, the distribution of material classes – given the value of another random variable – the partitions produced by segmentation algorithm. In the case of a perfect segmentation of the classes, H(class|seg) will be zero, as the segmentation perfectly reconstructs the class map. Thus, we prefer smaller values of H(class|seg). Our second measure of segmentation quality, the "impurity ratio," is the ratio of "impure" vs. "pure" segments with respect to the class map. A "pure" segment consists of pixels belonging to a single class, whereas an "impure" segment consists of pixels belonging to multiple classes. Because segment size can bias this score, we scale the impurity ratio for each segment by its size in pixels. As with H(class|seg), smaller impurity ratios are better.

We evaluate the quality of segmentations produced by each metric learning algorithm by segmenting spatially contiguous halves of each image. We sample 100 spectra from each class from the first half of the image (subsequently referred to as the "train" image), and use these points to train each metric learning algorithm. We then segment the train image and the remaining half of the image (the "test" image), using the metrics produced by LDA, ITML and the (baseline) Euclidean distance. Both the distance metric and the internal bias k (Equation 1) alter the size – and subsequently the quantity – of the resulting superpixels. To objectively compare results between several metrics, we must compare segmentations that produce a similar number of superpixels. To achieve this balance, we segment each image using a range of k values in $[10^{-3}, 10^{1}]$ and provide overall statistics for segmentations produced by each metric on that range. We chose this range because the number of superpixels produced by each metric followed a similar trend for all of the images we studied. We focus on segmentations that produce 200-1250 superpixels, as segmentations with few superpixels tend to inadequately capture morphological characteristics of the imagery we study, while segmentations with large quantities of superpixels are more sensitive to noise and insignificant differences in spectra. We ignore superpixels consisting of less than 50 pixels, as they tend to be unstable and noisy with respect to the training classes. Ignoring these small superpixels is done for evaluation purposes only, as it allows for a more consistent evaluation of the resulting segmentation maps.

4. CASE STUDY: CRISM IMAGERY

We examine three well-studied CRISM scenes: 3e12, 3fb9, and 863e (omitting the frt0000 catalog prefix). We use the Brown CRISM Analysis Toolkit [12] to perform radiometric correction and atmospheric calibration, and remove noisy bands in the extreme short and long wavelengths, leaving a total of 231 bands in the in the [1.06, 2.58] μm range for analysis. Our final preprocessing step is to normalize each spectrum by its Euclidean norm, to compensate for linear illumination effects [13]. Figure 1 shows the normalized mean spectra of the most pure, expert-labeled material samples for the classes in each image we consider. See [8] for further details regarding these images and their constituent material classes. Figures 2 and 3 give the H(class|seg) and impurity ratios, vs. the number of segments using each metric. LDA outperforms both the Euclidean metric and ITML, sometimes dramatically (e.g. on images 863e and 3fb9). The Euclidean metric performs worst, which is not surprising since it is more susceptible to noise that a learned metric will often suppress. ITML yields similar performance to the Euclidean distance for training images 3e12 and 3fb9, which is likely because the quantity of training samples is small for these two images – which consist of two and three material classes, respectively. On image 863e, with training consisting of 5 material classes, ITML approaches the performances of LDA. This is also reflected in the summary statistics per-image for each segmentation given in Table 1. Note that the performance improvements on testing data over training data on the 863e image are due to the fact that the test image contains a smaller number of Kaolinite (670) and Montmorillionite (93) pixels than in the training image, which are easily confused with other training classes (e.g., Kaolinite vs. FeMg Smectice). Figure 4 shows a set of resulting segmentation maps for which the Euclidean and LDA/ITML-learned metrics produced a comparable number of segments. Visually, the LDA-based segmentation produces segments that better match the underlying morphology of the

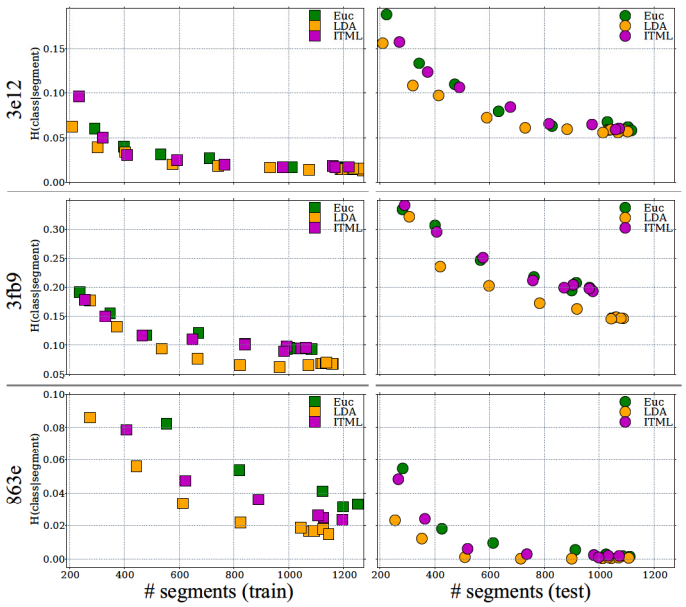


Fig. 2. H(class|seg) values for Euc (green), LDA (yellow) and ITML (magenta) segmentations vs. number of segments on training (left) and testing (right) images.

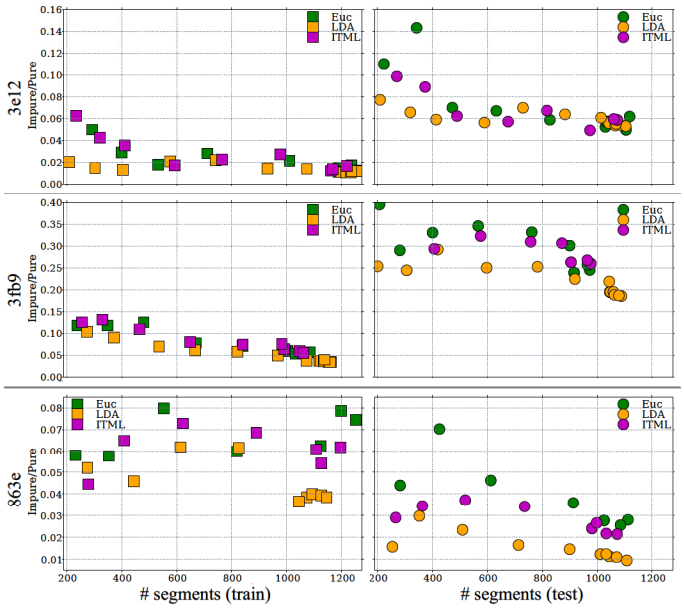


Fig. 3. Impurity ratios for Euc (green), LDA (yellow) and ITML (magenta) segmentations vs. number of segments on training (left) and testing (right) images.

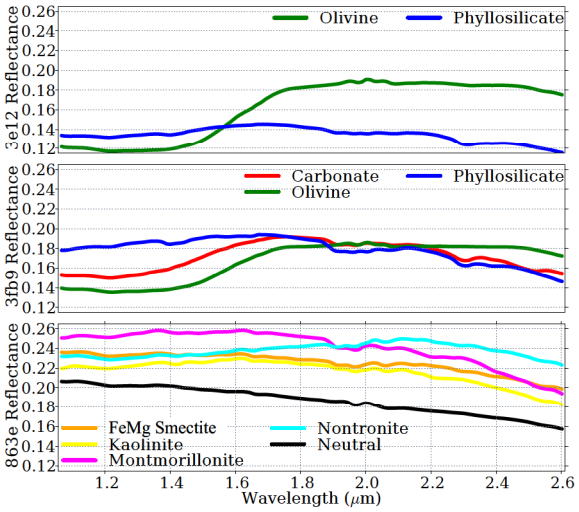


Fig. 1. Normalized mean spectra of samples from most pure material classes in images 3e12, 3fb9 and 863e. The "neutral" class in image 863e is a mostly featureless, dark material which is spectrally dissimilar from each of the other material species. Due to varying atmospheric and illumination conditions at capture time, and differences caused by atmospheric calibration, spectra belonging to the same material species may not have identical spectral representations in different images - e.g., the olivine spectra in image 3e12 vs. those in 3fb9.

image data. The Euclidean-based segmentation, and to a lesser degree, the ITML-based segmentation, both suffer from column striping artifacts as noisy bands are not well compensated for using these metrics. This is also reflected in the per-class purity percentages given in Table 2. Both learned metrics outperform the baseline, with LDA improving over the Euclidean metric for material classes FeMg Smectite, Montmorillonite and Nontronite. ITML gives comparable performance to LDA for most materials, but the gains are not as significant for the Montmorillionite and Nontronite classes.

H(class|seg)

Image	Euc	LDA	ITML
3e12	0.0169 / 0.0676	0.0148 / 0.0588	0.0191 / 0.0655
3fb9	0.0884 / 0.378	0.0497 / 0.242	0.0972 / 0.354
863e	0.0473 / 0.00403	0.0184 / 0.000584	0.031 / 0.00228

Impurity/Purity

Image	Euc	LDA	ITML
3e12	0.018 / 0.0619	0.0116 / 0.0573	0.02 / 0.0596
3fb9	0.0661 / 0.296	0.0368 / 0.195	0.0745 / 0.294
863e	0.0684 / 0.032	0.0398 / 0.0124	0.0611 / 0.0266

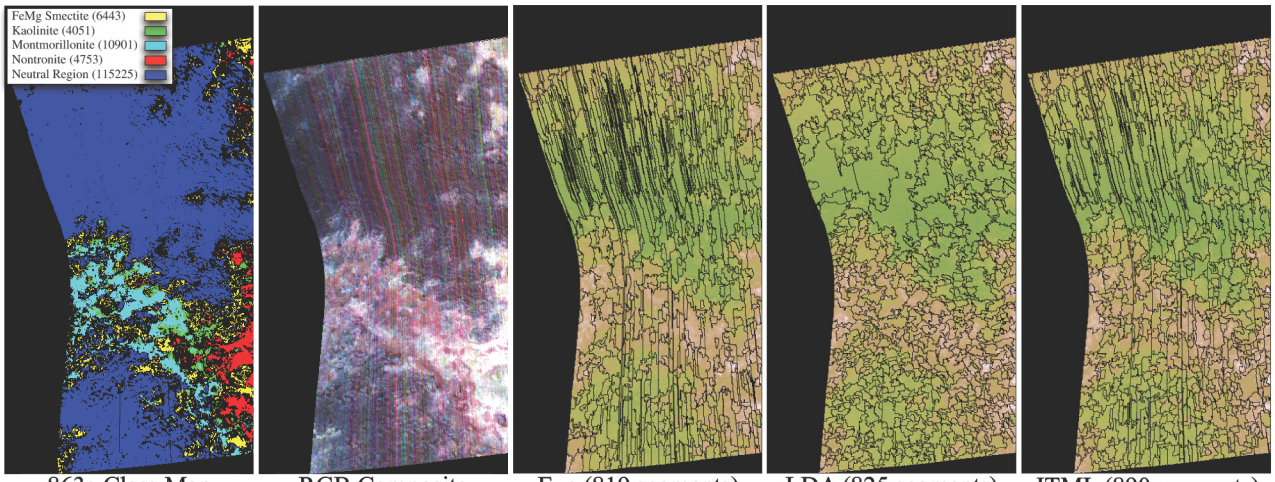
Table 1. Average H(class|seg) and impurity ratios for each image and similarity metric. Green and red fonts indicate the best and worst performing metrics, respectively.

Class (# pixels)	Euc	LDA	ITML
FeMg Smectite (6443)	26	49	48
Kaolinite (4051)	98	99	99
Montmorillonite (10901)	11	31	17
Nontronite (4753)	37	52	40
Neutral Region (115225)	97	99	98
Average	53	66	60

Table 2. Average pure pixels / segment for Euclidean, LDA and ITML-based segmentations of image 863e (Figure 4). Best and worst average per-class accuracy given in green and red font, respectively.

5. DISCUSSION AND FUTURE WORK

The superior performance of LDA over ITML on all three of our images is somewhat surprising, considering the simplicity of the



863e Class Map | RGB Composite | Euc (819 segments) | LDA (825 segments) | ITML (890 segments)

Fig. 4. Expert-labeled SAM class map for image 863e (left), RGB composite image (2nd from left), and the segmentation maps produced using Euclidean distance (center), LDA (2nd from right), and ITML (right). We overlay each segmentation map on a single color-stretched band (wavelength 1.25μm). The LDA-based segmentation is less susceptible to column striping artifacts, and better characterizes image morphology.

LDA projection in comparison to the more theoretically elegant optimization performed by ITML. An issue with ITML (as observed by Parameswaran et al. in [14]) is that the (global) metric is not optimized locally, which can cause problems with overfitting to multimodal data distributions. Regularized LDA does not suffer (at least, to the same degree) from such overfitting issues. Also, it may be necessary to use more samples per class to learn the metric using ITML. We expect additional training samples or alternative regularization schemes will likely yield improved results using ITML.

One avenue we are exploring is learning class structure across multiple, related images. We have developed an technique, Multi-Domain/Multi-Class LDA (MDMC-LDA) and a corresponding regularization scheme which allows LDA to exploit class structure local to individual images while simultaneously capturing class relationships common to other images with similar classes [15].

Acknowledgements: We thank Brown University and the CRISM team for the use of their CAT software package. Erzsébet Merenyi and Lukas Mandrake provided valuable advice and support. A portion of the work described in this manuscript was carried out at the Jet Propulsion Laboratory with support from the NASA AMMOS Multimission Ground Systems and Services office. Copyright 2011 California Institute of Technology. All Rights Reserved. U.S. government support acknowledged.

6. REFERENCES

[1] Y. Tarabalka, J. Chanussot, and J.A. Benediktsson, "Segmentation and classification of hyperspectral data using watershed transformation," *Pattern Recognition*, vol. 43, no. 7, pp. 2367–2379, 2010.

[2] A. Mohammadpour, O. Féron, and A. Mohammad-Djafari, "Bayesian segmentation of hyperspectral images," *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 735, pp. 541–548, 2004.

[3] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *Intl. J. Computer Vision*, vol. 59:2, September 2004.

[4] J Davis, B Kulis, P Jain, S Sra, and I Dhillon, "Information-theoretic metric learning," *Proceedings of the 24th international conference on Machine learning*, Jan 2007.

[5] J Goldberger, S Roweis, G Hinton, and R Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, Jan 2005.

[6] M.J Mendenhall and E Merényi, "Relevance-based feature extraction for hyperspectral images," *Neural Networks, IEEE Transactions on*, vol. 19, no. 4, pp. 658–672, 2008.

[7] S. Murchie et al., "CRISM (Compact Reconnaissance Imaging Spectrometer for Mars) on MRO (Mars Reconnaissance Orbiter)," *J. Geophys. Res*, vol. 112, no. E05, 2007.

[8] D. R. Thompson, L. Mandrake, M.S. Gilmore, and R. Castaño, "Superpixel Endmember Detection," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 11, pp. 4023–4033, 2010.

[9] R.A Fisher, "The statistical utilization of multiple measurements.," *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[10] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon, *Information Theoretic Metric Learning*, UT, Austin, http://www.cs.utexas.edu/users/pjain/itml/.

[11] Research Systems Inc, *ENVI 4.6 Users Guide*, 2008, 1196 pp.

[12] F. Morgan, F. Seelos, and S. Murchie, "Cat tutorial," in *CRISM Data Users Workshop, Lunar Planetary Sci. Conf.*, 2009.

[13] GW Pouch and DJ Campagna, "Hyperspherical direction cosine transformation for separation of spectral and illumination information in digital scanner data," *Photogrammetric Engineering and Remote Sensing*, vol. 56, no. 4, pp. 475–479, 1990.

[14] S Parameswaran..., "Large margin multi-task metric learning," *Proceedings of NIPS 2010*, 2010.

[15] David S. Hayden, Steve Chien, David R. Thompson, and Rebecca Castaño, "Using onboard clustering to summarize remotely sensed imagery," *IEEE Intelligent Systems*, vol. 25, pp. 86–91, 2010.