CONVEX FORMULATION FOR HYPERSPECTRAL IMAGE CLASSIFICATION WITH SUPERPIXELS

Yi Liu[a,b], Filipe Condessa[b,c], José Bioucas-Dias[b], Jun Li[d], Antonio Plaza[a]

[a]Dept. of Technol. of Comput. & Commun., Univ. of Extremadura, Cáceres, Spain
(e-mail: {yiliu, aplaza}@unex.es)
[b]Instituto de Telecomunicações, Instituto Superior Técnico Universidade de Lisboa, Lisboa, Portugal
(e-mail: {bioucas,filipe.condessa}@lx.it.pt)
[c]Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA
[d]School of Geography and Planning, Sun Yat-sen University, Guangzhou, China
(e-mail: lijun48@mail.sysu.edu.cn).

ABSTRACT

The superpixels provided by an unsupervised segmentation algorithm are sets of neighboring pixels homogeneous in some sense. Therefore it is very likely that, in a classification problem, most pixels in a superpixel belong to the same class, namely if the homogeneity criterion is compatible with the class statistics. Superpixels are, therefore, a powerful device to express spatial contextual information. However, the exploitation of superpixels in a principled way is not straightforward. Recent efforts attack this problem under a discrete optimization framework, by including regularization terms promoting consistence of the labels in the superpixels and computing approximate labelings with graph-cut algorithms. The well known hardness of integer optimization problems is a major limitation of this line of attack. In this paper, we introduce a new strategy, based on convex relaxation, to include the spatial information provided by superpixels in classification problems. The convex relaxation of an integer optimization problem opens a door to include extra information, such as spatial partitioning information given by over-segmented superpixels. The convex optimization problem thus obtained is solved by using SALSA algorithm. Experimental results with the ROSIS Pavia University dataset illustrate the effectiveness of the proposed framework.

***Index Terms*—** Hyperspectral image classification, superpixel, convex relaxation, SALSA, graph-cut algorithms, integer optimization

1. INTRODUCTION

In hyperspectral image classification, as in many other fields, spatial information holds the potential to improve the classification performance with respect to statistical accuracy, pixel consistency, object contour detection, specific target recognition, and so forth [1, 2, 3, 4, 5]. Spatial partitions, especially over-segmented superpixels, may be used to promote label consistence within superpixels, considering the usual huge correlation among the pixels within a given superpixel. This consistence is illustrated in Fig.1, where the boundaries between the scene objects are clearly preserved in the three displayed segmentations.

Fig. 1. Example of superpixels with multiple-scaled sizes.

1.1. Related work

Kohli et al. proposed in [3] a robust Potts model, which combines multiple segmentations in a principled manner based on higher order conditional random fields. In the hyperspectral field, superpixels from watershed transformation were used in [4] to bind the spatially clustered pixels with the same label, and a multiple spectral-spectral classification approach was designed to improve the classification performance in [6]. Recently, Zhang et al. [7] successfully introduced a superpixel based graphical model with high order potential for remote sensing image mapping.

Classification with spatial information is fundamentally an integer optimization problem and this view has led to the-state-of-the-art algorithms using optimization via graph-cuts. However, in spite of the great success of the integer formulations, the usual NP-hardness of integer optimization renders little flexibility with respect to the inclusion of the spatial priors, namely, those coming from over-segmentations (superpixels).

1.2. Contributions

Based on the convex relaxations of an integer classification optimization, and the spirit of [8, 9, 10], this paper introduces a segmentation methodology which is extremely flexible with

Thanks to (China Scholarship Council) CSC agency for funding. This work was partially supported by the Fundação para a Ciência e Tecnologia, Portuguese Ministry of Science and Higher Education, project UID/EEA/50008/2013.

respect to the inclusion of spatial prior information from superpixels, as well as other convex regularizers. The contributions are the following: 1) Expressing the information from over-segmented superpixels in the form of a convex regularizers. 2) An optimization algorithm to solve the convex relaxation using SALSA [11]. 3) Experimental evidence of the potential of the proposed methodology.

This work has strong connections with Bioucas-Dias et al. [8] and the work of the work of Condessa et al. [9]. There is, however, a major difference. Where the methodology presented in [8, 9] computes probabilities of labelings, to be then used elsewhere, namely in statistical inference problems, our objective, more in line with [10, 12], is to use convex relaxation to approximate the original discrete problem.

The paper is organized as follows: Section 2 provides the theoretical fundamentals of the presented methodology and describes the SALSA technique to solve the optimization problems. Section 3 presents experimental results, and Section 4 concludes the paper.

2. METHODOLOGIES

Let $\mathcal{S} \equiv \{1, 2, \cdots, n\}$ be a set of integers indexing the n pixels of an image, $\mathbf{x} \equiv [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ be a feature matrix holding d-dimensional image feature vectors. Let $\mathbf{y} \equiv (y_1, \cdots, y_n) \in \mathcal{L}^n$ be an image of class labels, termed segmentation or labeling, such that $y_i = k$ if and only if the label of pixel i is k.

Given the posterior probability $p_{\mathbb{Y}\|\mathbb{X}}(\mathbf{y}|\mathbf{x})$, the observation model $p_{\mathbb{X}\|\mathbb{Y}}(\mathbf{x}|\mathbf{y})$, and the prior probability $p_{\mathbb{Y}}(\mathbf{y})$, often an MRF, the maximum a posteriori (MAP) segmentation, or labeling, is given by

$$\begin{aligned}\hat{\mathbf{y}}_{MAP} &\in \arg\max_{\mathbf{y}\in\mathcal{L}^n} p_{\mathbb{Y}|\mathbb{X}}(\mathbf{y}|\mathbf{x}) \\ &= \arg\max_{\mathbf{y}\in\mathcal{L}^n} p_{\mathbb{X}|\mathbb{Y}}(\mathbf{x}|\mathbf{y}) p_{\mathbb{Y}}(\mathbf{y}). \qquad (1)\end{aligned}$$

Under the conditional independence assumption, we have

$$p_{\mathbb{X}|\mathbb{Y}}(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{n} p_{\mathbb{X}_i|\mathbb{Y}_i}(\mathbf{x}_i|y_i). \qquad (2)$$

In a supervised scenario, the class probabilities $p_{\mathbb{X}_i|\mathbb{Y}_i}(\cdot|y_i = k)$ for $k \in \mathcal{L}$ are known or learned from a training set. Having into consideration (2), we may then write

$$\begin{aligned}\hat{\mathbf{y}}_{MAP} &\in \arg\min_{\mathbf{y}\in\mathcal{L}^n} -\log(p_{\mathbb{X}|\mathbb{Y}}(\mathbf{x}|\mathbf{y}) p_{\mathbb{Y}}(\mathbf{y})) \\ &= \arg\min_{\mathbf{y}\in\mathcal{L}^n} \sum_{i=1}^{n} D_i(y_i) + \lambda U(\mathbf{y}),\end{aligned} \qquad (3)$$

where $D_i(y_i) = -\log p_{\mathbb{X}_i|\mathbb{Y}_i}(\mathbf{x}_i|y_i)$, $\lambda U(\mathbf{y}) = -\log p(\mathbf{y})$, and $\lambda > 0$ is a regularization parameter. The minimization of (3) is an integer optimization problem. In the case of Potts [13] prior and $K = 2$, the problem has exact solution obtained by mapping the problem into the computation of a min-cut on a suitable graph [14]. However, for $K > 2$, the optimization (3) is NP-hard [15] and, therefore, only approximations may be computed. To further complicate the use of integer formulations, the class of regularizers U that may be used in (3) is quite narrow; for example, it is not a simple task to include prior coming from superpixels into U.

2.1. Convex formulation

Let $\mathbf{z}_i = [z_{1i}, \ldots, z_{Ki}]^T \in \{0, 1\}^K$ be a "1-of-K" representation of y_i; that is, $(y_i = k) \Leftrightarrow [z_{il} = 0$ for $l \neq k$ and $z_{ik} = 1]$. Using this representation, optimization (3) may be written as

$$\hat{\mathbf{z}} \in \arg\min_{\mathbf{z}\in\{0,1\}^{K\times n}} \sum_{i=1}^{n} \mathbf{q}_i^T \mathbf{z}_i + \lambda\phi(\mathbf{z}) \qquad (4)$$

$$\text{s.t.:} \qquad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T, \qquad (5)$$

where $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$, $\mathbf{q}_i = [D(y_i = 1), \ldots, D(y_i = K)]^T$, $\phi(\mathbf{z}) = U(\mathbf{y})$, and $\mathbf{1}_p$ is a p-dimensional column vector of 1's.

As proposed in [10, 12], and also related with [8, 9], we relax the optimization (4) by replacing the discrete set $\{0, 1\}$ to the interval $[0, 1]$, obtaining the optimization

$$\hat{\mathbf{z}} \in \arg\min_{\mathbf{z}\in[0,1]^{K\times n}} \sum_{i=1}^{n} \mathbf{q}_i^T \mathbf{z}_i + \lambda\phi(\mathbf{z}) \qquad (6)$$

$$\text{s.t.:} \qquad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T, \qquad (7)$$

which is convex, provided that ϕ is a convex function. Although the solutions $\widehat{z}_{ki}$ yielded by (6) are mostly discrete [16], a few elements, mostly in the boundary of the classes, may not be in $\{0, 1\}$. In order to recover a complete discrete solution, we compute

$$\widehat{\mathbf{y}}_i = \arg\max_k \widehat{z}_{ki}, \quad i \in \mathcal{S}.$$

The formulation (6) yields excellent results when compared with the original integer formulation, as extensively illustrated in [16]. In addition, by a proper tailoring to the function ϕ, it allows to enforce a large variety of prior information. It is exactly this possibility that we explore in this paper.

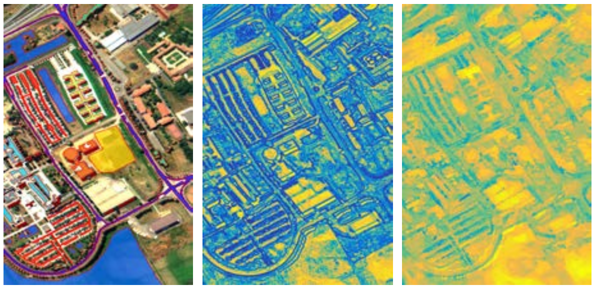
2.2. Superpixel and spatial total variation regularization

Herein we consider a regularizer with the structure $\lambda_1\phi_{TV} + \lambda_2\phi_{SP}$, where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters, ϕ_{TV} is a form of total variation (TV), thus promoting class regions with minimum boundary length [17], and ϕ_{SP} is associated to a one or more segmentations computed beforehand and promotes superpixels having the same label.

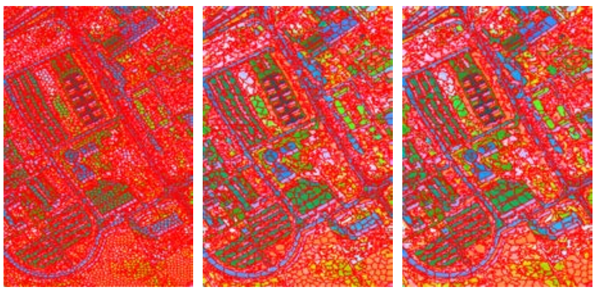
The formal structure of the ϕ_{TV} is

$$\phi_{TV}(\mathbf{z}) = \sum_{n\in\mathcal{S}} \sqrt{||\mathbf{D}_h\mathbf{z}[n]||^2 + ||\mathbf{D}_v\mathbf{z}[n]||^2}, \qquad (8)$$

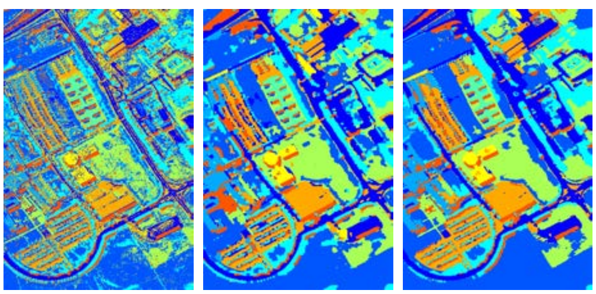
where $\mathbf{D}_h, \mathbf{D}_v : \mathbb{R}^{K\times n} \mapsto \mathbb{R}^{K\times n}$ are linear operators computing horizontal and vertical first order backward vector differences, respectively. The regularizer (8) is a form of vector total variation (VTV) [17] which in addition to promoting class boundaries of minimal length, it also promotes aligned edges across $\mathbf{z}$.



(a) RGB image and GT (b) Gradient of VTV (c) Gradient of graph TV



(d) Multiple-scaled spatial partitions



(e) MLR Classification (OA = 81.49%) (f) Classification of GC (OA = 93.26%) (g) Classification of CR (OA = 95.86%)

Fig. 2. ROSIS Pavia University data set and corresponding classification maps

In order to exploit the large correlation of spatial information that comes from over-segmented superpixels, we propose the use of the graph total variation

$$\phi_{SP}(\mathbf{z}) = \sum_c ||(\mathbf{A}_c - \mathbf{I})\mathbf{z}||_F^2, \quad (9)$$

where c indices the superpixelizations and $\mathbf{A}_c$ is an adjacency matrix of dimension $n \times n$ corresponding to the image partition imposed by the c-th superpixelization. There is a one to one correspondence between graph and superpixelization, subgraph and partition element, and node and image pixel. The adjacency matrix $\mathbf{A}_c$ corresponds to a disjoint union of fully connected subgraphs, where each subgraph corresponds to a partition element of the c-th superpixelization, and each node of the subgraph corresponds to a pixel inside the partition element. This means that each pixel is connected to all other pixels belonging to the same partition element, and to no pixel belonging to a different partition element. Thus, the minimization of $\mathbf{z}$ across the graph Laplacian $\mathbf{A}_c - \mathbf{I}$ promotes the minimization of the total variation of the vectors $\mathbf{z}_i$ across the pixels i belonging to the same partition element, for all partition elements belonging to the c-th superpixelization. Therefore, this regularizer promotes constant vectors $\mathbf{z}_i$ within the partition elements. In this work, we use fast superpixel generating algorithm SLIC [18] aiming to obtain multiple over-segmented spatial partitions.

The final optimization problem is then given by

$$\begin{aligned} &\hat{\mathbf{z}} \in \arg \min_{\mathbf{z} \in \mathbb{R}^{K \times n}} \sum_{i=1}^{n} \mathbf{q}_i^T \mathbf{z}_i + \lambda_1 \phi_{TV}(\mathbf{z}) + \lambda_2 \phi_{SP}(\mathbf{z}) \quad (10) \\ &\text{s.t.:} \qquad \mathbf{z} \geq 0 \qquad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T, \end{aligned}$$

where the constraint $\mathbf{z} \in [0,1]^{K \times n}$ was removed as is equivalent to $\mathbf{z} \geq 0$ and $\mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T$.

Optimization of (10) is convex, since (8) and (9) are both convex functions. To solve it, we use the Split Augmented Lagrangian Shrinkage (SALSA) methodology introduced in [11]. SALSA is essentially an instance of the alternating method of multipliers (ADMM) [19] designed to optimize sums of an arbitrary number of convex terms. As a proximal algorithm, the main ingredient of SALSA is the Moreau proximity operators [20] for each of the convex terms, which in the case of (10) are quadratic for the data term $\sum_{i=1}^{n} \mathbf{q}_i^T \mathbf{z}_i$ and for the graph total variation prior $\phi_{SP}(\mathbf{z})$, a vector soft thresholding for vector total variation prior $\phi_{TV}(\mathbf{z})$, projections on the positive orthant for the nonnegativity constraint, and projections on a simplex for the sum-to-one constraint. Optimization (10) has a solution as it is a convex problem defined on a compact set. The resulting optimization algorithm is very similar to the SALSA [11] and SegSALSA [8, 9], and is omitted in this paper due to space constraints.

3. EXPERIMENTAL RESULTS

We compare our introduced method with the-state-of-the-art graph-cut algorithm [15] via testing their performance in segmentation and classification with the benchmark ROSIS Pavia University data set [1], shown in Fig.2 (a), acquired using the ROSIS optical sensor over the urban area of the University of Pavia, Italy. The class models are learnt with the multinomial logistic regression (MLR) classifier, with a training set consisting of 50 pixels/class selected randomly. The negative logarithm supplied by MLR is used as the data term in (10). The classification obtained just with MLR has OA=81.49%. The corresponding classification MAP is shown in Fig.2 (e).

As explained in the previous section, we use two regularizers in the convex formulation: vector total variation (VTV) $\phi_{TV}(\mathbf{z})$ (8) and superpixels graph total variation (SP)

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Pavia_University_scene

$\phi_{SP}(\mathbf{z})$ (9) obtained from multiple segmentations using the SLIC algorithm with a varying parameters. The distance used in SLIC is computed from the first seven components of singular value decomposition (SVD). Fig.2 (b) shows the magnitude of the VTV gradient. Fig.2 (d), from left to right, shows the SLIC partitions obtained with three sets of parameters: size [6, 11, 13] and regularizers [0.015, 0.015, 0.010]. The graph gradient of multiple superpixelizations is shown in Fig.2 (c). From Fig.2 (b, c), we can observe that both priors of (10) contribute to an identification of the classification classes in a complementary way: VTV has a strong response on the boundaries whereas SP has a strong response inside the class regions.

The final labeling is obtained by solving (10) with SALSA, which was computed in approximately 2 minutes in a desktop PC (Intel Core i7-905 CPU@3.07GHz and 6GB RAM memory) using MATLAB R2015. The result of graph-cut (GC) is also obtained for comparison. Both maps are displayed in Fig.2 (f) (g) respectively. A few observations are in order. First, our introduced method shows advantage in the statistical overall accuracy (OA) with 95.86% over 93.29% of graph-cut, both of which improve greatly from the base OA of 81.49% of the MLR classifier. Second, it presents a huge advantage in retaining the spatial details, namely, the contours as well as structural details of land objects in the Pavia University scene. This observation is consistent with our initial expectation of graph total variation that exploits the large correlation of spatial information that comes from over-segmented superpixels. We remark that this advantage is not completely reflected in the OA because the training set does not contain pixels close to the class boundaries. Finally, it is remarkable that our introduced method is a very flexible and powerful scheme to exploit multiple spatial partitions, considering the fact that multiple partitions are generally more robust in describing the contours of land objects.

4. CONCLUSIONS

This paper proposed a new convex relaxation strategy aiming at flexibly appending extra priors, especially the spatial contextual information that comes from over-segmented superpixels into hyperspectral image labeling models. Two priors, the vector total variation and graph total variation of superpixels, are used to form a convex problem on a compact set whose solution was computed by SALSA. Experiments with real hyperspectral data illustrate the effectiveness of our proposed method, in comparison with the-state-of-the-art integer optimization algorithms via graph-cuts.

5. REFERENCES

[1] A. Schick, M. Bäuml, and R Stiefelhagen, “Improving foreground segmentations with probabilistic superpixel markov random fields,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012 CVPRW'2012*. IEEE, 2012, pp. 27–31.

[2] J. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, “Hyperspectral remote sensing data analysis and future challenges,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[3] P. Kohli L. Ladický and P. Torr, “Robust higher order potentials for enforcing label consistency,” *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[4] Y. Tarabalka, J. Chanussot, and J. Benediktsson, “Segmentation and classification of hyperspectral images using watershed transformation,” *Pattern Recognition*, vol. 43, no. 7, pp. 2367–2379, 2010.

[5] G. Zhang, X. Jia, and J. Hu, “Superpixel-based graphical model for remote sensing image mapping,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 5861–5871, 2015.

[6] Y. Tarabalka, J. Benediktsson, J. Chanussot, and J. Tilton, “Multiple spectral–spatial classification approach for hyperspectral data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4122–4132, 2010.

[7] G. Zhang, X. Jia, and N. Kwok, “Spectral-spatial based super pixel remote sensing image classification,” in *2011 4th International Congress on Image and Signal Processing (CISP)*. IEEE, 2011, vol. 3, pp. 1680–1684.

[8] J. Bioucas-Dias, F. Condessa, and J. Kovacevic, “Alternating direction optimization for image segmentation using hidden markov measure field models,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 90190P–90190P.

[9] F. Condessa, J. Bioucas-Dias, and J. Kovačević, “Supervised hyperspectral image segmentation: a convex formulation using hidden fields,” in *IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS'14)*, Lausanne, Switzerland, June 2014.

[10] T. Pock, A. Chambolle, D. Cremers, and H. Bischof, “A convex relaxation approach for computing minimal partitions,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2009 CVPR'2009*. IEEE, 2009, pp. 810–817.

[11] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, “An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems,” *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 681–695, 2011.

[12] C. Zach, D. Gallup, J. Frahm, and M. Niethammer, “Fast global labeling for real-time stereo using multiple plane sweeps.,” in *VMV*, 2008, pp. 243–252.

[13] S.Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag New York, Inc., 1995.

[14] D. Greig, B. Porteous, and A. Seheult, “Exact maximum a posteriori estimation for binary images,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 271–279, 1989.

[15] V. Kolmogorov, “Convergent tree-reweighted message passing for energy minimization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1568–1583, 2006.

[16] C. Nieuwenhuis, E. Töppe, and D. Cremers, “A survey and comparison of discrete and continuous multi-label optimization approaches for the potts model,” *International journal of computer vision*, vol. 104, no. 3, pp. 223–240, 2013.

[17] B. Goldluecke, E. Strekalovskiy, and D. Cremers, “The natural vectorial total variation which arises from geometric measure theory,” *SIAM Journal on Imaging Sciences*, vol. 5, no. 2, pp. 537–563, 2012.

[18] R. Achanta et al., “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] R. Glowinski, “On alternating direction methods of multipliers: a historical perspective,” in *Modeling, Simulation and Optimization for Science and Technology*, pp. 59–82. Springer, 2014.

[20] P. Combettes and V. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.