A GRAPH-BASED APPROACH FOR FEATURE EXTRACTION AND SEGMENTATION OF MULTIMODAL DATASETS

Geoffrey Iyer[1], *Jocelyn Chanussot*[1,2,3], *Andrea L. Bertozzi*[1]

1. University of California, Los Angeles
2. Univ. Grenoble Alpes, CNRS, GIPSA-lab, F-38000 Grenoble, France
3. Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland

ABSTRACT

In the past few years, graph-based methods have proven to be a useful tool in a wide variety of energy minimization problems [1]. In this paper, we propose a graph-based algorithm for feature extraction and segmentation of multimodal sets. By defining a notion of similarity that integrates information from each modality, we merge the different sources at the data level. The graph Laplacian then allows us to perform feature extraction and segmentation on the fused dataset. We apply this method in a practical example, namely the segmentation of optical and lidar images. The results obtained confirm the potential of the proposed method.

Index Terms— Image segmentation, multimodal image, data fusion, graph Laplacian, Nyström extension, graph cut minimization

1. INTRODUCTION

With the increasing availability of data we often come upon multiple datasets, derived from different sensors, that describe the same object or phenomenon. We call the sensors *modalities*, and because each modality represents some new degrees of freedom, it is generally desirable to use more modalities rather than fewer. For example, in the area of speech recognition, researchers have found that integrating the audio data with a video of the speaker results in a much more accurate classification [2, 3]. Similarly, in medicine, the authors of [4] and [5] fuse the results of two different types of brain imaging to create a final image with better resolution than either of the originals. However, correctly processing a multimodal dataset is not a simple task [6]. Even the naive method of analyzing each modality separately still requires clever thinking when combining the results, and this is rarely the optimal way to handle the data. In this paper, we will instead perform feature extraction on the full dataset, considering each modality simultaneously. After creating a feature space, we can then use any standard segmentation method to create a classification result.

Here we consider the case where each dataset contains the same number of elements, and these elements are co-registered (so the i-th point in one set corresponds to the i-th point in another). This often occurs in image processing problems, where the sets may be images of the same scene obtained from different sensors (as is the case in our experimental data), or taken at different times. This problem has also been addressed in [7, 8], although with different methods.

For notation, we label the sets, $X^1, X^2, \ldots, X^k$, with dimensions $d_1, d_2, \ldots, d_k$, and let

$$X = (X^1, X^2, \ldots, X^k) \subset \mathbb{R}^{n \times (d_1 + \cdots + d_k)}$$

be the concatenated dataset. Our method extracts features from the dataset by finding eigenvectors of the graph Laplacian, then uses standard data-segmentation algorithms on these features to obtain a final classification. In section 2 we give the general theory behind our method, and in 3 we show the results of the method applied to an optical/LIDAR dataset.

2. THE METHOD

2.1. Graph Laplacian

We approach this problem via graph-based methods. A more detailed survey of the theory can be found in [9]. Here we state only the results necessary to implement our algorithm.

2.1.1. The Graph Min-Cut Problem

We represent our dataset X using an undirected graph $G = (V, E)$. The nodes $v_i \in V$ of the graph correspond to elements of X, and we give each edge e_{ij} a *weight* $w_{ij} \geq 0$ representing the similarity between nodes v_i, v_j, where large weights correspond to similar nodes, and small weights to dissimilar nodes. This gives rise to a *similarity matrix* (also called the *weight matrix*)

$$W = (w_{ij})_{i,j=1}^{n}.$$

Since G is undirected, we require that $w_{ij} = w_{ji}$, which implies that W is a symmetric matrix. There are many different notions of "similarity" in the literature, and each has its own merits. In many applications, one defines

$$w_{ij} = -\exp\left(\|v_i - v_j\| / \sigma\right),$$

where σ is a scaling parameter. In this work we adapt this definition to apply to our multimodal dataset, as is explained in 2.3.

Once the weight matrix has been defined, the data clustering problem can be rephrased as a graph-cut-minimization problem of the similarity matrix W. Given a partition of V into subsets $A_1, A_2, \ldots, A_m$, we define the *ratio graph-cut*

$$\text{RatioCut}(A_1, \ldots, A_m) = \frac{1}{2}\sum_{i=1}^{m} \frac{W(A_i, A_i^c)}{|A_i|}. \quad (1)$$

Where

$$W(A, B) = \sum_{i \in A, j \in B} w_{ij},$$

and the $\frac{1}{2}$ is added to account for double-counting each edge. Heuristically, minimizing the ratio cut serves to minimize the connection between distinct A_i, A_j, while still ensuring that each set is of a reasonable size. Without the term $|A_i|$ term, the optimal solution often contains one large set and $m - 1$ small sets.

Solving the graph min-cut problem is equivalent to finding m indicator vectors $f_1, \ldots, f_m \in \mathbb{R}^n$ such that

$$f_{m,j} = \begin{cases} 1 & \text{if } x_j \in A_m \\ 0 & \text{else} \end{cases}.$$

It has been shown in [10] that explicitly solving this problem is an $O(|V|^{m^2})$ process. As this is unfeasible in most cases, we instead introduce the graph Laplacian along with an approximation of the minimization problem.

2.1.2. Graph Laplacian

After forming the weight matrix W, we define the graph Laplacian. For each node $v_i \in V$, define the *degree* of the node

$$d_i = \sum_j w_{ij}.$$

Intuitively, the degree represents the strength of a node. Let D be the diagonal matrix with d_i as the i-th diagonal entry. We then define the *graph Laplacian*

$$L = D - W. \quad (2)$$

For a thorough explanation of the properties of the graph Laplacian, see [11]. In our work we will use that L is symmetric and positive definite, as well as the following fact (proven in [9]).

Fact 2.1. For a given graph-cut $A_1, \ldots, A_m$, define the $f_1, \ldots, f_m$ as above, and have $h_j = f_j / \|f_j\|$. Let H be the $n \times m$ matrix whose columns are h_j. Then $H^T H = I$, and

$$\text{RatioCut}(A_1, \ldots, A_m) = \text{Tr}\left(H^T L H\right). \quad (3)$$

As explained in 2.1.1, we cannot solve this problem explicitly, so instead we relax the problem to allow entries of H to take on arbitrary real values. That is, we find

$$\text{argmin}_{H \in \mathbb{R}^{n \times m}} \text{Tr}\left(H^T L H\right) \quad \text{where } H^T H = I. \quad (4)$$

As L is symmetric and H is orthogonal, this problem is solved by choosing H to be the matrix containing the m eigenvectors of L corresponding to the m smallest eigenvalues. Using the eigenvectors H we define a map $X \to \mathbb{R}^m$. For each graph node $x_i \in X$ we get a vector $y_i \in \mathbb{R}^m$ given by the ith row of H. These y_i give the solution to the relaxed min-cut problem, as as such can be thought of as features extracted from the original dataset X.

To obtain a solution to the original min-cut problem, we must then perform some kind of classification on the y_i to create the indicator vectors $f_1, \ldots, f_m$ as described above. There are a large variety of such methods in the literature ([12, 13] are some examples). In section 3 we use k-means to segment the y_i, resulting in a well-known algorithm called *spectral clustering*. Although k-means is unlikely to give an optimal classification, it is quite easy to implement, and the final results are strong enough to give a proof-of-concept.

2.2. Nyström Extension

Calculating the full graph Laplacian is computationally intensive, as the matrix contains n^2 entries. Instead we use Nyström's extension to find approximate eigenvalues and eigenvectors with a heavily reduced computation time. See [14, 13, 15] for a more complete discussion of this method.

Let X denote the set of nodes of the complete weighted graph. We choose a subset $A \subset X$ of "landmark nodes", and have B its complement. Up to a permutation of nodes, we can write the weight matrix as

$$W = \begin{pmatrix} W_{AA} & W_{AB} \\ W_{BA} & W_{BB} \end{pmatrix}, \quad (5)$$

where the matrix $W_{AB} = W_{BA}^T$ consists of weights between nodes in A and nodes in B, W_{AA} consists of weights between pairs of nodes in A, and W_{BB} consists of weights between pairs of nodes in B. Nyström's extension approximates W as

$$W \approx \begin{pmatrix} W_{AA} \\ W_{BA} \end{pmatrix} W_{AA}^{-1} \begin{pmatrix} W_{AA} & W_{AB} \end{pmatrix}. \quad (6)$$

Here the error of approximation is determined by how well the rows of of W_{AB} span the rows of W_{BB}. This approximation is extremely useful, as we can use it to avoid calculating W_{BB} entirely. It is in fact possible to find $|A|$ approximate eigenvectors of W using only the matrices W_{AA}, W_{AB}. This results in a significant reduction in computation time, as we compute and store matrices of size at most $|A| \times |X|$, rather than $|X| \times |X|$.

In practice, the details of choosing A will not significantly affect the final performance of the algorithm. Although it is possible to choose specific "landmark nodes", in most applications (including ours) the elements of A are selected at random from the full set X. Furthermore, the amount of landmark nodes m can be chosen to be quite small without noticeably affecting performance. This makes Nyströms extension especially useful in application, as very little work is required to tune the parameters. In Section 3 we use $m = n^{\frac{1}{4}}$, and choosing a larger set A does not give a significant change in the error of approximation.

2.3. Edge Weights

To calculate the weight matrix W, we first scale our sets $X^1, \ldots, X^k$ to make distances in each set comparable. Let $X = (X^1, \ldots, X^k) \subset \mathbb{R}^{n \times (d_1 + \cdots + d_k)}$ be the concatenated dataset, and let $A \subset X$ be the collection of landmark nodes as in 2.2. For simplicity of notation, rearrange the entries of X so that $A = \{x_1, \ldots, x_m\}$. So $|A| = m$, and $m \ll n$. Then for $\ell = 1, \ldots, k$ define the scaling factor

$$\lambda_\ell = \text{stdev}\left(\left\|x_i^\ell - x_j^\ell\right\| \;;\; 1 \le i \le n,\; 1 \le j \le m\right) \quad (7)$$

For a graph node $x \in X$, we define

$$\|x\| = \max\left(\frac{\|x^1\|}{\lambda_1}, \cdots, \frac{\|x^k\|}{\lambda_k}\right). \quad (8)$$

Then define the weight matrix W (using the Nyström Extension), by

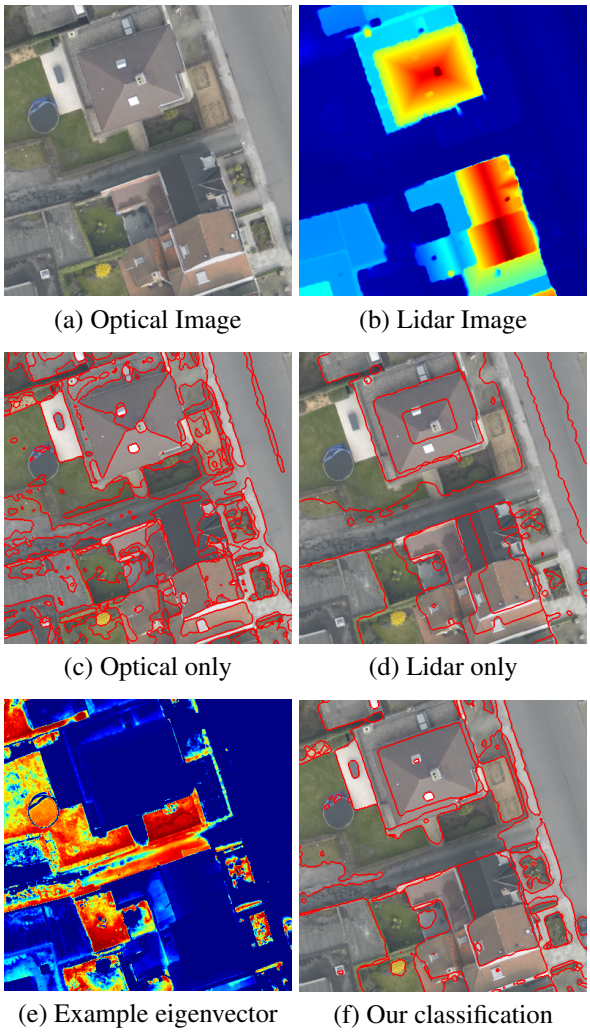
$$W = \begin{pmatrix} W_{AA} \\ W_{AB} \end{pmatrix} = (w_{ij})_{1 \le i \le n, 1 \le j \le m} \quad (9)$$

with $w_{ij} = \exp\left(-\|x_i - x_j\|\right)$.

Note that the $\|\cdot\|$ defined above is a norm on the concatenated dataset X. We specifically choose to use the maximum of the individual measurements to emphasize the unique information that each dataset brings. With this norm, two data points x_i, x_j are considered similar only when they are similar in each dataset.

3. EXPERIMENT

We test our algorithm on an optical/LIDAR dataset from the 2015 IEEE Data Fusion Contest [17] (figure 1a,b).



(a) Optical Image (b) Lidar Image

(c) Optical only (d) Lidar only

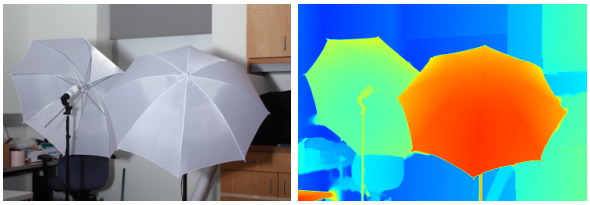
(e) Example eigenvector (f) Our classification

Fig. 1. DFC Data Experimental Results.

The data consists of an RGB image and an elevation map of a residential neighborhood in Belgium. We choose this particular scene because of the large amount of non-redundancy between the two images. The lidar data is effective at differentiating the roofs of the buildings from the adjacent streets, and the optical data is useful for segmenting the many different objects at ground-level. In figures 1c,d we show the results of spectral clustering performed using each modality separately. The issues with single-modality segmentation can be seen immediately, as both segmentations miss out on key features of the data.

In figure 1e,f we show the results of our method. 1e is one example eigenvector of the graph Laplacian. As explained in 2.1.2, this vector can be considered one feature of our dataset, and approximates a segmentation of the image into 2 groups. Notice how in this eigenvector the dark-grey street is highlighted, while both the light-grey

sidewalk (which is at the same elevation) and the nearby roof (which is the same color) are dark. This shows at the feature level that our algorithm is successfully using both the optical and the lidar data when determining what pixels can be considered similar. The difference shown in this example vector then causes the classification algorithm to separate those regions in the final result 1f. This last figure was obtained using a total of 12 eigenvectors (not pictured here), grouped into 5 classes.



(a) Optical Image (b) Lidar Image

(c) Example eigenvector (d) Our classification

Fig. 2. Umbrella Data Experimental Results

In fig 2 we show the results of our method applied to another optical/lidar set (found in [18]). Similar to the DFC set, the umbrella data serves as a good example because it cannot be easily analyzed using one modality alone. The umbrellas and the background walls are nearly the same shade of white, and can only be distinguised in the lidar data. Meanwhile, the different pieces of the background all lie at nearly the same depth, so can only be separated by color. As was the case with the DFC data, the final classification 2d can be understood by looking at the individual feature vectors. In 2c we show the vector responsible for separating the umbrellas from the background wall (using the lidar data), as well as from the black umbrella stands (using the RGB data).

For a given segmentation of an image, computing the graph-cut error as described in 2.1.1 is an $O(n^2)$ calculation, and requires the full weight matrix W. To avoid this, we instead measure the error of segmentation by how the data $X = (X^1, \dots, X^k) \subset \mathbb{R}^{n\times(d_1+\cdots+d_k)}$ varies within each class. More explicitly, we use the metric:

$$\text{Error} = \frac{1}{n} \sum_{\text{classes } \mathcal{C}} \sum_{x\in\mathcal{C}} \|x - \text{mean}\,(y \in \mathcal{C})\|\,. \qquad (10)$$

Where the norm $\|\cdot\|$ is the same as defined in 2.3.

Method	**Error**	**Time**
Our norm on concatenated set	0.40	9.2s
2-norm on concatenated set	0.41	9.4s
Intersection method	0.43	17.6s
Our norm on optical-only	0.83	8.1s
Our norm on lidar-only	0.75	7.9s
K-means on original data	0.75	4.1s

Table 1. DFC Data Quantitative Results

To test our method, we compare against a few other common methods. The results are given in Table 1. Unfortunately, due to space limitations, we cannot display the visual comparisons between the different algorithms. Instead we will briefly describe the methods used. The 2-norm on the concatenated set X still minimizes a graph-cut, but uses a slightly different norm to define the weight matrix

$$\|x\| = \sqrt{\frac{\|x^1\|^2}{\lambda_1} + \cdots + \frac{\|x^k\|^2}{\lambda_k}}, \qquad (11)$$

where the scaling factors λ_j are the same as in 2.3. The intersection method computes the full classification via spectral clustering on $X^1, \dots, X^k$ separately, then segments the data by intersecting the individual classifications.

4. CONCLUSIONS

In conclusion, graph-based methods provide a straightforward and flexible method of combining information from multiple datasets. By defining a weight map $\mathbb{R}^{n\times(d_1+\cdots+d_k)} \to \mathbb{R}_{\geq 0}$ with some reasonable norm-like properties, we can create the graph Laplacian of the data and extract features in the form of eigenvectors. These features can then be used as part of many different data-segmentation algorithms. For this paper, we use k-means on the eigenvectors as a simple-proof of concept. However this portion of our method could easily be replaced with a more in-depth approach, such as a Mumford-Shah model [12], or even a semi-supervised method such as [13].

Our next area of interest is the removal of the co-registration assumption. In section 3 our two images are of the same underlying scene, where pixels correspond exactly between images. We could not, for example, process two images taken from different angles. Our goal for the future is to remove this restriction and develop an algorithm that can be applied to a much larger variety of datasets.

This work was supported by NSF grant DMS-1118971, ONR grant N00014-16-1-2119, and NSF grant DMS-1417674.

5. REFERENCES

[1] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, Feb 2004. (document)

[2] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept 2003. 1

[3] Farnaz Sedighin, Massoud Babaie-Zadeh, Bertrand Rivet, and Christian Jutten, "Two Multimodal Approaches for Single Microphone Source Separation," in *24th European Signal Processing Conference (EUSIPCO 2016)*, Budapest, Hungary, Sept. 2016, pp. 110–114. 1

[4] X. Lei, P. A. Valdes-Sosa, and D. Yao, "EEG/fMRI fusion based on independent component analysis: integration of data-driven and model-driven methods," *J. Integr. Neurosci.*, vol. 11, no. 3, pp. 313–337, Sep 2012. 1

[5] S. Samadi, H. Soltanian-Zadeh, and C. Jutten, "Integrated analysis of eeg and fmri using sparsity of spatial maps," *Brain Topography*, vol. 29, no. 5, pp. 661–678, 2016. 1

[6] Dana Lahat, Tülay Adali, and Christian Jutten, "Challenges in Multimodal Data Fusion," in *22nd European Signal Processing Conference (EUSIPCO-2014)*, Lisbonne, Portugal, Sept. 2014, pp. 101–105. 1

[7] Guillaume Tochon, Mauro Dalla Mura, and Jocelyn Chanussot, *Segmentation of Multimodal Images Based on Hierarchies of Partitions*, pp. 241–252, Springer International Publishing, Cham, 2015. 1

[8] A. M. Ali and A. A. Farag, "A novel framework for n-d multimodal image segmentation using graph cuts," in *2008 15th IEEE International Conference on Image Processing*, Oct 2008, pp. 729–732. 1

[9] Ulrike von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007. 2.1, 2.1.2

[10] Olivier Goldschmidt and Dorit S. Hochbaum, "A polynomial algorithm for the k-cut problem for fixed k," *Mathematics of Operations Research*, vol. 19, no. 1, pp. 24–37, 1994. 2.1.1

[11] Bojan Mohar, "The laplacian spectrum of graphs," *Graph Theory, Combinatorics, and Applications*, vol. 2, pp. 871–898, 1991. 2.1.2

[12] Huiyi Hu, Justin Sunu, and Andrea L. Bertozzi, *Multi-class Graph Mumford-Shah Model for Plume Detection Using the MBO scheme*, pp. 209–222, Springer International Publishing, Cham, 2015. 2.1.2, 4

[13] Ekaterina Merkurjev, Tijana Kostic, and Andrea L Bertozzi, "An mbo scheme on graphs for classification and image processing," *SIAM Journal on Imaging Sciences*, vol. 6, pp. 1903–1930, October 2013. 2.1.2, 2.2, 4

[14] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik, "Spectral grouping using the nystrom method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, February 2004. 2.2

[15] J. T. Woodworth, G. O. Mohler, A. L. Bertozzi, and P. J. Brantingham, "Non-local crime density estimation incorporating housing information," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2028, 2014. 2.2

[16] Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik, *Spectral Partitioning with Indefinite Kernels Using the Nyström Extension*, pp. 531–542, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. 2.2

[17] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupre, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high-resolution lidar and rgb data: Outcome of the 2015 ieee grss data fusion contest #8211;part a: 2-d contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, Dec 2016. 3

[18] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proceedings of the 36th German Conference on Pattern Recognition*, september 2014. 3