Advanced Multimodal Information Processing

Geoffrey Iyer

1 Introduction

With the increasing availibility 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 information, 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 [1, 2]. Similary, in medicine, the authors of [3] and [4] 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. 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. Our goal is to create a general algorithm for feature extraction and data segmentation that can be applied to any multimodal dataset.

In the current state of the project, 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 is often the case 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. 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. Finally, in section 4 we discuss the extensions of the project that we hope to complete in France.

2 The Method

2.1 The Graph Min-Cut Problem

We represent our dataset X as a *similarity matrix*. That is, for each two data points x_i, x_j, we define a *weight* w_{ij} representing the similarity between the points. A large weight corresponds to very similar nodes, and a small weight to dissimilar nodes. There are many different choices of the weights w_{ij} in the literature, and each has its own merits. In most applications, one defines

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

where σ is a scaling parameter. In this work we adapt this definition to apply to our multimodal dataset. First we scale our sets $X^1, \ldots, X^k$ to make distances in each set comparable, then we define.

$$w_{ij} = \exp\left(-\max\left(\left\|x_i^1 - x_j^1\right\|, \ldots, \left\|x_i^k - x_j^k\right\|\right)\right).$$

Choosing to use the maximum of the individual values allows us to take advantage of the unique information in each dataset, as two points are only considered similar here when they are similar in every modality.

Moving forward, we aim to find a classification that groups pairs with high weight together, while also separating pairs with low weight. This goal is formalized as the Ratio Cut problem. Given a partition of X into subsets $A_1, A_2, \ldots, A_m$, we define the *ratio graph-cut*

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

where $W(A, B) = \sum_{i\in A, j\in B} w_{ij}$. Minimizing the ratio cut will give the classification we desire, as the $W(A_i, A_i^c)$ term penalizes partitions that separate elements with a large weight between them, while the $|A_i|$ term ensures that each segment of the final partition is of a reasonable size (without the $|A_i|$ term, the optimal solution often contains one large set and $m-1$ small sets). It has been shown in [5] that explicitly solving this problem is an $O(|V|^{m^2})$ process. As this is infeasible in most cases, we instead introduce the graph Laplacian to handle an approximation of the minimization problem.

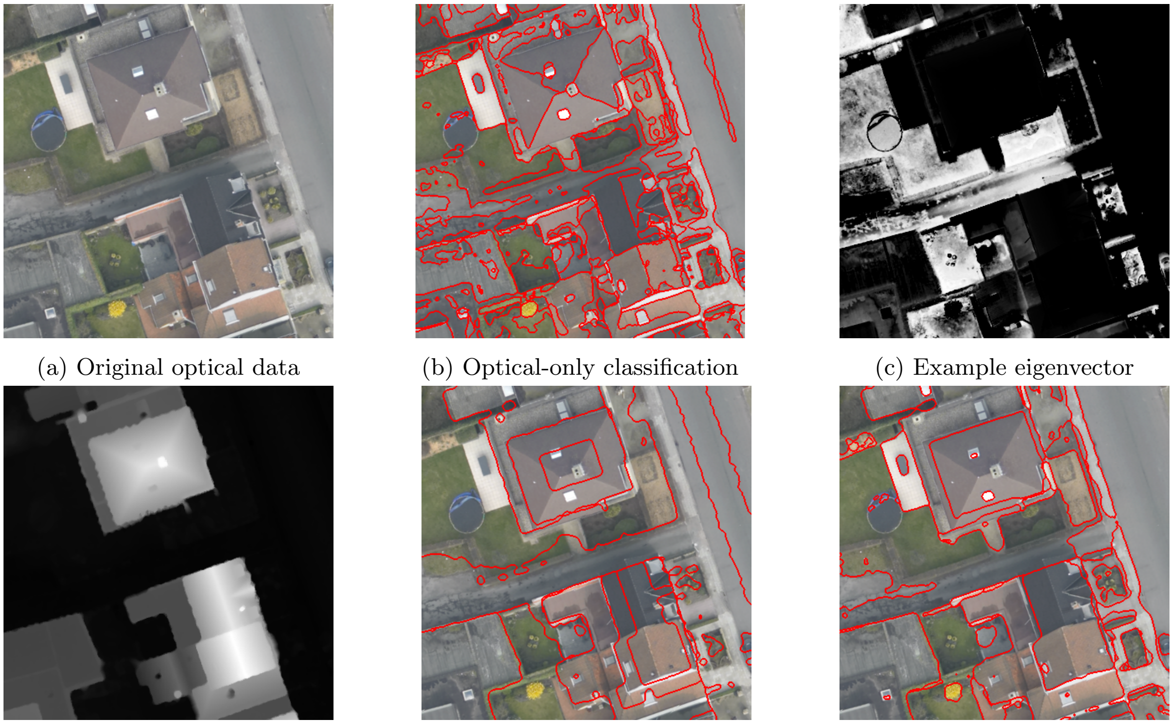
2.2 Graph Laplacian and Clustering

In [6] it is shown that the Ratio Cut problem can be approximately solved using the eigenvectors of the *graph Laplacian*, $L = D - W$. Here D is a diagonal matrix with entries $d_{ii} = \sum_j w_{ij}$. Each eigenvector represents a feature of the data, and if we let H be a matrix where the columns are eigenvectors, then the ith row of H represents the features of the data point x_i. We then get an approximate solution to the original min-cut problem by using any data clustering algorithm on these rows. In section 3 we use k-means to segment the row vectors, resulting in a well-known algorithm called *spectral clustering*.

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. Essentially, this consists of choosing a small number $m \ll n$ of columns of the weight matrix, and using some clever linear algebra to approximately solve the eigenproblem using only these columns. This results in a significant reduction in computation time, as we compute and store matrices of size at most $m \times n$, rather than $n \times n$. See [7], [8] for a more complete discussion of this method. In practice, m can be chosen to be quite small without creating significant error. In Section 3 we use $m = n^{\frac{1}{4}}$.

3 Experiment

We test our algorithm on an optical/LIDAR dataset from the 2015 IEEE Data Fusion Contest [9]. The data consists of an RBG image and an elevation map of a residential neighborhood in Belgium. Each picture contains roughly 160,000 pixels.



(a) Original optical data

(b) Optical-only classification

(c) Example eigenvector

(d) Original lidar data

(e) Lidar-only classification

(f) Our classification

Figure 1: Experimental results

In figure 1b and 1e we show the results of spectral clustering applied to each dataset individually, and the final results of our algorithm are pictured in 1f. These images show the importance of the multimodal approach, as the optical-only method is unable to differentiate the dark-grey roof from the adjacent street, and the lidar-only method cannot separate the white sidewalk from the green grass. In figure 1c we show an example eigenvector of the graph Laplacian. As explained in 2.2, this eigenvector represents one feature extracted from our dataset. Notice how 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 blacked out. This shows at the level of the feature vectors that our algorithm is successfully using both the optical and the lidar data when determining what pixels can be considered similar. The grayscale difference in the example feature vector then causes the classification algorithm to separate those regions in the final result 1f.

4 Future Work

Our current algorithm allows us to perform feature extraction and data segmentation on multimodal sets, as desired, but the assumption that the datasets are *co-registered* is quite restrictive. 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.

We approach this problem through the viewpoint of manifold alignment. Assuming that each dataset originates from some underlying manifold, we can compare the topologies of the different manifolds to obtain some information about the underlying source object. Often this is done by mapping each dataset into one common *latent space*, then analyzing this amalgam of data. The major obstacle to overcome is the inherent loss of information that comes with transferring the original data to the latent space. This technique has been used in [10, 11, 12], but for different purposes. These authors align the manifolds in order to take learning accomplished on one set and transfer it to another. Unforunately, these techniques will not directly solve our problem, since we specifically look to perform our analysis using each modality simultaneously, rather than working with the sets as individuals. Each modality has its own unique advantages and disadvantages, and a transfer of learned information implies that we ignore entirely the unique information found in the target set. Similar to our experiment in 3, we should expect to better feature extraction and segmentation results when using the datasets in tandem. Still, the underlying idea of looking at correlations between data points and using this to create a latent space serves as a good starting point for our research. So, moving forward, we aim to adapt these manifold alignment techniques to create a latent space that takes into account the information from each modality.

We are especially excited to continue this work in France, as it would give us the opportunity to collaborate with Christian Jutten of the GIPSA Lab in Grenoble. Jutten is an expert in multimodal data fusion [13], as well as the principal investigator in the CHESS (Challenges in extraction and separation of sources) group, which deals with multimodal source separation. This corresponds well to our research plan, as source separation is often the same thing as data classification. For example, in [2] they were able to use audio and visual information to successfully separate different human voice signals. Also, they developed a new approach to perform independent component analysis simultaneously over multiple datasets [14]. These kinds of advances all require correlating information across multiple modalities, which is exactly the topic of our research.

References

[1] 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*, 91(9):1306–1326, Sept 2003. 1

[2] 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)*, pages 110–114, Budapest, Hungary, September 2016. 1, 4

[3] 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.*, 11(3):313–337, Sep 2012. 1

[4] S. Samadi, H. Soltanian-Zadeh, and C. Jutten. Integrated analysis of eeg and fmri using sparsity of spatial maps. *Brain Topography*, 29(5):661–678, 2016. 1

[5] Olivier Goldschmidt and Dorit S. Hochbaum. A polynomial algorithm for the k-cut problem for fixed k. *Mathematics of Operations Research*, 19(1):24–37, 1994. 2.1

[6] Bojan Mohar. The laplacian spectrum of graphs. *Graph Theory, Combinatorics, and Applications*, 2:871–898, 1991. 2.2

[7] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2), February 2004. 2.2

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

[9] 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*, 9(12):5547–5559, Dec 2016. 3

[10] Hsiuhan Lexie Yang and Melba M. Crawford. Learning a joint manifold with global-local preservation for multitemporal hyperspectral image classification. In *IEEE International Geoscience and Remote Sensing Symposium*, 2013. 4

[11] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 4

[12] Yi-Ren Yes, Chun-Hao Huang, and Yu-Chiang Frank Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23(5):2009–2018, May 2014. 4

[13] Dana Lahat, Tülay Adali, and Christian Jutten. Challenges in Multimodal Data Fusion. In *22nd European Signal Processing Conference (EUSIPCO-2014)*, pages 101–105, Lisbonne, Portugal, September 2014. 4

[14] Dana Lahat and Christian Jutten. Joint Analysis of Multiple Datasets by Cross-Cumulant Tensor (Block) Diagonalization. In *9th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2016)*, Rio de Janeiro, Brazil, July 2016. 4