CLUSTERING TRAJECTORIES IN HETEROGENEOUS REPRESENTATIONS FOR VIDEO EVENT DETECTION

Wei-Cheng Wang* *Hsin-Wei Cheng*† *Chun-Rong Huang*† *Yen-Yu Lin**

*Research Center for Information Technology Innovation, Academia Sinica, Taiwan
†Department of Computer Science and Engineering, National Chung Hsing University, Taiwan

ABSTRACT

Trajectories have been shown to be robust and widely used in surveillance video event analysis. They encode spatial and temporal evidence simultaneously. Hence, clustering trajectories in a video can detect representative events. How to effectively represent trajectories is thus essential to video event detection. However, no a single representation of trajectories suffices in increasingly complex video analysis tasks. To address this issue, this paper presents a hierarchical clustering algorithm for grouping trajectories in multiple heterogeneous representations. It turns out that our method can not only group trajectories of highly similar events but also identify rare events from the dominant events. Experimental results show that our method can retrieve both dominant events and rare events compared with the state-of-the-art methods, leading to a better performance.

Index Terms— Video surveillance, event detection, multiple feature representations, trajectory clustering

1. INTRODUCTION

Trajectories are extracted by tracking spatially and temporally coherent keypoints in a video. The joint extraction of spatial layout and temporal motion has made trajectories crucial to various applications related video event analysis, such as [1, 2, 3, 4]. For detecting events in videos, supervised trajectory labeling is labor intensive and less practical since the learned trajectory models may not adapt themselves well to an unseen video due to the high diversity of videos. Therefore, in this work, we present an unsupervised method that jointly considers trajectories in multiple feature representations and can compensate for the absence of trajectory labeling.

The literature of video event detection is extensive. Vlachos et al. [5] proposed the longest common subsequence (LCSS) to measure the similarity between trajectories, and carry out trajectory retrieval. Naftel and Khalid [6] used self-organizing maps (SOM) to cluster trajectories in the forms of discrete Fourier transform (DFT) coefficients for grouping objects with similar activities. DFT coefficients are also adopted in [7] with a time-sensitive Dirichlet process mixture model (tDPMM) to cluster similar trajectories. Hu et al. [8] presented a clustering-based tracking method to hierarchically group spatially and temporally similar trajectories. In [9], trajectories are represented by principal component analysis (PCA) coefficients, and spectral clustering [10] is applied to trajectory grouping. Jung et al. [11] used trajectory clustering with motion histograms to achieve event detection. Jiang et al. [12] proposed to use hierarchical clustering and hidden Markov models (HMM) for unusual video event detection. Atev et al. [13] compared the performance of the longest common subsequence (LCSS) [5], dynamic time warping (DTW), and the modified Hausdorff distance with agglomerative clustering and spectral clustering for event detection. Zhou et al. [14] proposed a dynamic pedestrian-agents (MDA) based mixture model to unsupervisedly learn crowd behavior from videos. Nawaz et al. [15] clustered trajectories in urban scenes by mapping object trajectories on a reference plane with discrete wavelet transform (DWT) coefficients extracted from trajectories. Huang et al. [16] combined trajectory entropy descriptors (TED) and affinity propagation [17] to cluster trajectories of moving foreground objects. Bastani et al. [18] applied the Dirichlet process mixture model to incremental trajectory clustering.

Unlike existing unsupervised methods, our method can leverage heterogeneous trajectory representations to better discover both representative and unusual events without predefining the number of clusters. In our method, trajectories are compiled by using the scheme in [19] in advance. Complementary characteristics are extracted to yield the multiple feature representations of trajectories. Our method performs a two-stage clustering. At the first stage, trajectory clustering is applied based on each individual feature representation. At the second stage, the clustering is performed to divide trajectories into groups by fusing the multiple clustering results obtained at the previous stage. As shown in the experiments, our method achieves better accuracy compared with the state-of-the-art methods. In addition, our method can also identify trajectories of rare events from the dominant trajectory clusters, which is valuable for unusual event detection.

This work was supported in part by MOST104-2221-E-005-027-MY3 and MOST105-2221-E-001-030-MY2 from the Ministry of Science and Technology, Taiwan.

2. THE PROPOSED METHOD

Our method is introduced in this section. We first describe the multiple representations of trajectories and then specify a two-stage algorithm for multi-modal trajectory clustering.

2.1. Heterogeneous Trajectory Representations

For a given video, we extract a set of trajectories by using the method in [19]. For the ith trajectory T_i, we define it by

$$T_i = [\mathbf{p}_i(t_{s_i}) \ldots \mathbf{p}_i(t) \ldots \mathbf{p}_i(t_{e_i})] \in \mathbb{R}^{2(t_{e_i} - t_{s_i} + 1)}, \quad (1)$$

where $\mathbf{p}_i(t) = [x_i(t)\ y_i(t)] \in \mathbb{R}^2$ is the 2D spatial position of T_i in the tth frame, and t_{s_i} and t_{e_i} are the first and last appearing frames of T_i, respectively. In this work, we employ four heterogeneous representations for better trajectory descriptions, including the starting positions, ending positions, velocities, and accelerations.

The first trajectory representation of T_i is its starting position, namely

$$\mathbf{f}_i^1 = [x_i(t_{s_i})\ y_i(t_{s_i})] \in \mathbb{R}^2. \quad (2)$$

Similarly, its ending position serves as the second representation, i.e.,

$$\mathbf{f}_i^2 = [x_i(t_{e_i})\ y_i(t_{e_i})] \in \mathbb{R}^2. \quad (3)$$

Despite their simplicity, the first two feature representations record the spatial information, which is quite crucial to event identification.

Except for the spatial locations, distributions of velocities and accelerations provide temporal evidence to distinguish the motions of different trajectories. For example, a vehicle and a pedestrian move from the same starting position to the same ending position. To separate the trajectories on the vehicle from those on the pedestrian, trajectory velocities and accelerations are discriminative in this case. Nevertheless, the lengths of trajectories of the vehicle and the pedestrian may be different. Thus, how to compare trajectories of different lengths becomes an issue.

To solve this issue, the trajectory entropy descriptor (TED) [16] was proposed to characterize the velocity and acceleration distributions of trajectories with a unified length based on the entropy of velocities and accelerations. Our third trajectory representation $\mathbf{f}_i^3$ of T_i is the entropy of the velocity defined by

$$\mathbf{f}_i^3 = [E_{v_i^{x+}}\ E_{v_i^{x-}}\ E_{v_i^{y+}}\ E_{v_i^{y-}}\ E_{v_i^{+}}\ E_{v_i^{-}}] \in \mathbb{R}^6, \quad (4)$$

where $E_{v_i^{x+}}$ ($E_{v_i^{x-}}$) is the entropy of positive (negative) velocities in the horizontal direction, $E_{v_i^{y+}}$ ($E_{v_i^{y-}}$) is the entropy of positive (negative) velocities in the vertical direction, and $E_{v_i^{+}}$ ($E_{v_i^{-}}$) is the entropy of positive (negative) velocities in the diagonal direction. Similarly, the fourth trajectory representation $\mathbf{f}_i^4$ of T_i is the entropy based on accelerations defined below

$$\mathbf{f}_i^4 = [E_{a_i^{x+}}\ E_{a_i^{x-}}\ E_{a_i^{y+}}\ E_{a_i^{y-}}\ E_{a_i^{+}}\ E_{a_i^{-}}] \in \mathbb{R}^6. \quad (5)$$

All the elements in (5) are the same as those in (4) except all the velocity statistics are replaced by the acceleration statistics in the computation of entropy. Due to the space limit, please refer to [16] for the details.

2.2. Hierarchical based Clustering

A trajectory is described by different feature representations, including the positions at each time slot, the moving velocities, and the accelerations. Variations among different representations often make cross-representation comparison difficult. Consider a trajectory on a fast-moving vehicle. This trajectory is often much shorter than that on a slow-moving pedestrian with the same moving distance. To consider multiple representations simultaneously, a feasible way is to normalize each representation in advance and merge the normalized trajectories into a supervector for similarity measure. The dimensions and scales of each representation are inconsistent. Thus, representations with higher dimensions or larger variations dominate in similarity estimation.

To avoid the aforementioned problems, we propose to fuse multiple feature representations in the domain of clusters and present a two-stage clustering algorithm to carry it out. At the first stage, we get the clustering results by grouping similar trajectories for each trajectory representation. At the second stage, clustering is performed again to obtain the new groups of similar events under heterogeneous trajectory representations by fusing their cluster labels obtained at the previous stage. The details of the proposed clustering algorithm are given in the following.

The similarity between trajectories under each feature representation is evaluated prior to clustering. Specifically, the similarity between trajectories T_i and T_j under the kth trajectory representation is computed by

$$s^k(T_i, T_j) = e^{-\|\mathbf{f}_i^k - \mathbf{f}_j^k\|}, \text{ for } k \in \{1, ..., 4\}, \quad (6)$$

where $\mathbf{f}_i^k$ and $\mathbf{f}_j^k$ are the kth trajectory representations of T_i and T_j, respectively.

For each trajectory representation, we apply affinity propagation (AP) [17] to cluster all trajectories by taking their pairwise similarities, given in (6), as input. Based on the responsibility and availability of data points, AP produces the clustering result L^k for data under each representation k. Let n^k denote the number of resultant clusters. The clustering result L^k can be defined by

$$L^k = \{L_1^k, L_2^k, \ldots, L_n^k, \ldots, L_{n^k}^k\}, \text{ for } k \in \{1, ..., 4\}, \quad (7)$$

where L_n^k is the index set of data falling into the nth cluster.

Because different clusters represent different trajectory characteristics under each representation, we perform the enumerative combinatorics based on the clusters of all trajectory representations to obtain the fusion results. Here, a combination is a selection of the cluster label from each individual clustering result. Specifically, the combination label L_ℓ, the ℓth element in the combination set of the fusion results, is defined as

$$L_\ell = (\ell_1,\ \ell_2,\ \ell_3,\ \ell_4), \tag{8}$$

where $1 \leq \ell_k \leq n^k$, for $k \in \{1, ..., 4\}$, is the cluster label selected from the clustering result under the kth representation. Thus, the total number ω of clusters of our hierarchical based clustering method is

$$\omega = \prod_{k=1}^{4} n^k, \tag{9}$$

where n^k is the number of clusters yielded under the kth representation.

Based on the combination labels, trajectories are reassigned to new groups where a group will contain the trajectories of the same labels in all individual clustering results. In our paper, we consider four heterogeneous trajectory representations including the starting position $\mathbf{f}_i^1$, ending position $\mathbf{f}_i^2$, TED of velocities $\mathbf{f}_i^3$, and TED of accelerations $\mathbf{f}_i^4$. We cluster trajectories under each representation individually, and fuse all the clustering results. It is easier to understand the two-stage clustering through an example. The trajectories are grouped to 2 clusters with respect to $\mathbf{f}_i^1$, say trajectory groups starting at the right and the left sides of the frame, respectively. Trajectories are grouped to 3 clusters with respect to $\mathbf{f}_i^2$, say trajectory groups ending at the top, the middle, and the bottom sides of the frame, respectively. The feature representation $\mathbf{f}_i^3$ divides trajectories to 4 clusters. The feature representation $\mathbf{f}_i^4$ divides trajectories to 5 clusters. The total number ω of clusters is 120 $(= 2 \times 3 \times 4 \times 5)$ by combining the 2 clusters of $\mathbf{f}_i^1$, 3 clusters of $\mathbf{f}_i^2$, 4 clusters of $\mathbf{f}_i^3$, and 5 clusters of $\mathbf{f}_i^4$. Note that if a trajectory has a different label combination with respect to other trajectories, our method will reserve the group for the trajectory instead of merging it to other dominant groups. As a result, trajectories of rare and unusual events can be reserved in our method. Certain combination labels may contain no trajectories, which implies that these kinds of events do not exist in the video.

3. EXPERIMENTAL RESULTS

In the experiments, one indoor (`hall`) and three outdoor surveillance videos (`sidewalk`, `street` and `QMUL`) from [19, 20] are employed to serve as the testbed. Detailed information of the videos is shown in Table 1. The ground truth of events was labeled manually according to the starting positions, ending positions, velocities, and accelerations of foreground objects. To evaluate the performance of clustering results, clustering accuracy and normalized mutual information (NMI) [21] were adopted. We compared our method with the state-of-the-art methods using different trajectory representations and clustering algorithms, including [5, 7, 8, 9, 13, 16].

Table 1. Resolutions and frame numbers of the four videos.

	Resolution	# of Frames
hall [19]	320 × 240	66,771
sidewalk [19]	320 × 240	104,864
street [19]	320 × 240	79,449
QMUL [20]	360 × 288	90,000

Table 2. Performance comparison in accuracy (%).

	hall	sidewalk	street	QMUL
Vlachos et al. [5]	49.593	26.538	41.431	29.673
Hu et al. [7]	**54.878**	38.462	42.373	38.663
Hu et al. [8]	43.415	29.000	31.959	31.909
Bashir et al. [9]	42.480	31.962	32.467	32.612
Atev et al. [13]	40.366	33.192	32.599	31.357
TED [16] + AP	21.138	32.308	20.527	12.404
TED [16] + SC	44.919	34.000	34.237	28.890
Ours	53.659	**52.308**	**50.094**	**46.316**

The comparison results are shown in Table 2 and Table 3 for accuracy and NMI, respectively. Our method achieves the best performance on videos `sidewalk`, `street` and `QMUL`, and the second best accuracy on video `hall`. In the outdoor videos, both pedestrians and vehicles appear at the same frames. Using features related to velocities and accelerations can better separate trajectories on pedestrians from those on vehicles, even if these trajectories have the same moving directions. The competing methods such as ”TED + AP” and ”TED + SC”, which combine the trajectory entropy descriptor (TED) [16] with affinity propagation (AP) [17] and spectral clustering (SC) respectively [22], do not perform well. The main reason is that concatenating normalized heterogeneous features to a supervector often leads to the scale problem. In addition, the position information is lost during clustering. Similar results can be observed in Table 3 for NMI.

Fig. 1(a) shows the clustering results on the `hall` video. The starting positions of the trajectories are marked in red. The first three figures in Fig. 1(a) show the most dominant three trajectory clusters moving from the main gate to the elevator, from the main gate to the left stair, and from bottom to the left stair, respectively. The last two columns display two individual events which are not merged to others clusters because their starting and ending positions are different from those of other trajectories. By individually considering het-

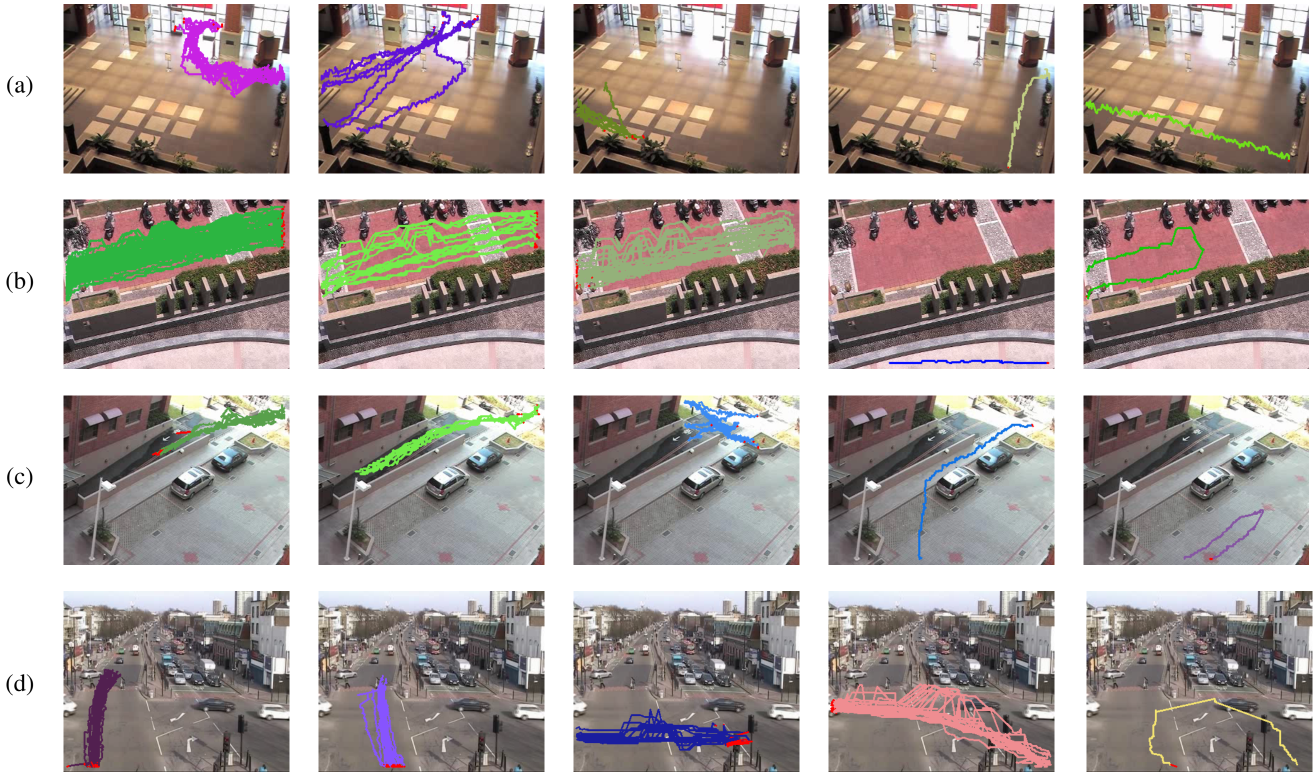

Fig. 1. Five trajectory clusters, one in each column, detected by our approach on four videos, including (a) `hall`, (b) `sidewalk`, (c) `street`, and (d) `QMUL`. See the text for the details.

Table 3. Performance comparison in NMI (%).

	hall	sidewalk	street	QMUL
Vlachos et al. [5]	86.584	63.802	83.120	76.033
Hu et al. [7]	86.804	73.688	83.672	84.890
Hu et al. [8]	82.039	57.254	68.992	67.980
Bashir et al. [9]	78.781	57.798	67.725	68.940
Atev et al. [13]	84.081	60.271	76.675	76.714
TED + AP [16]	65.199	57.864	65.037	67.244
TED + SC [16]	84.691	64.086	74.041	74.288
Ours	**87.698**	**77.523**	**87.506**	**87.811**

erogeneous trajectory representations, our method can separate these distinctive trajectories from the rest. As shown in Fig. 1(b), the trajectories in the first column are extracted from pedestrians, and the trajectories in the second column are extracted from motorcycles and bicycles. Although these two groups have similar starting and ending positions, our method can distinguish them referring to the clustering results related to velocity- and acceleration-based features. The fourth column displays an unusual event walking on the non-sidewalk area. The fifth column gives a trajectory resulting from tracking failure. Although these events are rare, our method can successfully identify them from dominant groups.

The first three columns in Fig. 1(c) show the dominant groups of events moving near the entrance of the underground parking lot. The fourth column illustrates a trajectory of a pedestrian moving straight and then changing his moving direction. The fifth column shows a miss-tracked trajectory. The first four columns of Fig. 1(d) give the dominant groups of events. Because our method considers the starting and ending positions individually, vehicles of different lanes can be separated to two groups as shown in the first two columns. The fifth column shows a trajectory of a vehicle making a U-turn. Such a rare and interesting event is detected by our method.

4. CONCLUSIONS

We propose a novel hierarchical clustering algorithm that groups trajectories in heterogeneous feature representations for event discovery. The proposed method leverages complementary evidence extracted from diverse features, and effectively and efficiently divides trajectories into clusters of high quality. It turns out that our method can not only group trajectories to discovery dominant events but also detect rare events. Our method is evaluated and compared to state-of-the-art methods, and achieves superior results in the experiments. In the future, we will apply the proposed method to applications, such as anomaly detection and object tracking, where high-quality trajectory clusters are appreciated.

5. REFERENCES

[1] Y. H. Lai and C. K. Yang, "Video object retrieval by trajectory and appearance," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 6, pp. 1026–1037, June 2015.

[2] L. Brun, A. Saggese, and M. Vento, "Dynamic scene understanding for behavior analysis based on string kernels," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1669–1681, Oct 2014.

[3] R. Laxhammar and G. Falkman, "Online learning and sequential anomaly detection in trajectories," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1158–1173, June 2014.

[4] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1336–1349, Aug 2014.

[5] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *18th Intl. Conf. on Data Engineering*, 2002, pp. 673–684.

[6] A. Naftel and S. Khalid, "Motion trajectory learning in the dft-coefficient feature space," in *Fourth IEEE Intl. Conf. on Computer Vision Systems (ICVS'06)*, Jan 2006, pp. 47–47.

[7] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang, "An incremental dpmm-based method for trajectory clustering, modeling, and retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1051–1065, May 2013.

[8] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. on Image Processing*, vol. 16, no. 4, pp. 1168–1181, April 2007.

[9] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Realtime motion trajectory-based indexing and retrieval of video sequences," *IEEE Trans. on Multimedia*, vol. 9, no. 1, pp. 58–65, Jan 2007.

[10] C. Ding and X. He, "Linearized cluster assignment via spectral ordering," in *21st Intl. Conf. on Machine Learning*, New York, NY, USA, 2004, ICML '04, ACM.

[11] C. R. Jung, L. Hennemann, and S. R. Musse, "Event detection using trajectory clustering and 4-d histograms," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1565–1575, Nov 2008.

[12] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. on Image Processing*, vol. 18, no. 4, pp. 907–913, April 2009.

[13] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," *IEEE Trans. on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 647–657, Sept 2010.

[14] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2012, pp. 2871–2878.

[15] T. Nawaz, A. Cavallaro, and B. Rinner, "Trajectory clustering for motion pattern extraction in aerial videos," in *IEEE Intl. Conf. on Image Processing*, Oct 2014, pp. 1016–1020.

[16] W.-Y. Huang, W.-C. Wang, C.-L. Chang, W.-A. Wang, and C.-R. Huang, "Trajectory clustering using affinity propagation with trajectory entropy descriptor," in *Intl. Conf. on Industrial Application Engineering*, Mar. 2016, pp. 525–531.

[17] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[18] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric bayesian activity mining and analysis from surveillance video," *IEEE Trans. on Image Processing*, vol. 25, no. 5, pp. 2089–2102, May 2016.

[19] C.-R. Huang, P.-C. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, "Maximum a posteriori probability estimation for online surveillance video synopsis," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1417–1429, Aug 2014.

[20] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong, "Stream-based joint exploration-exploitation active learning," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2012, pp. 1560–1567.

[21] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, Nov 2011.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug 2000.