GEOMETRIC-GUIDED LABEL PROPAGATION FOR MOVING OBJECT DETECTION

Jiun-Yu Kao[12*] *Dong Tian*[1] *Hassan Mansour*[1] *Anthony Vetro*[1] *Antonio Ortega*[2]

[1] Mitsubishi Electric Research Labs (MERL),
201 Broadway, Cambridge, MA 02139, USA
[2] Department of Electrical Engineering, University of Southern California,
3740 McClintock Ave., Los Angeles, CA 90089, USA

ABSTRACT

Moving object segmentation in video has uses in many applications and is a particularly challenging task when the video is acquired by a moving camera. Typical approaches that rely on principal component analysis (PCA) tend to extract scattered sparse components of the moving objects and generally fail in extracting dense object segmentations. In this paper, a novel label propagation framework based on *motion vanishing point* (MVP) analysis is proposed to address the challenges. A weighted graph is constructed with image pixels as nodes and the MVP-guided approach is used to define the graph weights. Label propagation is then performed by incorporating the graph Laplacian. In addition, a PCA result is used to initialize the foreground/background labels. Experiments on the Hopkins data set of outdoor sequences captured by a hand-held moving camera demonstrate that the proposed label propagation method outperforms state-of-the-art PCA and spectral clustering methods for a dense segmentation task. Moreover, the framework is capable of correcting mislabeled foreground pixels and thus does not require accurate initial label assignment.

Index Terms— Motion segmentation, foreground / background separation, motion vanishing point, robust principal component analysis, label propagation

1. INTRODUCTION

Foreground/background (FG/BG) segmentation is the process of separating moving foreground objects from an independent background scene in a video. This is essential for analyzing the moving targets and helps achieve object detection, object segmentation and scene understanding, which is reflected by a vast amount of literature on motion segmentation. However, such segmentation is still a challenging task when there exists ego-motion due to moving camera, which is the focus of this work.

Motion segmentation algorithms can be roughly categorized into statistical techniques, algebraic decomposition techniques and spectral clustering techniques. Statistical approaches alternate between assigning trajectories to subspaces and refitting subspaces to their assigned points as in for example RANSAC methods in [1][2] and EM algorithm in [3].

Algebraic decomposition approaches such as GPCA [4][5] formulate motion segmentation as a problem of subspace separation. Robust PCA is another algebraic decomposition - based alternative with the background scene being modelled as a low-dimensional subspace. It has been shown to successfully segment the foreground objects from the background [6][7] when the camera is stationary. When there exist camera motions, global motion estimation and compensation need to be done first in order for the low rank structure hold for the background scene [8]. Although the segmentation is usually successful, the extracted foreground component highlights the edges of moving objects rather than whole objects, specifically when the image batch size is small, e.g. a batch of two consecutive frames. Also, small background objects, such as electric poles in a traffic scene, have a good chance being identified as foreground due to the limitation in accuracy of global motion compensation.

Spectral clustering-based approaches, regarded as the state-of-the-art in motion segmentation, first use local information to compute the pairwise similarity between keypoint trajectories, from which an affinity matrix is generated. Then spectral clustering [9] is used to cluster the trajectories into independent subspaces. One example is sparse subspace clustering (SSC) which tries to represent each trajectory as a sparse linear combination of the other trajectories [10]. The coefficients associated with each trajectory are used to compute an affinity matrix, which is then used for spectral clustering. The advantage of spectral clustering is the ability to separate the trajectories based on underlying manifolds. However, no semantic meaning is associated to the clusters and the correct choice for the number of clusters is not obvious. As in spectral clustering-based approaches, we make use of a graph derived from an affinity matrix; however, instead of directly applying clustering on the graph spectral domain, we propose to propagate an initial set of semantic labels over

*This work was done when Jiun-Yu Kao worked at MERL.

the similarity graph, which not only achieves better manifold separation but also directly provides a semantic interpretation for the resulting manifolds.

In this paper, we extend the use of affinity matrix computed via the motion vanishing point analysis method [11] into a label propagation framework, where the similarity graph serves as a geometric constraint among moving objects and is used to propagate a set of initial labels. The initial FG/BG labels are generated by adopting the factorized robust matrix completion (FRMC) algorithm [8] which decomposes a group of video frames into a low rank component corresponding to the background scene and a sparse component corresponding to the foreground moving objects. One novel contribution of this paper lies in the proposed label propagation, which combines the advantages of FRMC and spectral clustering. Furthermore, initially mislabeled components are corrected based on neighboring motion information using our proposed framework.

The remainder of this paper is organized as follows. The concept of motion vanishing point and how to compute affinity matrix is briefly reviewed in Section 2. Then the MVP-guided label propagation framework that joins FRMC and graph spectral clustering is proposed in Section 3. Section 4 completes the framework by providing a way to generate initial FG/BG labels via FRMC. Experiments are conducted in Section 5 with conclusions in Section 6.

2. MVP AFFINITY COMPUTATION

In order to capture the manifold geometry for the proposed label propagation scheme, we adopted the *motion vanishing point* analysis in [11] and use it to compute the affinity.

To illustrate the concept of a motion vanishing point, assume two points lie on the same moving object under a 3D coordinate frame, with two corresponding 3D motion vectors. Due to the perspective effect, these two motion vectors will intersect at a motion vanishing point at ∞ in the 3D world. When projected onto the image plane, the two corresponding motion vectors in the image plane will then intersect at a point on the image plane as well, which is denoted as a *motion vanishing point*. We utilize the fact that motion vectors of all points that belong to the same rigid object will share the same motion vanishing point.

Consequently, as shown in Fig. 1, we can define the distance between two motion vectors associated with a pair of pixels as the distance between their corresponding motion vanishing points,

$$d_{p,ij} = \|R_i - R_j\| = \sqrt{(R_{xi} - R_{xj})^2 + (R_{yi} - R_{yj})^2} \quad (1)$$

An undirected graph $G = (V, E)$ is then constructed which consists of a collection of vertices $V = \{1, 2, ..., N\}$ connected by a set of edges $E = \{(i, j, w_{ij})\}$,$i, j \in V$ where (i, j, w_{ij}) denotes the edge between vertices i and j

Fig. 1. **Left**: Example frame from a car driving sequence in [12]. **Center**: Motion vanishing point image for that frame. **Right**: Representation point of a MV and perspective distance between a pair of MVs.

having weights w_{ij}. Vertex i corresponds to pixel i. As for the selection of edge set E, a sparse and regular connectivity is considered where each pixel is only connected to its 4 spatially neighbouring pixels. Finally, the affinity matrix $\mathbf{W} = \{w_{ij}\}$ of the graph is computed as $w_{ij} = e^{-\beta d_{p,ij}} + \epsilon$ where $\beta = 25$ and $\epsilon = 0$ in this work.

3. PROPOSED GUIDING LABEL PROPAGATION

From Section 2, a graph G and its associated affinity matrix $\mathbf{W}$ is built based on geometric analysis for motion vanishing points, where w_{ij} reflects the pairwise similarity between pixel i and j in terms of the corresponding motions. In contrast to [11], where the affinity is directly used for spectral clustering, this paper proposes extending the use of the affinity into a label propagation framework. In such a way, the initial labels could be propagated to all pixels under the guidance of graph G, which is capable of maintaining the geometric constraint determined by the motion vanishing point analysis. The proposed framework can be used with alternative initial label assignment methods as long as they provide reasonable starting point. Section 4 will describe the initialization method we choose in this work.

3.1. MVP-guided label propagation

Let $\mathbf{f}$ denote the vector of labels that are to be assigned and suppose, without loss of generality, that the first l entries in $\mathbf{f}$ are assigned binary labels 0 or 1. We refer to the subset of $\mathbf{f}$ with known labels as $\mathbf{f}_l$. The remaining entries in $\mathbf{f}$ indexed $l + 1, \cdots, N$ are without labels. We refer to the unlabeled subset as $\mathbf{f}_u$. One efficient way to propagate labels on the similarity graph is implemented through an iterative procedure such as in [13], where the propagation is started from nodes with initial labels in their neighborhood and the process is repeated until convergence is reached. Furthermore, it was demonstrated in [14] that such an iterative algorithm could be converted to solving a graph regularization problem. In this paper, we formulate an MVP-guided graph regularization problem in the proposed label propagation framework.

Denote the output estimated labeling for the whole set of pixels by $\hat{\mathbf{f}} = [\hat{\mathbf{f}}_l^T, \hat{\mathbf{f}}_u^T]^T$. Our objective is to leverage consistency among the input labels and smoothness with respect

to the geometric structure of the output labels. Therefore, we aim to minimize the difference between the output labels and initially labeled input, i.e. minimize with respect to $\hat{\mathbf{f}}$ the sum

$$\sum_{j=1}^{l} (\hat{f}_j - f_j)^2 = \|\hat{\mathbf{f}}_l - \mathbf{f}_l\|^2. \quad (2)$$

Here we avoid strict equality in order to allow for a correction of the initial labeling according to the geometric smoothness defined by the graphical structure.

Since the underlying data manifold should be smooth, we wish to promote the smoothness of the estimated labels with respect to the geometry of the graph structure. Thus, we introduce the following regularization term

$$\begin{aligned} &\frac{1}{2}\sum_{i,j=1}^{N} W_{ij}(\hat{f}_i - \hat{f}_j)^2 \\ &= \frac{1}{2}\left(2\sum_{i=1}^{N}\hat{f}_i^2\sum_{j=1}^{N}W_{ij} - 2\sum_{i,j=1}^{N}W_{ij}\hat{f}_i\hat{f}_j\right) \quad (3) \\ &= \hat{\mathbf{f}}^T(\mathbf{D}-\mathbf{W})\hat{\mathbf{f}} \\ &= \hat{\mathbf{f}}^T\mathbf{L}\hat{\mathbf{f}} \end{aligned}$$

where $\mathbf{D}$ is the degree matrix and $\mathbf{L} \triangleq \mathbf{D} - \mathbf{W}$ is the combinatorial Laplacian matrix associated with G.

Therefore, we formulate the label propagation problem as the following box constrained minimization problem

$$\hat{\mathbf{f}} = \arg\min_{\hat{\mathbf{f}}} \|\hat{\mathbf{f}}_l - \mathbf{f}_l\|^2 + \lambda\hat{\mathbf{f}}^T\mathbf{L}\hat{\mathbf{f}}, \quad \text{s.t. } \mathbf{0} \le \hat{\mathbf{f}} \le \mathbf{1} \quad (4)$$

where λ is a regularization parameter.

3.2. Proposed solution

Problem (4) can be recast as the constrained least-squares problem as follows,

$$\begin{aligned} \hat{\mathbf{f}} = \arg\min_{\hat{\mathbf{f}}} &\left\| \begin{pmatrix} \mathbf{f}_l \\ \mathbf{0}_{N\times 1} \end{pmatrix} - \begin{pmatrix} \mathbf{I}_l\ \mathbf{0}_{l\times(N-l)} \\ \lambda\mathbf{L} \end{pmatrix} \hat{\mathbf{f}} \right\| \quad (5) \\ &\text{subject to } \mathbf{0} \le \hat{\mathbf{f}} \le \mathbf{1} \end{aligned}$$

which can be readily solved by a multitude of available optimization toolboxes, e.g., MATLAB Optimization Toolbox.

The estimated labels $\hat{\mathbf{f}}$ assign a continuous valued label between 0 and 1 for every node (i.e. pixel). The larger the $\hat{f}_j$ is, the more likely pixel j belongs to the moving foreground in that picture.

The selection of λ plays an important role in controlling the tradeoff between graph-based smoothness and sample consistency. Fig. 2 shows the impact of choosing different values of λ. We observe that, with a smaller λ, the consistency to the initial labels is more favoured than the smoothness on the graph, and thus has less capability of correcting those initially mislabelled pixels, e.g. the electric pole behind the car. Conversely, using larger λ will favour to maintain the smoothness of estimated labels over the graph and it is inclined to adjust those mislabelled initial pixels. In this work, we choose λ to be 10^3 for all the following experiments considering the initial label quality. However, it is worth mentioning that the selection of lambda may depend on the the means by which the graph is constructed and/or the scheme used for initial label assignment; the approach used in this work is discussed in the next section. For instance, if the initial FG/BG label assignment is quite reliable, then λ should be assigned a smaller value.

Fig. 2. **From left to right:** Solve (5) with $\lambda = 10$, $\lambda = 100$ and $\lambda = 1000$ on *cars1* sequence.

4. INITIAL FG/BG LABEL GENERATION

To complete the proposed MVP-guided label propagation framework, we need an approach to assign initial labels. Here we adopt a factorized version of typical PCA approach as in [8], which solves the following problem,

$$\min_{L,R,S} \frac{1}{2}\|L\|_F^2 + \frac{1}{2}\|R\|_F^2 + \mu\|S\|_1 \text{ s.t. } b = \mathcal{A}(LR^T + S) \quad (6)$$

where LR^T represents the factorized low rank component, and S denotes the sparse component. The above optimization problem can be solved with the scheme proposed in [8]. And the component S of the solution is used to determine whether a pixel belongs to the foreground or background.

Here we propose to generate a set of supervised labels of FG/BG using the following method. For an $m \times n$ image, we define the label vector for an image as $\mathbf{f}$, where $\mathbf{f} \in \{0, 1, u\}^{mn}$. For a pixel j in the image, $f_j \in \{0, 1, u\}$ respectively represents the j-th pixel to be labelled as background, labelled as foreground, or unlabeled. Given the sparse component S from FRMC and a threshold value T, we have $f_j = 1$ if $S_j \ge T$ where S_j is the j-th element of S. In other words, we label those pixels with large values in the sparse component as the foreground similar to the example shown in Fig. 3. Notice that since FRMC generally outputs scattered pixels, these foreground labels are also scattered. It is worth noting that simply thresholding S does not necessarily provide a foreground map since some background objects can still have large values in S.

As for the generation of background labels, any heuristic which selects a subset Γ of pixels out of the set of pixels without sparse components, i.e. $\{j : S_j \neq 0\}$, can be used. For

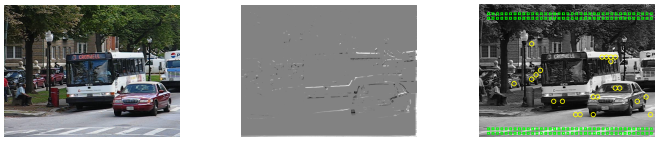

Fig. 3. **Left**: Original picture. **Center**: Sparse component S output from FRMC, $S \in [0, 255]$. **Right**: Initlal label assignment: Squares: Backgrounds. Circles: Foregrounds.

example, the pixels belonging to the n columns and rows at the frame boundaries with no sparse components can be used as shown in Fig. 3. In a more advanced scenario, texture or motion information may also be included as criterion to select the subset Γ. In any case, given a Γ, we have $f_j = 0,\ \forall j \in \Gamma$, which act as the background label.

5. EXPERIMENTS AND DISCUSSIONS

5.1. Experimental setup

We apply the proposed framework to Hopkins 155 Dataset [15]. Ten traffic sequences, *cars1-cars10*, in the dataset that consist of vehicles moving on the street while the scenes are taken by a hand-held moving camera are used. We encode the raw video sequences using the H.265/HEVC test model [16] with default encoding settings. Motion vectors with quarter pixel accuracy are extracted, such that, every motion vector corresponds to a 4×4 pixel block.

For the performance comparison, we first show the segmentation provided by the FRMC algorithm as a representative algebraic decomposition-based approach. The second benchmark is generated using graph spectral clustering on the same affinity computed utilizing the MVP concept as in [11].

5.2. Experimental results and discussions

Fig. 4 shows our experimental results on five sequences. First, we observe that our proposed method is able to significantly adjust those pixels mislabeled as foreground in the background scene compared to using only FRMC, such as the electric pole and non-moving cars. This advantage occurs naturally due to our label propagation scheme because the initially mislabeled pixels are adjusted via the smoothness constraint among their neighbors with large affinity, i.e. experiencing the same motion. Moreover, our proposed method removes all scattered labels that are visible in the FRMC output. It is also clear that the interiors of the moving objects can be compactly labelled as foreground with our method while with FRMC alone only the edges are labelled as foreground.

On the other hand, our proposed framework also outperforms the clustering-based method of [11] in several ways. First, clustering-based methods usually cannot provide semantic meanings to the resulting clusters. However, in our method, as we have the labels generated from sparse and low rank component of data, the resulting separated manifolds

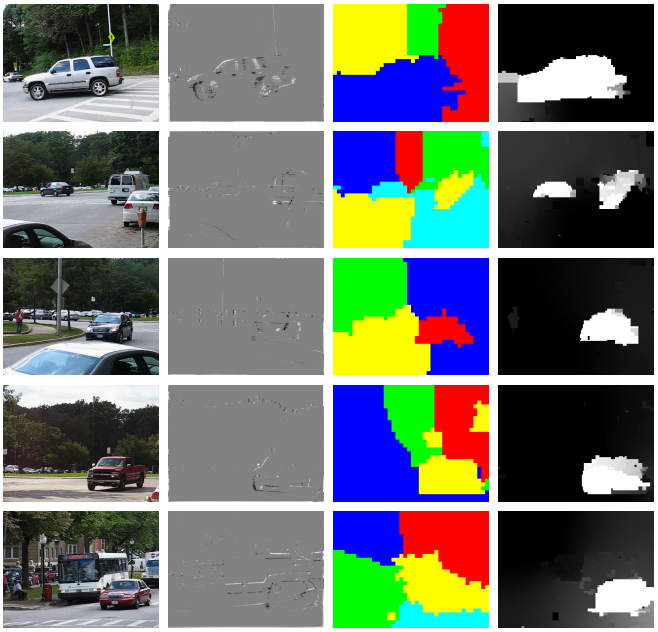

Fig. 4. **From left to right:** Original image, result of using only FRMC as in [8], result of spectral clustering with same affinity defined as in Section 2 [11] and result of our proposed label propagation with FRMC.

possess semantic meaning as FG/BG. Furthermore, the selection for the number of clusters is not obvious. Using a small cluster number (such as 2) will not be able to accommodate the variances introduced by the noise or camera motion and thus lead to a failed segmentation. For the benchmark shown here, the number of clusters is set to be the number of different motions in the scene plus 2. Although the foreground objects are generally successfully separated, it has a problem with over-segmenting the background. Last but not least, the foreground objects detected via our approach have sharper boundaries compared to simple spectral clustering. We attribute this to correctly detecting the boundaries of foreground pixels from FRMC by the proposed initial label generation scheme.

6. CONCLUSION

In this work, a geometrically-guided label propagation framework is proposed to segment moving foreground from background. Our approach performs well with scenes containing significant global motion, such as scene acquired by hand-held moving cameras. An initial label assignment is first realized by a conventional PCA-based method, then a geometric constraint is implemented through the graph affinity that is later embedded into the label propagation scheme. Experiments show that the proposed framework outperforms the PCA method and a spectral-only clustering method. Applying the approach on sequences with even faster camera motion is subject to future work.

7. REFERENCES

[1] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[2] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept. 2009, pp. 1219–1225.

[3] A. Gruber and Y. Weiss, "Multibody factorization with uncertainty and missing data using the em algorithm," in *Computer Vision and Pattern Recognition, 2004*, June 2004, vol. 1, pp. 707–714.

[4] R. Vidal and R. Hartley, "Motion segmentation with missing data using powerfactorization and gpca," in *Computer Vision and Pattern Recognition, 2004*, June 2004, vol. 2, pp. 310–316.

[5] K. Kanatani, "Motion segmentation by subspace separation and model selection," in *Computer Vision, 2001. Proceedings. Eighth IEEE International Conference on*, 2001, vol. 2, pp. 586–591 vol.2.

[6] J. Wright, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems 22*, 2009.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[8] H. Mansour and A. Vetro, "Video background subtraction using semi-supervised robust matrix completion," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6528–6532.

[9] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. 2001, pp. 849–856, MIT Press.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[11] D. Tian, J.-Y. Kao, H. Mansour, and A. Vetro, "Graph spectral motion segmentation based on motion vanishing point analysis," in *2015 IEEE International Workshop on Multimedia Signal Processing (accepted)*, Oct. 2015.

[12] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proceedings of the 10th European Conference on Computer Vision: Part I*, 2008, pp. 44–57.

[13] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., Carnegie Mellon University, 2002.

[14] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, The MIT Press, 1st edition, 2010.

[15] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition, 2007. IEEE Conference on*, June 2007, pp. 1–8.

[16] I.-K. Kim, K. McCann, K. Sugimoto, B. Bross, W. Han, and G. Sullivan, "High efficiency video coding (hevc) test model 10 (hm10) encoder description," Tech. Rep., JCTVC-L1002, Joint Collaborative Team on Video Coding (JCT-VC), 2013.