Low-Rank Spatio-Temporal Video Segmentation

Alasdair Newson
alasdairnewson@gmail.com
Mariano Tepper
http://www.marianotepper.com.ar
Guillermo Sapiro
http://www.ee.duke.edu/faculty/guillermo-sapiro

Duke University
Durham, NC, USA

Recently, a great deal of interest has been generated by the technique known as *Robust Principle Component Analysis* (RPCA) of Candès et al. [1], which addresses the problem of separating a matrix into a low-rank and a sparse component. This very general formulation can be used for tasks such as background estimation in videos and face recognition. In the case of background estimation, the low-rank matrix models the background, and the sparse matrix corresponds to the foreground. A considerable drawback of this approach is its poor robustness to local lighting conditions. If lighting conditions vary locally, one of two things may happen. Either the method incorporates the lighting variation into the foreground, which is clearly undesirable, or the rank of the background model is allowed to increase. Unfortunately, this second option means that the *true* foreground is likely to become included in the background, especially for objects which are static for a short while. Here, we propose to model the background as a *piece-wise low-rank* matrix. In this manner, it will be possible to extract several localised models which correspond to coherent lighting conditions. However, for this we need to *segment* the input video into such coherent regions.

We refer to this problem as a *low-rank spatio-temporal video segmentation*. We present an algorithm to address this segmentation problem, based on region merging and spectral clustering techniques. We show that by carrying out a local RPCA in each region, the results of foreground/background separation are greatly improved, in comparison with both the standard RPCA and several other well-known background estimation techniques.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ represent an input video, in matrix form. Each frame contains m pixels, and there are a total of n frames in our video. The goal of RPCA is to decompose $\mathbf{X}$ as $\mathbf{X} \approx \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ is the low-rank matrix and $\mathbf{S}$ is the sparse matrix. Unfortunately, the rank of a matrix is a non-convex function, so a surrogate function, the *nuclear norm* is used. Thus, the background/foreground separation problem may be formulated as follows:

$$\min_{\mathbf{L},\mathbf{S} \in \mathbb{R}^{m \times n}} \tfrac{1}{2} \|\mathbf{X} - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda_* \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \tag{1}$$

where $\|\mathbf{L}\|_* = \sum_i \sigma_i(\mathbf{L})$ is the nuclear norm of $\mathbf{L}$ and $\sigma_i(\mathbf{L})$ is the i^{th} singular value of $\mathbf{L}$. The scalars λ_* and λ are optimisation parameters, $\|\cdot\|_F$ is the Frobenius matrix norm and $\|\cdot\|_1$ is the ℓ^1 matrix norm, which induces sparsity in the foreground matrix.

To segment $\mathbf{X}$ into different regions where the low-rank requirement is respected, we start by creating a regular 3D grid, which we denote with Ω, on the video domain. Each Ω_i corresponds to a rectangular cuboid of video information. We then create an undirected, weighted graph where each node represents a region Ω_i, and a node is connected with a 6-connectivity to the regions around it. Our goal will be to cluster this graph using spectral clustering techniques. The main challenge here is to design a cost function which shows how "coherent" two regions are in terms of their low-rank background representation.

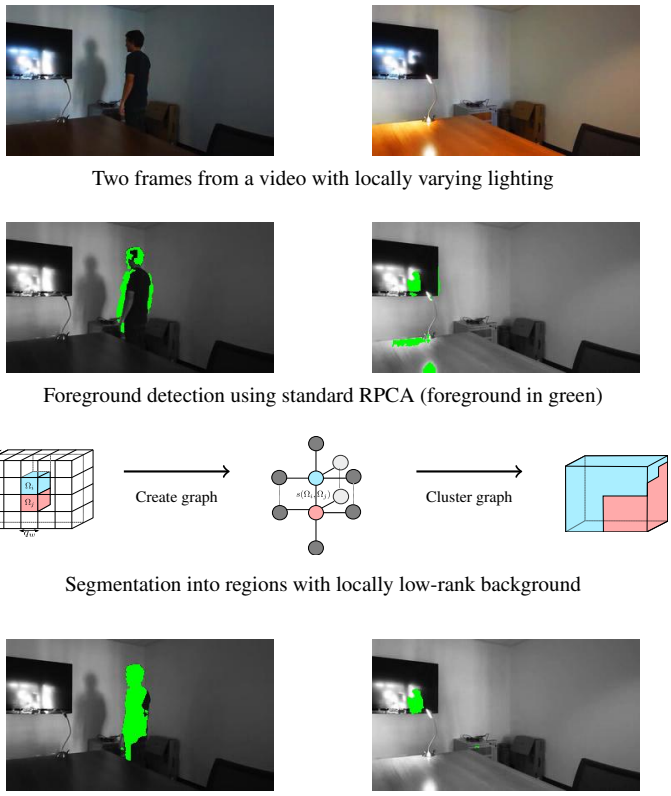
More formally, consider two regions to merge, Ω_i and Ω_j. We wish to see whether it is better to decompose the regions separately or jointly. The decomposition of Ω_i will be $\Omega_i \approx \mathbf{L}_i + \mathbf{S}_j$, and similarly for Ω_j. Our first observation is that it is easier to compare the coherence of the decompositions resulting from a *rank-constrained* version of Equation (1):

$$\{\mathbf{L}_i, \mathbf{S}_i\} = \underset{\mathbf{L},\mathbf{S}}{\operatorname{argmin}} \tfrac{1}{2} \|\mathbf{X}_i - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda \|\mathbf{S}\|_1 \quad \text{subject to} \quad \operatorname{rank}(\mathbf{L}) \leq r. \tag{2}$$

The comparisons are made clearer because the λ_* parameter is removed and replaced with one which is more easily interpretable, the maximum rank of each local model, r. Once the decompositions of Ω_i, Ω_j and $\Omega_{i \cup j}$ are obtained in this manner, we can calculate the cost of merging the two regions. Let $e_i = \|\mathbf{X}_i - \mathbf{L}_i - \mathbf{S}_i\|_F^2$ be the quadratic error of the decomposition of Ω_i, and similarly for Ω_j and $\Omega_{i \cup j}$. Our cost function is:

$$d(\Omega_i, \Omega_j) = \frac{|e_i + e_j - e_{i \cup j}|}{\phi_{i \cup j}}. \tag{3}$$

where $\phi_{i \cup j}$ is a scalar. Once we have established the cost of merging two regions, we convert it into a *similarity* cost, and cluster the resulting graph using robust spectral clustering techniques [5].



Two frames from a video with locally varying lighting

Foreground detection using standard RPCA (foreground in green)

Segmentation into regions with locally low-rank background

Foreground detection using (proposed) local RPCA

Figure 1: Illustration and results of the algorithm

Figure 1 illustrates the problems caused by locally varying lighting conditions: either the foreground is merged into the background (second row, left), or the global (standard) RPCA is not able to represent local lighting changes (second row, right). This is corrected by segmenting the video, and carrying out a local RPCA in each region. We compare our algorithm qualitatively and quantitatively with respect to several algorithms of the literature [2, 3, 4] and find greatly improved performance in challenging situations.

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 2011.

[2] L. Maddalena and A. Petrosino. A Fuzzy Spatial Coherence-based Approach to Background/Foreground Separation for Moving Object Detection. *Neural Computing & Applications*, 2010.

[3] C. Stauffer and W. E. L. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. In *CVPR*, 1999.

[4] J. Yao and J-M. Odobez. Multi-Layer Background Subtraction Based on Color and Texture. In *CVPR*, 2007.

[5] L. Zelnik-Manor and P. Perona. Self-tuning Spectral Clustering. In *Advances in Neural Information Processing Systems*, 2004.