Video Segmentation with Global Optimality Guarantees

Brendon Anderson

17 May 2019

Abstract

In this report, we consider the problem of moving object detection in a video. Specifically, we study the video's decomposition into foreground and background layers by formulating the segmentation as a nonconvex optimization problem. Although the resulting problem is more computationally tractable than its commonly employed convex relaxation, it is not generally solvable to global optimality. In spite of this limitation, we derive conditions on the video data under which the uniqueness and global optimality of the foreground segmentation are guaranteed. We illustrate these novel optimality criteria through synthetically generated video examples.

1 Introduction

One of the most fundamental problems in the field of computer vision is that of video segmentation. In this line of research, the typical goal is to recognize moving objects in a video in order to extract the foreground and recover the background. See for example, Figure 1. Popular traditional approaches to solve this problem include mixtures of Gaussians (MOG) [1, 2], which offer simple models, and neural networks [3], which perform well and provide computational efficiency [4]. Recently, more attention has been placed on approaches based on robust principal component analysis (RPCA), which model the video as the sum of low-rank and sparse matrices. Perhaps the most notable of these methods is Principle Component Pursuit (PCP) introduced in the seminal paper by Candès [5]. Although the convexified approach in PCP provides conditions under which exact recovery of the sparse components is guaranteed, its use of lifted variables results in scalability and computational hindrances.

Figure 1: A video frame (left) and its segmented foreground (right). [6]

In order to tackle large-scale segmentation problems, nonconvex formulations such as robust nonnegative matrix factorization (RNMF) have been proposed [7, 8, 9]. These nonconvex approaches to low-rank matrix factorization often permit parallelization, lending themselves to lowered computational cost and scalability to larger problems [10]. Furthermore, the nonnegative nature of grayscale pixel values is explicitly enforced in modern methods like RNMF, unlike many of the more traditional techniques. Although RNMF has been empirically shown to have performance on par with the popular PCP method, previous works have only considered local optimality of the solutions to the nonconvex RNMF problem, often solved for by alternating over its subproblems which are convex in the variables separately [7, 8].

In this work, we aim to supplement the strong empirical and computational properties of video segmentation via nonconvex RNMF by filling this gap in global optimality guarantees. Guarantees such as these are necessary in a variety of safety-critical video segmentation applications, such as autonomous driving [11]. We approach this problem by exploiting new results on the benign landscape of the rank-1 RNMF problem [6]. Under this framework, we propose criteria under which the video segmentation is guaranteed to be unique and globally optimal.

The remainder of this report is structured as follows. In Section 2 we describe the problem and introduce our common terminology and notation. In Section 3, we show that the problem can be simplified to one in which the moving objects consist of elementary shapes. Then, in Sections 4 and 5, we derive conditions on video data under which global optimality guarantees can be made. Finally, we showcase possible extensions to this work and make our concluding remarks in Section 6.

2 Problem Statement

Consider a video defined by a sequence of frames given by the matrices $X^{(k)} \in \mathbb{R}^{d_m \times d_n}$, for $k \in K \coloneqq \{1, 2, \ldots, d_f\}$. By defining the *pixel set* as $\Pi = \{1, 2, \ldots, d_m\} \times \{1, 2, \ldots, d_n\}$, the pixels of a grayscale video are given by

$$X_{ij}^{(k)} \in [1, 256], \quad (i, j) \in \Pi, \ k \in K.$$

Vectorizing each frame of the video, we form the data matrix

$$X = \begin{bmatrix} \operatorname{vec} X^{(1)} & \operatorname{vec} X^{(2)} & \cdots & \operatorname{vec} X^{(d_f)} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

where $m = d_m d_n$ and $n = d_f$. Let us also define the *measurement set* as $\Omega = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$.

We choose to model the video as the sum of two components. The first component is chosen to be a nonnegative rank-1 matrix, used to capture the relatively static behavior of the video's background. The second component is a sparse matrix, taken to represent the dynamic foreground (i.e. the moving object). Under this model, we seek the decomposition

$$X \approx uv^\top + S, \tag{1}$$

where $u \in \mathbb{R}^m_+$ and $v \in \mathbb{R}^n_+$, and $S \in \mathbb{R}^{m \times n}$ is sparse. This can be solved for through the following nonconvex, nonnegative l_1-minimization problem, commonly termed *robust nonnegative matrix factorization* (RNMF):

$$\begin{array}{ll} \text{minimize} & \|X - uv^\top\|_1 \\ \text{subject to} & u \in \mathbb{R}^m_+, \ v \in \mathbb{R}^m_+. \end{array} \tag{2}$$

Note that we enforce nonnegativity of the optimization variables u and v, yielding natural interpretations as the video's nominal background pattern and its associated scalings in each frame, respectively.

Under the decomposition (1), we define the video's *background set* and *foreground set* as $B = \{(h,k) \in \Omega : S_{hk} = 0\}$ and $F = \Omega \backslash B$, respectively. Accordingly, two bipartite graphs can be introduced, the *background graph* $\mathcal{G}_{m,n}(B)$ having edge set B, and the *foreground graph* $\mathcal{G}_{m,n}(F)$ having edge set F. The first vertex set of each graph corresponds to pixel numbers: $V_u = \{1, 2, \ldots, m\}$. The second vertex set associates with frame numbers: $V_v = \{m+1, m+2, \ldots, m+n\}$. A toy example of these graphs is given below.

Example 2.1. Suppose a video has frames given by $X^{(1)} = \begin{bmatrix} 1 & 256 \\ 256 & 256 \end{bmatrix}$, $X^{(2)} = \begin{bmatrix} 256 & 1 \\ 256 & 256 \end{bmatrix}$, and $X^{(3)} = \begin{bmatrix} 256 & 256 \\ 256 & 1 \end{bmatrix}$, where elements of 256 represent background. Then the data matrix is

$$X = \begin{bmatrix} 1 & 256 & 256 \\ 256 & 256 & 256 \\ 256 & 1 & 256 \\ 256 & 256 & 1 \end{bmatrix},$$

and the foreground and background sets are $F = \{(1,1), (3,2), (4,3)\}$ and $B = \{1,2,3,4\} \times \{1,2,3\} \backslash F$, respectively. The corresponding graphs are shown in Figure 2.

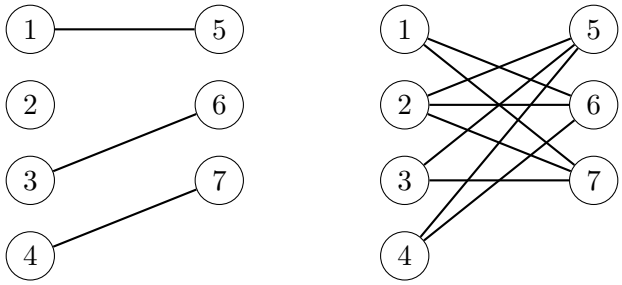

Figure 2: Example graphs $\mathcal{G}_{m,n}(F)$ (left), and $\mathcal{G}_{m,n}(B)$ (right).

□

Now, a remarkable property of the nonconvex problem (2) is that, under certain conditions on the problem data, the optimization landscape is *benign*, i.e. there are no spurious local minima, and the global minimum is unique [6]. This permits the use of simple local search algorithms to solve (2) to global optimality. The sufficient conditions for benign landscape are:

$$\text{Connectivity:} \quad \mathcal{G}_{m,n}(B) \text{ is connected.} \tag{3}$$

$$\text{Identifiability:} \quad \delta(\mathcal{G}_{m,n}(B)) > \frac{48}{c^2} \kappa(w^*)^4 \Delta(\mathcal{G}_{m,n}(F)). \tag{4}$$

In these expressions, we denote the optimal solution of (2) as $w^* = (u^*, v^*)$, the condition number (maximum element divided by minimum element) of a vector in the positive orthant as $\kappa(\cdot)$, maximum degree of a graph as $\Delta(\cdot)$, and minimum degree of a graph as $\delta(\cdot)$. The value c is a constant that depends on problem data which will be discussed in more detail later.

The problem we address is as follows: *when do videos satisfy the conditions (3) and (4) for benign landscape of the optimization problem (2)?* In other words, we would like to determine

conditions on the size, shape, and speed of a moving object which provide theoretical guarantees for the unique and globally optimal foreground segmentation of a video. We begin by showing the problem can be simplified to one with elementary foreground shapes through the notion of object embedding.

3 Object Embedding

In this section, we consider two videos with identical backgrounds. We are interested in the case that the moving object of one video can be completely covered by the moving object of the other video in each frame. The question we would like to address is, *if the video with the larger moving object satisfies the conditions (3) and (4) for benign landscape of (2), does the video with the smaller object also satisfy these conditions?* To answer this question precisely, let us start with the following definition.

Definition 3.1 (Embedding)**.** Consider two videos $\mathcal{O}$ and $\mathcal{R}$ having the same background $uv^\top$, i.e. $X_\mathcal{O} = uv^\top + S_\mathcal{O}$ and $X_\mathcal{R} = uv^\top + S_\mathcal{R}$. We say that object $F_\mathcal{O}$ is *embedded* in object $F_\mathcal{R}$ if the foreground of video $\mathcal{O}$ is a subset of that of video $\mathcal{R}$ in every frame; if

$$F_\mathcal{O}^{(k)} \subseteq F_\mathcal{R}^{(k)} \text{ for all } k \in K,$$

where $F_\mathcal{O}^{(k)} = \{(i,j) \in \Pi : (S_\mathcal{O})_{hk} \neq 0,\ h = (j-1)d_m + i\}$, and similarly for $F_\mathcal{R}^{(k)}$. □

We are now in a position to show that the answer to our earlier question is affirmative. We prove these implications in the following two propositions.

Proposition 3.2 (Embedded connectivity)**.** *If object $F_\mathcal{O}$ is embedded in object $F_\mathcal{R}$ and video $\mathcal{R}$ satisfies the connectivity condition (3), then video $\mathcal{O}$ also satisfies the connectivity condition.*

Proof. Since $F_\mathcal{O}$ is embedded in $F_\mathcal{R}$, we have for all $k \in K$ that $F_\mathcal{O}^{(k)} \subseteq F_\mathcal{R}^{(k)}$, which implies $\Pi \setminus F_\mathcal{R}^{(k)} \subseteq \Pi \setminus F_\mathcal{O}^{(k)}$. This shows the background of video $\mathcal{R}$ is a subset of that of video $\mathcal{O}$, i.e. $B_\mathcal{R}^{(k)} \subseteq B_\mathcal{O}^{(k)}$ for all $k \in K$, which gives $B_\mathcal{R} \subseteq B_\mathcal{O}$. Therefore, we have that $\mathcal{G}_{m,n}(B_\mathcal{R})$ is a spanning subgraph of $\mathcal{G}_{m,n}(B_\mathcal{O})$. Since $\mathcal{G}_{m,n}(B_\mathcal{R})$ is connected by our assumption, so must be $\mathcal{G}_{m,n}(B_\mathcal{O})$, as desired. □

Proposition 3.3 (Embedded identifiability)**.** *If object $F_\mathcal{O}$ is embedded in object $F_\mathcal{R}$ and video $\mathcal{R}$ satisfies the identifiability condition (4), then video $\mathcal{O}$ also satisfies the identifiability condition.*

Proof. Since $F_\mathcal{O}$ is embedded in $F_\mathcal{R}$, we have for all $k \in K$ that $F_\mathcal{O}^{(k)} \subseteq F_\mathcal{R}^{(k)}$, which implies $F_\mathcal{O} \subseteq F_\mathcal{R}$. Hence, the maximum degrees of the foreground graphs satisfy $\Delta(\mathcal{G}_{m,n}(F_\mathcal{O})) \leq \Delta(\mathcal{G}_{m,n}(F_\mathcal{R}))$. Similarly, we have that $B_\mathcal{R} \subseteq B_\mathcal{O}$, and therefore the minimum degrees of the background graphs satisfy $\delta(\mathcal{G}_{m,n}(B_\mathcal{R})) \leq \delta(\mathcal{G}_{m,n}(B_\mathcal{O}))$. Combining these inequalities with the identifiability inequality for video $\mathcal{R}$ yields

$$\delta(\mathcal{G}_{m,n}(B_\mathcal{O})) \geq \delta(\mathcal{G}_{m,n}(B_\mathcal{R})) > \frac{48}{c^2}\kappa(w^*)^4\Delta(\mathcal{G}_{m,n}(F_\mathcal{R})) \geq \frac{48}{c^2}\kappa(w^*)^4\Delta(\mathcal{G}_{m,n}(F_\mathcal{O})),$$

showing video $\mathcal{O}$ also satisfies the identifiability condition. □

It is clear that Propositions 3.2 and 3.3 are independent of the size, shape, and speed of a moving object. This allows us to restrict the rest of our analysis to videos with moving objects of elementary shapes, since a more complicated object may always be embedded into a larger object

which covers it. In the case that the larger, simpler object is found to satisfy the conditions (3) and (4), the results of this section show the embedded object can be extracted to unique global optimality. Therefore, we will focus on rectangular moving objects for the remainder of this report, for convenience.

4 Conditions for Connectivity

Remark. In this section, we aim to derive necessary and sufficient criteria for a video to satisfy the connectivity condition (3). Alongside these conditions, we provide toy examples to clearly demonstrate the main ideas being presented. These examples use synthetic video data generated using MATLAB. In each example, we solve the optimization problem (2) using stochastic gradient descent, initialized at a point $u \in \mathbb{R}^m$ randomly chosen from the uniform distribution, and at $v = \mathbb{1}_n$.

Let us start by defining the notion of connected backgrounds.

Definition 4.1 (Background connectivity)**.** Given a video where each frame has associated background pixel set

$$B^{(k)} = \{(i,j) \in \Pi : S_{hk} = 0,\ h = (j-1)d_m + i\},\ k \in K,$$

the video is said to have a *connected background* if the following two conditions are satisfied:

1. $\cup_{k\in K} B^{(k)} = \Pi$.
2. $B_1 \cap B_2 \neq \emptyset$ for any $B_1 = \cup_{k\in K_1} B^{(k)}$ and $B_2 = \cup_{k\in K_2} B^{(k)}$ such that $K_1 \cup K_2 = K$. □

We now show that having a connected background is equivalent to the video's background graph $\mathcal{G}_{m,n}(B)$ being connected; videos with connected backgrounds satisfy the connectivity condition (3). This is useful, since we will use Definition 4.1 to derive simple and intuitive necessary conditions a video must satisfy in order to have a connected background (and therefore to satisfy the connectivity condition). Afterwards, we prove a sufficient condition for background connectivity which we claim is likely satisfied for nearly any video in practice.

Proposition 4.2 (Connectivity equivalence)**.** *A video's associated background graph, $\mathcal{G}_{m,n}(B)$, is connected if and only if the video has a connected background.*

Proof. The proof will proceed via contrapositive argument. We will first prove necessity.

[Necessity.] Suppose a video does not have a connected background. Then one of the two following cases must hold:

1. $\cup_{k\in K} B^{(k)} \neq \Pi$.
2. There exists $B_1 = \cup_{k\in K_1} B^{(k)}$ and $B_2 = \cup_{k\in K_2} B^{(k)}$, where $K_1 \cup K_2 = K$, such that $B_1 \cap B_2 = \emptyset$.

Assume the first case holds. Then there exists a pixel $(i_0, j_0) \in \Pi$ such that $(i_0, j_0) \notin B^{(k)}$ for any $k \in K$. Therefore, we have $S_{h_0 k} \neq 0$ where $h_0 = (j_0 - 1)d_m + i_0$, which implies

$$(h_0, k) \notin B \text{ for all } k \in K.$$

This shows that vertex $h_0 \in V_u$ has no incident edges in $\mathcal{G}_{m,n}(B)$, and therefore the graph is disconnected.

Now assume the second case holds. We first note that $K_1 \cap K_2 = \emptyset$, since otherwise B_1 and B_2 cannot be disjoint. Now, $B_1 \cap B_2 = \emptyset$ implies that for all pixels $(i,j) \in \Pi$, either $(i,j) \in B_1$ and $(i,j) \notin B_2$, or $(i,j) \in B_2$ and $(i,j) \notin B_1$, or $(i,j) \notin B_1$ and $(i,j) \notin B_2$. In the trivial case that some pixel (i_0, j_0) is neither an element of B_1 nor B_2, then $\cup_{k\in K} B^{(k)} = B_1 \cup B_2 \neq \Pi$, and the first case above shows the graph $\mathcal{G}_{m,n}(B)$ is disconnected. For pixels $(i,j) \in B_1$, we have $(i,j) \notin B^{(k)}$ for any $k \in K_2$, and therefore $S_{hk} \neq 0$ where $h = (j-1)d_m + i$, which implies

$$(h,k) \notin B \text{ for all } k \in K_2.$$

This shows that vertex $h \in V_u$ is not adjacent to vertex $m + k \in V_v$ for any $k \in K_2$. Similarly, one can show that for any $(i,j) \in B_2$, the corresponding vertex $h \in V_u$ is not adjacent to vertex $m + k \in V_v$ for any $k \in K_1$. Since $B_1 \cap B_2 = \emptyset$ and $K_1 \cap K_2 = \emptyset$, the bipartite graph $\mathcal{G}_{m,n}(B)$ contains at least two connected components, defined by the edge sets $\mathcal{E}_1 \subseteq \{(h, m+k) : h = (j-1)d_m + i,\ (i,j) \in B_1,\ k \in K_1\}$ and $\mathcal{E}_2 \subseteq \{(h, m+k) : h = (j-1)d_m + i,\ (i,j) \in B_2,\ k \in K_2\}$. Therefore, the graph is disconnected.

[Sufficiency.] Suppose a video's associated background graph, $\mathcal{G}_{m,n}(B)$, is disconnected. Then one of the two following cases must hold:

1. There exists a vertex with no incident edges.
2. Every vertex has at least one incident edge.

Assume the first case holds. Then either the isolated vertex corresponds to a pixel number h_0 or to a frame number k_0. If vertex $h_0 \in V_u$ is isolated, then $(h_0, k) \notin B$ for all $k \in K$. This implies $S_{h_0 k} \neq 0$ and therefore $(i_0, j_0) \notin B^{(k)}$ for any $k \in K$, where

$$(i_0, j_0) = \left(h_0 - \left\lfloor \frac{h_0}{d_m} \right\rfloor d_m, \left\lfloor \frac{h_0}{d_m} \right\rfloor + 1 \right).$$

[This formula comes from the one-to-one correspondence between a pixel (i,j) and its pixel number h through the vectorization of a given video frame.] Thus, $(i_0, j_0) \notin \cup_{k\in K} B^{(k)}$ which implies $\cup_{k\in K} B^{(k)} \neq \Pi$. Hence the video does not have a connected background. On the other hand, if vertex $k_0 \in K$ is isolated, then $(h, k_0) \notin B$ for all $h \in V_u$. This implies $S_{hk_0} \neq 0$ and therefore $(i,j) \notin B^{(k_0)}$ for any $(i,j) \in \Pi$. Thus, $B^{(k_0)} = \emptyset$. Define $B_1 = B^{(k_0)}$ and $B_2 = \cup_{k\in K\setminus\{k_0\}} B^{(k)}$. Then $B_1 \cap B_2 = \emptyset$, so again the video does not have a connected background.

Now assume the second case from above holds. Then the graph contains at least two nontrivial connected components. Therefore, the set B, which defines the edge set of the graph, can be partitioned as $B = Q_1 \cup Q_2$, where $Q_1 = \{(h,k) \in \Omega : S_{hk} = 0,\ h \in H_1,\ k \in K_1\}$ and $Q_2 = \{(h,k) \in \Omega : S_{hk} = 0,\ h \in H_2,\ k \in K_2\}$ are nonempty, such that $H_1 = V_u \setminus H_2$ and $K_1 = K \setminus K_2$. Now, define $B_1 = \cup_{k\in K_1} B^{(k)}$. This gives

$$\begin{aligned} B_1 &= \cup_{k\in K_1} \{(i,j) \in \Pi : S_{hk} = 0,\ h = (j-1)d_m + i\} \\ &= \{(i,j) \in \Pi : S_{hk} = 0,\ h = (j-1)d_m + i,\ k \in K_1\}. \end{aligned}$$

Now, from the partitions Q_1 and Q_2 we see that a frame $k \in K_1$ has $S_{hk} = 0$ only for pixel numbers $h \in H_1$. Thus, B_1 can be written equivalently as

$$B_1 = \{(i,j) \in \Pi : S_{hk} = 0,\ h = (j-1)d_m + i,\ k \in K_1,\ h \in H_1\}.$$

Similarly, it can be shown that by defining $B_2 = \cup_{k\in K_2} B^{(k)}$, we obtain

$$B_2 = \{(i,j) \in \Pi : S_{hk} = 0,\ h = (j-1)d_m + i,\ k \in K_2,\ h \in H_2\}.$$

Since $H_1 \cap H_2 = \emptyset$ and $K_1 \cap K_2 = \emptyset$, we immediately see that $B_1 \cap B_2 = \emptyset$. Therefore, the video does not have a connected background. □

Proposition 4.2 shows that the connectivity of the graph $\mathcal{G}_{m,n}(B)$ is entirely dictated by whether or not a video has a connected background. Therefore, we can use the notion of background connectivity to derive intuitive and meaningful criteria a video should satisfy in order to meet the connectivity condition (3).

4.1 Necessary Conditions for Connectivity

From Definition 4.1, we develop three necessary conditions for background connectivity of a video, which are intuitively interpretable in terms of properties of the video (i.e. properties of pixels and frames). These necessary conditions give simple methods for showing when a video does not have a connected background, in which case no guarantees on the global optimality of the minimization (2) can be made.

Proposition 4.3 (Pixel connectivity; necessary)**.** *If a video has a connected background, then each pixel is a background pixel in at least one frame.*

Proof. Suppose there exists a pixel $(i_0, j_0) \in \Pi$ which is not a background pixel for any frame $k \in K$. Then $(i_0, j_0) \notin B^{(k)}$ for all $k \in K$. Thus, $(i_0, j_0) \notin \cup_{k\in K} B^{(k)}$ which implies $\cup_{k\in K} G^{(k)} \neq \Pi$, and therefore the video does not have a connected background. □

Proposition 4.3 shows that if any single pixel remains as part of the foreground throughout the video's duration, we cannot guarantee benign landscape of (2). This makes sense intuitively: if part of the background remains obscured throughout the video's entirety, it appears implausible to guarantee unique and globally optimal recovery of that part of the background. We demonstrate these ideas in the example that follows.

Example 4.4. Consider two videos with white rectangles oscillating across a gradient-like background, as shown in Figure 3. In the first video, the rectangle is oriented vertically so that every background pixel is unobscured in at least one frame. In the second video, the rectangle traverses the same trajectory, but since it is oriented horizontally, part of the background remains obscured throughout the video's duration. Proposition 4.3 shows that the horizontal rectangle video does not have a connected background. Therefore, we immediately know that the horizontal rectangle video does not satisfy the sufficient conditions for benign landscape of (2), and therefore there are no guarantees regarding the uniqueness and global optimality of its resulting decomposition. We see this issue manifest itself in a poor background extraction, as shown in the right column of Figure 3. On the other hand, the vertical rectangle video does have a connected background (see Proposition 4.9).

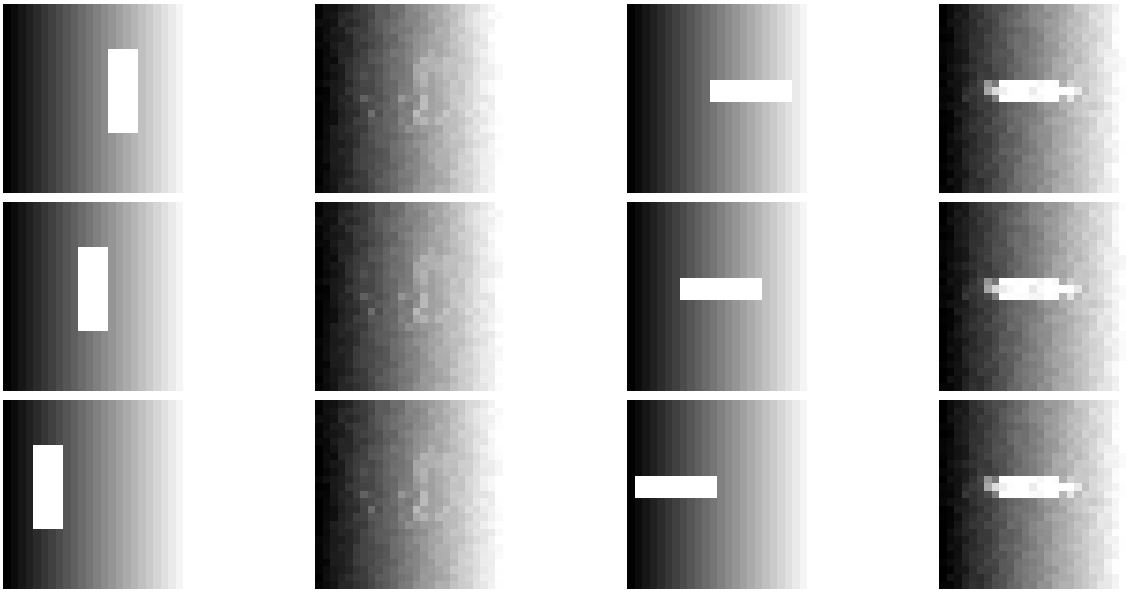

Figure 3: Three frames of the vertical rectangle video (left), its extracted background (middle-left), horizontal rectangle video (middle-right), and its extracted background (right). In this example, the connectivity condition (3) is satisfied by the vertical rectangle video, but not by the horizontal rectangle video.

□

Proposition 4.5 (Frame connectivity; necessary)**.** *If a video has a connected background, then each frame contains at least one background pixel.*

Proof. Suppose there exists a frame $k_0 \in K$ which contains no background pixels, i.e. $B^{(k_0)} = \emptyset$. Then the video's background pixel sets can be partitioned as $B_1 = B^{(k_0)} = \emptyset$ and $B_2 = \cup_{k \in K \setminus \{k_0\}} B^{(k)}$. Thus, $B_1 \cap B_2 = \emptyset$, and therefore the video does not have a connected background. □

Proposition 4.5 can be interpreted as the requirement that the object can at no point cover the entirety of the frame. This matches intuition, since a moving object surely cannot be uniquely segmented from its background in these types of frames. Consider the following example.

Example 4.6. Consider again two videos with white rectangles oscillating across a gradient-like background, as shown in Figure 4. In the first video, the rectangle remains a constant size so that at no point is the background completely covered by the foreground. In the second video, the object traverses the same trajectory, but at just one instant in time the foreground is enlarged to cover the entire background (i.e. the middle frame). Proposition 4.5 shows that the enlarged size video does not have a connected background. Similar to Example 4.4, we conclude that the enlarged size video has no guarantees for the uniqueness and global optimality of its decomposition. As shown in the right column of Figure 4, this issue can result in lower quality background extraction.

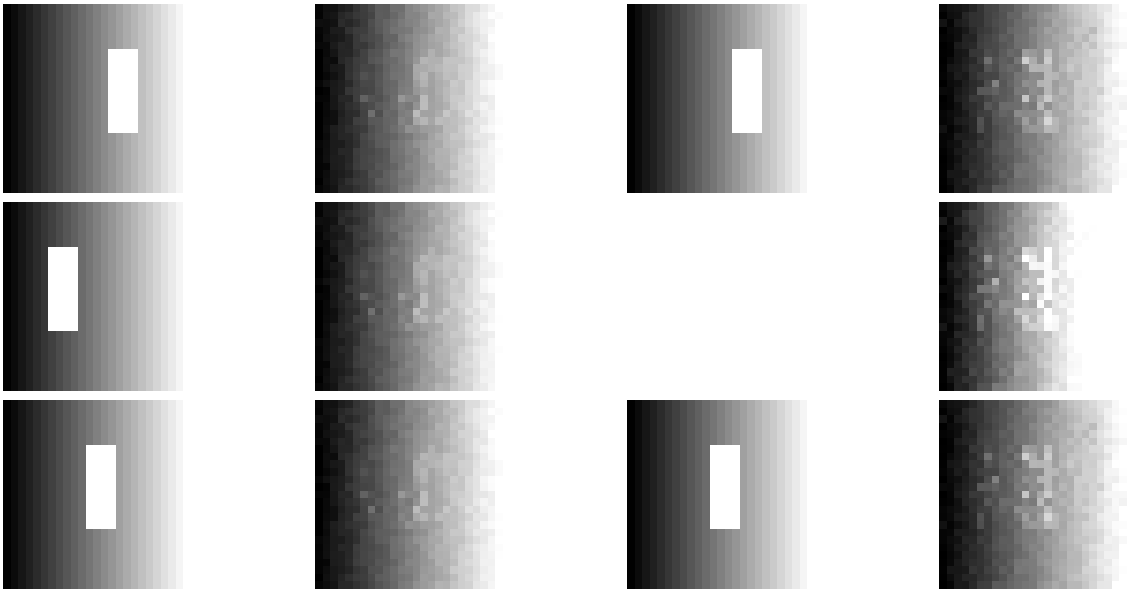

Figure 4: Three frames of the constant size video (left), its extracted background (middle-left), enlarged size video (middle-right), and its extracted background (right). In this example, the connectivity condition (3) is satisfied by the constant size video, but not by the enlarged size video.

□

Proposition 4.7 (Object size; necessary)**.** *If a video has a connected background, then there are at most* $d_m d_n d_f - (d_m d_n + d_f - 1)$ *foreground pixels in the data matrix* X*.*

Proof. Since the background graph $\mathcal{G}_{m,n}(B)$ has $m + n = d_m d_n + d_f$ vertices and is connected, the number of edges $|B|$ is at least $d_m d_n + d_f - 1$. Therefore, $|F| = mn - |B| \leq d_m d_n d_f - (d_m d_n + d_f - 1)$. □

Though Proposition 4.7 gives a simple upper bound on the object's relative size, necessary for unique and globally optimal decomposition in (2), perhaps its most interesting implication comes from the following corollary.

Corollary 4.8. *As the video resolution and number of frames increase, the size of recognizable objects increases.*

Proof. Since there can be at most $d_m d_n d_f - (d_m d_n + d_f - 1)$ foreground pixels across all frames of the video, the maximum relative size of the object can be expressed as

$$p_{\max} \coloneqq \frac{d_m d_n d_f - (d_m d_n + d_f - 1)}{d_m d_n + d_f - 1}.$$

As the resolution of the video increases, $d_m d_n \to \infty$, and therefore

$$\lim_{d_m d_n \to \infty} p_{\max} = d_f - 1.$$

Furthermore, as the length of the video increases, $d_f \to \infty$, and therefore

$$\lim_{\substack{d_m d_n \to \infty \\ d_f \to \infty}} p_{\max} = \infty.$$

Thus, we see that the maximum permissible ratio of foreground pixels to background pixels increases to with the video's resolution and number of frames, as desired. □

Interestingly, the maximum relative object size $p_{\max}$ also shows us that with $d_f = 1$ frame (i.e. a single picture), the largest recognizable object size decreases to $p_{\max} = 0$. On the other hand, with $d_m d_n = 1$ (i.e. a single pixel resolution), the largest recognizable object again decreases to $p_{\max} = 0$. In other words, we cannot recognize moving objects with only one frame, even with infinite resolutions, and we also cannot recognize objects with only one pixel, even with infinitely many frames. Both of these observations align with the restrictions on video properties one would expect.

4.2 Sufficient Conditions for Connectivity

The necessary conditions derived in Section 4.1 are most useful in determining when the global optimality guarantees for (2) *fail* to hold. In this section, we reverse the implications to derive a simple and relatively relaxed sufficient condition for ensuring the graph $\mathcal{G}_{m,n}(B)$ is connected.

Proposition 4.9 (Common background pixel; sufficient)**.** *Suppose each pixel of a video is a background pixel in at least one frame. If any single pixel is a background pixel in all frames of the video, then the video has a connected background.*

Proof. Since each pixel in the video is assumed to be a background pixel in at least one frame, there exists a $k_0 \in K$ such that $(i, j) \in B^{(k_0)}$ for all $(i, j) \in \Pi$. Therefore, $B^{(k_0)} = \Pi$ which implies $\cup_{k \in K} B^{(k)} = \Pi$, so the video satisfies the first condition for background connectivity.

Now, suppose there exists a pixel $(i_0, j_0) \in \Pi$ such that (i_0, j_0) is a background pixel in all frames of the video. Furthermore, suppose we partition the background pixels as $B_1 = \cup_{k \in K_1} B^{(k)}$ and $B_2 = \cup_{k \in K_2} B^{(k)}$, where K_1 and K_2 are any two arbitrary subsets of K such that $K_1 \cup K_2 = K$. Since $(i_0, j_0) \in B^{(k)}$ for all $k \in K$, it must be that $(i_0, j_0) \in B_1$ and $(i_0, j_0) \in B_2$, and therefore $B_1 \cap B_2 \neq \emptyset$. Since B_1 and B_2 are arbitrary partitions, the video satisfies the second condition for background connectivity. Thus, the video has a connected background. □

The sufficient condition given in Proposition 4.9 is relaxed in the sense that many videos satisfy the property of having at least one common background pixel among all frames. These common background pixels are often found in the corners of a video, away from the "action" of the moving object. Therefore, with the *a priori* knowledge that a single pixel remains unobscured by the moving object throughout the duration of the video, the connectedness of the video's background (and therefore the connectedness of $\mathcal{G}_{m,n}(B)$) comes at only the price of ensuring that no single pixel is obscured by foreground throughout the video's entirety. This is demonstrated in the next example.

Example 4.10. Consider the large white rectangle translating across the gradient-like background shown in Figure 5. As seen, the moving object obscures a large portion of the background at certain points in time. However, the object traverses a sufficient amount such that every background pixel is unobscured at least once. Furthermore, the bottom portion of the background remains unobscured through the video's entirety. Therefore, by Proposition 4.9, we expect the video to have a connected background. Computing the algebraic connectivity of the graph $\mathcal{G}_{m,n}(B)$, we find

$$\lambda_2(L_B) \approx 18.3,$$

i.e. the Fiedler eigenvalue of the Laplacian matrix is greater than zero. Therefore, the graph is connected [12].

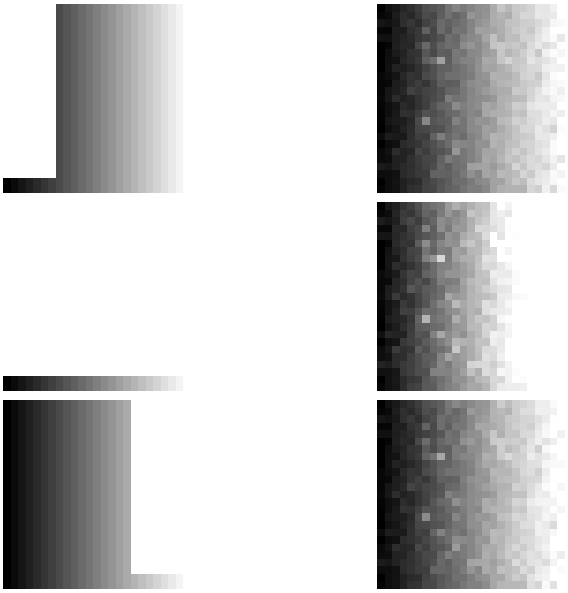

Figure 5: Original video (left) and extracted background (right) in three different frames. In this example, the connectivity condition (3) is satisfied.

□

In the previous example we see that although the background graph is connected (as predicted by Proposition 4.9), there is still some performance degradation during the time the object covers most of the background (i.e. the middle frame). [The "flare up" in brightness is the performance degradation we speak of.] It is important to note that although the connectivity condition (3) is found to be satisfied in this example, the identifiability condition (4) may not be. Therefore, in order to guarantee the extraction we obtain is unique and globally optimal, we must derive sufficient conditions for the identifiability inequality (4). This motivates the next section.

5 Conditions for Identifiability

Recall the identifiability condition (4). The goal of this section is to determine what properties a video and its moving object must possess in order to satisfy this condition. We make the following assumption.

Assumption 5.1. We assume the object F is a $p_m \times p_n$ rectangle, and that at least one frame contains the entire object. Furthermore, we assume that the object moves with a constant speed at $\dot{x} > 0$ pixels per frame horizontally and $\dot{y} > 0$ pixels per frame vertically. We also assume that the number of frames it takes an object to traverse a horizontal distance of $x \geq 0$ pixels at a speed of $\dot{x}$ pixels per frame is rounded to the nearest integer, denoted $[\frac{x}{\dot{x}}] \in \mathbb{Z}_+$, and similarly for vertical movements. We define $p_f = \min\{[\frac{p_n}{\dot{x}}], [\frac{p_m}{\dot{y}}]\}$. [Under these assumptions, the value $p_f + 1$ is the number of frames needed for the object to uncover all of the background it once obscured.] We assume the video is sufficiently long such that no part of the background remains obscured in all frames; $d_f > p_f$. We also assume that the video background is colored black in each frame; $uv^\top = \mathbb{1}_m \mathbb{1}_n^\top$. [This assumption on background color is for convenience, as certain constants can be taken as unity in this scenario. Otherwise, the results that follow are scaled by a suitable constant.] □

We now prove the primary result of this section.

Proposition 5.2 (Rectangle identifiability; necessary and sufficient)**.** *Suppose a video satisfies Assumption 5.1. Then the video satisfies the identifiability condition (4) if and only if*

$$\begin{aligned} d_f &> \max\{49p_f, 48p_mp_n + p_f\}, \\ p_mp_n &< \min\left\{\frac{1}{49}d_md_n, d_md_n - 48p_f\right\}. \end{aligned}$$

Proof. As seen in the identifiability condition (4), there are four values to analyze: the condition number $\kappa(w^*)$, the parameter c, the maximum degree $\Delta(\mathcal{G}_{m,n}(F))$, and the minimum degree $\delta(\mathcal{G}_{m,n}(B))$. We will first divide the proof into four separate computations, each dedicated to one of these values, then combine the final results at the end.

[Condition number.] By Assumption 5.1, the video's background is given by $uv^\top = \mathbb{1}_m\mathbb{1}_n^\top$, and therefore the solution to the problem (2) is $w^* = (u^*, v^*) = \mathbb{1}_{m+n}$. Hence, we have for the condition number

$$\kappa(w^*) = \frac{w^*_{\max}}{w^*_{\min}} = 1, \tag{5}$$

where $w^*_{\max}$ and $w^*_{\min}$ are the maximum and minimum elements of w^*, respectively.

[Parameter c.] Notice that the identifiability condition (4) depends on a parameter c. This parameter is defined in [6] to be a value in the interval $(0, 1]$ such that the following holds:

$$\bar{S}_{hk} + w^*_h w^*_k > c{w^*_{\min}}^2, \tag{6}$$

where

$$\bar{S} = \begin{bmatrix} 0_{m\times m} & S \\ S^\top & 0_{n\times n} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}.$$

Now, since $\bar{S}_{hk} + w^*_h w^*_k = \bar{X}_{hk}$, where $\bar{X}$ is the symmetrized matrix of grayscale video data (see [6]), it is clear that $\bar{S}_{hk} + w^*_h w^*_k \geq 1$ always, and therefore taking $0 < c < 1/(w^*_{\min})^2$ yields a constant which makes (6) satisfied. This amounts to taking $c < 1$ under Assumption 5.1. Note that the identifiability condition (4) relaxes as c increases, and therefore we would ideally take c to be as large as possible while still satisfying (6). Hence, we can choose

$$c \uparrow 1. \tag{7}$$

[Maximum degree.] Consider the foreground graph $\mathcal{G}_{m,n}(F)$ and let us denote the degree of a vertex in $\mathcal{G}_{m,n}(F)$ as $\deg(\cdot, F)$. We first note that $\deg(h, F),\ h \in V_u = \{1, 2, \ldots, m\}$, exactly equals the number of frames in which pixel h appears as foreground. Since the object has a width of p_n pixels and moves horizontally at $\dot{x}$ pixels per frame, the number of frames in which any single pixel can be considered as foreground is no more than $[\frac{p_n}{\dot{x}}]$. Similarly, the object has a height of p_m pixels and moves vertically at $\dot{y}$ pixels per frame, so the number of frames in which a given pixel can be foreground is also no more than $[\frac{p_m}{\dot{y}}]$. By Assumption 5.1, $d_f > p_f = \min\{[\frac{p_n}{\dot{x}}], [\frac{p_m}{\dot{y}}]\}$, so one of these two bounds is active. Hence, the maximum number of frames in which a single pixel appears as foreground is

$$\max\{\deg(h, F) : h \in V_u\} = \min\left\{\left[\frac{p_n}{\dot{x}}\right], \left[\frac{p_m}{\dot{y}}\right]\right\} = p_f.$$

Next, we note that $\deg(m+k, F),\ (m+k) \in V_v = \{m+1, m+2, \ldots, m+n\}$, exactly equals the number of foreground pixels in frame k. By Assumption 5.1, at least one frame contains the entire object, and therefore the maximum number of foreground pixels in any given frame is

$$\max\{\deg(m+k, F) : (m+k) \in V_v\} = p_m p_n.$$

Therefore, we find that the maximum degree of the foreground graph becomes

$$\Delta(\mathcal{G}_{m,n}(F)) = \max\{p_f, p_m p_n\}. \tag{8}$$

[Minimum degree.] Consider the background graph $\mathcal{G}_{m,n}(B)$ and let us denote the degree of a vertex in $\mathcal{G}_{m,n}(B)$ as $\deg(\cdot, B)$. Since F and G are complements with respect to Ω, we have that $\mathcal{G}_{m,n}(F)$ and $\mathcal{G}_{m,n}(B)$ are bipartite complements of one another. Hence, it must be that

$$\begin{aligned} |V_u| &= \deg(k, F) + \deg(k, B), \\ |V_v| &= \deg(h, F) + \deg(h, B), \end{aligned}$$

for all $h \in V_u$ and $k \in V_v - m$. This, together with the analysis of the foreground graph above, yields

$$\begin{aligned} \min\{\deg(h, B) : h \in V_u\} &= d_f - p_f, \\ \min\{\deg(k, B) : (m+k) \in V_v\} &= d_m d_n - p_m p_n. \end{aligned}$$

Therefore, we find that the minimum degree of the background graph becomes

$$\delta(\mathcal{G}_{m,n}(B)) = \min\{d_f - p_f, d_m d_n - p_m p_n\}. \tag{9}$$

[Final results.] Substituting (5), (7), (8), and (9) into the identifiability condition (4), we obtain

$$\min\{d_f - p_f, d_m d_n - p_m p_n\} > 48 \max\{p_f, p_m p_n\}. \tag{10}$$

Expanding the minimum and maximum, we find that this is equivalent to the following system of inequalities:

$$\begin{aligned} d_f &> 49 p_f, \\ d_f &> 48 p_m p_n + p_f, \\ p_m p_n &< \frac{1}{49} d_m d_n, \\ p_m p_n &< d_m d_n - 48 p_f. \end{aligned}$$

Now, rearranging terms and rewriting the inequalities in terms of minimums and maximums, we find that this system of equalities becomes

$$\begin{aligned} d_f &> \max\{49 p_f, 48 p_m p_n + p_f\}, \\ p_m p_n &< \min\left\{\frac{1}{49} d_m d_n, d_m d_n - 48 p_f\right\}. \end{aligned} \tag{11}$$

Hence, we have shown that, under Assumption 5.1, the identifiability condition (4) becomes (10) which is equivalent to the proposed set of inequalities (11). Hence, the conditions we provide relating the video length to the size and speed of the object are seen to be necessary and sufficient, as desired. □

Proposition 5.2 provides us necessary and sufficient conditions on the size and speed of a rectangular object, as well as the length of the video, in order to guarantee the satisfaction of the identifiability condition (4). As one's intuition might predict, smaller rectangles and longer videos relax these conditions, indicating that videos with small moving objects and many frames are inherently easier to achieve globally optimal foreground segmentation. Together with Proposition 4.9, we can provide deterministic guarantees that the optimization problem (2) used to decompose a video has benign landscape, and that the resulting decomposition is unique and globally optimal.

6 Future Work and Conclusions

Though the work in this report provides a fundamental set of conditions useful in checking for global optimality of video foreground segmentation, there remains a handful of open extensions to this line of research. For instance, the conditions we provide for identifiability might be generalized to videos with object speeds and sizes which vary with time, nonlinear object trajectories, and different background and foreground colors. Additionally, the toy examples we provide are seen to align closely with expected behavior according to our analysis. However, running simulations on open-source test sets would yield a more realistic demonstration of these optimality guarantees on actual video data.

With these extensions in mind, we recapitulate the novelties in our study on extracting a video's moving object from its background via robust nonnegative matrix factorization. Although the optimization problem of interest is nonconvex, it exhibits benign landscape under certain criteria. We exploit this fact to develop conditions under which the video segmentation is unique and globally optimal. We derive these global optimality guarantees in terms of intuitive and meaningful parameters, such as the size and speed of the moving object, as well as the length of the video. Furthermore, toy examples are given to illustrate the need for these criteria, and in what scenarios the problem's benign landscape breaks down.

Acknowledgements

A special thanks is in order to my advisor, Professor Somayeh Sojoudi, for her assistance and guidance on this project. Additionally, I would like to thank Salar Fattahi for providing me with a baseline stochastic gradient descent code to work off of.

References

[1] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, pp. 246–252, IEEE, 1999.

[2] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219–237, 2008.

[3] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht, "A neural network approach to bayesian background modeling for video object segmentation.," in *VISAPP (1)*, pp. 474–479, 2006.

[4] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer science review*, vol. 11, pp. 31–66, 2014.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, 2011.

[6] S. Fattahi and S. Sojoudi, "Exact guarantees on the absence of spurious local minima for non-negative robust principal component analysis," *arXiv preprint arXiv:1812.11466*, 2018.

[7] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, no. 2, pp. 192–200, 2011.

[8] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "Mahnmf: Manhattan non-negative matrix factorization," *arXiv preprint arXiv:1207.3438*, 2012.

[9] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Computer Science Review*, vol. 23, pp. 1–71, 2017.

[10] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *arXiv preprint arXiv:1809.09573*, 2018.

[11] M. S. Ramanagopal, C. Anderson, R. Vasudevan, and M. Johnson-Roberson, "Failing to learn: autonomously identifying perception failures for self-driving cars," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3860–3867, 2018.

[12] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.