

Simultaneous Multi-frame MAP Super-Resolution Video Enhancement using Spatio-temporal Priors

Sean Borman and Robert L. Stevenson

Department of Electrical Engineering, University of Notre Dame
 Notre Dame, IN 46556, USA
 rls@nd.edu

ABSTRACT

A simultaneous multi-frame super-resolution video reconstruction procedure, utilizing spatio-temporal smoothness constraints and motion estimator confidence parameters is proposed. The ill-posed inverse problem of reconstructing super-resolved imagery from the low resolution, degraded observations is formulated as a statistical inference problem and a Bayesian, maximum a-posteriori (MAP) approach is utilized for its approximate solution. The inclusion of motion estimator confidence parameters and temporal constraints result in higher quality super-resolution reconstructions with improved robustness to motion estimation errors.

1. INTRODUCTION

In recent years super-resolution (SR) enhancement of video has attracted increasing attention [1–4]. Estimation theoretic approaches using the Bayesian maximum *a-posteriori* (MAP) framework for solving the SR problem have achieved high-quality reconstructions [2, 5] of single frame images from sequences of noisy, under-sampled, low resolution (LR) observed video. Existing techniques effect video enhancement by applying the single frame technique to a shifting subset of the observed frames. A new MAP based SR enhancement procedure is proposed which estimates several unknown SR frames simultaneously. The proposed approach enables the inclusion of spatial *as well as temporal* constraints on the reconstructed super-resolution sequence, yielding improved imagery.

1.1. Problem statement

Given a short video sequence consisting of p observed LR frames with N_1 rows by N_2 columns of square pixels, estimate temporally corresponding super-resolved images with dimensions qN_1 rows by qN_2 columns, $q \in \mathbb{N}$.

2. OBSERVATION MODEL

In this section the imaging observation model which relates the unknown SR image sequence to the observed LR sequence is developed using a linear observation model based on [2]. SR and LR images are represented as lexicographically ordered vectors $\{\mathbf{z}^{(k)}\}_1^p$ and $\{\mathbf{y}^{(l)}\}_1^p$ over an index set $P = \{1, 2, \dots, p\}$.

2.1. Temporally coincident observation model

Consider the observation of the LR frame $\mathbf{y}^{(l)}$ from the unknown, but temporally coincident SR frame $\mathbf{z}^{(l)}$. It is assumed that the frames are related by a *known* (possibly spatially invariant) PSF matrix $\mathbf{A}^{(l,l)}$ with additive noise accounting for measurement errors and uncertainties in the knowledge of the PSF matrix,

$$\mathbf{y}^{(l)} = \mathbf{A}^{(l,l)} \mathbf{z}^{(l)} + \mathbf{n}^{(l,l)}, \quad l \in P. \quad (1)$$

In (1) the dimensions of $\mathbf{y}^{(l)}$ and $\mathbf{n}^{(l,l)}$ are $N_1 N_2 \times 1$, $\mathbf{z}^{(l)}$ is $q^2 N_1 N_2 \times 1$ and $\mathbf{A}^{(l,l)}$ is $N_1 N_2 \times q^2 N_1 N_2$. The temporally coincident observation model (1) is illustrated (excluding the additive noise terms) in Figure 1. The structure of the linear observation equation is typical in image restoration problems but for the unequal dimensions of the unknown and observed vectors. Since the dimensions of the unknown vectors $\mathbf{z}^{(l)}$ are larger by a factor of q^2 than the observations $\mathbf{y}^{(l)}$, it is clear that (1) is a highly under-determined system of equations. Multiple solutions $\mathbf{z}^{(l)}$ satisfying (1) exist, that is, $\mathbf{A}^{(l,l)}$ has a non-trivial nullspace (singular).

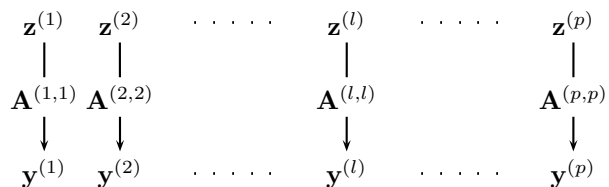


Figure 1: Temporally coincident observation model.

In (1) the additive noise is modeled as independent, zero mean Gaussian noise with known per-pixel variances $\sigma_j^{(l,l)^2}$. The observed frames $\{\mathbf{y}^{(l)}\}_1^p$ are known with high confidence since they derive from direct observation of the scene via the known observation matrix $\mathbf{A}^{(l,l)}$. The variances $\sigma_j^{(l,l)^2}$ may therefore be chosen according to the imaging system SNR. Given the assumption of independence, the observation noise pdf may be expressed as,

$$\mathcal{P}(\mathbf{n}^{(l,l)}) = \frac{1}{(2\pi)^{\frac{N_1 N_2}{2}} |\mathbf{K}^{(l,l)}|} \exp\left\{-\frac{1}{2} \mathbf{n}^{(l,l)T} \mathbf{K}^{(l,l)-1} \mathbf{n}^{(l,l)}\right\} \quad (2)$$

where $\mathbf{K}^{(l,l)} = \text{diag}(\sigma_1^{(l,l)^2}, \sigma_2^{(l,l)^2}, \dots, \sigma_{N_1 N_2}^{(l,l)^2})$. Furthermore, it is assumed that the observation noise is independent over the observed frames $\{\mathbf{y}^{(l)}\}_1^p$.

2.2. Temporally non-coincident observation model

To obtain additional constraints for estimating the SR sequence, unknown SR frames $\{\mathbf{z}^{(k)}\}_1^p$ are related to temporally non-coincident observed LR frames. For each $\mathbf{z}^{(k)}$ assume it is possible to find a *motion compensating* observation matrix $\mathbf{A}^{(l,k)}$ and vector $\mathbf{n}^{(l,k)}$ such that

$$\mathbf{y}^{(l)} = \mathbf{A}^{(l,k)} \mathbf{z}^{(k)} + \mathbf{n}^{(l,k)}, \quad l \in P \setminus \{k\}. \quad (3)$$

The dimensions of the matrix and vectors remain the same as in (1), however fundamental differences exist between (1) and (3) concerning the functions of the matrix $\mathbf{A}^{(l,k)}$ and the additive noise $\mathbf{n}^{(l,k)}$. In (1), $\mathbf{A}^{(l,l)}$ models only the effects of spatial degradations inherent in the imaging system, whereas in (3), $\mathbf{A}^{(l,k)}$ accounts for imaging degradations *and* compensates for motion occurring between frames k and l . In (1), the additive noise term $\mathbf{n}^{(l,l)}$ accounts for observation noise and sensor PSF uncertainties. In the temporally non-coincident observation model, $\mathbf{n}^{(l,k)}$ accounts for these effects, as well as for pixels which are not observable in $\mathbf{y}^{(l)}$ from motion compensation of $\mathbf{z}^{(k)}$. Similar to (1), the noise term $\mathbf{n}^{(l,k)}$ in (3) is described by an independent, zero mean Gaussian random vector with per-pixel variances $\sigma_j^{(l,k)^2}$, however two specific cases are identified:

1. Pixels in $\mathbf{y}^{(l)}$ which are *unobservable* from motion compensation of $\mathbf{z}^{(k)}$.

These pixels contribute no information towards determining the unknown SR image $\mathbf{z}^{(k)}$. The corresponding row of $\mathbf{A}^{(l,k)}$ contains all zeroes, while the corresponding entry in $\mathbf{n}^{(l,k)}$ contains the unobservable pixel value. Since this value is cannot be determined, the noise variance for the unobservable pixel is assigned the value $+\infty$. This ensures that no weight is given to the pixels in $\mathbf{y}^{(l)}$ which are unobservable from $\mathbf{z}^{(k)}$ when estimating $\mathbf{z}^{(k)}$.

2. Pixels in $\mathbf{y}^{(l)}$ which are *observable* from motion compensation of $\mathbf{z}^{(k)}$.

These pixels contribute valuable constraints for determining $\mathbf{z}^{(k)}$. It is well recognized [2] that reliable sub-pixel motion estimation is essential for high quality SR image reconstruction. A measure of the reliability of the motion estimates is therefore incorporated into the reconstruction by appropriate choice of the observed pixel noise variances $\sigma_j^{(l,k)^2}$. In this work the noise variances of the motion compensated observations are proportional to the mean absolute difference (MAD) associated with the motion vector obtained from the hierarchical block matching motion estimation procedure [2]. Blocks for which the MAD is large typically represent less reliable motion estimates and therefore corresponding to large observation noise variances. Similarly, small MAD corresponds to reliable motion estimates and small observation variance.

As in (1) it is assumed that the observation noise is independent over the observed frames $\{\mathbf{y}^{(l)}\}_1^p$. The choice of an independent Gaussian random vector model for errors committed in motion estimation and for the values of the unobservable pixels, though arguably over-simplistic, nevertheless ensures computational tractability of the optimization problem discussed in Section 3. Under these assumptions, the noise pdf for $k, l \in P, l \neq k$ is given by

$$\mathcal{P}(\mathbf{n}^{(l,k)}) = \frac{1}{(2\pi)^{\frac{N_1 N_2}{2}} |\mathbf{K}^{(l,k)}|} \exp\left\{-\frac{1}{2} \mathbf{n}^{(l,k)T} \mathbf{K}^{(l,k)-1} \mathbf{n}^{(l,k)}\right\}. \quad (4)$$

The motion compensating observation equation (3) is sufficiently general to represent a wide variety of scene motion and imaging degradations by suitable choice of the entries of $\mathbf{A}^{(l,k)}$ which may be found by linearization of the continuous optical correspondence field and discretization of the continuous image intensities. The practical difficulty lies in the fact that $\mathbf{A}^{(l,k)}$ must be estimated from the observed (degraded) LR sequence $\{\mathbf{y}^{(l)}\}_1^p$.

2.3. Combined observation model

Figure 2 illustrates the combined temporally coincident and non-coincident observation model relating the LR sequence $\{\mathbf{y}^{(l)}\}_1^p$ and a single SR image $\mathbf{z}^{(k)}$.

The p equations relating $\{\mathbf{y}^{(l)}\}_1^p$ to $\mathbf{z}^{(k)}$ may be expressed in vector notation as

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(p)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{(1,k)} \\ \mathbf{A}^{(2,k)} \\ \vdots \\ \mathbf{A}^{(p,k)} \end{bmatrix} \mathbf{z}^{(k)} + \begin{bmatrix} \mathbf{n}^{(1,k)} \\ \mathbf{n}^{(2,k)} \\ \vdots \\ \mathbf{n}^{(p,k)} \end{bmatrix}. \quad (5)$$

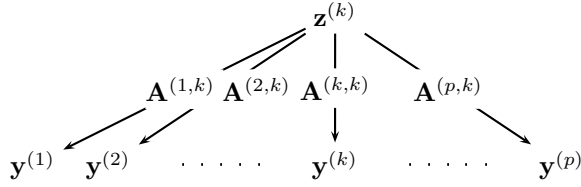


Figure 2: Combined observation model for $\mathbf{z}^{(k)}$.

These p equations may be written compactly as

$$\mathbf{y} = \mathbf{A}^{(*,k)} \mathbf{z}^{(k)} + \mathbf{n}^{(*,k)}. \quad (6)$$

Combining the temporally coincident and non-coincident observation equations (1) and (3) for all $\{\mathbf{z}^{(k)}\}_1^p$ yields p^2 equations relating the SR sequence $\{\mathbf{z}^{(k)}\}_1^p$ to the observed LR images $\{\mathbf{y}^{(l)}\}_1^p$. Following the notation in (6) this relationship may be written as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \vdots \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{(*,1)} & & & \\ & \mathbf{A}^{(*,2)} & & \\ & & \ddots & \\ & & & \mathbf{A}^{(*,p)} \end{bmatrix} \begin{bmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \\ \vdots \\ \mathbf{z}^{(p)} \end{bmatrix} + \begin{bmatrix} \mathbf{n}^{(*,1)} \\ \mathbf{n}^{(*,2)} \\ \vdots \\ \mathbf{n}^{(*,p)} \end{bmatrix}. \quad (7)$$

This may be expressed in the classic form

$$\mathbf{Y} = \mathbf{AZ} + \mathbf{N}. \quad (8)$$

Under the noise model assumptions in Sections 2.1 and 2.2 the pdf of the noise \mathbf{N} is given by

$$\mathcal{P}_{\mathbf{N}}(\mathbf{N}) = \frac{1}{(2\pi)^{\frac{p^2 N_1 N_2}{2}} |\mathbf{K}|} \exp\left\{-\frac{1}{2} \mathbf{N}^T \mathbf{K}^{-1} \mathbf{N}\right\}. \quad (9)$$

$\mathbf{K} = \text{diag}(\mathbf{K}^{(1,1)}, \mathbf{K}^{(2,1)}, \dots, \mathbf{K}^{(p,1)}, \mathbf{K}^{(1,2)}, \dots, \mathbf{K}^{(p,p)})$ and is itself diagonal.

3. BAYESIAN ESTIMATION OF SR IMAGES

Bayesian maximum *a-posteriori* (MAP) estimation is used to estimate the super-resolved images $\{\mathbf{z}^{(k)}\}_1^p$ given the observed data $\{\mathbf{y}^{(l)}\}_1^p$. The MAP estimate $\hat{\mathbf{Z}}_{\text{MAP}}$ maximizes the *a-posteriori* probability $\mathcal{P}(\mathbf{Z}|\mathbf{Y})$ given by:

$$\begin{aligned} \hat{\mathbf{Z}}_{\text{MAP}} &= \arg \max_{\mathbf{Z}} \{\mathcal{P}(\mathbf{Z}|\mathbf{Y})\} \\ &= \arg \max_{\mathbf{Z}} \left\{ \frac{\mathcal{P}(\mathbf{Y}|\mathbf{Z}) \mathcal{P}(\mathbf{Z})}{\mathcal{P}(\mathbf{Y})} \right\} \end{aligned} \quad (10)$$

The second expression in (10) results from the application of Bayes' theorem. Taking logarithms and noting that the maximum independent of $\mathcal{P}(\mathbf{Y})$, yields

$$\hat{\mathbf{Z}}_{\text{MAP}} = \arg \max_{\mathbf{Z}} \{\log \mathcal{P}(\mathbf{Y}|\mathbf{Z}) + \log \mathcal{P}(\mathbf{Z})\}. \quad (11)$$

It remains to determine the form of the likelihood function $\mathcal{P}(\mathbf{Y}|\mathbf{Z})$ and the prior $\mathcal{P}(\mathbf{Z})$.

3.1. The likelihood function

Given the observation equation (8) the conditional probability is determined by the noise pdf (9) as $\mathcal{P}(\mathbf{Y}|\mathbf{Z}) = \mathcal{P}_{\mathbf{N}}(\mathbf{Y} - \mathbf{AZ})$, so that

$$\mathcal{P}(\mathbf{Y}|\mathbf{Z}) = \frac{1}{(2\pi)^{\frac{p^2 N_1 N_2}{2}} |\mathbf{K}|} \exp\left\{-\frac{1}{2} (\mathbf{Y} - \mathbf{AZ})^T \mathbf{K}^{-1} (\mathbf{Y} - \mathbf{AZ})\right\}. \quad (12)$$

3.2. Spatio-temporal prior

A motivation for this work is the potential for the inclusion of more powerful *a-priori* constraints via simultaneous estimation of the SR frames. In single frame SR reconstruction the prior $\mathcal{P}(\mathbf{Z})$ typically includes only *spatial* smoothness constraints. Effecting *simultaneous* multi-frame enhancement enables the inclusion of *spatio-temporal* constraints in the prior. This is achieved using a motion-trajectory compensated, Markov random field model (MRF), in which the Gibbs distribution energy is dependent on pixel variation along the motion trajectory as well as the usual spatial interactions. The prior is chosen to be convex to ensure a computationally tractable optimization. The pdf $\mathcal{P}(\mathbf{Z})$ of the MRF is, according to the Hammersley-Clifford theorem [6], given by a Gibbs distribution. For the development that follows, the term in the exponent of the Gibbs distribution is expanded as

$$\mathcal{P}(\mathbf{Z}) = \frac{1}{k_p} \exp\left\{-\frac{1}{\beta} \sum_{c \in \mathcal{C}} \rho_{\alpha}(\partial_c \mathbf{Z})\right\}. \quad (13)$$

In (13), k_p is a normalizing constant known as the partition function, β is the MRF "temperature" parameter (inspired by the Gibbs distribution in thermodynamics). The summation is over the set of all "cliques" \mathcal{C} with ∂_c computing local spatio-temporal activity. The non-linear spatial activity penalizing function $\rho_{\alpha}(x)$ is the Huber function [7],

$$\rho_{\alpha}(x) = \begin{cases} x^2 & |x| \leq \alpha \\ 2\alpha|x| - \alpha^2 & |x| > \alpha \end{cases}. \quad (14)$$

The clique structure determines the spatial and temporal interactions. Five clique types are divided into two classes:

1. Spatial activity is computed using finite difference approximations to second directional derivatives (vertical, horizontal and two diagonal directions) in each SR image $\{\mathbf{z}^{(k)}\}_1^p$. In particular,

$$\begin{aligned} \partial_1 \mathbf{z}^{(k)} &= z_{n_1-1, n_2}^{(k)} - 2z_{n_1, n_2}^{(k)} + z_{n_1+1, n_2}^{(k)}, \\ \partial_2 \mathbf{z}^{(k)} &= z_{n_1, n_2-1}^{(k)} - 2z_{n_1, n_2}^{(k)} + z_{n_1, n_2+1}^{(k)}, \\ \partial_3 \mathbf{z}^{(k)} &= \frac{1}{2} z_{n_1+1, n_2-1}^{(k)} - z_{n_1, n_2}^{(k)} + \frac{1}{2} z_{n_1-1, n_2+1}^{(k)}, \\ \partial_4 \mathbf{z}^{(k)} &= \frac{1}{2} z_{n_1-1, n_2-1}^{(k)} - z_{n_1, n_2}^{(k)} + \frac{1}{2} z_{n_1+1, n_2+1}^{(k)}. \end{aligned} \quad (15)$$

At pixels on the image boundaries where it is not possible to compute the above second derivatives, appropriate first derivatives are substituted.

2. Temporal smoothness constraints are imposed on the reconstructed SR sequence through finite difference approximation to the second temporal derivative along the motion trajectory as,

$$\partial_5 \mathbf{z}^{(k)} = z_{n_1+\delta_1, n_2+\delta_2}^{(k-1)} - 2z_{n_1, n_2}^{(k)} + z_{n_1+\Delta_1, n_2+\Delta_2}^{(k+1)}. \quad (16)$$

The tuples (δ_1, δ_2) and (Δ_1, Δ_2) represent the motion vectors into the previous and subsequent frames respectively. In the case of pixel coverings / uncoverings, where it is not possible to find the forward or backward motion vectors, the appropriate first order difference is utilized. In the case where neither forward or backward motion information is available, the temporal activity measure is not included in the computation of the MRF energy. These choices ensure that pixel values in the temporally adjacent frames are unaffected by the covered / uncovered pixel, thereby assigning no penalty to temporal discontinuities where they are known to exist. Where correspondences are known to exist, the temporal derivatives may be weighted proportional to the reliability of the motion estimates, thereby imposing a robust temporal smoothness constraint. At the sequence temporal boundaries, appropriate first temporal differences are utilized.

3.3. Objective function

Substituting the likelihood term (12) and the prior (13) into (11), removing constants independent of \mathbf{Z} gives the objective function

$$\begin{aligned} \hat{\mathbf{Z}}_{\text{MAP}} &= \arg \max_{\mathbf{Z}} \\ &\left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{AZ})^T \mathbf{K}^{-1}(\mathbf{Y} - \mathbf{AZ}) - \frac{1}{\beta} \sum_{c \in \mathcal{C}} \rho_{\alpha}(\partial_c \mathbf{Z}) \right\} \\ &= \arg \min_{\mathbf{Z}} \\ &\left\{ \frac{1}{2}(\mathbf{Y} - \mathbf{AZ})^T \mathbf{K}^{-1}(\mathbf{Y} - \mathbf{AZ}) + \frac{1}{\beta} \sum_{c \in \mathcal{C}} \rho_{\alpha}(\partial_c \mathbf{Z}) \right\}. \quad (17) \end{aligned}$$

3.4. Optimization

Determining the MAP estimate $\hat{\mathbf{Z}}_{\text{MAP}}$ thus requires the minimization of the objective function in (17). By the choice of a strictly convex prior in Section 3.2, existence and uniqueness of a global minimum, corresponding to MAP estimate $\hat{\mathbf{Z}}_{\text{MAP}}$ in (17) is assured. As a result, efficient gradient descent optimization methods [8] may be employed to yield

the MAP estimate $\hat{\mathbf{Z}}_{\text{MAP}}$. Despite the large number of unknowns, convexity ensures that the optimization remains tractable.

4. EXAMPLES

Tests were performed using the algorithm proposed in this paper on a short sequence of a landmark at the University of Notre Dame. Ten monochrome frames of $N_1 = 120$ rows by $N_2 = 160$ columns are used as input data. A frame from the original data sequence is shown in Figure 3.

Motion occurring between all pairs of frames of the observation sequence are computed using the hierarchical subpixel motion estimator described in [2] yielding 1/4 pixel resolution local translation motion estimates for each low resolution pixel. Reconstruction is performed with $q = 4$ resulting in reconstructed images of 480 rows by 640 columns. The reconstructed frame in temporal correspondence with the original image in Figure 3 is shown in Figure 5. Additionally, a zero-order hold image of the original frame is presented for comparison in Figure 4.

Evaluation of the video sequences demonstrate subtle but noticeable improvements in reconstruction quality resulting from the inclusion of both motion estimation confidence parameter via the covariance matrix \mathbf{K} as well as by the inclusion of the temporal cliques in the prior. The latter is impossible to demonstrate without viewing the actual video reconstruction.



Figure 3: Original frame from dome sequence.



Figure 4: Zero order hold.



Figure 5: Super-resolution reconstruction.

5. CONCLUSIONS AND EXTENSIONS

A simultaneous multi-frame super-resolution video reconstruction procedure utilizing Bayesian Maximum *A Posteriori* estimation with spatio-temporal smoothness prior and motion estimator confidence parameters is presented. The technique exhibits improved robustness to motion estimation errors through the inclusion of reliability measures for motion estimates as well as by imposing temporal constraints on the reconstructed image sequence.

The observation noise covariance matrix \mathbf{K} was chosen to be diagonal. With minimal additional complexity it is possible to generalize \mathbf{K} to an arbitrary positive definite covariance matrix, thereby allowing more realistic motion estimation error modeling.

6. REFERENCES

- [1] Y. Nakazawa, T. Saito, T. Sekimori, and K. Aizawa, "Two approaches for image-processing based high resolution image acquisition," in *Proceedings of the IEEE International Conference on Image Processing*, Austin, TX, 1994, vol. III, pp. 147–151.
- [2] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.
- [3] B. C. Tom and A. K. Katsaggelos, "An Iterative Algorithm for Improving the Resolution of Video Sequences," in *Visual Communications and Image Processing*, Orlando, FL, Mar. 1996, vol. 2727 of *Proceedings of the SPIE*, pp. 1430–1438.
- [4] A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Superresolution Video Reconstruction with Arbitrary Sampling Lattices and Nonzero Aperture Time," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1064–1076, Aug. 1997.
- [5] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP Registration and High-Resolution Image Estimation Using a Sequence of Undersampled Images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.
- [6] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society B*, vol. 36, pp. 192–236, 1974.
- [7] R. L. Stevenson, B. E. Schmitz, and E. J. Delp, "Discontinuity Preserving Regularization of Inverse Visual Problems," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 3, pp. 455–469, Mar. 1994.
- [8] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Non-linear Programming: Theory and Algorithms, Second Edition*, John Wiley & Sons, 1993.