Robust Segmentation of Highly Dynamic Scene with Missing Data

Yinhui ZHANG†, Member, Zifen HE††, Nonmember, and Changyu LIU††, Member

SUMMARY Segmenting foreground objects from highly dynamic scenes with missing data is very challenging. We present a novel unsupervised segmentation approach that can cope with extensive scene dynamic as well as a substantial amount of missing data that present in dynamic scene. To make this possible, we exploit convex optimization of total variation beforehand for images with missing data in which depletion mask is available. Inpainting depleted images using total variation facilitates detecting ambiguous objects from highly dynamic images, because it is more likely to yield areas of object instances with improved grayscale contrast. We use a conditional random field that adapts to integrate both appearance and motion knowledge of the foreground objects. Our approach segments foreground object instances while inpainting the highly dynamic scene with a variety amount of missing data in a coupled way. We demonstrate this on a very challenging dataset from the UCSD Highly Dynamic Scene Benchmarks (HDSB) and compare our method with two state-of-the-art unsupervised image sequence segmentation algorithms and provide quantitative and qualitative performance comparisons.

key words: image sequence segmentation, convex optimization, total variation, conditional random field

1. Introduction

Scene segmentation is the task of extracting salient foreground objects from backgrounds, such as segmenting pedestrians in video sequences. Image sequence segmentation has been widely used in many vision problems, such as tracking [1] and multi-view geometry [2]. In practice, video sequence segmentation is a non-trivial task, since real-world data may contain highly dynamic background in the scene and may also be contaminated by a large amount of missing data.

Early approaches to image sequence segmentation aim at annotating [3]–[5] the presence of foreground objects at given frame locations by a user. These approaches are labor-intensive as they require hand-segmented training data and so would not scale to rapid processing of image sequences. Moreover, these methods are also limited in that results heavily depend on the selection of the seeds. Recently, Badrinarayan On et al. [18] use patch cross-correlation to construct temporal tree structured models which permits efficient and exact inference. This method achieves good results on both foreground/background and multiclass video segmentation problems, but the algorithm is semi-supervised, which also requires ground truth annotation of key frame(s) as input to the framework.

To address this issue, background subtraction is introduced by exploiting an adaptive mixture technique [7] to learn an explicit background model. Instead of learning background models based on dozens of frames in most conventional background subtraction methods, Barnich et al. [8] build a background model with a single frame of video sequence. To circumvent the problem of limited variances of Gaussian Mixture Model (GMM), Haines et al. [9] propose to use Dirichlet process Gaussian mixture models to estimate background distributions and input them into a model learning process for continuous update as scene changes. These methods are typically based on the strong assumption that the backgrounds are static, or changing slowly, which is not valid for subtraction of highly dynamic backgrounds.

For automatically segmenting dynamic scenes, it is beneficial to go beyond pure pixel-level background subtraction approaches to patch-level hierarchical segmentation methods. Grundmann et al. [11] combine hierarchical cues by constructing a tree of spatio-temporal segmentation. This approach allows subsequent choosing from varying levels of granularity. Although good for handling hierarchical appearance cues, a strong limitation of this method is that it does not solve the foreground segmentation task on its own. Papazoglou et al. [6] propose a Fast Object Segmentation (FOS) algorithm which attempts to build dynamic appearance models of the object and background under the assumption that they change smoothly over time. An advantage of this approach is that it may be possible to handle spatio-temporal cues on image patches such as color and location in the labeling refinement stage. But rely initialization of inside foreground object points only on motion boundaries tends to produce substantial amount of false-positive seeds, especially in the case of highly dynamic scenes.

Alternative approaches in image sequence segmentation focus primarily on ranking of object proposals. Endres et al. [12] propose to generate bag of regions based on seeds and rank them using structured learning. Lee et al. [14] use consistent appearance and motion to rank hypothesis groups among object-like regions. On the other hand, Ma et al. [13] introduce intra and inter frame constraints in a weighted re-
gion graph so as to find maximum weight cliques. Segmentation is performed using graph cuts and simple color cues, and the regions are ranked through classification based on gestalt cues with a simple diversity model in [16]. Most recently, object models [15] are built based on the primary object hypothesis regions. While the object proposal facilitates a clique-level object hypotheses, in practice, all the above methods suffer from the high complexity of choosing the object proposals in cases of even moderate video size due to highly redundant segmentations for the different regions of the input image sequences.

Rather than estimating object proposals, our approach guides segmentation with a combination of convex optimization of total variation energy function with a conditional random field. The proposed approach coupled segmenting foreground objects while inpainting the highly dynamic scene with a variety of missing data rates. We thoroughly evaluate each stage of the process and demonstrate that it can generalize well across sequences of a typical dynamic scene benchmark.

2. Convex Optimization of Total Variation

In this section we describe the underlying total variation energy formulation for our proposed convex optimization based approach. The motivation of this work is to exploit the efficiency and accuracy of total variational optimization of dynamic images with missing data in order to provide smoothed images in a fully unsupervised manner.

2.1 Total Variation Model

We cast reconstruction of depleted images with missing data as a convex optimization of total variation. In particular, holes and missing pixels in the raw images are filled by minimizing total variation. This is reasonable because total variation optimization aims at preserving piecewise smooth areas and sharp edges in the images. For this purpose, the total variation energy function is given as

$$\operatorname{argmin} \left\| x \right\|_{TV} \quad \text{s.t. } \left\| Mx - y \right\|_2 \leq \epsilon$$

(1)

where $x$ is the optimally reconstructed image with inpainted regions, $y$ is the observation of the input image at each time slice and $M$ is a nonempty subset of $\mathbb{R}^2$, representing the inpainting matrix used to generate missing data in the original frames.

2.2 ADMM Based Convex Optimization

Convex optimization has been shown to provide efficient algorithms for computing reliable solutions in a broadening spectrum of applications. Once a problem is formulated as a convex optimization problem, it can then be solved very reliably and efficiently, because any local minimum must be a global minimum. In this section, we present an Alternating Direction Method of Multipliers (ADMM) based convex optimization method which is tailored to coupled image inpainting and contrast improvement featuring highly dynamic backgrounds.

The total variation problem in Eq. (1) is further decomposed into two individual functions so as to yield an easily implementable splitting-coordination optimization algorithm. We denote the objective function in the proposed convex optimization problem with $f(x) = \left\| x \right\|_{TV}$, in which the discrete total variation operator $\left\| \cdot \right\|_{TV}$ aims at preserving piecewise smooth areas and sharp edges of object hypothesis regions. Similarly, we denote with $g(y) = \iota_{C}(y)$ the indicator function of the set $C$ defined by $\| Mx - y \|_2 \leq \epsilon$, which returns 0 when the condition $y \in C$ is satisfied and $+\infty$ otherwise. In fact, the indicator function $\iota_{C}$ belongs to $\Gamma_{0}(\mathbb{R}^N)$, which is the class of lower semicontinuous convex functions from $\mathbb{R}^N$ to $[0, +\infty]$ such that $\text{dom} f \neq \Phi$.

Denoting $\mathcal{L}_\gamma$ the augmented Lagrangian of index $\gamma \in [0, +\infty)$, the ADMM algorithm consists in minimizing $\mathcal{L}_\gamma$ over $x$, then over $y$, and then applying a proximal operator step with respect to the Lagrange multiplier $z$ [17]. The proximity operator $\text{prox}_{\mathcal{L}_\gamma}$ is a natural extension of the notion of a projection operator onto a convex set, which has very attractive properties that make it particularly well suited for iterative minimization algorithms.

3. Conditional Random Field

Once the image has been inpainted by ADMM based convex optimization, our goal is to extract the foreground object out of the image sequences, yielding a binary foreground-background segmentation in every frame. We adopt the volumetric conditional random field to formulate the segmentation problem as superpixel labeling over a spatiotemporal graph $\mathcal{G}$, in which the hidden nodes corresponding to the unknown labels. The using of superpixels to represent reconstructed images enables greatly reduced computational complexity as well as memory usage. The resulting optimization problem is to find the joint label assignment $L$ for all superpixels in the reconstructed image sequences. Similarly to other segmentation works [6], [19], we define an energy function as

$$E(L) = \lambda_1 \sum_{i \in \mathcal{G}} V_i(l_i) + \lambda_2 \sum_{i \in \mathcal{G}} \sum_{j \in \mathcal{N}_p(i)} W_{ij}(l_i, l_j) + \sum_{i \in \mathcal{G}} \sum_{j \in \mathcal{N}_p(i)} W_{ij}(l_i, l_j)$$

(2)

where the labeling $L = \{l_i \in \mathcal{G}\}$ of all superpixels represents a segmentation of the video. $V_i$ denotes the unary potentials representing the data term. $V_{ij}$ and $W_{ij}$ are the pairwise potentials representing the smoothing term over spatial and temporal constraints, respectively. $\lambda_1$ and $\lambda_2$ are tradeoff parameters and both are fixed as 0.3 throughout the
experiments. \( N_s \) and \( N_t \) represent the spatial and temporal neighborhood systems, respectively. In the spatial neighborhood domain, two superpixels are connected if they are adjacent. Two superpixels are temporally connected if at least one pixel of a superpixel moves into another superpixel in the subsequent frame in terms of the motion cues through optical flow. We employed the instantiations of the energy function in Eq. (2) proposed by Papazoglou et al. in [6] for characterizing the unary and pairwise potentials. In particular, the unary potential term measures superpixel appearance relative to the appearance models, which are computed from a probabilistic model for the colors of background and foreground pixels using Gaussian Mixture Models (GMM). In addition, the pairwise potentials are instantiated as standard Potts potentials by combining spatial distance with color difference in both spatial and temporal dimension. In this way, the binary valued energy function in Eq. (2) with submodular pairwise potentials can be exactly minimized by graph cuts.

4. Experiments

In this section, we aim to demonstrate the robustness of the proposed approach and will make a case for its use to deal with the likely event of missing data in the dynamic scenes. In such task, the first step is to generate missing data in the input image sequence, which are then used as input for the subsequent convex optimization of total variation steps. To this end, we generate for each frames with missing data as given by a depletion mask \( M \). Specifically, the mask \( M \) for each frame has the same size as the original image and each entry of the mask \( M_{ij} \) returns a binary value denoting whether the pixel located at \((i, j)\) can be observed. The entry values of the inpainting masks are generated randomly for each frames.

We use the ADMM based convex optimization of total variation as in Eq. (1) to revert the images with missing data. The maximum number of ADMM iterations was fixed to 100 and a stop criteria was fixed at 10\(^{-5}\) with respect to the variations of reconstructed images from consecutive iterations. The proposed segmentation algorithm is then applied to the reconstructed images by ADMM convex optimization. We use the \textit{birds} sequence from UCSD highly dynamic scene benchmark [20] to test the proposed algorithm.

The first column of Fig. 1 shows the original images from \textit{birds} sequence, in which highly dynamic water scene can be observed. The second row illustrates the contaminated frames with 10\% missing data and the third row shows the inpainting result by ADMM convex optimization of total variation. We can find that the missing pixels are filled by ADMM convex optimization. In addition, we observe that the inpainted images are more likely to yield areas of object instances with improved grayscale contrast by comparing them with the original images. The images with improved contrast facilitate segmenting foreground objects from dynamic backgrounds.

The plot in Fig. 2 provides quantitative comparison between reconstruction process of different missing data rates in terms of the number of iterations and energy values. Our empirical results in Fig. 2 suggest that the ADMM based convex optimization of total variation algorithm gradually converges into a steady state, in which the solutions have comparable energy values for images with different missing data rates. Also it can be seen that the algorithm’s efficiency decrease with increasing the rate of missing data.

To find foreground object hypotheses, we estimate foreground seeds by utilizing point-in-polygon[10] algorithm. The foreground hypotheses are used to compute unary and pairwise potentials in the energy function, which is then minimized by graph cut algorithm to find binary segmentation of the image sequence. Figure 3 shows segmentation results of frames \#1, \#10, \#20 and \#30 using the proposed algorithm on images with different data missing rates (10, 30, 50, and 70 percent). These results are typical for the described algorithm. Qualitatively we found that the performance is stable over the entire image sequence. We further compare our work to two state-of-the-art image sequence segmentation algorithms: Fast Object Segmentation (FOS) [6] and Video Object Segmentation (VOS) [15]. Table 1 summarizes the performances of VOS, FOS and our method in terms of averaged Area Under Curve (AUC) of segmentation with respect to ground truth on the \textit{birds} sequence with different missing data rate. It can be concluded from Table 1 that, by combining conditional random field with convex optimization based total variation functional minimization, the robustness of the proposed algorithm is guaranteed against missing data, say nearly 90 percent segmentation accuracy for up to 70 percent missing data in highly dynamic scenes.

This result clearly illustrates the limitations of alterna-
We employ the ing and CRF based segmentation in the proposed method, optimization for detecting foreground object instances plays approaches. In this approach, ADMM based convex optimization framework for coupled image sequence inpainting and segmental variation with conditional random field yields an optimal scene dynamics as well as substantial amount of missing data. Besides, it is worth noting that with the ADMM based inpainting process plays the key role in segmenting dynamic scene with a variety of missing data rates and paves the way to better solutions of high-level vision problems.

**5. Conclusion**

We present an unsupervised segmentation method for image sequence segmentation that is applicable to extensive scene dynamics as well as substantial amount of missing data rates. Our combination of convex optimization of total variation with conditional random field yields an optimal framework for coupled image sequence inpainting and segmentation that compares favorably with two prior leading approaches. In this approach, ADMM based convex optimization for detecting foreground object instances plays an important role, facilitating stable appearance and motion based minimization of conditional random field. To conclude, our image sequence segmentation method efficiently extracts salient foreground regions from the highly dynamic scene with a variety of missing data rates and paves the way to better solutions of high-level vision problems.

**Acknowledgments**

This work was supported by the National Science Foundation of China (NSFC) under Grants 61461022 and 61302173. Yinhui Zhang was supported by the China Scholarship Council (File No. 20120835035) while studying as a visiting scholar at University of Miami.

**References**
