Video Segmentation as a Distributed Convex Optimization Problem using Primal Decomposition

Animesh Garg* ANIMESH.GARG@BERKELEY.EDU
Jeff Mahler* JMAHLER@BERKELEY.EDU
Shubham Tulsiani* SHUBHTULS@BERKELEY.EDU
Department of EECS, UC Berkeley, CA 94720

Abstract

Getting exact video segmentations for tracking and recognition is a challenging problem. A majority of existing methodstrack but provide a bounding box rather than a an exact foreground mask for the object. For real worl applications of perception, like robotics, the silhoutte of the object perhaps even pose need to be known for hope of success in manipulation tasks.

We propose a method in this study which formulated the problem of video segmentation as a Markov random field. However solving such a large graph to global optimality may be computationally expensive. Hence we propose a distributed method using Primal decomposition.

1. Introduction

The problem of video segmentation is of interest for many areas. (**??**) and (**?**) have looked at the problem of modelling the MRF in terms of energies. The solution strategy they use is Dual decomposition but without integer programming.

Our study explores the use of state-of-the-art integer program solvers. Modelling integers allows us to capture more rich features in video which are usually not directly put in current models. From a given video sequence, and user intialized object(s) of interest, the aim is to track the region(s) of interest through the subsequent image frames in the video. Majority of other methods which address the problem provide locally optimal solutions. Such an approach though successful in some applications requires a substantial amount of human intervention at several points in the solution such as in cases of occlusion, change in pose, shape and color and in extreme cases object (or a part of object) egresses the frame and re-enters later.

This is a draft version for EE227-BT Fall 2013 Project.
* denotes equal contribution.

2. Problem Formulation

NOTATIONS AND VARIABLES

- We denote the video volume by I. A pixel in I is indexed by its location in space as well as time and is denoted by I_{ijt}
- We wish to recover a complete segmentation of the video into foreground and background. This labelling is captured by the variable X where $X_{ijt} \in \{0, 1\}$
- The time continuity between frames in a video implies that any pixel in a given frame corresponds to some pixel in the next frame. We capture this notion by a weak correspondence between a pixel and its neighbors in the next frame. The correspondence weights for a pixel are denoted by W_{ijt}^{ab} $(a, b \in \{-h, .., h\})$ i.e we define a correspondence weight variable between each pixel and the $(2h+1)X(2h+1)$ grid surrounding it in the next frame.
- By $N_s(i, j, t)$, we denote the indices of the pixels in the spatial neighborhood of the pixel (i, j, t)
- We define pseudo-variables $U, V, \overline{U}, \overline{V}$ for notational convenience. The variables (U, V) capture the average motion direction of a pixel between consecutive frames in X, Y directions respectively. We also denote the average direction of motion of the neighborhood of a pixel by $(\overline{U}_{ijt}, \overline{V}_{ijt})$. These pseudo variables are defined in terms of the previously defined variables as follows -

$$U_{ijt} = \sum_{a,b \in \{-h,..,h\}} a W_{ijt}^{ab} \tag{1}$$

$$V_{ijt} = \sum_{a,b \in \{-h,..,h\}} b W_{ijt}^{ab} \tag{2}$$

$$(\overline{U}_{ijt}, \overline{V}_{ijt}) = \frac{1}{|N_s(i,j,t)|} \sum_{Y \in N_{s(i,j,t)}} (U_Y, V_Y) \quad (3)$$

OBJECTIVE

$$\min_{X,W} \ \lambda_1 A(X,I) + \lambda_2 S(X) + \lambda_3 T(X,W) \quad (4)$$
$$+\lambda_4 F(W,I) + \lambda_5 C(W) + \lambda_6 M(W)$$

subject to $W \geq 0, \forall(i,j,t) X_{ijt} \in \{0,1\}, \sum_{a,b} W_{ijt}^{ab} = 1$ and $\forall t | \sum_{i,j} X_{ijt} - \sum_{i,j} X_{ij(t+1)} | \leq \sigma \sum_{i,j} X_{ijt}$

The objective function comprises of various penatly terms which are explained below. The last constraint specifies that the number of foreground pixels in do not change rapidly between consecutive frames.

APPEARANCE MODEL $A(X,I)$

Given the initial user labelled segmentation X', we can form a foreground model and a corresponding penalty function $f_{I,X'}$ for a pixel's label given its value. We then define the unary potential as follows -

$$A(X,I) = \sum_{i,j,t} f_{I,X'}(X_{ijt}, I_{ijt}) \quad (5)$$

SPATIAL LABELLING COHERENCE $S(X)$

We want to drive the system towards a labelling where neighbouring pixels have similar labels. The spatial labelling coherence term defined below encapsulates this.

$$S(X) = \sum_{i,j,t} \sum_{Y \in N_s(i,j,t)} |X_{ijt} - X_Y| \quad (6)$$

TEMPORAL LABELLING COHERENCE $T(X,W)$

For a given pixel, the corresponding pixel in the next frame should also have the same label. We formalize this notion using the penalty function below.

$$T(X,W) = \sum_{i,j,t} \sum_{a,b \in \{-h,..,h\}} W_{ijt}^{ab} |X_{ijt} - X_{i+a,j+b,t+1}| \quad (7)$$

FLOW SIMILARITY $F(W,I)$

For each pixel, the corresponding pixel in the next frame should be similar. This is enforced by the flow similarity defined below.

$$F(X,I) = \sum_{i,j,t} \sum_{a,b \in \{-h,..,h\}} W_{ijt}^{ab} |I_{ijt} - I_{i+a,j+b,t+1}| \quad (8)$$

FLOW CONTINUITY $C(W)$

The direction of movement of pixels is continuous over a small spatial neighbourhood. We therefore penalize rapid variations in flow as follows-

$$C(W) = \sum_{i,j,t} |U_{ijt} - \overline{U}_{ijt}| + |V_{ijt} - \overline{V}_{ijt}| \quad (9)$$

MOMENTUM CONTINUITY $M(W)$

It also needs to be enforced that the velocity of a pixel and its corresponding pixel in the next frame do not vary rapidly. This is ensured by the momentum continuity terms defined below

$$M(W) = \sum_{i,j,t} \sum_{a,b \in \{-h,..,h\}} W_{ijt}^{ab} (|a - \overline{U}_{i+a,j+b,t+1}| + |b - \overline{V}_{i+a,j+b,t+1}|) \quad (10)$$

3. Experiments and Metrics

As presented in the model in Section **??** we have a complete optimization model with several integer variables for foreground-background labels.

We will test the performance of our solution on the Berkeley Motion Segmentation Dataset as provided by (**?**). The dataset has 26 video sequences with pixel-accurate segmentation annotation of moving objects. A total of 189 frames are annotated.

We will evaluate results from our approach and compare the performance with that of (**?**), (**?**) and (**?**) on this dataset.

Furthermore multiple decoupling strategies will implementation and compared, like decoupling time frames v/s decoupling in space. Finally a dual decomposition method with also be explored and compared qualitatively with (**?**).

We plan on completing the implementation in MATLAB with the use of CVX and CPLEX optimization libraries.