Video Segmentation using Combinatorial Optimization

Animesh Garg, Santanu Dey and James M. Rehg

***Abstract*—This is the optimization model for implementation in Computer Vision domain for Video Segementation.**

***Index Terms*—Video Segmentation, Optical Flow, Combinatorial Optimization.**

I. Introduction

In this document, we have explain a combinatorial optimization over pixel space in each frame of the video sequence for segmentation over video volume.

II. Modelling Methodology

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 after a certain time.

Hence, the formulation of the problem in the video segmentation domain needs to take into account the following constraints of a generic video sequence:

1) Spatial attribute coherence of every pixel in its neighborhood.
2) Temporal attribute coherence of every pixel with itself in the adjacent frames.
3) Temporal motion coherence of every pixel with its corresponding pixel in adjacent frames.
4) The neighborhood of *i*th pixel under consideration should also have a similar temporal motion coherence as pixel *i*.

Let us define $\mathcal{G}$, a set of pixel sites in the video volume. Furthermore, we consider three consecutive frames in time *t*, $t+1$ and $t+2$ wherein pixel indices in each frame is denoted by subscripts by *i*,*j* and *k* respectively. Further, subscripts *f* and *b* are used to denote foreground and background regions respectively. $N(i)$ is defined as the spatial neighborhood of pixel with index *i*, i.e. in same time frame t. The size of the set $N(i)$ directly affects in adding combinatorial complexity to the model. $\mathcal{T}(i)$ is defined as the temporal neighborhood of pixel *i* in time t which is set of pixels in the next time frame where pixel *i* can move to, such that *j* $\epsilon\mathcal{T}(i)$ and *k* $\epsilon\mathcal{T}(j)$.

A label set $\mathcal{L}$ is defined as the set of all possible configurations a pixel is allowed to take in the video volume. As discussed in [2], if we allow a movement of M pixels on either side in both x and y axes in every time frame then the cardinality of label space,

$$n(\mathcal{L}) = 2\left[(2M+1)^2\right]^2$$

because for every pixel *i* in time t , a choice of $(2M+1)^2$ pixel sites is possible in each of successive frames $t+1$ and $t+2$. And further in case of binary labeling, either foreground or background, pixel *i* in time t could either take an attribute f or *b*.

The resulting optimization problem is to find a label assignment $\mathcal{L}$ for all pixels in the video volume so as to minimize the objective function modelled as total energy over all assignment operations,

$min \quad \sum_i (c_{fi}x_i + c_{bi}(1-x_i)) \; + \; \xi\sum_i\sum_{u\epsilon N(i)} z_{iu} \; + \; \sum_i\sum_j\sum_k \delta_{ijk}y_{ijk} + \lambda\sum_i\sum_{u\epsilon N(i)}\alpha_{iu} + \sum_i\sum_j w_{ij}$

s.t.

$$z_{iu} \geqslant x_i - x_u \qquad \forall i\epsilon\mathcal{G}, \forall u\epsilon N(i) \tag{1}$$

$$z_{iu} \geqslant x_u - x_i \qquad \forall i\epsilon\mathcal{G}, \forall u\epsilon N(i) \tag{2}$$

$$\alpha_{iu} \;\geqslant\; \sum_j\sum_k d_{ijk}y_{ijk} - \sum_j\sum_k d_{ujk}y_{ujk}$$

$$\forall i\epsilon\mathcal{G}, \forall u\epsilon N(i), \forall j\epsilon\mathcal{T}(i), \forall k\epsilon\mathcal{T}(j) \tag{3}$$

$$\alpha_{iu} \geqslant \sum_j\sum_k d_{ijk}y_{ujk} - \sum_j\sum_k d_{ijk}y_{ijk}$$

$$\forall i\epsilon\mathcal{G}, \forall u\epsilon N(i), \forall j\epsilon\mathcal{T}(i), \forall k\epsilon\mathcal{T}(j) \tag{4}$$

$$z_{ij} \geqslant x_i - x_j \qquad \forall i\epsilon\mathcal{G}, \forall j\epsilon\mathcal{T}(i) \tag{5}$$

$$z_{ij} \geqslant x_j - x_i \qquad \forall i\epsilon\mathcal{G}, \forall j\epsilon\mathcal{T}(i) \tag{6}$$

$$w_{ij} \geqslant \sum_k y_{ijk} + z_{ij} - 1 \tag{7}$$

Animesh Garg is with the School of Industrial Systems and Engineering, Georgia Institute of Technology, Atlanta, USA, e-mail: garg.animesh@gatech.edu.

Santanu Dey is with the School of Industrial Systems and Engineering, Georgia Institute of Technology, Atlanta, USA, e-mail: santanu.dey@isye.gatech.edu.

James M. Rehg is with Robotics and Intelligent Machines Center, College of Computing, Georgia Institute of Technology, Atlanta, USA, e-mail: rehg@cc.gatech.edu.

$$\sum_{j}\sum_{k} y_{ijk} = 1 \qquad \forall i \epsilon \mathcal{G}, \forall j \epsilon \mathcal{T}(i), \forall k \epsilon \mathcal{T}(j) \tag{8}$$

$$\sum_{i}\sum_{k} y_{ijk} = 1 \qquad \forall i \epsilon \mathcal{G}, \forall j \epsilon \mathcal{T}(i), \forall k \epsilon \mathcal{T}(j) \tag{9}$$

$$\sum_{i}\sum_{j} y_{ijk} = 1 \qquad \forall i \epsilon \mathcal{G}, \forall j \epsilon \mathcal{T}(i), \forall k \epsilon \mathcal{T}(j) \tag{10}$$

$$y_{ijk} \leqslant \sum_{k'} y_{jkk'} \qquad \forall k' \epsilon (t+3) \tag{11}$$

$$z_{ij}, z_{iu} \epsilon \{0,1\}$$

$$x_{i,} y_{ijk} \epsilon \{0,1\}$$

In the aforementioned formulation, objective has been modelled as a sum of costs associated with asssignment of labels to each pixel while under the constraints of a generic video sequence.

A. Explanation of objective function, variables and constraints

The first term in the objective function,

$$\sum_{i} \left[c_{fi} x_i + c_{bi} (1 - x_i) \right],$$

represents the *Appearance Model*. It consists of the unary cost of assignment of background or foreground model to each pixel site with index *i*. x_i is a binary variable which takes a value 1 when a pixel with index *i* is assigned to foreground, *f*, and 0 other wise. c_{fi} and c_{bi} is the cost of assigning a pixel site to foreground and background respectively. c_{fi} and c_{bi} measures the appearance similarity relative to foreground/background color models and also across temporal dimensions. One suggested approach could be Guassian Mixture Models as in [2].

The second term in the objective function,

$$\xi \sum_{i} \sum_{u \epsilon N(i)} z_{iu},$$

represents the *Spatial Attribute Coherence.* We model the energy as a constant penalty term ξ which is applied if any of the pixel sites in the neighborhood $N(i)$ if *i* have a different label assignment than that of pixel *i.* z_{iu} is a binary variable which takes 1 when a pixel *u* in neighborhood $N(i)$ *takes* a label other than that of pixel *i.* The constraint, $z_{iu} = \|x_i - x_u\|$ is modeled in equations1 and 2.

The third term in the objective function,

$$\sum_{i}\sum_{j}\sum_{k} \delta_{ijk} y_{ijk},$$

represents *Temporal Appearance Similarity and Motion Coherence.* The energy in this term represents that if pixel *i* is in time frame *t,* moves to pixel *j* in time $t+1$ and then to pixel k in time $t+2$, then a binary variable y_{ijk}takes on value 1 and 0 otherwise. This results in a cost of δ_{ijk}. Herein δ is the problem specific cost matrix calculated for every possible set of movements $\{i, j, k\}$ over the video sequence such that i, j and k are pixel indices in consecutive time frames *t,* $t+1$ and $t+2$. Furthermore it is also taken into account that every pixel i in frame t moves to only one correspoding pixel in consecutive periods $t+1$ and $t+2$. This constraint is modeled in equation 8.

The fourth term in the objective function,

$$\lambda \sum_{i} \sum_{u \epsilon N(i)} \alpha_{iu},$$

represents the *Spatial Displacement Coherence*. This term enforces that the neighborhood $N(i)$of pixel i also moves in the same manner as i. λ is a fixed cost applied whenever one of the members $u \epsilon N(i)$ does not match that of pixel i. $\alpha_{iu}$is a variable which represents the magnitude of distance of $u \epsilon N(i)$ in time t corresponding to pixel i in time t, when $u \epsilon N(i)$ in time t does not take the same label as corresponding pixels for i in time $t+1$ and $t+2$, i.e. j and k. This would in simpler terms mean that y_{ijk} and $y_{ujk}$are not same. It should be noted that *j* and *k* in y_{ujk} is such that j_u has the same spatial relationship with $j_i$as that of *u* with *i.* The resulting constraint,

$$\alpha_{iu} = \left\| \sum_{j}\sum_{k} d_{ijk} y_{ijk} - \sum_{j}\sum_{k} d_{ujk} y_{ujk} \right\| \quad \forall i \epsilon \mathcal{G}, \forall u \epsilon N(i),$$

has been modeled in equation 3 and 4. d_{ijk} is a constant distance metric for the specific problem which accounts for the distance between the pixel locations i, j and k both temporally and spatially. This set of constraints also account for the realistic fact that the the label of i should be affected to a higher degree by pixels in its neighborhood which are closer to i.

The fifth term in the objective function,

$$\sum_{i}\sum_{j} \left(\sum_{k} y_{ijk} \right) z_{ij},$$

represents the *Temporal Attribute Coherence*. Here the variable y_{ijk} is same as above. A new binary variable z_{ij} is introduced to account for difference in attribute assignments of a pixel in time t and $t+1$. This would mean that z_{ij} is 1 when x_i is not same as x_j. The resulting constraint $z_{ij} = \|x_i - x_j\|$ is modeled in equations 5 and 6. It is worth noting that the last term in the objective function,

$$\sum_{i}\sum_{j} \left(\sum_{k} y_{ijk} \right) z_{ij},$$

introduces non-linearity in an otherwise linear energy function. However, we have reduced this to linear form by introduction of another artificial variable w_{ij} and constraint 7 and the last term in objective function in this case will be modified to $\sum_i \sum_j w_{ij}$.

Equations 8,9,10 and11 enforces the constraint for optical flow, where each pixel is always passed onto corresponding pixels in the next frame. Hence for every y_{ijk} , there exists a k' in time $t+3$, such that $y_{jkk'} = 1$.

REFERENCES

[1] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *focs*, page 14. Published by the IEEE Computer Society, 1999.

[2] David Tsai, Matthew Flagg, and James M. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC'10*, pages 1–11, 2010.