Foreground segmentation in video by background tracking via guided superpixel flow.

Juan Perez Rua
juanmanuel.perezrua@technicolor.com
Tomas Crivelli
tomas.crivelli@technicolor.com
Patrick Perez
patrick.perez@technicolor.com

Technicolor
Rennes, FR

Abstract

Among the state of the art segmentation methods for objects in video sequences, point trajectories based ones stand for its performance and reliability, even when only sparse trajectories are known because of computational reasons. In the other hand, for the problem of extracting out a preselected object in still frames, max-flow min-cut based approaches have demostrated to be a powerful tool. We propose to mix these two ideas for reliable object segmentation through video. Moreover, in this paper we introduce the superpixel flow method for tracking of background regions, rather than sparse points. We show how this extra information can be used to expand the max-flow min-cut techniques for an efficient foregorund-background segmentation. Experiments on several video sequences demostrate the characteristics of the proposed method.

1 INTRODUCTION

Superpixels and over segmentation techniques became a widely used pre-processing stage for a large number of machine vision applications, after the original concept was introduced [1]. Superpixels are traditionally used as performance booster for several other techniques. However, it is still mostly related to single frame processing [1][10][11]. In the search for consistency in superpixel labeling through video, some authors have proposed different techniques, which go from simple extension to supervoxels[9][11], to more complicated approaches [8]. These approaches, nonetheless, usually require a global processing and knowledge of all (or several of) the video frames beforehand. One of the contributions of this work is a superpixel matching technique which assumes a flowlike behavior in the image sequences (natural video), and propose an application for improving object segmentation in videos.

For some kind of video sequences and computer vision applications the superpixel labelling usually loses coherency between frames, even when computed as supervoxels. This means superpixel matching techniques offer some interest and can still be explored. Some work have been done towards a superpixel based image comparison using the Earth Mover's

Distance, by taking super pixels as bins of a global histogram [7]. The label propagation or superpixel flow can be achieved with this technique by selecting the superpixel in the second frame that maximize the flow from each superpixel in the first frame. However, we look at the problem of superpixel propagation between a pair of frames, in terms of the indexes of the already computed superpixelization in both of them.

By taking into account superpixels computed separately in images we can open the advantages of streaming like approaches, so the video process can be performed by processing only two frames at a time, saving memory and moving towards a more time friendly approach. This propagation means finding the superpixel labels in the second frame that correspond to a given label in the first frame. This matching, however, has to comply with a set of constraints. Firstly, two correspondent superpixels should be similar in terms of some appearance feature, which most likely depends on the way the superpixelization was performed (color, texture, shape). Also, the superpixel propagation should maintain certain global homogeneity (at least for superpixels that belong to the same object). This is because the superpixel propagation or matching is not completely a one-to-one combinatorial problem, but more a âĂIJsuperpixel flowâĂİ. In this sense, it seems natural that the problem of superpixel propagation could be solved with a discrete energy minimization procedure. If the size compactness of the superpixels is maintained, it actually seems to share some of the properties of the optical flow problem, with the difference that the smoothness is usually a very strong constraint for the last one. The strength of this smoothness prior relies not only in the nature of the problem, but also because it gives better cues towards an easy-to- minimize global approach.

2 PROBLEM DEFINITION

The objective of the superpixel propagation is therefore to find the best labeling L for every superpixel p (with $L_p \in 0,1,...N-1$) between a pair of frames (I_1,I_2), but holding a flow-like behavior. Thus, the superpixelization should maintain certain regularity in size within a

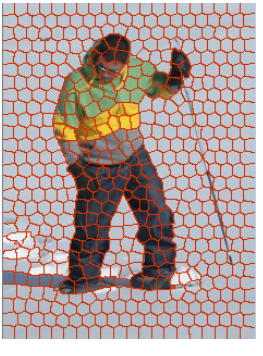

Figure 1: SLIC super segmentation in the Snow Shoes sequence.

single frame. Some super pixel techniques can cope with this requirement [9][10]. For the experiments presented in this work, we prefer the SLIC (Fig. 1) method [9], which usually gives good results in homogeneity of the superpixelization. The proposed steps to solve the propagation problem assume this requirement is hold. For other kind of the techniques, other approaches should be followed.

2.1 Energy Formulation

Inspired by a large number of optical flow and stereo techniques [7][12][13], the superpixel propagation can be modeled with pairwise Markov Random Fields. If the matching is performed with MAP inference, its posterior probability is:

$$P(L|I_0,I_1) = \prod_{p\in\Omega} \mathrm{e}^{-D_p(L_p;I_0,I_1)} \prod_{p,q\in\mathcal{N}_p} \mathrm{e}^{-S_{p,q}(L_p;L_q)} \tag{1}$$

With L the set of labels of the super pixels in I_0, that match with those in I_1. $\mathcal{N}_p$ is a neighborhood of the superpixel p, which defines its adjacency. Given this posterior probability, the equivalent energy function can be directly obtained by extracting the negative logarithm of the posterior,

$$E(L) = \sum_{p\in\Omega} -D_p(L_p;I_0,I_1) + \sum_{p,q\in\mathcal{N}_p} -S_{p,q}(L_p,L_q) \tag{2}$$

The terms D, and S in 2 stand for data and spatial as they are popularly known in the MRF literature. The first one determines how accurate is the labeling in terms of consistency of the measured data (Color, Shape, etc.). In the usual optical flow counterpart of this equation, the data term corresponds to the pixel brightness [7][5]. However, as superpixels are a set of similar (or somehow homogenous) pixels, an adequate color based feature can be a low binned histogram or its average color. So it can be written more precisely as

$$D_p(L_p;I_0,I_1) = \rho(h(p),h(p')) \tag{3}$$

Where $h(p)$ and $h(p')$ are the histograms of the super- pixel p and its correspondent superpixel in the second frame (I_1). *rho* can be replaced by the Bhattacharyya distance. Note that the low binned histogram or average color gives certain robustness against noise, and slowly changing colors between frames. The spatial term is a penalty function for horizontal and vertical changes of the vectors that have origin in the centroid of the super-pixel of the first frame and end in the centroid of the super-pixel of the second frame.

$$S_{p,q}(L_p,L_q) = \lambda(p)\sqrt{\frac{|u_{p_c}-u_{q_c}|}{\|p_c-q_c\|} + \frac{|v_{p_c}-v_{q_c}|}{\|p_c-q_c\|}} \tag{4}$$

$$\text{where, } \lambda(p) = (\rho(h(p),h(p')))^2$$

In 4 the operator ρ has the same meaning as in the data term 3. The histograms distance is nonetheless computed between super pixels p and q, which belong to the same neighborhood. The super pixels centroids are noted as q_c and p_c, and u and v are the horizontal and vertical changes between centroids. This term is usual in the MRF formulation and has a smoothing effect in superpixels that belong to the same object. It has to be observed that when two superpixels are different, thus, more probable to be in different semantic groups within the image, the term λ allows them to have matches that do not hold the smoothness prior with the same strength. It has to be noted that the proposed energy function is highly non-convex and robust.

2.2 Energy Minimization

A fair amount of work had been dedicated to discrete optimization techniques in computer vision, leading to a couple of well-defined and widely tested approaches to solve the pairwise MRF of labels [3][4]. However, some of the approaches restrict the construction of the spatial term, and/or enforce limitations in the number of labels [3]. Because of the high amount of possible labels for each element in the proposed approach, the use of the Fusion Moves [7] technique seems to be well suited. This algorithm employs the Quadratic Pseudo- Boolean Optimization (QPBO) graph-cut, to combine incremental sets of proposal labeling, resulting a semi- globally-optimal solution [4]. Thus, the minimization starts by proposing a set of possible solutions, and iteratively merge them with the QPBO graph-cut technique.
The possible solutions that can be given depend on the kind of problem that is intended to be solved. For example, in stereo super-pixel matching, some assumptions related to the cameras organization can be made to generate solutions. In a more generic sense, other assumptions can be made towards option generation. For instance, for a given super-pixel in the initial frame, in the second frame the corresponding matching would be the most similar one in terms of color, or the most similar un terms of shape, or the spatially closer super-pixel. More proposal solutions can be added by defining a neighborhood in the second frame and select random pairs from every neighborhood of every super-pixel in the first frame. This is more suitable for problems where the images are extracted from the same video sequence. It is interesting to notice that different assumptions for this neighborhood can lead to a technique for generic image based retrieval, where the total cost of matching can be used as metric.

3 EXPERIMENTAL RESULTS

The Fig. 2 shows some examples of superpixel matching with the presented method. It can be seen that the matching performs well even in difficult cases, like the hands in the top row. It has to be noted as well that even in superpixels where there is a lack of texture, there is correct matching. This seems to be the effect of enforcing the regularization between superpixels that are close, but are also similar to each other.

Moreover, unlike most of the optical flow methods, superpixel flow extends naturally for more distant frames. The Fig. 3 shows results for larger separations between frames, without tweaking or adjusting any parameters. For this case, however, the matches in the textureless part of the scene are mostly invalids. Though this is expected because because of the aperture problem and heavy occlusions.

3.1 Superpixel propagation for object segmentation in videos

The GrabCut algorithm proposed in [14], offers a good deal in terms of background-foreground separation from user interaction. A technique like this, however, performs very well in still images, but it may not be well adapted for sequential videos, The GrabCut works by implementing an iterative graph cut based minimization to separate regions according to appearance information, that can be extracted from the user interaction. This interaction, however, could be minimized in videos, given the extra information that offers the flow of the sequence. Some authors had approached the GrabCut or similar graph based segmentation techniques in sequential videos, to propagate a consistent segmentation [15]. However, some

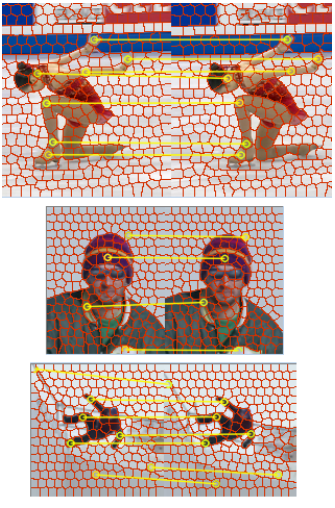

Figure 2: The yellow lines show selected super-pixel matching between pairs of subsequent frames in a video with the proposed method. The video frames go from right to left.

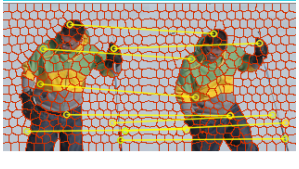

Figure 3: The yellow lines show selected superpixel matching between a pair of distant frames in the Snow Shoes sequence.

more work on reducing user interaction given the extra flow-like information that video sequences offer is still needed. We propose to combine the presented super pixel propagation as an automatic method to initialize the Grabcut (or similar) algorithm to perform object segmentation through frames in a video sequence.

The idea is to track (or more exactly, match) super pixels that are labeled as foreground, thanks to an object tracker initialization. Thus, the super pixels that are initially outside the ROI, can be propagated through the sequence, and if they fall into the ROI of the next frame, they can be safely labeled as foreground again. The process is repeated for any labeled superpixel through the video. Having several labeled superpixels can reduce widely the necessity for user interaction in subsequent frames. Thus, to perform object segmentation in a full video sequence, the required user interaction would only be the initial bounding box. Moreover, a fully automatic approach can be obtained if a reliable object detector is available.

Fig. 4 shows how the initialization of a tracker and a superpixelization provides information for better background-foreground separation. The background-foreground models are updated as the frames go on, giving more robustness for sequential propagation of the segmentation. The method is tested in the Walking Couple sequence, by allowing only a small amount of iterations in the graph based segmentation. Observe how the contour in the manâĂŹs head is correctly delineated when another personâĂŹs head occludes part of it. In this case, the super-pixels that belong to the womanâĂŹs face were correctly propagated and thus, labeled as background.

In order to understand the effect of including super- pixel propagation in a video sequence for object segmentation, some results are shown in the Fig. 5. For these experiments only one iteration is allowed in both grab-cuts initialized with only the tracker, and the one performed with the super-pixel propagation. Observe that in general, the contour delineated is usually better in terms of precision and stability for the later one.

References

[1] J. Malik and X. Ren, Learning a classification model for segmentation, *Computer Vision, International Conference*, 2003.

[2] S. Boltz; F. Nielsen and S. Soatto, Earth mover distance on superpixels*International Conference on Image Processing*, 2010.

[3] E. Boros; P. Hammer and G. Tavares, Preprocessing of unconstrained quadratic binary optimization *RUTCOR*, 2010

[4] E. Boros and P. Hammer, Pseudo-boolean optimization*Discrete applied Mathematics.*, 2002

[5] B. Horn and B. Schunck, Determining Optical Flow*Artificial Intelligence*, 1981

[6] H. Ishikawa and P. Bouthemy, Multimodal estimation of discontinous optical flow using Markov random fields*TPAMI.*, 1993

[7] V. Lempitsky, S. Roth and C. Rother, Fusion Flow: Discrete-Continuos optimization for optical flow estimation*Computer Vision and Pattern Recognition*, 2008

Figure 4: Segmentation through the sequence âĂIJWalking CoupleâĂİ (Yellow contour) initialized in the manâĂŹs head.

Figure 5: Face segmentation in the âĂIJAmelie RetroâĂİ and the âĂIJSnow shoesâĂİ sequences in three different frames. For each group, the Top Row: One-iteration window-based grabcut; and the Bottom Row: One-iteration grabcut with super pixel propagation.

[8] M. Reso and J. Jachalsky, Temporally Consistent Superpixels*International Conference Computer Vision.*, 2011

[9] R. Achanta; A. Shaji; K. Smith; Aurelien Lucchi; P. Fua and S. Susstrunk SLIC Superpixels compared to state of the art superpixel methods*Discrete applied Mathematics.*, 2002

[10] F. Perbet and A. Maki, Homogeneus superpixels from random walks*MVA.*, 2011

[11] C. Xu and J.J. Corso. Evaluation of super-voxel methods for early video proccesing. *Computer Vision and Pattern Recognition*. 2012.

[12] A. Shekhovtsov, I. Kovtun and V. Hlavac. Efficient MRF deformation model for non-rigid image matching. *Computer Vision and Pattern Recognition*. 2007.

[13] J. Sun, N.N Shen and H.Y. Shum. Stereo matching using propagation belief. *TPAMI*. 2003.

[14] C. Rother, V. Kolmogorov and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *SIGGRAPH*. 2004.

[15] L. Yang, Y. Guo, X. Wu and X. Wang. A new video segmentation approach: Grabcut in local window. *Soft Computing and Pattern Recognition*. 2011.