SeamSeg: Video Object Segmentation using Patch Seams

S. Avinash Ramakanth and R. Venkatesh Babu
Video Analytics Lab, SERC,
Indian Institute of Science, Bangalore, India.
avinashrs@ssl.serc.iisc.in, venky@serc.iisc.in

Abstract

In this paper, we propose a technique for video object segmentation using patch seams across frames. Typically, seams, which are connected paths of low energy, are utilised for retargeting, where the primary aim is to reduce the image size while preserving the salient image contents. Here, we adapt the formulation of seams for temporal label propagation. The energy function associated with the proposed video seams provides temporal linking of patches across frames, to accurately segment the object. The proposed energy function takes into account the similarity of patches along the seam, temporal consistency of motion and spatial coherency of seams. Label propagation is achieved with high fidelity in the critical boundary regions, utilising the proposed patch seams. To achieve this without additional overheads, we curtail the error propagation by formulating boundary regions as rough-sets. The proposed approach out-perform state-of-the-art supervised and unsupervised algorithms, on benchmark datasets.

1. Introduction

Video object segmentation divides a video into component objects, by spatially segmenting objects in every frame *i.e*., the aim of object segmentation is to group pixels in a video into spatio-temporal regions that exhibit coherence in both appearance and motion [9]. The general problem of video object segmentation becomes ill-posed because, i) the number and types of objects in a video are unknown, ii) videos do not generally consist of a single scene, iii) the background in a video is not always well-behaved, since variations in background cannot be modelled accurately.

Hence to well-define the problem of video object segmentation, existing methods make two major assumptions viz., 1) given video is composed of a single scene or action. 2) The object being segmented is present across all frames [6, 11, 13, 19]. In addition to this, segmentation requires prior knowledge of the object to be segmented. Since semantic object detection is itself an ill-posed problem and because definition of correct objects varies between different problems, existing approaches use one of the following three criterion to circumvent object detection, i) label key-frames which could be either initial frame or multiple frames [2, 6, 8, 19] or ii) perform over segmentation (*e.g*. based on super-pixels) and with user intervention, obtain the final object segmentation [5, 18, 21] or iii) assume the focus of video is a single object [11, 13, 22]. Apart from the initial object detection problem object segmentation is a complex problem since it needs to handle problems such as, abrupt object motion, motion blur, compression artifacts, illumination and pose change of objects, non-rigid object structures, occlusions, camera motion etc.

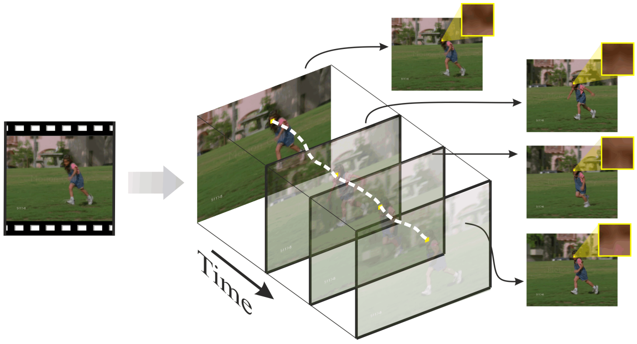


Figure 1: Video seams are used to capture the motion of objects across frames. As can be observed, the path shown in white connects patches across frames, to efficiently transfer object labels.

In this paper, we begin with user-defined object labels at the beginning of a video and propagate these labels using video seams. Existing approaches that deal with label transfer have focused on establishing mappings between frames via optic flow algorithms [8, 9] or long term point trajectories [5, 12]. However these methods have not been able to achieve satisfactory results for semantic label propagation [2, 7]. The shortcomings of these approaches include lack of occlusion handling, high cost of multilabel MAP inference, sparsity of robust mappings and label drift caused by rounding errors. These issues have led to the use of label inference over short overlapping time windows as opposed to a full length video volume [6, 19].

Our label propagation scheme is motivated by Avidan *et al.*'s work [1], which utilised connected paths of low-energy in images to re-size images. These seams were further extended by Rubinstein *et al.* [17] for video retargeting. An illustration of how the proposed seams compare with existing formulations is shown in Fig. 1 and 2. In the existing image or video resizing approaches, seams minimise energy at pixel level by connecting pixels which minimise a chosen energy function. In our proposed approach, we adapt seams to connect $p \times p$ patches across frames, such that the distance between these patches is minimised with an additional constraint that seams in coherent regions move coherently. To minimise the energy function accurately, we adapt approximate nearest neighbour algorithm to compute a mapping between two frames, thus forming seams by connecting patches across frames. Since seams minimise energy across all $p \times p$ patches temporally, every pixel in a frame is contained in p^2 seams with corresponding labels. The final label for each pixel is assigned by examining the probability distribution of all p^2 labels. To decide the labels of each pixel, we make use of rough sets, by which we estimate if a pixel is in the positive, negative or boundary region. A pixel belongs to the positive or negative region, if it belongs or not to a label set respectively. A pixel is in the boundary region if decision about pixel belonging to either positive or negative regions cannot be taken reliably with available information.

To sum up, the current approach combines video seams, approximate nearest neighbour fields (ANNF) and rough sets to perform video object segmentation. In the next section we will present a brief overview of existing video segmentation techniques, followed by, detailed explanation of video seams, ANNF maps and rough sets in section 3. The proposed approach is described in section 4, followed by experiments in section 5. Setion 6 concludes with a brief note on future directions.

2. Related Work

Video object segmentation can be broadly classified into following two categories:
1) *Unsupervised segmentation*, aims at autonomously grouping pixels in a video, which are visually and motion-wise consistent. Recent techniques, as summarised by Xu *et al.* [21], have been inspired by super-pixels in images and focus on merging image super-pixels, based on motion consistency to form space-time super-pixels [5, 18, 21]. In cases with clear boundaries between objects and background, the super-voxels are semantically meaningful, however in real world videos, the results are over-segmented and require additional knowledge, for example in the form of human intervention, to achieve object level segmentation. The second widely developed unsupervised segmentation algorithms start with a goal of detecting the primary object in a video and to delineate it from the background in all frames [11, 13, 22]. As it is evident from the formulation, this approach requires the full video or at least a bunch of frames to analyse and work with the assumption that there is only a single object present throughout the video. Recently, Zhang et al. [22] merged both super-pixels and foreground object segmentation to obtain accurate unsupervised object segmentation.
2) In *Semi-Supervised segmentation*, first frame or key frames are user labelled and the object is segmented in the remaining frames. Badrinarayanan *et al.* [2] proposed a probabilistic graphical model for propagating labels in video sequences, which used multi-frame labelling, typically at the start and end of the video. Expectation maximisation is used to propagate labels across frames along with random forest classifiers. Similarly, Budvytis *et al.* [6] also used mixture of trees graphical model for video segmentation. This approach transferred labels provided by the used, via a tree structured temporal linkage between super-pixels from the first to the last frame of a video sequence.

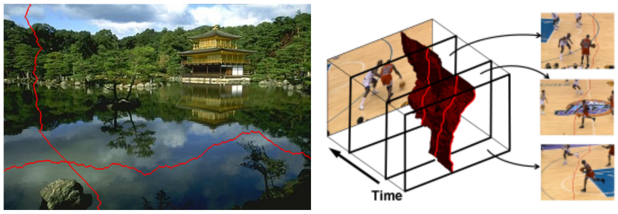
Fathi *et al.* [8] proposed an algorithm for video segmentation using harmonic functions and an incremental self-training approach, which iteratively labels the least uncertain frame and updates similarity metrics. Using active learning models for providing guidance to user on what to annotate in order to improve labelling efficiency. This method was able to achieve accurate pixel level object segmentation. Another off-line algorithm was proposed by Tsai *et al.* [19] using multi-label Markov Random Fields. Segmentation was achieved by finding the minimum energy label assignment across frames. One of the drawbacks with off-line segmentation methods is the large memory requirement. In comparison to such off-line methods the proposed approach is based on a sequential processing and has very low memory requirements.

3. Background

This section starts with an introduction of seams for videos and computation of seams using modified ANNF maps. A brief introduction of rough sets is also provided before proceeding to explain the proposed approach in next section.

3.1. Seams in a Video

Avidan *et al.* proposed seam carving [1] to perform image resizing/retargeting by introducing the concept of seams in an image. The aim of seam carving is to reduce size of an image by removing seams from an image instead of removing a row or column of data. Seams are connected paths of low energy in an image [1], *i.e.* the sum of gradients along



(a) (b)

Figure 2: Comparison of seams. (a) Seams used for image retargeting [1] (b) Seams used for video retargeting [17]. The proposed seams are shown in Fig. 1.

the seam path is minimised. this happens since the energy function for retargeting, is conventionally based on spatial gradient of image. In other words seams conventionally pass through planar regions and avoid major edges when retargeting to preserve the salient regions in an image. Similar idea of minimising gradient energy is also used for video retargeting [17], where the seams are connected both spatially and temporally. The seams in an image are of size $[width, 1]$ or $[1, height]$ when reducing height and width respectively. Whereas the seams in a video are of size $[width, 1, f]$ or $[1, height, f]$ when reducing height and width respectively. Here, 'f' indicates the total number of frames in the video, since for video retargeting, seams are connected both spatially and temporally. Figure 2 illustrates how seams connect regions of low energy when retargeting images and videos, in comparison, Fig. 1 shows the proposed seams which minimise energy temporally.

When adapting seams to video segmentation, we encounter the following drawbacks, with conventional formulation:

1) *Energy function:* In retargeting, a gradient based energy function is used since the objective is to avoid removing salient objects. This any major edges, while in segmentation, an object is defined by its boundaries and textures, and seams which avoid such regions are not helpful. Hence we modify the energy function, as explained in section 4, eq. (3), such that it captures the motion of objects, instead of energy function based on gradients.

2) *Seam sizes:* The seams in videos span the width or height of the frames, while an object is a sub-region in a frame and does not span the whole width/height. In other words, for seams to be able to capture the motion of object, they need to be at object level. To handle this, we define seams as paths of size $p \times p \times f$ which connect patches across frames. Here, p is patch size and f is total number of frames in a video.

3) *Seam coherency:* When retargeting, there is no relationship between seams, a reason for avoiding relationship between seams is that removing excessive information from adjacent locations creates artifacts after resizing. On the other hand, adjacent seams within an object must be coherent to accurately model object motion.

4) *Connectivity:* While retargeting, though a seam is connected both spatially and temporally, only the 8-connected neighbourhood of the pixel contained in the seam is considered for propagating/connecting the seam. When modelling object motion, a patch need not always be overlapping, *i.e.* motion more than one pixel also needs to be captured efficiently.

3.2. Approximate Nearest Neighbour Field

To efficiently propagate labels using seams, one criterion is that patches along a seam must be similar. This objective is in line with approximate nearest neighbour field algorithms, like PatchMatch [3], FeatureMatch [15, 16] and Coherency Sensitive Hashing [10]. The aim of such algorithms is: "For a pair of images (target and source), for every $p \times p$ patch in the target image, find the closest patch in the source image (minimum Euclidean distance, or any other appropriate measure)." The optimisation function in the existing ANNF map algorithms is based solely on the patch distance, and to improve the accuracy of mapping between two images, the coherency of images is exploited *i.e.*, if two patches are similar in a pair of images then their neighbouring patches will also be similar.

In computing the ANNF mapping between two frames in a video, from image I_t to image I_{t-1}, the energy function is defined as the Euclidean distance between patches:

$$E'_{i,j,t}(x, y) = ||I_t(i, j)_p - I_{t-1}(x, y)_p||_2 \qquad (1)$$

$I_t(i, j)_p$ be a $p \times p$ patch at (i, j) in image I_t, which maps to $p \times p$ patch at (x, y) in image I_{t-1}, $I_{t-1}(x, y)_p$, if E' is minimised, and we denote this mapping as $I_t(i, j)$:

$$I_t(i, j) = (x, y) \Leftrightarrow \underset{x,y}{arg\,min}\; E'_{i,j,t}(x, y) \qquad (2)$$

For computing video seams, we adapt existing ANNF computation methods [15, 16]. In these methods E' is approximately minimised, by searching for nearest neighbour among all patch in I_{t-1}, for every patch in image I_t. To speed up the search every patch in image I_t and I_{t-1} are represented with lower dimension features. The lower-dimension feature representation for each patch, helps in using fast nearest-neighbour algorithms, like kd-tree. To approximate a $p \times p$ patch, the colour information is captured by using the mean of R(ed), G(reen) and B(lue) channels, the direction information is captured using mean of x-, y-gradient, the first two frequency components of Walsh Hadamard bases [4] and the maximum value of the patch. The advantage of these features is that they are extremely efficient to compute using integral images [20]. Furthermore, lower dimension representation using these proposed features, is much more accurate for computing ANNF maps, in comparison to standard dimension reduction techniques like PCA and random projection [15].

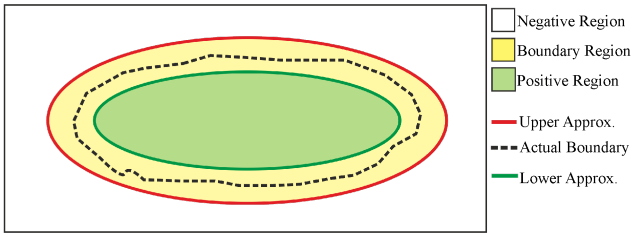


Figure 3: The boundary of a label set is formalised by rough sets.

3.3. Rough Sets

In label propagation and video object segmentation, label sets are conventionally modelled as crisp sets, *i.e.* if p is any pixel in image I and X is an object label.

$$\forall\, p \in I,\ p \in X \; or \; p \in X'$$

In such crisp sets, there is no modelling for the boundary entities, *i.e.* crisp sets do not model $\{p \in I \mid p \notin X \text{ and } p \notin X'\}$. These are boundary pixels which cannot be confidently classified to a label set or outside the label set, *i.e.* these are pixels which may belong to a label set but the confidence of belonging to the set is low. The problem with formulating labels as crisp sets, arises when the boundaries are mis-labelled and propagated. To handle such mis-labelling, conventional labelling techniques warrant further optimisation to handle boundary pixels much more accurately. To provide an objective form of analysing these low confidence entities without any additional information/optimisation, we make use of Rough sets as proposed by Pawlak [14], and illustrated in Fig. 3.

In a rough set, the lower approximation or positive region $\underline{P}X$, is a union of all the entities which definitely belongs to the target label set X, *i.e.* an entity will unambiguously belong to a given label set if it belongs to $\underline{P}X$.

$$\underline{P}X = \{p \mid p \subseteq X\}$$

Similarly, the upper approximation $\overline{P}X$, is union of all entities which have non-empty intersection with the target set, *i.e.* union of all entities that may possibly belong to the target set forms $\overline{P}X$.

$$\overline{P}X = \{p \mid p \cap X \neq \emptyset\}$$

Thus, the set $\mathbb{U} - \overline{P}X$ constitutes the negative region, containing all entities that can be definitely ruled out as members of target label set.

$$\forall\, p \in I,\ p \in X' \;\Leftrightarrow\; p \in \mathbb{U} - \overline{P}X$$

In other words, an entity within the upper approximation is a possible member of the target label set and an entity within the lower approximation definitely belongs to the target label set. The boundary region, given by $< \overline{P}X - \underline{P}X >$, consists of the entities that can neither be ruled in, nor ruled out as members of the target label set X.

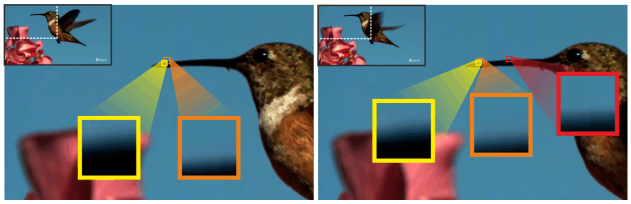


Figure 4: Difference between E' and E. E' does not require connectivity or coherency of seams, *i.e.* orange patch in image I_t (on left) can match to red patch in image I_{t-1} (on right). On the other hand E ensures connectivity of orange patch across time. This is further enforced by coherency *i.e.* seams through orange and yellow patches should flow together.

4. Algorithm

As discussed in previous section, we make use of video seams to propagate labels temporally for object segmentation. A seam flows from patch in frame 't-1' to patch in frame 't' when the following energy function, E, is minimised.

$$\begin{aligned} E_{i,j,t}(x,y) = \sigma_1 * ||I_t(i,j)_p - I_{t-1}(x,y)_p||_2 \\ +\ \sigma_2 * ||(i,j) - (x,y)||_2\ +\ \sigma_3 \\ * \sum_{\delta,\epsilon} ||I_t(i,j) - I_t(i+\delta, j+\epsilon)||_2 \end{aligned} \qquad (3)$$

Here, I_t is a frame at time 't' and I_{t-1} is a frame at 't-1'. $E_{i,j,t}(x,y)$ represents the energy between a $p \times p$ patch at (i,j) in I_t and patch at (x,y) in I_{t-1}. As can be compared, between E' and E, the additional terms in eq. (3), enforce connectivity and coherency of seams, as illustrated in Fig. 4.

In conventional ANNF mapping, eq. (1), there is no $||(i,j)-(x,y)||_2$ term, *i.e.* a patch in image I_t can match to any patch in image I_{t-1} as long as E' is minimised. But in a video there exists connectivity of patches across frames, and to capture this relation, we penalise the energy function based on how far is the matching patch in frame 't-1' from current location in frame 't'. Also to ensure adjacent seams flow coherently, we introduce $||I_t(i,j) - I_t(i+\delta, j+\epsilon)||_2$, which minimises the neighbourhood incoherence. This coherency term is not used in conventional ANNF mapping, since more the incoherence in mapping between two images, the better is the accuracy of ANNF mapping [10]. It is required that adjacent pixels move coherently, while propagating labels, hence this term is introduced to minimise

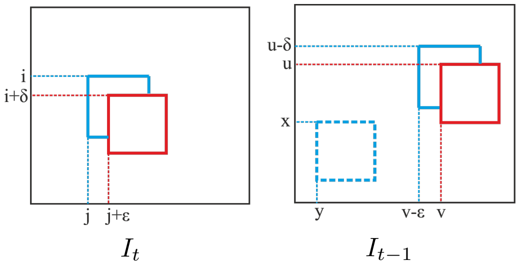


Figure 5: Initially $I_t(i,j) = (x,y)$. By checking for seam coherency, if $E_{i,j,t}(u-\delta, v-\epsilon) < E_{i,j,t}(x,y)$ then $I_t(i,j) = (u-\delta, v-\epsilon)$, since this mapping provides a better minima for the energy function 'E'.

in-coherency of mapping, to accurately capture object motion.

Conventionally, the energy across a seam is minimised, by piece-wise minimisation of the associated energy function. To connect seams across a video, we minimise E for every patch, between every adjacent frame. That is, a seam flows from patch (i,j) in frame I_t to patch at (x,y) in I_{t-1} if E is minimised. This minimisation is performed in two stages. For the first stage σ_3 is set to 0, *i.e.* we optimise only for patch similarity and connectivity and in the second stage we optimise for coherency of seams as well.

A $p \times p$ patch in an image is approximated to an 8 dimensional feature formed by concatenating the mean of R(ed), G(reen), B(lue) channels, average x and y gradients, first two frequency components of Walsh-Hadamard bases and the maximum value of the patch. These features were shown to capture the colour and direction information of a $p \times p$ patch, with better accuracy than conventional dimension reduction approaches [15], while being extremely computationally efficient, since these features can be computed using 'integral images'. In addition, to incorporate the connectivity of seams, we add the x and y co-ordinates weighted by σ_2/σ_1, to form a ten-dimension feature. For every patch in I_t, kd-tree is used to search for patches in I_{t-1}, which approximately minimises eq. (3).

The second stage optimises the mapping between frames by taking into account the coherency of seams. The mapping obtained from earlier stage is used as initialisation, and for every $p \times p$ patch in I_t, the mapping in neighbourhood is considered to improve the energy function E. Suppose after first stage,

$$I_t(i,j) = (x,y)$$

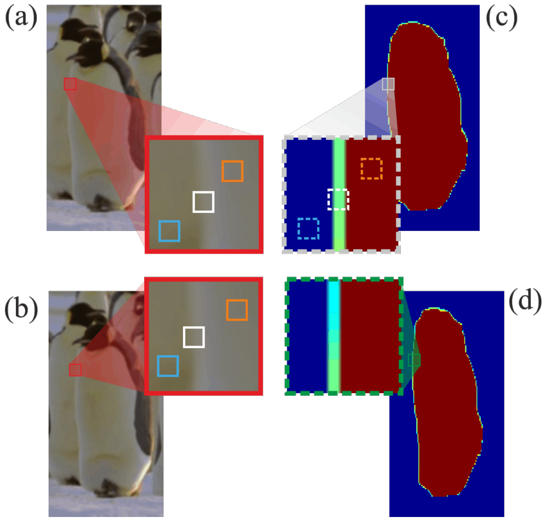


Figure 6: Transfer of labels across frames. (a) frame at t-1, (b) frame at t, (c) is previous label set, (d) shows inferred label set for current frame.

That is, E is approximately minimised between $p \times p$ patch at (i,j) in I_t and $p \times p$ patch at (x,y) in I_{t-1}. Consider a mapping in the neighbourhood of (i,j),

$$\begin{aligned}&\text{Say } I_t(i+\delta, j+\epsilon) = (u,v),\\ &\text{if } E_{i,j}(u-\delta, v-\epsilon) < E_{i,j}(x,y),\\ &\text{then } I_t(i,j) = (u-\delta, v-\epsilon).\end{aligned}$$

This is illustrated in Fig. 5, and it can be observed that this stage of optimisation helps in capturing coherently moving pixels of the object.

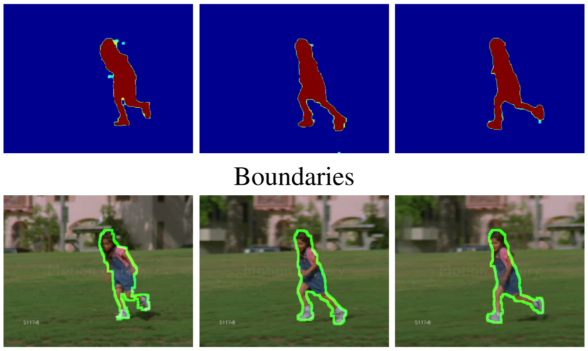
Every pixel in image I_t, $I_t(i,j)_1$, is contained in $p \times p$ different seams. Hence each pixel at (i,j) has p^2 different labels associated with it, denoted by the set $\{L_{i,j,t}\}$. For a label X, we define a pixel to belong to the positive or negative regions, if:

$$\begin{aligned}I_t(i,j)_1 \in \underline{P}X &\Leftrightarrow |L_{i,j}|_X >= \alpha * p^2\\ I_t(i,j)_1 \in \mathbb{U} - \overline{P}X &\Leftrightarrow |L_{i,j}|_{X'} >= \alpha * p^2\end{aligned}$$

where, $|L_{i,j}|_X$ is the number of elements in the candidate label set $\{L_{i,j}\}$, belonging to label X. The boundary pixels are thus defined as, all pixels in image A which do not fall in the positive or negative regions.

$$\begin{aligned}I_t(i,j)_1 \in <\overline{P}X - \underline{P}X> \quad &\Leftrightarrow\\ &I_t(i,j)_1 \notin \mathbb{U} - \overline{P}X \ \&\ I_t(i,j)_1 \notin \underline{P}X\end{aligned}$$

Figures 6 and 7 show flow of labels across frames, and it can be observed that boundary pixels are captured accurately with this formulation. Figure 6(a) shows frame at t-1 and Fig. 6(b) shows frame at t. The connectivity of seams is shown with orange seam in object region, white on the boundary and blue in the background region. As can be


Boundaries

Segmentations

Figure 7: Figure illustrates how rough sets are able to capture the boundary regions accurately. Here red denotes object, blue denotes background and the rest is boundary. For best viewing please zoom in on a digital display.

seen, orange and blue seams are within positive and negative regions respectively, and the same labels are propagated. On the other hand, seam through white patch cannot be classified into either of these label sets and hence is propagated as boundary label. To provide the final segmentation, these boundary sets are divided in the middle and shown as contours as illustrated in Fig. 7.

5. Experiments

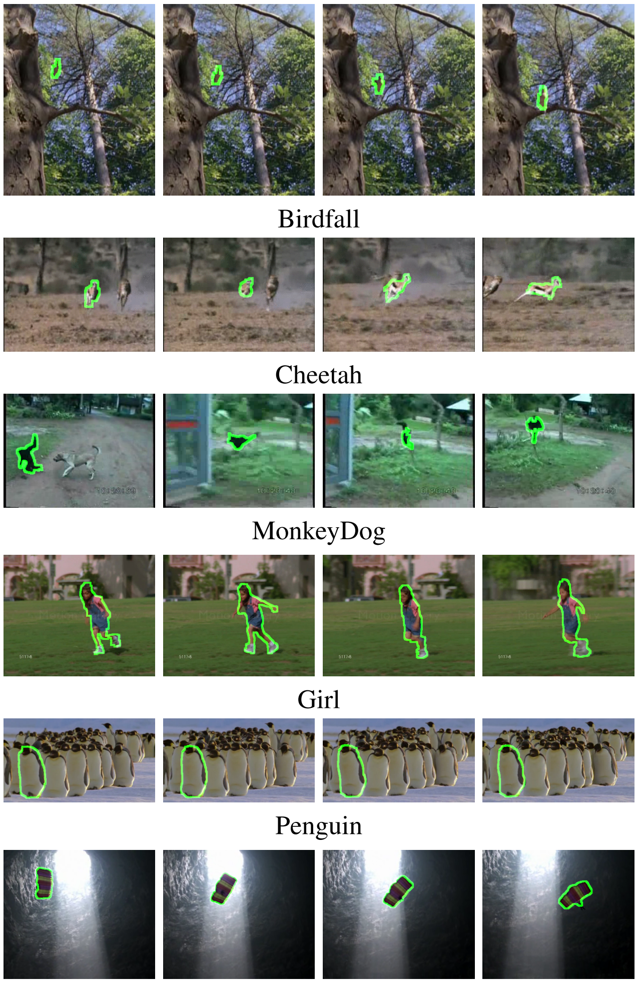
The proposed algorithm is implemented in MATLAB, with C/C++ implementations for critical/slow functions. All the experiments are executed on an Intel i7, 3.4GHz processor with 8GB RAM.

For segmenting a video into its component objects, we begin with a pixel level labelling, either using available ground truth or user-defined labels. In case of object and background, the label set contains object labelled as +1, and background as -1. In formulating the lower and upper approximation sets, α is set at 0.8, *i.e.* if the label set $L_{i,j,t}$ contains more than 80% elements of one label, then it is assigned to that particular label. The label of a boundary pixel is the sum of all the labels of seams which contain that particular pixel. The boundary region label, thus propagated belongs to $(-1, 1)$. To provide an output for segmentation in each frame, the boundary region is divided in the middle and drawn as output contour. The three constituents of the energy function, viz. patch similarity, coherency and connectivity are given equal weights, *i.e.* $\sigma_1 = \sigma_2 = \sigma_3 = 1$, and the patch size p is defaulted to 4.

Quantitative Evaluation

In this section, we show comparison of proposed approach against various state-of-the-art methods on the SegTrack dataset [19]. The results shown are taken from the respective author's publications. The error measure shown in table 1, is the average number of pixels mis-labelled per frame. The error is defined as $e(S) = \frac{|XOR(S,GT)|}{F}$, where S is the segmentation output, GT is the ground-truth segmentation and F is the total number of frames. We have used the standard first frame ground truth provided in SegTrack dataset as initialization, so that the comparison with existing methods will be fair. Segtrack database is fairly complex, with different foreground objects and cluttered background, on which the proposed approach performances better than existing methods, indicating the robustness of the proposed approach.

The proposed approach out-performs existing state-of-the-art methods in 4 out of 6 videos as well as giving the lowest overall error. In the other two videos, the proposed approach performs comparable to the state-of-the-art. The marginal fall in performance could be attributed to the small


Birdfall

Cheetah

MonkeyDog

Girl

Penguin

Parachute

Figure 8: Results on SegTrack database. For best viewing please zoom in and view on a digital display.

	SeamSeg (Proposed)	**Tsai *et al.*** [19]	**Lee *et al.*** [11]	**Ma *et al.*** [13]	**Budvytis** ***et al.* [6]**	**Fathi *et al.*** [8]	**Zhang *et al.*** [22]
Birdfall	186	252	288	189	508	342	**155**
Cheetah	**535**	1142	905	806	855	711	633
Girl	**761**	1304	1530	1698	1200	1206	1488
MonkeyDog	**358**	563	521	472	412	598	365
Parachute	249	235	**201**	221	296	251	220
Penguin	**355**	1705	136285	-	1736	1367	-
Mean	**407.3**	866.8	23288.3(689)	677.2	834.5	745.8	572.2

Table 1: Comparison of average number of error pixels per frame for proposed approach against state-of-the-art methods, on the SegTrack dataset.

size of the object, especially for birdfall sequence. In the parachute sequence the person on the parachute causes the error. In the first frame the person is labelled as background and in initial frames the proposed approach classifies the person as background. After a few frames the person is part of object and once he leaves the parachute, he is still classified as object by proposed approach. Though the person is accurately captured even with extreme similarity with the background, the error creeps up due to difference from ground truth as shown in Fig. 8.

It can also be noted that the performance of proposed approach on the penguin sequence is much superior to all the existing methods. The existing methods perform poorly in this video since the motion of penguins is similar as well as the regions are visually very similar. In foreground estimation methods [11, 22], the group of penguins is classified as object of interest and thus produce very high error. In other methods, due to high confusion between adjacent regions, both visually as well as motion wise, the error is comparatively higher. This confusion arising from cluttered background is handled effectively by the proposed approach, as can be observed in both penguin and cheetah sequences.

Qualitative Evaluation

Results obtained using proposed approach on the SegTrack database are shown in Fig. 8. As can be observed, the segmentation accuracy of the proposed approach is very high in all sequences even under cluttered background, like Cheetah, Birdfall and Penguin sequences. The proposed seams inherently handle scale (object size) variations efficiently. The scale variation is handled through splitting and merging of seams *i.e.* as the object grows, seams split (additional seams are generated adaptively), and when objects shrink multiple seams merge together. This scale handling capability can be observed in Birdfall, Cheetah and MonkeyDog sequences where the size of the object varies widely. The proposed approach is also able to handle fast motions like in MonkeyDog, Parachute and Yuna Kim sequences. To evaluate the performance of proposed approach on longer sequences, we experiment with the Yuna Kim sequence, and results for same are shown in Fig. 9. This complex and lengthy sequence demonstrates the robustness of proposed approach for fast object motion and partial occlusion. This shows, the capability of proposed approach to handle accurate label propagation in longer sequences while curtailing the error in boundary regions.

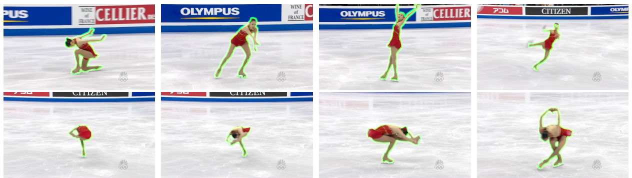
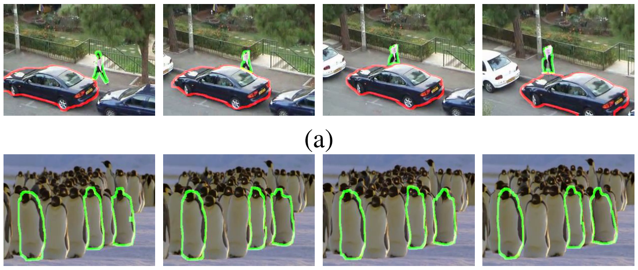


Figure 9: Performance on a long video sequence, *Yuna Kim*, with extreme deformations and motion blur. Frames shown - 6, 19, 37, 86, 124, 128, 190, 247.



(a)

(b)

Figure 10: Multi-object segmentation. (a) shows segmentation with partial overlap of various object. (b) shows multi-object segmentation with no overlap.

The proposed approach can also handle multi-object segmentation efficiently as illustrated in Fig. 10. To handle multi-object segmentation, all other object labels are treated as negative class with respect to one object, *i.e.* when propagating labels for a particular object all other object labels are treated as part of background.

To further illustrate the robustness of proposed approach,

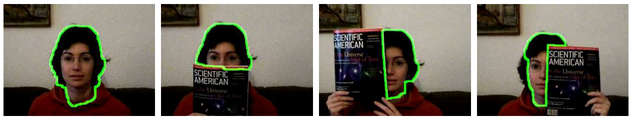


Figure 11: Segmentation with partial occlusion.

we show results for segmentation with partial occlusions in Fig. 11. As can be observed in face sequence, even with heavy occlusion, the proposed approach accurately segments the object in subsequent frames.

Advantages and Limitations

The major advantages of proposed approach are:

- By modelling a video as connected seams, and performing piece-wise minimisation, we alleviate huge memory and computational requirement needed by time window based video volume processing.
- By formulating label sets as rough sets, there is no need to do extra optimisations to handle boundary pixels, and boundaries can thus be modelled with existing information alone.
- The proposed approach is computationally efficient, taking less than a minute for processing the MonkeyDog video (frame size of 320 × 240 pixels, with 71 frames), on a Intel i7, 3.4GHz processor with 8GB RAM.

The limitations of proposed approach are:

- Label transfers for newly uncovered ambiguous object or background regions can be further improved to reduce error. This limitation can be observed in the 'Girl' sequence where the hand is not visible in the first frame and hence is not segmented in subsequent frames.
- Though the proposed approach handles partial occlusions effectively, complete occlusions are difficult to handle, since only two frames are considered for seam propagation. To handle complete occlusions, future work can extend the seams propagation beyond two frames.

6. Conclusions

Video object segmentation and Seam Carving are not related problems, but the concept of seams (connected paths of low energy) are very useful for object segmentation. Hence the novelty of proposed approach stems from adapting an image retargeting concept to propagate labels in video. Furthermore, to generate seams in a video, an appropriate energy function is proposed, which is minimised using recent developments in ANNF map computation. The confidence measure obtained from the patch based ANNF computation is used for identifying the ambiguous (boundary) regions efficiently. We utilise rough sets to handle ambiguity in label transfer, and thus curtail the error propagation. The major challenge of video object segmentation is the accurate labelling of boundary pixels. The proposed approach is better equipped to handle the boundary pixels compared to existing methods as shown experimentally. The proposed algorithm takes less than a second for processing each frame and achieves state-of-the-art results on the publicly available SegTrack dataset. Further, the proposed approach is suitable for long-term video segmentation since the rough set formulation curtails the error propagation from the ambiguous boundary regions. In summary, the novelty of proposed approach comes from uniquely combining incongruous concepts of seam carving, ANNF maps and rough sets to perform video object segmentation.

Acknowledgement

This work was supported by Joint Advanced Technology Programme, Indian Institute of Science, Bangalore, India.

References

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3), 2007. 2, 3

[2] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *CVPR*, 2010. 1, 2

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3), 2009. 3

[4] G. Ben-Artzi, H. Hel-Or, and Y. Hel-Or. The gray-code filter kernels. *TPAMI*, 29(3), 2007. 3

[5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 1, 2

[6] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Mot - mixture of trees probabilistic graphical model for video segmentation. In *BMVC*, 2012. 1, 2, 7

[7] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. *ACM Trans. Grap.*, 21(3), 2002. 1

[8] A. Fathi, M. F. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *BMVC*, 2011. 1, 2, 7

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 1

[10] S. Korman and S. Avidan. Coherency sensitive hashing. In *ICCV*, 2011. 3, 4

[11] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 1, 2, 7

[12] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011. 1

[13] T. Ma and L. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012. 1, 2, 7

[14] Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5), 1982. 4

[15] S. A. Ramakanth and R. V. Babu. FeatureMatch: An efficient low dimensional patchmatch technique. ICVGIP, 2012. 3, 5

[16] S. A. Ramakanth and R. V. Babu. FeatureMatch: A general ANNF estimation technique and its applications. *IEEE Trans. Image Proc.*, 2014. 3

[17] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. In *ACM SIGGRAPH*, 2008. 2, 3

[18] J. Shi, K. Fragkiadaki, and G. Zhang. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 1, 2

[19] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label MRF optimization. *IJCV*, 100(2), 2012. 1, 2, 6, 7

[20] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 3

[21] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012. 1, 2

[22] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013. 1, 2, 7