International Journal of Latest Trends in Engineering and Technology
Vol.(15)Issue(1), pp.019-024
DOI: http://dx.doi.org/10.21172/1.151.04
e-ISSN:2278-621X

VIDEO OBJECT EXTRACTION USING OPTIMIZED SMOOTHED DIRICHLET PROCESS MULTI-VIEW DEEP REINFORCEMENT LEARNING WITH IMPROVED ADAPTIVE MODIFIED SHAPE PRIOR MRF

G.S.Gowri[1], P. Ponmuthuramalingam[2]

Abstract- Video object extraction (VOE) using Optimized smoothed Dirichlet Process Multi-view learning (OsDPMVL) segmentation and Label refinement using improved adaptive Shape Prior Modified Markov Random Field (IASMMRF) was proposed for extracting most exact Objects from videos. Moreover, Contour tracking and Contour prediction using the Teh–Chin algorithm enhanced the accuracy of VOE. OsDPMVL has used a fixed feature extractor to encode key spatial feature points in images as feature representations for Multi View learning. The learned feature representation for sample efficient domain adaptation is used in the segmentation. However, a fixed feature extractor cannot predict camera motion and masks along with rotation and translation values for moving objects in the image. The various vision tasks such as image classification, object detection and semantic segmentation has shown impressive performance using the deep learning technology. In this paper, Deep Reinforcement Learning (DRL) is used in OsDPMVL to segment objects along with the consideration of camera motion and masks related to the rotation of moving objects in the image. DRL that automatically identify moving objects and utilize the appropriate information for action selection in video sequences. The learned model is then used as the rule of learning for the moving objects. The proposed work is named as Video Object Extraction Using Optimized Smoothed Dirichlet Process Multi-View Deep Reinforcement Learning with Improved Adaptive Modified Shape Prior MRF (OsDPMVDRL-IASMMRF-VOE). The experimental results show that the proposed OsDPMVDRL-IASMMRF-VOE approach gives better object extraction results than OsDPMVL- IASMMRF in terms of Accuracy, Precision and Recall.
Key words : Deep Reinforcement learning, Markov Random Field, Video object extraction, graph cut, Convolution Neural Network.

1. INTRODUCTION

Video object segmentation is the process of masking pixels into a specific class of objects in videos. The unsupervised video segmentation does not provide any manual annotation for object segmentation while semi supervised segmentation provides information about objects of interest in the first frame of a video for any application like summarization and action recognition. The semi supervised segmentation is generally provided high segmentation accuracy then unsupervised segmentation. In semi supervised segmentation, the class of target objects is not assumed that fully known prior. A temporal label propagation problem is solved with partially known information by using spatio-temporal graph structures [1] like a Markov Random Field (MRF) model.

A video segmentation method [2] that integrates Markov random field based contour tracking and graph-cut image segmentation. The contour tracking propagates the shape of the target object, whereas the graph-cut refines the shape and improves the accuracy of video segmentation. However, the drawback arises when the target object has very fast silhouette variation. If the boundary of the target is quite straight in the previous frame but seriously arched in the current frame, the image segmentation algorithm may not correctly extract the boundary of the target object, because the real boundary has moved beyond the unlabeled region.

A spatiotemporal Markov Random Field (MRF) model was defined over pixels to handle this problem [3]. To solve challenging tasks like classification, segmentation and object detection Convolution Neural Network (CNN) has been used, which will give good performance. Unlike conventional MRF models, the spatial dependencies among pixels in this model are encoded by a Convolution Neural Network. Specifically, for a given object, the probability of labeling to a set of spatially neighboring pixels can be predicted by a CNN trained for this specific object. As a result, higher-order, richer dependencies among pixels in the set can be implicitly modeled by CNN. With temporal dependencies established by optical flow, the resulting MRF model combines both spatial and temporal cues for tackling video object segmentation. However, performing inference in the MRF model is very difficult due to the very high order dependencies.

Optimized smoothed Dirichlet Process Multi-view learning with improved adaptive Modified Markov Random Field for Video Object Extraction (OsDPMVL-IASMMRF-VOE) [4] with contour track was proposed to solve the problem of object

[1] Department of Computer Science, Govt. Arts College,Coimbatore,Tamilnadu,India
[2] RJD of Collegiate Education, Coimbatore,Tamilnadu,India

silhouette variation between the frame sequences efficiently. The Teh–Chin algorithm has been used with OsDPMVL-IASMMRF for predicting the contour of the objects in the current frame the previous segmented frame. The contour tracking propagates the shape of the target object, whereas the OsDPMVL-IASMMRF segmentation refined the object boundary and the shape for enhancing the accuracy of video segmentation.

Deep learning tries to represent high-level notions in data using deep networks of supervised / unsupervised learning algorithms, to facilitate study from multiple levels of notions. Nowadays, deep learning has got huge attraction by educational institutions and industrial products like Googles translator, Image search engine, Apple's Siri, Microsoft's Bing voice search,etc.

In order to obtain an efficient video segmentation and object extraction, OsDPMVL-IASMMRF is enhanced with Deep Reinforcement Learning (DRL) to support high order dependencies and independencies of objects between frame sequences more effectively. DRL learn the dependencies and independencies of objects in terms of Dirichlet parameters used in OsDPMVL from training videos. Then the learned model is used to predict the target objects from entire frame sequences globally. Thus the OsDPMVDRL- IASMMRF-VOE approach provides better results for Video object segmentation and extraction than any other methods in terms of accuracy, precision and recall.

2. LITERATURE SURVEY

Most of the recent research work in Video Object Segmentation have mainly based on the semi-supervised setting. A spatiotemporal automatic object tracking and segmentation [5] was proposed by formulating a multi-label Markov Random Field (MRF) graph over neighbor pixels. This algorithm consumed more memory for video object segmentation consumed. Video segmentation using Spatio-temporal graph structures over image patches across frames [6] was proposed to connect the paths by low energy model while protecting the main image contents. Though this method handle partial occlusions, difficult to handles complete occlusion.

The memory and running time of segmentation was overcome by a novel parallel out-of-core algorithm and a clip-based processing approach was proposed [7]. The long video shots were processed fairly efficiently by using these techniques. The clip-based processing could be used for segmenting streaming videos of subjective length, in case a distinct level of the segmentation hierarchy is sufficient. The algorithm proposed for achieving highly accurate segmentation [8] by simultaneously processing video segmentation and optical flow estimation. A multilevel spatio-temporal objective function used for recomposing the segmentation iteratively until reaching the expected quality of segmentation.

An appearance or object based segmentation method was proposed to handle large displacement motion of the objects in the videos [9]. The temporal connections were established using regular spatio-temporal lattices, optical flow, or other similar techniques like nearest neighbour fields to infer the labels for subsequent frames. The appearance based long range connections segments large displacement objects more efficiently.

A fully connected spatiotemporal graph was built over the object was proposed [10] for handling long range connections. A similarity term is computed into a Euclidian space is computationally more efficient to optimize object segmentation. The fully connected nature of the graph implies information exchange between both spatially and temporally distant object proposals[11], which helps to handle critical cases like long range motion and occlusion of objects.

3. PROPOSED METHODOLOGY

3.1 Video Object Extraction using OsDPMVL- IASMMRF

This model, the contour prediction of the current frame in the video, the Teh Chin algorithm is adopted for extracting the object contour from the previous segmented frame.

The equation of posterior energy in [5] is used to separate the object and non object pixels in the frame sequences.

The adaptive MRF approach is decreased the posterior energy through assuming each vertex in the contour Ct−1 as a node in an adaptive MRF system. The hidden state of every node corresponds to the motion vertex of its vertex. For instance, if the search range for every vertex is 5 X 5 after that every node has a state selected from a 25 element set I=(l1,l2,…….l25) for indicating its motion vector. The posterior energy is described as follows,

$$E(Dt|It,Ct-1)=\omega LL+\omega GG+\omega FF+\Omega ss \quad (1)$$

In the above equation, the first three terms represent the energy and the last term denotes the link energy of the adaptive MRF system. L is used for computing the block difference among the current frame and the previous frame. G denotes the gradient of the RGB channel of image. F is used for penalizing the large vertex motion and constrains the contour velocity. S is related to all spatial information of image.

3.2 Video Object Extraction using OsDPMVDRL- IASMMRF

A novel spatio-temporal OsDPMVL-IASMMRF model is proposed for video object segmentation problem. The innovation of the proposed model is that the spatio-temporal potentials are encoded by CNN trained for objects of interest, so higher-order dependencies among pixels can be modeled to enforce the holistic segmentation of object instances

The total energy in our model is defined as follows

$$E(x) = \sum_{i,j\in\square_T} E_L(x_i, x_i) + \sum_{i,j\in\square_T} E_G(x_i, x_i) + \sum_{i,j\in\square_T} E_F(x_i, x_i) + \sum_{i\in v} E_S(x_i, x_i)) \quad (2)$$

where E_L , E_G , E_F are the energies related to spatial and temporal dependencies, E_S is the energy related to spatial dependencies. NT refers to the set of all spatial and temporal associations while S refers to the set of all spatial cliques. The concrete definitions are as follows.

The set of spatial and temporal associates NT is formed using partial sparse optical flow, such that each pixel is only linked to pixels in neighboring frames when the motion estimation is consistent enough. A forward backward consistency checking is used to filter reliable motion vectors one-step temporal dependencies can be further extended to k-step temporal dependencies by directly calculating optical flow between a frame and the frame that is k frames away. The value of k is used as a value of 2 in our model, which means for a certain frame t, all the subsequent links are reputed with t-2, t-1, t+1 and t+2, finally almost 4 temporal neighbors flow is found for each pixel. The flows are removed during forward and backward consistency checking. The temporal energy function is defined as

$$E_t(x_i, x_j) = \theta_t w_{ij}(x_i, x_j)^2 \quad (3)$$

where θ_t a balancing parameter is for this term θ and w_{ij} is a data-dependent weight to measure the confidence of the temporal connection between variables (x_i, x_j). The energy encourages a temporally consistent labeling when the temporal connection is confident.

For spatial dependencies, all the pixels in a frame are defined as a clique, in which the classifying for each pixel depends on all other pixels in the same frame. In order to build a spatial energy function defined over all the pixels in a frame, an energy function f(.) is defined that can review the quality of a given mask xc as a whole. Perfectly, it is easy to construct the function f(.) if the ground-truth mask x*c of an input mask xc is given. For example, we can define f(.),

$$f(x_c) = ||x_c - x_c^*||_2^2 \quad (4)$$

this provides lower energies to masks that are more similar to the ground-truth mask. However, x*c is unknown and indeed what we need to solve for. We here resort to a feed-forward CNN to approximate x*c and define f(.) as follows

$$f(x_c) = ||x_c - g_{CNN}(x_c)||_2^2 \quad (5)$$

where gCNN(.) is a mask refinement CNN that accepts as input a given mask xc and outputs a refined mask. Note that the operator gCNN(.) is a feed-forward pass of a CNN. Intuitively, the above description allocates lower energy to a mask whose mapping through gCNN(.) is more similar to itself. With a well-trained gCNN(.) that can dependably refine a coarse mask to a better one and keep a good mask unchanged, the function f(.) could assign better masks lower energies. Fortunately, it is shown that such a CNN can be trained in a two-stage manner using the first frame of a given video and performs very reliably during the inference for the following frames.. The spatial energy is defined in Eq. (2) as

$$E_s(x_c) = \theta_s f(x_c) \quad (6)$$

where θ_s is a balancing parameter for this term.

The description about the spatial energy in equation (6) has a powerful expressive than classical energy functions. This higher order energy function insisting on label uniformity in pre-segmented regions.

4. RESULT AND DISCUSSION

4.1 Dataset description

The Brown Pelican video [12] is used as dataset for VOE. The frame size of the video is 1174 × 1086 .The video clipping time is 160 secs. The total number of frames available in this video is 600.

Input image sequences are extracted from Brown Pelican video. The sample input video frame and the segmented output of existing and proposed system is shown Figure 1.These segmentation processes separate an object of interest from the remaining image. The accuracy, precision and recall are the metrics used to measure the performance of VOE. The ground truth objects are matched with extracted objects. The number of pixels correctly extracted belonging to objects is True positive and the number of pixels not extracted belongs to objects is True Negative. The number of pixels not extracted not belongs to objects is False Positive. The number of pixels extracted not belongs to objects is False Negative.

The Video Object Segmentation and Extraction is conducted for OsDPMVL- IASMMRF and OsDPMVDRL- IASMMRF to evaluate the performance in terms of Accuracy, Precision and Recall.

Original Video Frame	
OsDPMVL- IASMMRF	
OsDPMVDRL- IASMMRF	

Figure 1 Comparison Results of Segmentation Images

4.2 Accuracy

Accuracy has calculated the proportion of true positives and true negatives among the total number of features clustered.

$$\text{Accuracy} = \frac{(\text{True positive} + \text{True negative})}{(\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative})}$$

Table 1. VOE Accuracy (in %)

Number of Frames	OsDPMVL-IASMMRF-VOE	OsDPMVDRL- IASMMRF-VOE
200	93.4	96.3
300	94.6	97.2
400	95.3	98.4
600	96.8	99.3

In Table 1 and Figure 2, the Accuracy of VOE extraction for proposed sDPMVDRL- IASMMRF and OsDPMVL-IASMMRF are mentioned. The segmentation Accuracy of proposed work is high for any number of frame sizes. The Accuracy is improved from 2% to 5% than the existing approach. The Reinforcement Learning capability of CNN for each energy functions increases the VOE accuracy level than the existing approach.

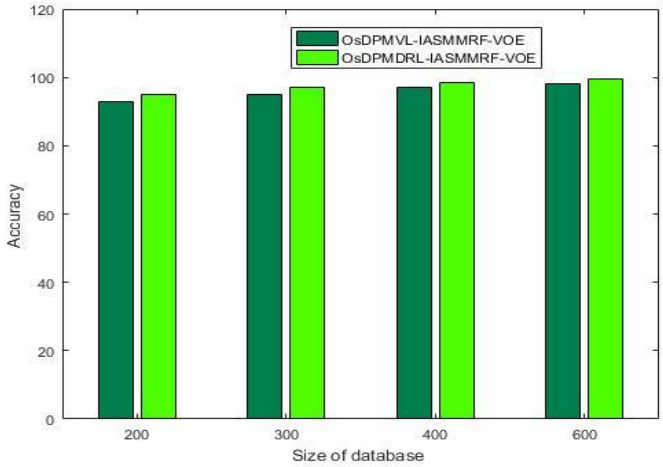

Figure 2 Accuracy Comparisons for VOE

4.3 Precision

Precision value is calculated according to the clustering level at true positive prediction, false positive.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Table 2. VOE Precision

Number of Frames	OsDPMVL-IASMMRF-VOE	OsDPMVDRL- IASMMRF
200	0.943	0.966
300	0.945	0.977
400	0.957	0.986
600	0.966	0.991

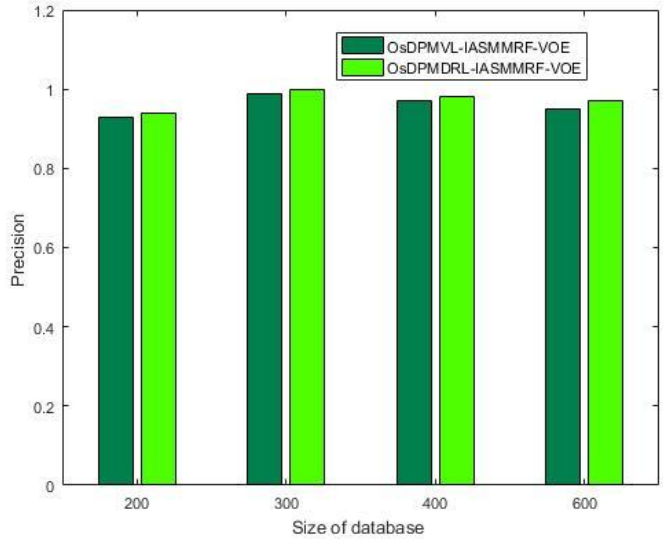

Figure 3 Precision Comparisons for VOE

Table 2 and Figure 3 represents the results of the proposed sDPMVDRL- IASMMRF and OsDPMVL- IASMMRF in terms of Precision. In Figure 3, the Precision of VOE extraction for proposed sDPMVDRL- IASMMRF and OsDPMVL-IASMMRF are represented in the y-axis. The segmentation Precision of the proposed work is high for any number of frame sizes.

4.4 Recall

Recall value is calculated according to the clustering level at true positive prediction, false negative.

$$Recall = \frac{True\ Positive}{(True\ positive + False\ negative)}$$

Table 3. VOE Recall

Number of Frames	OsDPMVL-IASMMRF-VOE	OsDPMVDRL- IASMMRF
200	0.934	0.966
300	0.949	0.977
400	0.954	0.986
600	0.989	0.997

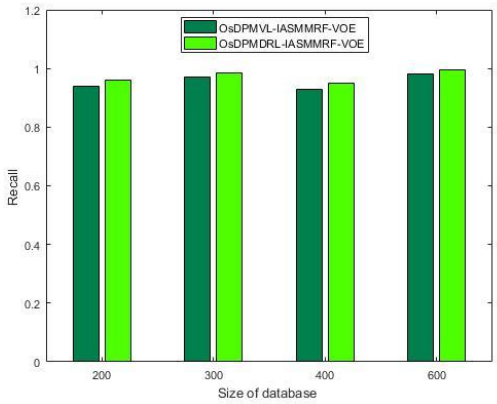

Figure 4 Recall Comparisons for VOE

Table3 and Figure 4 represent the results of the proposed sDPMVDRL- IASMMRF and OsDPMVL- IASMMRF in terms of Recall. In Figure 4, the Recall of VOE extraction for proposed sDPMVDRL-IASMMRF and OsDPMVL-IASMMRF are represented in the y-axis. The segmentation Recall of the proposed work is high for any number of frame sizes.

5. CONCLUSION

In this paper, a spatio-temporal MRF model for video objects segmentation has been proposed. By performing inference in the MRF model, an algorithm has been developed that alternates between a temporal fusion operation and a mask refinement feed-forward CNN, progressively inferring the results of video object segmentation has been developed. The efficiency of the proposed work is demonstrated through widespread experiments on challenging datasets through performance metrics like Accuracy, Precision and Recall.

6. REFERENCES

[1] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In CVPR,2010. 1, 2
[2] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In ICCV, 2015.
[3] Chung, C. Y., and Chen, H. H., "Video object extraction via MRF-based contour tracking", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 20, Issue. 1, January 2010, pp. 149-155.
[4] B.Linchao B ,Baoyuan Wu and L .Wei . CNN in MRF: Video Object Segmentation via Inference in A CNN-Based Higher-Order Spatio-Temporal MRF.
[5] 5. G. S. Gowri, Dr. P. Ponmuthuramalingam. Video object extraction using optimized smoothed dirichlet process multi-view learning with improved adaptive modified Markova random field .Volume :7,Issue :4 , 2018 ,pages 2598-2602
[6] D.Tsai, M. Flagg, and J. M.Rehg. Motion coherent tracking with multi-label mrf optimization. BMVC, 2010.
[7] S. Avinash Ramakanth and R. Venkatesh Babu. Seamseg: Video object segmentation using patch seams. In CVPR, 2014.
[8] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In CVPR,2010.
[9] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In CVPR, 2016
[10] L. Bao, Q. Yang, and H. Jin. Fast edge-preserving patchmatch for large displacement optical flow. IEEE Trans. On Image Processing, 12(23):4996–5006, 2014. 3
[11] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In ICCV, 2015
[12] http://arma.sourceforge.net/vb100/