PAPER

Nonlinear Regression of Saliency Guided Proposals for Unsupervised Segmentation of Dynamic Scenes

Yinhui ZHANG[†a)], ***Member***, **Mohamed ABDEL-MOTTALEB**[††,†††b)], ***and*** **Zifen HE**[†c)], ***Nonmembers***

SUMMARY This paper proposes an efficient video object segmentation approach that is tolerant to complex scene dynamics. Unlike existing approaches that rely on estimating object-like proposals on an intra-frame basis, the proposed approach employs temporally consistent foreground hypothesis using nonlinear regression of saliency guided proposals across a video sequence. For this purpose, we first generate salient foreground proposals at superpixel level by leveraging a saliency signature in the discrete cosine transform domain. We propose to use a random forest based nonlinear regression scheme to learn both appearance and shape features from salient foreground regions in all frames of a sequence. Availability of such features can help rank every foreground proposals of a sequence, and we show that the regions with high ranking scores are well correlated with semantic foreground objects in dynamic scenes. Subsequently, we utilize a Markov Random Field to integrate both appearance and motion coherence of the top-ranked object proposals. A temporal nonlinear regressor for generating salient object support regions significantly improves the segmentation performance compared to using only per-frame objectness cues. Extensive experiments on challenging real-world video sequences are performed to validate the feasibility and superiority of the proposed approach for addressing dynamic scene segmentation.

key words: *video object segmentation, salient object-like proposal, nonlinear regressor, dynamic scene, random forest*

1. Introduction

In contrast to interactive [12] or semi-supervised [13] video segmentation, unsupervised video object segmentation addresses the problem of automatically assigning labels for objects in an unannotated video sequence, which is an important task with many potential applications in computer vision and pattern recognition. Unfortunately, the performance of the state-of-the-art video object segmentation algorithms tend to drop significantly in the case of dynamic scenes. Several factors such as motion ambiguity, appearance variations, background clutter and camera motion make the segmentation of objects from dynamic scenes extremely challenging. The major bottleneck lies in the inherent difficulty of inferring temporal consistent object proposals in video sequences.

It is well recognized that the use of only appearance features is inefficient to characterize consistent object proposals due to ineffective learning of discriminative models and their sensitivity to both foreground appearance variations and background scene dynamics. Overall, many object classes cannot be adequately described by color or texture cues alone and, indeed, even the same object in a video sequence might comprise several parts with different appearances.

To some extent the problems of appearance modeling can be alleviated by formulating object hypothesis or trajectory clustering in the temporal domain using motion cues. Motion based inside mapping and trajectory clustering usually utilize optical flow [5] or point trajectory [23], [24] to impose consistency constraints over time, which could range from two consecutive frames to hundreds of frames. Aside from the well-known challenges associated with optical flow (e.g., boundary ambiguity) and spectral clustering of point trajectory (e.g., model selection), these motion segmentation methods lack explicit mid-level information (e.g., size and shape) specific to foreground features during object hypothesis estimation. Consequently, directly applying the motion cues to video object segmentation inevitably results in a fairly sparse and ambiguous trajectory embedding especially near occlusion and disocclusion areas due to camera motion or nonhomogeneous articulation.

Proposal ranking approaches [7]–[10] explicitly leverage mid-level information about object-like features conveyed by a number of foreground hypotheses and hence can address the sparsity and occlusion problems caused by temporal tracking of key points. However, most of these approaches attempt to explore proposal ranking at intra-frame level using objectness cues derived in the spatial domain. As a consequence, the proposals generated in each frame generally have difficulty in remaining consistent over time. Another major drawback of these approaches is the complicated hypothesis generating models that further restricts the real-world applicability of proposal ranking.

In this paper we propose a nonlinear regression scheme for video object segmentation. The main assumption of the proposed segmentation approach is that the foreground objects should be visually salient against their dynamic scenes. Our main idea is to leverage saliency guided object-like features to train a random forest regressor to enable temporal consistent hypothesis of foreground proposals. Our main contributions with respect to previous work can be summa-

Manuscript received July 24, 2015.
Manuscript revised October 5, 2015.
Manuscript publicized November 6, 2015.

[†]The authors are with Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, 727 Jingming South Road, Chenggong, Kunming 650500, China.
[††]The author is with Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables, FL 33146, USA.
[†††]The author is with Effat University, Jeddah, Saudi Arabia.
a) E-mail: yinhui_z@163.com
b) E-mail: mottaleb@miami.edu
c) E-mail: zyhhzf1998@163.com
DOI: 10.1587/transinf.2015EDP7295

rized as follows:

(1) We present an automatic approach to characterize inter-frame features of saliency guided proposals across a video sequence. The integration of both appearance and shape features enables temporal consistent ranking of foreground regions in an unsupervised manner.

(2) We show that a combination of the top ranked proposals with spatio-temporal constraints leads to a more accurate video object segmentation of real-world dynamic scenes compared to other competing algorithms.

The rest of the paper is organized as follows. We begin by reviewing related work on video object segmentation in Sect. 2. In Sect. 3, we generate descriptive features of saliency guided hypotheses by leveraging saliency signature in the frequency domain. In Sect. 4, we detail our approach for saliency guided nonlinear regression scheme to deal with scene dynamics. We propose to use a random forest based nonlinear regression technique to learn and predict inter-frame features within all frames of a sequence. Our system leverages the learned ranking scores, along with a Markov Random Field to integrate spatio-temporal constraints of the top ranked regions. Subsequently, in Sect. 5, experimental results are illustrated. Finally, we draw conclusions in Sect. 6.

2. Related Work

Unsupervised video object segmentation methods can be roughly divided into three categories: background subtraction, proposal ranking and motion trajectory.

2.1 Background Subtraction

Approaches in this category take into account classic background subtraction techniques [1], [15] or make use of dynamic background texture modeling [2] for unsupervised segmentation. Barnichm et al. [3] determine whether a pixel belongs to the background by comparing its current value with past ones and propagate the value of background pixels into a background subtraction model. Haines et al. [4] use Dirichlet process Gaussian mixture models to estimate background distributions and use them as input to a model learning process for continuous update as scene changes. These methods are typically based on the strong assumption that the dynamic backgrounds are changing slowly, which is not the case for highly dynamic scenes. The large variations in the appearance and locations of objects that may appear in the image sequence remain a concern for these object segmentation approaches. Grundmann et al. [6] combine hierarchical cues by constructing a tree of spatio-temporal segmentation. This approach allows for subsequent selection of granularity from varying levels. Although good for handling multiscale appearance cues, a strong limitation of this method is that it does not solve the foreground segmentation task on its own due to the oversegmentation of the scene. Unlike other saliency based segmentation methods such as [28] and [29] wherein foreground objects are extracted from images, our approach focus on video object segmentation in spatiotemporal domain. Moreover, the main difference between our approach and the discriminant center-surround spatiotemporal saliency proposed in [30] is that, in addition to appearance stimulus, our approach incorporates both appearance and shape geometry into descriptive features to generate more reliable foreground hypotheses in dynamic scenes.

2.2 Proposal Ranking

Another category of recent work in video object segmentation focuses primarily on a ranking mechanism of object proposals, which uses the notion of objectness and a selection mechanism. Endres et al. [7] propose to generate bag of regions based on seeds and rank them using structured learning. Lee et al. [8] use consistent appearance and motion to rank hypothesis of object-like regions. Whereas Ma et al. [9] introduce a weighted region graph to find maximum weight cliques. Alternatively, in [17] segmentation is performed using graph cuts and simple color cues, and the regions are ranked through classification based on gestalt cues with a simple diversity model. Most recently, object models [10] are built based on the primary object hypothesis regions. Our approach falls in this category as it also estimates object-like foreground regions. However, unlike most proposal ranking techniques that are limited to intra-frame object hypothesis, we are able to find temporal consistent object proposals via a nonlinear regressor across a video sequence. Moreover, instead of learning random forest using training set with ground truth annotations, we employ descriptive features derived from saliency guided mappings to efficiently rank foreground hypotheses in an unsupervised manner.

2.3 Motion Trajectory

Papazoglou et al. [5] propose a Fast Object Segmentation (FOS) algorithm which attempts to build dynamic appearance models of the object and background under the assumption that they change smoothly over time. An advantage of this approach is that it may be possible to handle spatio-temporal cues on image patches such as color and location in the labeling refinement stage. But relying in the initialization of inside foreground object points only on motion boundaries tends to produce a large number of false-positive seeds, especially in case of highly dynamic scenes. Opposed to classical two-frame optical flow, point trajectories that span hundreds of frames are less susceptible to short term variations that hinder separating different objects [24]. Instead of using the motion boundaries by optical flow between two consecutive frames, point trajectory approaches employ spectral cluster [24] or boundary discontinuities of embedding density [23] between neighboring trajectories. Similarities between each pair of trajectories are derived from the maximal motion difference between them over all frames. The underlying assumption for point tra-

jectory [23], [24] is that the video sequence covers a long time span, which makes this approach not suitable for short clips. Our method, instead, does not attempt to cluster trajectory points and does not assume any kind of relative motion between foreground and background. Moreover, the selection of the number of clusters or the dimensionality of embedded spaces is nontrivial as it could easily confuse sparse trajectory clustering algorithms, causing over- or under-segmentation of embedded trajectories.

3. Features of Salient Hypotheses

We will first introduce the saliency mapping scheme at superpixel level that forms the foundation of our feature extraction. Then we detail our descriptive features generated from salient superpixel regions.

3.1 Salient Hypotheses

Due to the highly dynamic and complex nature of real-world scenes, a set of object candidates are initially generated in order to form a pool of hypotheses using object intrinsic features. Unfortunatelly, an image sequence may contain multiple object instances and we would like to compute a mapping to highlight the most salient foreground regions for each frame. Intuitively, saliency characterizes some locations of a scene that appear to an observer to stand out relative to their neighboring parts. Moreover, we should derive a soft decision on the support of foreground regions to maintain local coherence of saliency mappings at pixel level. For this purpose, our saliency mapping function builds on the saliency signature [21] scheme, which is defined in the Discrete Cosine Transform (DCT) domain and works well for image segmentation with medium sized objects. Although there are more sophisticated visual attention modeling methods [22] to estimate saliency mappings, this frequency domain technique suffices in providing satisfactory results.

The original video sequence is denoted as $\mathcal{V} = \{I^t\}_{t=1}^{T}$, in which I^t denotes the t-th frame. Let $\mathcal{S}(i)$ denote the saliency signature operator for all pixels $i \in \mathcal{V}$. To capture local coherency of saliency mappings, we first break down the video sequence into superpixels. Specifically, given the video $\mathcal{V}$, let $\mathcal{R} = \{r_j\}_{j=1}^{K}$ be a partitioning of $\mathcal{V}$ with K regions, each of which is associated with an unique region index j. The partitioning of a video sequence can easily be computed by common superpixel or supervoxel methods. With the partitioned video $\mathcal{R}$, we then define the local consistent saliency map as the mean saliency signature over each region:

$$\bar{\mathcal{S}}(r_j) = \left\{ \frac{1}{|r_j|} \sum_i \mathcal{S}(i) \middle| i \in r_j \right\} \tag{1}$$

where $|r_j|$ is the number of pixels in the j-th region. Once the local consistent saliency map is computed, the perceptual saliency guided foreground hypothesis, which returns a soft decision $\mathcal{H}(r_j)$ for each region, can be evaluated by thresholding $\bar{\mathcal{S}}$, *i.e.* $\mathcal{H}(r_j) = 1$ if $\bar{\mathcal{S}}(r_j) \geq \tau$ and $\mathcal{H}(r_j) = 0$, otherwise. For simplicity, we abbreviate $\mathcal{H}(r_j)$ by $\mathcal{H}_j$.

3.2 Descriptive Features

Descriptive features are then extracted from saliency guided hypothesis regions. The descriptive features $\mathcal{X} \in \mathbb{R}^{m \times n}$ are defined as the concatenation of m feature vectors, and n is the dimension of a feature vector. Formally, the feature set of a video sequence is given by $\mathcal{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_j; \ldots; \mathbf{x}_m]$, and $\mathbf{x}_j \in \mathbb{R}^n$.

In this paper, the descriptive features and their corresponding soft labels constitute a training set $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{H}$, which will be used to train a nonlinear regressor. More specifically, the training data is identified by $\{(\mathbf{x}_j, \mathcal{H}_j) : j = 1, \ldots, m\}$, where $\mathcal{H}_j \in \{0, 1\}$. Note that in our case, the saliency guided foreground hypothesis is regarded as a soft classifier that projects input training data pairs $(\mathbf{x}_j, \mathcal{H}_j)$ into positive ($\mathcal{H}_j = 1$) or negative ($\mathcal{H}_j = 0$) samples.

To build feature vectors for random forest regression, we combine multiple cues that are commonly used in mid-level proposal selection [18] for image segmentation. Unfortunately, generating proposals using uniformly distributed seeds on regular grids of an image risk missing the true correspondences due to inevitable flaws caused by the diversification scheme via maximal marginal relevance measure. In contrast, we consider object-like proposals together with the perceptually salient description, which allows us to distinguish the most promising proposals from a list of hypotheses.

Specifically, for a region r_j of a partition $\mathcal{R}$, the feature vector $\mathbf{x}_j$ encodes both the appearance and the shape geometry of the region and consists of two components:

Appearance: we incorporate color information into the feature vector as this is one of the most important cues returned by the saliency signature operator and certain object appearances tend to stand out from their surroundings in dynamic scenes. We compute the average pixel color in r_j and characterize the color features in CIELab color space.

Shape geometry: we use area and perimeter of the region; relative position, aspect ratio and area of the bounding box; the area balance (defined as the minimum area divided by the maximum area of the regions); normalized perimeter (defined as the perimeter divided by the squared root of the area); area ratio (defined as the area of the region divided by that of the bounding box); contour strength; minimum and maximum ultrametric contour map; thresholds that cause appearance or disappearance of the proposals [19].

The aforementioned features are concatenated and will be used in the ranking of the salient regions through random forest regression. Through this process, superpixel regions whose appearance and shape geometry are comparable to the salient object-like hypotheses in the video sequence will result in top ranked proposals.

4. Inter-Frame Regression of Descriptive Features

Directly using the mean saliency for hypotheses ranking tends to generate inconsistent proposals across different frames, as the mean saliency signature and the corresponding descriptive features are independently estimated at intra-frame level. For this reason, nonlinear regression of saliency regions is employed to impose temporal consistent constraints. In this section we detail our approach for saliency guided nonlinear regression scheme to deal with scene dynamics. Our goal is to find temporally consistent proposals in salient foreground to obtain a set of candidate object proposals across all frames of a video sequence.

4.1 Forest Model Training

We train a random forest similar to the randomized tree algorithms in [26], [27]. Each tree tr in a forest $\mathcal{F}$ is trained independently on a random subset of the training set $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{H}$ according to the saliency guided foreground hypothesis. By concatenating features in all frames of a video sequence, the feature vectors $\mathcal{X}$ with their corresponding soft labels $\mathcal{H}$ are used to train the random forest regressor $\mathcal{F}$, which is composed of a collection of N trees. A tree $tr \in \mathcal{F}$ is characterized by a binary split function $\psi(\mathbf{x}) : \mathcal{X} \rightarrow \{0, 1\}$. The role of the split function is to decide whether a sample feature $\mathbf{x}$ should be forwarded to the left sub-tree or to the right sub-tree according to an information gain criterion. Formally, the split function ψ of a tree tr is selected from a randomly generated set Ψ in a way to minimize the classic information gain [20] as follows:

$$\psi^* = \arg\min_{\psi \in \Psi} |\mathcal{D}_l^{\psi}| \cdot \mathbb{H}(\mathcal{D}_l^{\psi}) + |\mathcal{D}_r^{\psi}| \cdot \mathbb{H}(\mathcal{D}_r^{\psi}) \tag{2}$$

where $\mathbb{H}(\cdot)$ denotes the entropy of class distributions of the left and right data splitted by ψ. Subsequently, the tree's left and right sub-trees are recursively grown with their respective training data $\mathcal{D}_l^{\psi}$ and $\mathcal{D}_r^{\psi}$. In our case, the termination criterion of the training process is based on the maximum number of trees. We continue growing the forest until the maximum number, N, of trees are assembled in the random forest model $\mathcal{F}$.

4.2 Random Forest Ranking

Nonlinear regression through random forest is subsequently carried out to obtain temporal consistent foreground regions across a video sequence. For each unknown feature vector $\mathbf{x} \in \mathcal{X}$, the learned forest predicts a ranking score $y \in \mathcal{Y}$ by routing $\mathbf{x}$ through the forest to a leaf, where the prediction is taking place. Formally, given $\mathcal{F}$, the ranking score of the feature vector $\mathbf{x}$ extracted from a region is defined as:

$$\hat{y} = \sum_{tr \in \mathcal{F}} \mathbb{I}[p_{tr}(\mathbf{x}|tr) = y] \tag{3}$$

where $p_{tr} \in \mathcal{Y}$ denotes the prediction for a feature vector $\mathbf{x}$ obtained from tree $tr \in \mathcal{F}$, and $\mathbb{I}[\cdot]$ is the indicator function. In our case, the features with the top k ranking scores are returned by the nonlinear regressor, which correspond to the top k ranked proposals r_j.

4.3 MRF Refinement

The outputs of the regressor are the regions ranked by random forest prediction during the unsupervised nonlinear regression. Once we have the top ranked region proposals r_j, we can use a Markov random field (MRF) to further refine the foreground labels of every frame of the video sequence. As in [5], to assign labels $\mathcal{L} = \{l_i^t\}_{i,t}$ to a video sequence, we minimize an energy function to refine the proposals:

$$\begin{aligned} E(\mathcal{L}) = & \sum_{t,i} A_i^t(l_i^t) + \alpha_1 \sum_{t,i} L_i^t(l_i^t) \\ & + \alpha_2 \sum_{(i,j) \in \mathcal{E}_s} V_{ij}^t(l_i^t, l_j^t) \\ & + \alpha_3 \sum_{(i,j,t) \in \mathcal{E}_t} W_{ij}^t(l_i^t, l_j^{t+1}) \end{aligned} \tag{4}$$

where A denotes the appearance term. More specifically, A is defined as the negative log-likelihood of a superpixel i to have label l_i given the foreground and background Gaussian Mixture Models (GMMs). The location prior L of a superpixel i denotes the percentage of pixels that are inside the foreground hypothesis. $(i, j) \in \mathcal{E}_s$ indexes the edges of the MRF graphical model in spatial domain, where i and j correspond to superpixels in every frame. Similarly, $(i, j, t) \in \mathcal{E}_t$ indexes the edges of the MRF model in temporal domain. The potential term V encodes Euclidean distance and color contrast between superpixels on spatial edges in $\mathcal{E}_s$. The pairwise term W encodes the percentage of overlap and color contrast between superpixels connected by temporal edges in $\mathcal{E}_t$. Note that $\mathcal{E}_t$ is derived from temporal connections of consecutive frames by optical flow [11], [14]. The main difference between the proposed method and [5] is that, instead of estimating foreground seeds by inside mapping of motion boundaries, we initialize the GMMs of appearance model and the location term by the top ranked proposals r_j returned by forest regression.

5. Experimental Results

The main goal of our experiments is to verify that the nonlinear regressor, which is automatically trained using features from saliency guided foreground hypotheses, is likely to generalize better and attain higher segmentation accuracy than alternative algorithms, especially tested on videos from unconstrained real-world dynamic scenes. To this end, we provide qualitative and quantitative evaluations on the test set of the Freiburg-Berkeley Motion Segmentation (FBMS) [24] dataset publicly available at [25], which is composed of 30 video clips captured in challenging real-world dynamic scenes. The sequences show wide variation in scale and comprise non-rigid shape articulation as

Fig. 1 Example images from the FBMS dataset used in our experiments.

well as dramatic camera motion. To reduce the computational requirements, the maximum length of each video sequence is limited to 100 frames and each frame is sampled every 5 pixels. In our experiments, the test set has a total of 2552 frames, among them every 20th frame comes with ground truth and 367 annotated frames are used for quantitative evaluation. Figure 1 shows example images from the FBMS dataset.

For all video sequences, we utilized the SLIC algorithm [16] to compute superpixels because of its low computational cost. Throughout our experiments, the size of each superpixel is fixed at 10 pixels and the regularizer is fixed at 0.2. The length of each feature vector $\mathbf{x}_j$ is $n = 19$, which consists of three appearance features and sixteen shape geometry features of a region r_j. We have observed a range of 1620~3000 feature vectors being concatenated from different frames in a video. As in traditional random forest regression approach, the feature importance can be directly calculated in the training process. In our case, the area balance feature is most important among the different features, whose relative importance attains 13.7% among nineteen features. The training is terminated when the maximum $N = 50$ trees are assembled in the random forest model.

For a quantitative evaluation of the results, we use the standard mean Area Under Curve (AUC) to evaluate the segmentation performance with respect to ground truth (GT), which is equal to the area under the Receiver Operating Characteristic (ROC) curve. For each video sequence, the AUC of the segmentation at each annotated frame is obtained. These per-frame AUC values are then averaged to produce the mean average accuracy for each of the sequences.

5.1 Random Forest Ranking

We start by providing ranking results by random forest regression. In order to evaluate the ranking performance of forest regression, we compare ROC curves of top four proposal regions. Initially, the soft label $\mathcal{H}_j$ of a region r_j is fixed at $\tau = 0.55$. The ROC curves and its AUC results are reported in Fig. 2. As can be seen from this figure, the proposal with the maximum ranking score significantly outperforms the three other proposal results on the FBMS dataset. This result coincides with our modeling intuition and confirms that the appearance and shape features with high ranking scores are well correlated with ground truth object descriptors. This is very encouraging, as we expect that tuning optimal soft threshold and encoding temporal constraints would provide even better segmentation performance of top ranked proposals.

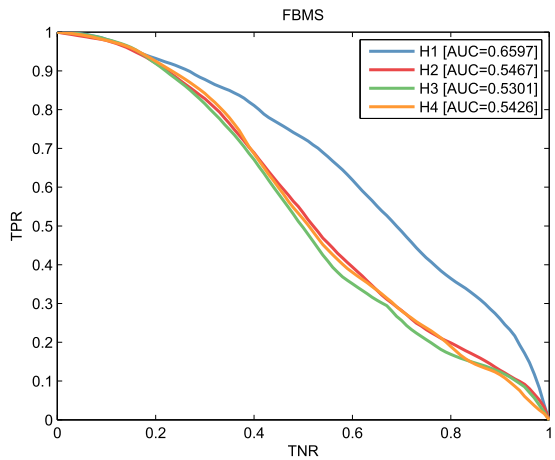

Fig. 2 Comparison of ROC curves of top four ranking proposals (denoted as H1-H4) on the FBMS dataset ($\tau = 0.55$).

5.2 Parameter Tuning

To quantify the effect of the parameter τ on random forest regression, we measure segmentation accuracy on the FBMS dataset for different values of τ from 0.2 to 0.7 with step 0.1 and report AUC values of the top ranked proposals in Fig. 3. We can observe that the performance is maximized for $\tau = 0.4$, at which the top ranked proposals attain performance measure $AUC = 71.29\%$.

5.3 Top Proposals Refinement

We utilized the MRF model to further refine the top ranked proposals returned by the random forest. The appearance and location terms in the energy function are derived from the top ranked proposals estimated at $\tau = 0.4$. We use the same experimental setup as [5] to infer the optimal labels of a video sequence and compare the proposed approach

Table 1 Comparison of segmentation accuracies in terms of Correct Pixels of the proposed method and VOS.

Video	camel01	cars1	cars4	cars5	cars10	cats01	cats03	cats06	dogs01	dogs02
VOS	6219	8340	**10913**	**11899**	3688	8140	2188	1869	4272	5592
Ours	**8562**	**10035**	8099	9750	**8457**	**10745**	**6481**	**2576**	**7198**	**7154**
Video	farm01	giraffes01	goats01	horses02	horses04	horses05	lion01	marple2	marple4	marple6
VOS	4441	5330	3183	3291	2627	3900	**10740**	**3795**	**3692**	1410
Ours	**8290**	**7762**	**5104**	**7113**	**3885**	**5500**	6981	1853	1980	**2196**
Video	marple7	marple9	marple12	people1	people2	people03	rabbits02	rabbits03	rabbits04	tennis
VOS	1433	2900	2456	8557	**11555**	**7682**	**3034**	4023	5093	4677
Ours	**2885**	**3062**	**4207**	**8634**	10117	5614	2812	**5939**	**6438**	**6641**

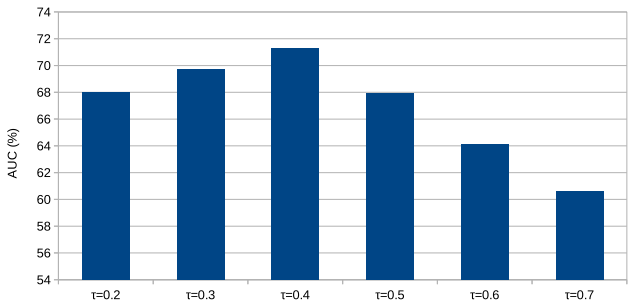

Fig. 3 Impact on segmentation accuracy of the top ranked proposals for varying τ on the FBMS dataset.

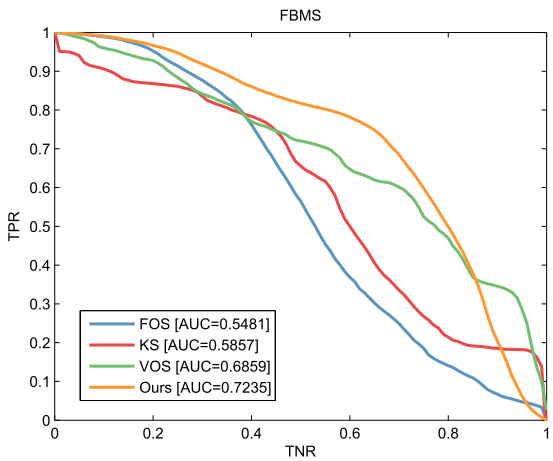

Fig. 4 Comparison between our segmentation method and three state-of-the-art approaches on the FBMS dataset: directed acylic graph based Video Object Segmentation (VOS) [10], Key-Segments (KS) [8] and Fast Object Segmentation (FOS) [5].

to three state-of-the-art video object segmentation methods: directed acylic graph based Video Object Segmentation (VOS) [10], Key-Segments (KS) [8] and Fast Object Segmentation (FOS) [5]. The segmentation performance of the proposed method, KS, VOS and FOS are shown in Fig. 4.

As can be seen from the figure, our method outperforms other competing approaches in terms of the AUC averaged over all 30 video sequences of the FBMS dataset. Compared with VOS, KS and FOS, the segmentation accuracy in terms of AUC is improved by 5.48%, 23.53% and 32.00%, respectively. These improvements indicate that our approach is very successful in correctly localizing the objects and therefore obtaining temporal consistent segmentation across the video sequences. It is worthwhile to point out that the significant improvement over FOS mainly comes from two factors: (1) Our foreground seeds are estimated with the top ranked proposals, which avoids the false-positive hypothesis of inside mapping. (2) Our top ranked proposals encodes appearance and shape geometry cues across frames.

In Fig. 5, we present qualitative segmentation results from our method on 30 video sequences of the FBMS dataset. For evaluating the segmentation quality of each video sequence, we report results obtained with the metric of Correct Pixels (CP): $\frac{1}{T}\sum_t(|I^t| - |\mathrm{XOR}(\mathcal{L}^t, GT^t)|)$. The segmentation accuracies in terms of CP on every video sequence of the FBMS dataset are reported in Table 1. As can be seen from this table, compared with the VOS algorithm, which is the most similar approach to our work, the proposed method achieves higher performance on 22 out of the 30 video sequences. More specifically, the average CP of VOS and our method over the 30 video sequences are 5231 and 6202, respectively. This result again shows that the proposed inter-frame regression of descriptive features derived from salient support regions enables temporal consistent video object segmentation compared to other competing methods.

6. Conclusion

In this paper, we have presented an unsupervised approach for object segmentation in video clips of dynamic scenes. To this end, we have formulated a foreground hypothesis as a random forest based nonlinear regression framework to associate a ranking score with each proposed region across the video sequence. Both appearance and shape features of saliency guided proposals are effectively incorporated into the regression framework. We integrate the appearance and location information into a Markov random field, which facilitates further refinement of the top ranked proposals. Extensive experiments on a challenging dynamic scene video dataset demonstrate the feasibility and superiority of the proposed segmentation approach. In the future, we intend to study how features of this regressor can be extended by including, e.g., HOG and MBH features of foreground hypotheses at multiscales.

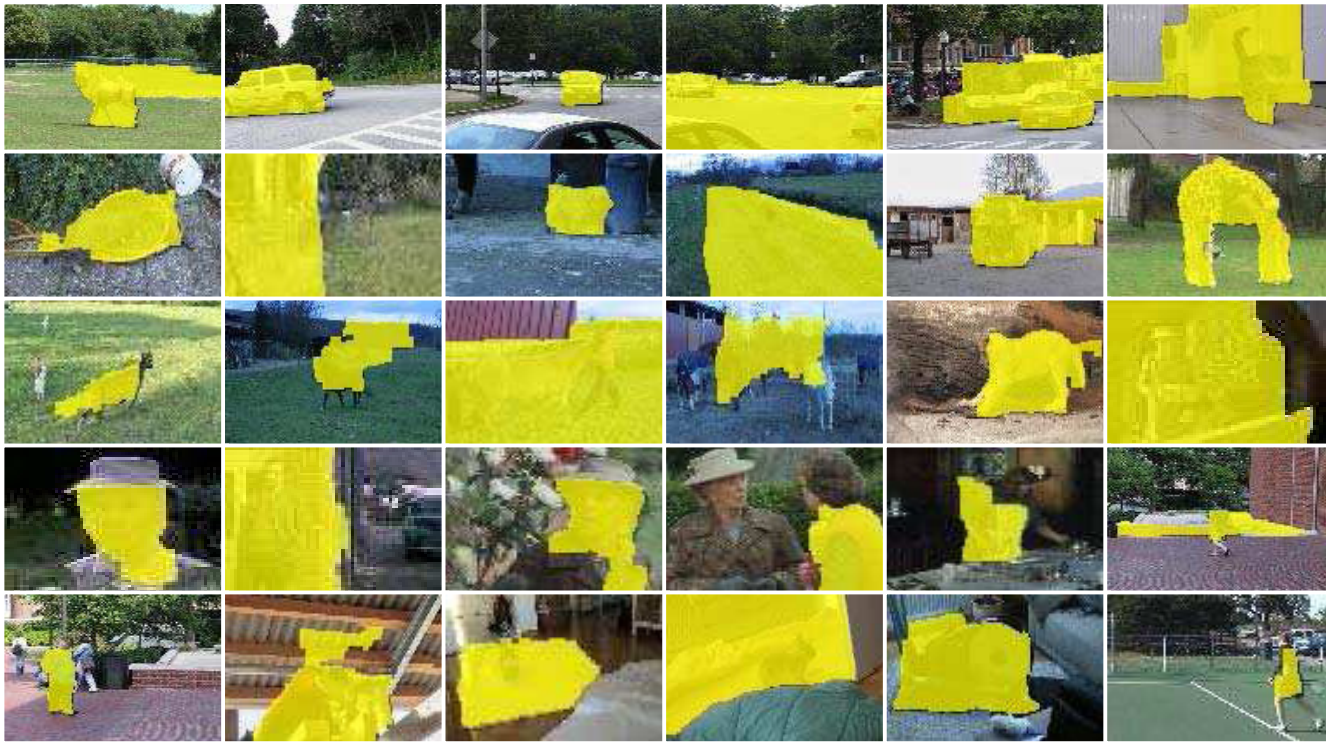

Fig. 5 Qualitative segmentation results obtained by the proposed method on 30 sequences of the FBMS dataset.

Acknowledgments

This work was supported by the National Science Foundation of China (NSFC) under Grants 61461022 and 61302173. Yinhui Zhang was supported by the China Scholarship Council (File No. 201208535035) while studying as a visiting scholar at University of Miami.

References

[1] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp.637–663, 1999.

[2] X. Ren, T.X. Han, and Z. He, "Ensemble video object cut in highly dynamic scenes," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1947–1954, 2013.

[3] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," IEEE Trans. Image Process., vol.20, no.6, pp.1709–1724, 2011.

[4] T.S.F. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.4, pp.670–683, 2014.

[5] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," IEEE International Conference on Computer Vision (ICCV), pp.1777–1784, 2013.

[6] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2141–2148, 2010.

[7] I. Endres and D. Hoiem, "Category independent object proposals," European Conference on Compter Vision (ECCV), pp.575–588, 2010.

[8] Y.J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," IEEE International Conference on Computer Vision (ICCV), pp.1995–2002, 2011.

[9] T. Ma and L.J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.670–677, 2012.

[10] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.628–635, 2013.

[11] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.3, pp.500–513, 2011.

[12] B.L. Price, B.S. Morse, and S. Cohen, "LIVEcut: Learning-based interactive video segmentation by evaluation of multiple propagated cues," IEEE International Conference on Computer Vision (ICCV), pp.779–786, 2009.

[13] V. Badrinarayanan, I. Budvytis, and R. Cipolla, "Semi-supervised video segmentation using tree structured graphical models," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.11, pp.2751–2764, 2013.

[14] B.K.P. Horn and B.G. Schunck, "Determining optical flow," Artifical Intelligence, vol.17, no.1-3, pp.185–203, 1981.

[15] H.J. Chang, H. Jeong, and J.Y. Choi, "Active attentional sampling for speed-up of background subtraction," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2088–2095, 2012.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels," Technical Report, EPFL, 2010.

[17] J. Carreira and C. Sminchisescu, "Constrained parametric min cuts for automatic object segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3241–3248, 2010.

[18] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmen-

tation using constrained parametric min-cuts," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.7, pp.1312–1328, 2012.
[19] P. Arbelaez, J. Pont-Tuset, J.T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.328–335, 2014.
[20] L. Breiman, "Random forests," Mach. Learn., vol.45, no.1, pp.5–32, 2001.
[21] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.1, pp.194–201, 2012.
[22] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.1, pp.185–207, 2013.
[23] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1846–1853, 2012.
[24] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.6, pp.1187–1200, 2014.
[25] http://lmb.informatik.uni-freiburg.de/resources/datasets/
[26] P. Kontschieder, S.R. Bulo, M. Pelillo, and H. Bischof, "Structured labels in random forests for semantic labelling and object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.10, pp.2104–2116, 2014.
[27] N. Payet and S. Todorovic, "Hough forest random field for object recognition and segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.5, pp.1066–1079, 2013.
[28] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes," International Journal of Computer Vision, vol.111, no.2, pp.153–170, 2015.
[29] S. Kang, H. Lee, J. Kim, and J. Kim, "Automatic image segmentation using saliency detection and superpixel graph cuts," Robot Intelligence Technology and Applications 2012, Advances in Intelligent Systems and Computing, vol.208, pp.1023–1034, Springer, Berlin, Heidelberg, 2013.
[30] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.1, pp.171–177, 2010.

Yinhui Zhang received the Ph.D. degree in the direction of image segmentation from Kunming University of Science and Technology, Kunming, China, in 2010. He is currently an associate professor with the Department of Mechanical Engineering, Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology. His research interests include image processing and computer vision.

Mohamed Abdel-Mottaleb received the Ph.D. degree in computer science from the University of Maryland, College Park, in 1993. He joined the University of Miami in 2001. Currently, he is a Professor and Chairman of the Department of Electrical and Computer Engineering. His research focuses on 3-D face and ear biometrics, dental biometrics, visual tracking, and human activity recognition. Prior to joining the University of Miami from 1993 to 2000, he was with Philips Research, Briarcliff Manor, NY, where he was a Principal Member of the Research Staff and a Project Leader. At Philips Research, he led several projects in image processing and content-based multimedia retrieval. He represented Philips in the standardization activity of ISO for MPEG-7, where some of his work was included in the standard. He holds 22 U.S. patents and more than 30 international patents. He published more than 100 journal and conference papers in the areas of image processing, computer vision, and content-based retrieval. He is an editorial board member for the Pattern Recognition journal. He is an IEEE fellow since January 2011.

Zifen He received the Ph.D. degree in the direction of digital image halftoning from Kunming University of Science and Technology, Kunming, China, in 2013. She is currently an associate professor with the Department of Packaging Engineering, Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology. Her main areas of research are image processing and computer vision.