# HUMAN OBJECT EXTRACTION IN VIDEOS USING MOTION SALIENCY

Seemanthini.K
Department of Computer Science and Engineering
R.V.College of Engineering, Bangalore, India
E-mail:be.outstanding@gmail.com

Under the guidance of :
Mr.Vinay Hegde
Assosiate Professor
Department of Computer Science and Engineering
R.V.College of Engineering, Bangalore, India
E-mail:vinayhegde@rvce.edu.in

***Abstract*— This paper presents a saliency-based human object extraction (HOE) framework. The proposed framework aims to automatically extract foreground human object without any user interaction or the use of any training data. In the proposed method, The coarse foreground extraction is obtained by using the motion and the edge information of an object. Then, the human object is extracted by using the horizontal/ vertical filling scheme based on the coarse foreground extraction. The proposed method integrate these feature models into a unified framework via a Conditional Random Field (CRF), and this CRF can be applied to video object segmentation and further video editing and retrieval applications.Experimental results shows that the proposed human object extraction has good performance in sensitivity, specificity, spatial accuracy and execution time.**

***Index Terms* — Conditional Random Field (CRF), Human Object Extraction (HOE), Visual Saliency,Foreground Extraction.**

IJERT

## I. INTRODUCTION

Extraction of Objects (human) in video sequences is very important in many aspects of multimedia applications. Video object extraction is an important key technology for content based video coding, representation, indexing, and retrieval. Video object extraction can be described as a method of extracting the foreground object from each frame of a video sequence.Video object extraction requires consistent object labeling throughout the video sequence, where the consistent object labeling would be color, texture, motion, spatial-temporal structure etc.Video object extraction can also be applied to some interesting and potential applications, such as video surveillance, digital watermarking, behavior analysis of sport video and advanced story retrieval.

Human can easily determine the subject of interest in a video, even though that subject is presented in an unknown or cluttered background or even has never been seen before. With the complex cognitive capabilities exhibited by human brains, this process can be interpreted as simultaneous extraction of both foreground and background information from a video. Many researchers have been working toward closing the gap between human and computer vision.

Many methods have been proposed for video object extraction. Generally, these methods can be roughly classified into two types: Background construction-based video object extraction and foreground extraction –based video object extraction. In background construction – based video object extraction, the background information is first constructed. Then, an initial video object is obtained based on the difference between the background and the current frame.Finally, a video object in the successive frame can be obtained by using object tracking or background substraction. Background construction- based video object extraction can be keep object tracking with fast moving objects. Furthermore, its computational cost is low and its implementation is easy.

In foreground extraction-based video object extraction, temporal information, spatial information, or temporal – spatial information is first used to obtain an initial video object. Then, the video object in the successive frame can be obtained by using motion information, change information and other feature information. In contrast to background construction-based video object extraction, foreground extraction- based video object extraction can obtain accurate video object boundaries for low moving objects.

However, background construction-based video object extraction is difficult to obtain good background information when the moving object exists in the first frame of the video sequence. That is, good background information has not yet been constructed in the first frame. Therefore, foreground extraction- based video object extraction is suitable for video object extraction when the moving object exists in the first frame of the video sequence.

In this paper, a foreground extraction-based method using motion information is proposed to obtain good human video object extraction in the whole video sequence. In the proposed method, the motion information of the video object is obtained using the angle-module rule. Then, a coarse foreground extraction is obtained by using motion information. Next, the human video object is extracted using the horizontal / vertical filling scheme based on the coarse foreground extraction and fine foreground extraction.

Finally, the conditional random field is applied for video object segmentation.

A conditional random field is applied to automatically determines the label (foreground or background) of each pixel based on the observed models. With the ability to preserve both spatial and temporal consistency. A conditional random field is applied to effectively combine the saliency induced features, which allows us to deal with unknown pose and scale variations of the foreground object (and its articulated parts). Based on the ability to preserve both spatial continuity and temporal consistency in the proposed HOE framework, experiments on a variety of videos verify that our method is able to produce quantitatively and qualitatively satisfactory HOE results.

However, without any prior knowledge on the subject of interest or training data, it is still very challenging for computer vision algorithms to automatically extract the foreground object of interest in a video.

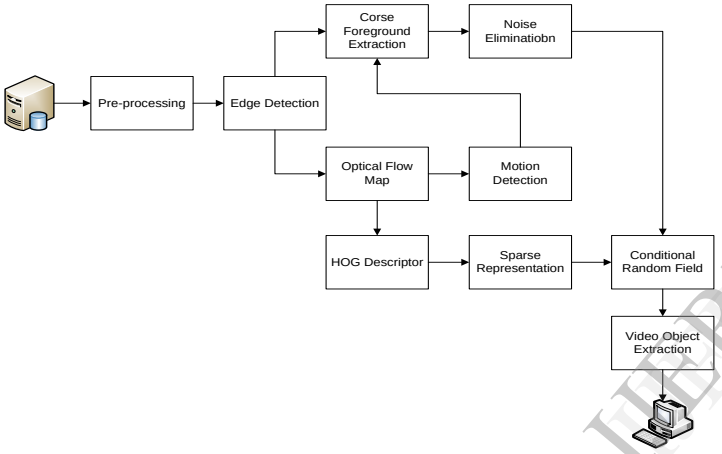


fig 1:Overview of the proposed system

## II. RELATED WORK

Several methods were proposed to detect object parts rather than the entire object. For example, Nevatia et al and Davis et al Both decomposed an object shape model in a hierarchical way to train object part detectors. These detectors are used to describe all possible configurations of the object of interest. Gorelick and Basri collected a set of object silhouette exemplars. To extract the object of interest, the authors over-segmented the input image and determined the segments which best matched the associated templates. To deal with multiple human instances with large pose deformations, Niebles et al applied a human body detector on each frame, and their detection results were refined by pose density estimation function and probability diffusion between adjacent frames. Recently in the authors further utilized template matching between the result produced by pedestrian detectors and a set of upright human pose templates, which is to simultaneously regularize and reduce the search space of possible object model configurations. However, these part-based methods typically assume that the object categories are known in advance, and they need to collect the object part templates to design each part detector.

Besides the above approaches, graph-based methods have been shown to be effective for foreground object segmentation. Using such methods, an image is typically represented by a graph, in which each observed node indicates an image pixel and the associated hidden node corresponds to its label. By determining the cost between adjacent hidden nodes using color, motion, etc. information, one can segment the foreground object by dividing the graph into disjoint parts while minimizing the total cost. Previous work such as and focused on an interactive scheme and required users to manually provide the ground truth label information.

*Our Contributions:*

This paper aims at automatically extracting foreground objects in videos which are captured by freely moving cameras. Instead of assuming that the background motion is dominant and different from that of the foreground, we relax this assumption and allow foreground objects to be presented in freely moving scenes. We advance both visual and motion saliency information across video frames, and a CRF model is utilized for integrating the associated features for HOE (i.e., visual saliency, shape, foreground/background color models, and spatial/temporal energy terms). From our quantitative and qualitative experiments, we verify that our HOE performance exhibits spatial consistency and temporal continuity, and our method is shown to outperform state-of the- art unsupervised HOE approaches. It is worth noting that, our proposed HOE framework is an unsupervised approach, which does not require the prior knowledge (i.e., training data) of the object of interest nor the user interaction for any annotation.

## III. AUTOMATIC OBJECT MODELING AND EXTRACTION

Most existing unsupervised HOE approaches assume the foreground objects as outliers in terms of the observed motion information, so that the induced appearance, color, etc. features are utilized for distinguishing between foreground and background regions. However, these methods cannot generalize well to videos captured by freely moving cameras as discussed
earlier. In this work, we propose a saliency-based HOE framework which learns saliency information in both spatial (visual) and temporal (motion) domains. By advancing conditional random fields (CRF), the integration of the resulting features can automatically identify the foreground object without the need to treat either foreground or background as outliers. Fig. 1 shows the proposed VOE framework, and we now detail each step in the following subsections.


(a) (b) (c)

IJERT

Fig. 2. Example of visual saliency calculation. (a) Original video frame.(b) Visual saliency of (a) derived by (1). (c) Visual saliency of (a) refined by (2).

*A. Determination of Visual Saliency*

To extract visual saliency of each frame, we perform image segmentation on each video frame and extract color and contrast information. In our work, we advance Turbopixels proposed by for segmentation, and the resulting image segments (superpixels) are applied to perform saliency detection. The use of Turbopixels allows us to produce edge preserving superpixels with similar sizes, which would achieve improved visual saliency results as verified later. For the $k$th superpixel $r_k$ , we calculate its saliency score $S(r_k)$ as follows:

$$S(r_k) = \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i)D_r(r_k, r_i) \approx \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)D_r(r_k, r_i) \quad (1)$$

where $D_s$ is the Euclidean distance between the centroid of $r_k$ and that of its surrounding superpixels $r_i$ , while $\sigma_s$ controls the width of the kernel. The parameter $\omega(r_i)$ is the weight of the neighbor superpixel $r_i$ , which is proportional to the number of pixels in $r_i$ . Compared to [27], $\omega(r_i)$ can be treated as a constant for all superpixels due to the use of Turbopixels (with similar sizes). The last term $D_r(r_k, r_i)$ measures the color difference between $r_k$ and $r_i$ , which is also in terms of Euclidean distance.

As suggested by [22], we consider the pixel $i$ as a salient point if its saliency score satisfies $S(i) > 0.8 * \max(S)$, and the collection of the resulting salient pixels will be considered as a salient point set. Since image pixels which are closer to this salient point set should be visually more significant than those which are farther away, we further refine the saliency $\hat{S}(i)$ for each pixel $i$ as follows:

$$\hat{S}(i) = S(i) * (1 - \text{dist}(i)/\text{dist}_{\max}) \quad (2)$$

where $S(i)$ is the original saliency score derived by (1), and $\text{dist}(i)$ measures the nearest Euclidian distance to the salient point set. We note that $\text{dist}_{\max}$ in (2) is determined as the maximum distance from a pixel of interest to its nearest salient point within an image, thus it is an image-dependent constant. An example of visual saliency calculation is shown in Fig. 2.

*B. Extraction of Motion-Induced Cues*

*1)Determination of Motion Saliency:* We now discuss how we determine the motion saliency, and how we extract the associated cues for HOE purposes. Unlike prior works which assume that either foreground or background exhibits dominant motion, our proposed framework aims at extracting motion salient regions based on the retrieved optical flow

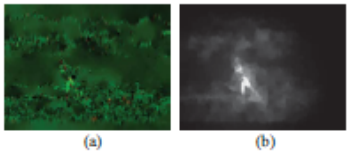


Fig. 3. Motion saliency calculated for Fig. 2. (a) Calculation of the optical flow. (b) Motion saliency derived from (a).

information. To detect each moving part and its corresponding pixels, we perform dense optical-flow forward and backward propagation [28] at each frame of a video. A moving pixel $q_t$ at frame $t$ is determined by

$$q_t = \hat{}q_{t,t-1} \bigcap \hat{}q_{t,t+1} \quad (3)$$

where $\hat{}q$ denotes the pixel pair detected by forward or ackward
optical flow propagation. We do not ignore the frames which result in a large number of moving pixels at this stage as [13], [14] did, and thus our setting is more practical for real-world videos captured by freely-moving cameras.

After determining the moving regions, we propose to derive the saliency scores for each pixel in terms of the associated optical flow information. Inspired by visual saliency approaches like [27], we apply our proposed algorithms in (1) and (2) on the derived optical flow results to calculate the motion saliency $M(i, t)$ for each pixel $i$ at frame $t$, and the saliency score at each
frame is normalized to the range of [0, 1] (see Fig. 3 for example). It is worth noting that, when the foreground object exhibits significant movements (compared to background), its motion will be easily captured by optical flow and thus the corresponding motion salient regions can be easily extracted. On the other hand, if the camera is moving and thus results in remarkable background movements, the proposed motion saliency method will still be able to identify motion salient regions (associated with the foreground object), as verified later by our experiments. Compare Figs. 1(a) and (b), we see that the motion saliency derived from the optical flow has a better representative capability in describing the foreground regions than the direct use of the optical flow does. Another example is shown in Fig. 3, in which we observe that the foreground object (the surfer) is significantly more salient than the moving background in terms of motion. From the above discussions, we consider motion saliency as an important and supplementary information for identifying foreground objects.

*2) Learning of Shape Cues:* Although motion saliency allows us to capture motion salient regions within and across video frames, those regions might only correspond to moving parts of the foreground object within some time interval. If we simply assume the foreground should be near the high motion saliency region as the method in [13] did, we cannot easily identify the entire foreground object. Since it is typically observed that each moving part of a foreground object forms a complete sampling of the entire

IJERT

foreground object (e.g., same assumption is made in [5], [6], [13], [14]), we advance part-based shape information induced by motion cues for characterizing the foreground object.

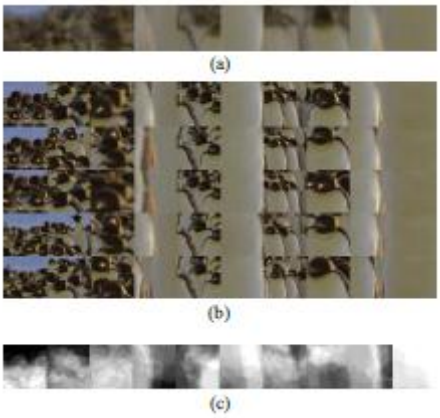


Fig. 4. Visualization of sparse shape representation. (a) Example codewords for sparse shape representation. (b) Corresponding image patches (only top 5 matches shown). (c) Corresponding masks for each codeword.

To describe the motion salient regions, we convert the motion saliency image into a binary output and extract the shape information from the motion salient regions. More precisely, we first binarize the aforementioned motion saliency *M(i, t)* into Mask*(i, t)* using a threshold of 0.25. We divide each video frame into disjoint 8 × 8 pixel patches. For each image patch, if more than 30% of its pixels are with high motion saliency (i.e., pixel value of 1 in the binarized output), we compute the histogram of oriented gradients (HOG) descriptors with 4 × 4 = 16 grids for representing its shape information. To capture scale invariant shape information, we further downgrade the resolution of each frame and repeat the above process. We choose the lowest resolution of the scaled image as a quarter of that of the original one. We note that a similar setting for scale invariance has also been applied in [29] when extracting the HOG descriptors.

Since the use of sparse representation has been shown to be very effective in many computer vision tasks [30], we learn an over-complete codebook and determine the associated sparse representation of each HOG. Now, for a total of $N$ HOG descriptors calculated for the above motion-salient patches $\{\mathbf{h}n, n = 1, 2, \ldots, N\}$ in a $p$-dimensional space, we construct an over-complete dictionary $\mathbf{D}p \times K$ which includes $K$ basis vectors, and we determine the corresponding sparse coefficient $\alpha n$ of each HOG descriptor. Therefore, the sparse coding problem can be formulated as

$$\min_{\mathbf{D},\boldsymbol{\alpha}} \_ \_ 1 \sum_{n=1}^{N} 1 \|\mathbf{h}_n - \mathbf{D}\boldsymbol{\alpha}_n\| 22+ \lambda\|\boldsymbol{\alpha}_n\|1 \quad (4)$$

where $\lambda$ balances the sparsity of $\alpha n$ and the $l2$-norm reconstruction error. We use the software developed by [31] to solve the above problem. Fig. 4(a) shows example basis vectors
(codewords) in terms of image patches. We note that each codeword is illustrated by averaging image patches with the top 15 $\alpha n$ coefficients (see examples illustrated in Fig. 4(b), in which only the top 5 matches are shown). To alleviate the possible presence of background in each codeword $k$, we combine the binarized masks of the top 15 patches using the corresponding weights $\alpha n$ to obtain the map $Mk$. As a

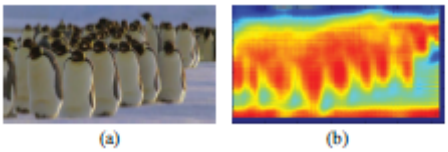


Fig. 5. Shape likelihood reconstructed by sparse shape representation. (a) Original frame. (b) Shape likelihood.

result, the moving pixels within each map (induced by motion saliency) has non-zero pixel values, and the remaining parts of that patch are considered as static background and thus are zeroes. Fig. 4(c) shows example results for each codeword shown in Fig. 4(a).
After obtaining the dictionary and the masks to represent the shape of foreground object, we use them to encode all image patches at each frame. This is to recover non-moving regions of the foreground object which does not have significant motion and thus cannot be detected by motion cues. For each image patch, we derive its sparse coefficient vector $\alpha$, and each entry of this vector indicates the contribution of each shape codeword. Correspondingly, we use the associated masks and their weight coefficients to calculate the final mask for each image patch. Finally, the reconstructed image at frame $t$ using the above maps $Mk$ can be denoted as foreground shape likelihood
$\_ X$
$S$
$t$, which is calculated as follows:

$$\_ XSt = \_ n \in ItK\_k=1 (\alpha n,k \cdot Mk) \quad (5)$$

where $\alpha n,k$ is the weight for the $n$th patch using the $k$th codeword.
Fig. 5 shows an example of the reconstruction of a video frame using the motion-induced shape information of the foreground object. We note that $\_ XS\ t$ serves as the likelihood of foreground object at frame $t$ in terms of shape information.

*3) Learning of Color Cues:* Besides the motion-induced shape information, we also extract both foreground and background color information for improved VOE performance. According to the observation and the assumption that each moving part of the foreground object forms a complete sampling of itself, we cannot construct foreground or background color models simply based on visual or motion saliency detection results at each individual frame; otherwise, foreground object regions which are not salient in terms of visual or motion appearance will be considered as background, and the resulting color models will not be of sufficient discriminating capability. In our work, we utilize the shape likelihood $\wedge X^S_t$ obtained from the previous step, and we threshold this likelihood by 0.5 to

IJERT

determine the candidate foreground ($FS_{shape}$) and background ($BS_{shape}$) regions. In other words, we consider color information of pixels in *FS*shape for calculating the foreground color GMM, and those in *BS*shape for deriving the background color GMM.Once these candidate foreground and background regions
are determined, we use Gaussian mixture models (GMM) *GC f* and *GCb* to model the RGB distributions for each model. The parameters of GMM such as mean vectors and covariance matrices are determined by performing an Fig. 6. CRF for foreground object segmentation.expectation-maximization (EM) algorithm. Finally, we integrate both foreground and background color models with visual saliency and shape likelihood into a unified framework for HOE.

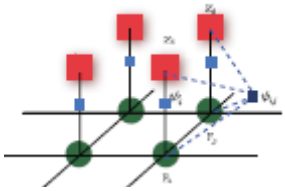

Fig. 6. CRF for foreground object segmentation.

## IV. CONDITION RANDOM FIELD FOR HOE

### *A. Feature Fusion via CRF*

Utilizing an undirected graph, conditional random field (CRF) [32] is a powerful technique to estimate the structural information (e.g. class label) of a set of variables with the associated observations. For video foreground object segmentation, CRF has been applied to predict the label of each observed pixel in an image *I* [13], [14]. As illustrated in Fig. 6, pixel *i* in a video frame is associated with observation $z_i$ , while the hidden node $F_i$ indicates its corresponding label (i.e. foreground or background). In this framework, the label $F_i$ is calculated by the observation $z_i$ , while the spatial coherence between this output and neighboring observations $z_j$ and labels $F_j$ are simultaneously taken into consideration. Therefore, predicting the label of an observation node is equivalent to maximizing the following posterior probability function

$$p(F|I,\psi) \alpha \exp\left\{-\left(\sum_{i\epsilon I}(\psi_i) + \sum_{i\epsilon I,\, j\epsilon \text{Neighbor}}(\psi_{i,j})\right)\right\} \quad (6)$$

where $\psi_i$ is the unary term which infers the likelihood of $F_i$ with observation $z_i$ . $\psi_{i,j}$ is the pairwise term describing the relationship between neighboring pixels $z_i$ and $z_j$, and that between their predicted output labels $F_i$ and $F_j$ . Note that the observation *z* can be represented by a particular feature, or a combination of multiple types of features (as our proposed framework does).

To solve a CRF optimization problem, one can convert the above problem into an energy minimization task, and the object energy function *E* of (6) can be derived as

$$E = -\log(p)$$

$$= \sum_{i\epsilon I}(\psi_i) + \sum_{\substack{i\epsilon I \\ j\epsilon \text{Neighbor}}}(\psi_{i,j})$$

$$= E_{\text{unary}} + E_{\text{pairwise}}. \quad (7)$$

In our proposed VOE framework, we define the shape energy function *ES* in terms of shape likelihood $\hat{X}^S_t$ (derived by (5)) as one of the unary terms

$$E^S = -w_S \log(\hat{X}^S t). \quad (8)$$

In addition to shape information, we need incorporate visual saliency and color cues into the introduced CRF framework. As discussed earlier, we derive foreground and background color models for VOE, and thus the unary term *EC* describing color information is defined as follows:

$$E^C = w^c(E^{CF} - E^{CB}) \quad (9)$$

Note that the foreground and background color GMM models $G_{C\,f}$ and $G_{Cb}$ (discussed in Section III-B) are utilized to derive the associated energy terms *ECF* and *ECB*, which are calculated as

$$\begin{cases} E^{CF} = -\log\left(\sum_{i\epsilon I} G^C{}_f(i)\right) \\ E^{CB} = -\log\left(\sum_{i\epsilon I} G^C{}_b(i)\right) \end{cases}$$

As for the visual saliency cue at frame *t*, we convert the visual saliency score $\hat{S}t$ derived in (2) into the following energy term $E^V$ :

$$E^V = -w^v \log(\hat{S}t) \quad (10)$$

We note that in the above equations, parameters $w_S$ , $w_C$, and $w_V$ are the weights for shape, color, and visual saliency cues, respectively. These weights control the contributions of the associated energy terms of the CRF model for performing HOE. It is also worth noting that, Liu and Gleicher [13] only considers the construction of foreground color models for HOE. As verified by [14], it can be concluded that the disregard of background color models would limit the performance of HOE, since the only use of foreground color model might not be sufficient for distinguishing between foreground and background regions. In the proposed HOE framework, we now utilize multiple types of visual and motion salient features for HOE, and our experiments will confirm the effectiveness and robustness of our approach on a variety of real-world videos.

### *B. Preserving Spatio-Temporal Consistency*

In the same shot of a video, an object of interest can be considered as a compact space-time volume, which exhibits smooth changes in location, scale, and motion across frames. Therefore, how to preserve spatial and temporal consistency within the extracted foreground object regions across video frames is a major obstacle for HOE. Since

IJERT

there is no guarantee that combining multiple motion-induced features would address the above problem, we need to enforce additional constraints in the CRF model in order to achieve this goal.

*1) Spatial Continuity for HOE:* When applying a pixel-level prediction process for VOE (like ours and some prior HOE methods do), the spatial structure of the extracted foreground region is typically not considered during the HOE process. This is because that the prediction made for one pixel is not related to those for its neighboring ones. To maintain the spatial consistency for the extracted foreground object, we add a pairwise term in our CRF framework. The introduced pairwise term $E_{i,j}$ is defined as

$$E_{i,j} = \sum_{i \in I, \; j \in \text{Neighbor}} |F_i - F_j| \times \left(\lambda_1 + \lambda_2 \left(\exp\left(-\frac{\|z_i - z_j\|}{\beta}\right)\right)\right)$$


Fig. 7. (a) Original frame. Example HOE results (b) with and (c) without imposing the temporal consistency term for CRF.

Note that $\beta$ is set as the averaged pixel color difference of all pairs of neighboring pixels. In (11), $\lambda 1$ is a data-independent Ising prior to smoothen the predicted labels, and $\lambda 2$ is to relax the tendency of smoothness if color observations $z_i$ and $z_j$ form an edge (i.e. when $\|z_i - z_j\|$ is large). This pairwise term is able to produce coherent labeling results even under low contrast or blurring effects, and this will be verified later in Section V.

*2) Temporal Consistency for VOE:* Although we exploit both visual and motion saliency information for determining the foreground object, the motion-induced features such as shape and foreground/background color GMM models might not be able to well describe the changes of foreground objects across videos due to issues such as motion blur, compression loss, or noise/artifacts presented in video frames. To alleviate this concern, we choose to propagate the foreground/background shape likelihood and CRF prediction outputs across video frames for preserving temporal continuity in our HOE results. To be more precise, when constructing the foreground and background color GMM models, the corresponding pixel sets *FS* and *BS* will not only be produced by the shape likelihood *FS*shape and *BS*shape at the current frame, those at the previous frame (including the CRF prediction outputs $\hat{F}$foreground and $\hat{F}$background) will be considered to update *FS* and *BS* as well. In other words, we update foreground and background pixel sets *FS* and *BS* at frame $t + 1$ by

$$\begin{cases} FS_{t+1} = FS\text{shape}(t+1) \cap \_FS\text{shape}(t) \_ \hat{F}\text{foreground}(t) \\ BS_{t+1} = BS\text{shape}(t+1) \_ BS\text{shape}(t) \_ \hat{F}\text{background}(t) \end{cases} \quad (12)$$

where $\hat{F}$foreground$(t)$ indicates the pixels at frame $t$ to be predicted as foreground, and $FS$shape$(t)$ is the set of pixels whose shape likelihood is above 0.5 as described in Section III.B3.Similar remarks apply for $\hat{F}$background$(t)$ and $BS$shape$(t)$. We show an example in Fig. 7 to verify the use of such temporal
terms when updating our VOE model.
Finally, by integrating (8), (9), (10), and (11), plus the introduced terms for preserving spatial and temporal information, the objective energy function (7) can be updated as

$$\begin{aligned} E &= E\text{unary} + E\text{pairwise} \\ &= (E^S + E^{CF} - E^{CB} + E^V) + Ei, j \\ &= E^S + E^C + E^V + Ei, j \; . \end{aligned} \quad (13)$$

To minimize (13), one can apply graph-based energy minimization techniques such as max-flow/min-cut algorithms. When the optimization process is complete, the labeling function output $F$ would indicate the class label (foreground or background) of each observed pixel at each frame, and thus the HOE problem is solved accordingly.

## V. EXPERIMENTAL RESULTS

In this section, we conduct experiments on a variety of videos. We first verify the integration of multiple types of features for HOE, and show that it outperforms the use of a particular type of feature. We also compare our derived saliency maps and segmentation results to those produced by other saliency based or state-of-the-art supervised or unsupervised HOE methods. Both qualitative and quantitative results will be presented to support the effectiveness and robustness of our proposed method.


Fig. 8. VOE results using different feature cues (the CRF pairwise term is considered for all cases for fair comparisons).

IJERT

## VI. CONCLUSION

In this paper, we proposed an automatic HOE approach which utilizes multiple motion and visual saliency induced features, such as shape, foreground/background color models, and visual saliency, to extract the foreground objects in videos.We advanced a CRF model to integrate the above features, and additional constraints were introduced into our CRF model for preserving both spatial continuity and temporal consistency when performing HOE. Compared with state-of-the-art unsupervised HOE methods, our approach was shown to better model the foreground object due to the fusion of multiple types of saliency-induced features. A major advantage of our proposed method is that we do not require the prior knowledge of the object of interest (i.e., the need to collect training data), nor the interaction from the users during the segmentation progress. Experiments on a variety of videos with highly articulated objects or complex background presented verified the effectiveness and robustness of our proposed method.

## VII. REFERENCES

[1] Wei-Te Li, Haw-Shiuan Chang, Kuo-Chin Lien, Hui-Tang Chang, and Yu-Chiang Frank Wang, "Exploring Visual and Motion Saliency for Automatic Video Object Extraction", *IEEE transactions on image processing, vol. 22, no. 7, july 2013*

[2] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–618, Apr. 2010.

[3] Y. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*,vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[4] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[5] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 53–60.

[6] P. Yin, A. Criminisi, J. M. Winn, and I. A. Essa, "Bilayer segmentation of webcam videos using tree-based classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 30–42, Jan. 2011.

[7] X. Bai and G. Sapiro, "A geodesic framework for fast interactive image and video segmentation and matting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[8] M. Gong and L. Cheng, "Foreground segmentation of live videos using locally competing 1SVMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2105–2112.

[9] T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of Gaussians for foreground detection—A survey," *Recent Patents Comput. Sci.*, vol. 3, no. 3, pp. 219–237, 2008.

[10] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 185–204, 2009.

[11] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 1. Oct. 2003, pp. 44–50.

[12] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," in *Proc.9th Eur. Conf. Comput. Vis.*, 2006, pp. 628–641.

[13] F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit.*, Jun. 2009, pp. 320–327.

[14] K.-C. Lien and Y.-C. F. Wang, "Automatic object extraction in single concept videos," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011,pp. 1–6.

[15] M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 1142–1149.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[17] P. Harding and N. M. Robertson, "Visual saliency from image features with application to compression," *Cognit. Comput.*, vol. 5, no. 1, pp. 76–98, 2012.

[18] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[19] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[20] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[21] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010,pp. 2653–2656.

[22] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[23] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[24] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Int. Conf. Multimedia*, 2003,pp. 374–381.

[25] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2368–2375.

[26] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J. Dickinson,"TurboPixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[27] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[28] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 optical flow," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[29] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models,"*IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[30] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2009.

[32] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Data*. SanMateo, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[33] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. Brit. Mach. Vis. Conf.*, 2010,pp. 1–11.

[34] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE Int. Conf. Multimedia Expo*,Jun.–Jul. 2009, pp. 638–641

[35] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph based video segmentation," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit.*, Jun. 2010, pp. 2141–2148.

[36] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011,pp. 1995–2002.

[37] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 575–588.

IJERT