International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 05 (May 2015), PP.08-23

Automatic Foreground object detection using Visual and Motion Saliency

Dr.M.V.L.N.Raja Rao[1] M.Nagaraju[2], Y.K.Viswanadham[3,]
Ch. Amala[4], B.Revathi[5]
[1]Professor & HoD, IT Dept, GEC [2]Assistant Professor, IT Dept, GEC
[3]Associate Professor, IT Dept, GEC [4]Assistant Professor, ECE Dept, GEC
[5]Assistant Professor, IT, Dept

Abstract:- This paper presents a saliency-based video object extraction (VOE) framework. The proposed framework aims to automatically extract foreground objects of interest without any user interaction or the use of any training data (i.e., not limited to any particular type of object). To separate foreground and background regions within and across video frames, the proposed method utilizes visual and motion saliency information extracted from the input video. 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 VOE framework, experiments on a variety of videos verify that our method is able to produce quantitatively and qualitatively satisfactory VOE results.

Keywords:- Conditional random field (CRF), video object extraction (VOE), visual saliency.

I. INTRODUCTION

At a Glance, 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. 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. As a result, if one needs to design an algorithm to automatically extract the foreground objects from a video, several tasks need to be addressed. 1) Unknown object category and unknown number of the object instances in a video. 2) Complex or unexpected motion of foreground objects due to articulated parts or arbitrary poses. 3) Ambiguous appearance between foreground and background regions due to similar color, low contrast, insufficient lighting, etc. conditions. In practice, it is infeasible to manipulate all possible foreground object or background models beforehand. However, if one can extract representative information from either foreground or background (or both) regions from a video, the extracted information can be utilized to distinguish between foreground and background regions, and thus the task of foreground object extraction can be addressed. As discussed later in Section II, most of the prior works either consider a fixed background or assume that the background exhibits dominant motion across video frames. These assumptions might not be practical for real world applications, since they cannot generalize well to videos captured by freely moving cameras with arbitrary movements.

In this paper, we propose a robust video object extraction (VOE) framework, which utilizes both visual and motion saliency information across video frames. The observed saliency information allows us to infer several visual and motion cues for learning foreground and background models, and a conditional random field (CRF) 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, our VOE framework exhibits promising results on a variety of videos, and produces quantitatively and qualitatively satisfactory performance. The remainder of this paper is organized as follows. Section II reviews recent works on video object extraction and highlights the contributions of our method. Details of our proposed VOE framework are presented in Sections III and IV. Section V shows our empirical results on several types of video data, and both qualitative and quantitative results are While we focus on VOE problems for single concept videos (i.e., videos which have only one object category of interest presented), our proposed method is able to deal with multiple object instances (of the same type) with pose, scale, etc. variations.

The remainder of this paper is organized as follows. Section II reviews recent works on video object extraction and highlights the contributions of our method. Details of our proposed VOE framework are presented in Sections III and IV. Section V shows our empirical results on several types of video data, and both qualitative and quantitative results are presented to support the effectiveness and robustness of our method. Finally, Section VI concludes this paper.

II. RELATED WORK

In general, one can address VOE problems using supervised or unsupervised approaches. Supervised methods require prior knowledge on the subject of interest and need to collect training data beforehand for designing the associated VOE algorithms. For example, Wu and Nevatia [1] and Lin and Davis [2] both decomposed an object shape model in a hierarchical way to train object part detectors, and these detectors are used to describe all possible configurations of the object of interest (e.g. pedestrians). Another type of supervised methods require user interaction for annotating candidate foreground regions. For example, image segmentation algorithms proposed in [3], [4] focused on an interactive scheme and required users to manually provide the ground truth label information. For videos captured by a monocular camera, methods such as Criminisi *et al*., Yin *et al*. [5], [6] applied a conditional random field (CRF) maximizing a joint probability of color, motion, etc. models to predict the label of each image pixel. Although the color features can be automatically determined from the input video, these methods still need the user to train object detectors for extracting shape or motion features. Recently, researchers proposed to use some preliminary strokes to manually select the foreground and background regions, and they utilized such information to train local classifiers to detect the foreground objects [7], [8]. While these works produce promising results, it might not be practical for users to manually annotate a large amount of video data.

On the other hand, unsupervised approaches do not train any specific object detectors or classifiers in advance. For videos captured by a static camera, extraction of foreground objects can be treated as a background subtraction problem. In other words, foreground objects can be detected simply by subtracting the current frame from a video sequence [9], [10]. However, if the background is consistently changing or is occluded by foreground objects, background modeling becomes a very challenging task. For such cases, researchers typically aim at learning the background model from the input video, and the foreground objects are considered as outliers to be detected. For example, an autoregression moving average model (ARMA) that estimates the intrinsic appearance of dynamic textures and regions was proposed in [11], and it particularly dealt with scenarios in which the background consists of natural scenes like sea waves or trees. Sun *et al*. [12] utilized color gradients of the background to determine the boundaries of the foreground objects. Some unsupervised approaches aim at observing features associated with the foreground object for VOE. For example, graph-based methods [13], [14] identify the foreground object regions by minimizing the cost between adjacent hidden nodes/pixels in terms of color, motion, etc. information. More specifically, one can segment the foreground object by dividing a graph into disjoint parts whose total energy is minimized without using any training data. While impressive results were reported in [13], [14], these approaches typically assume that the background/camera motion is dominant across video frames. For general videos captured by freely moving cameras, these methods might not generalize well (as we verify later in experiments). Different from graph-based methods, Leordeanu and Collins [15] proposed to observe the co-occurrences of object features to identify the foreground objects in an unsupervised setting. Although promising results under pose, scale, occlusion, etc. variations were reported, their approach was only able to deal with rigid objects (like cars).

Since Itti *et al*. [16] first derived the visual saliency of a single image, numerous works have been proposed to extract the saliency information of images for the tasks of compression, classification, or segmentation. For example, Harding and Robertson [17] demonstrate that the visual saliency can be utilized to improve image compression ratio by combining SURF features and task-dependent prior knowledge. Unlike compression or classification problems which might utilize task or object category information for deriving the associated saliency, general saliency detection or image segmentation tasks are solved in an unsupervised setting. For example, based on spectrum analysis, Hou and Zhang [18] utilized the spectral residual as saliency information, while Guo *et al*. [19] advanced the phase part of the spectrum together with Quaternion Fourier Transform for saliency detection. Liu *et al*. [20] considered contrast information and color histogram of different image regions in multiple scales to detect local and global image saliency. Achanta and Süsstrunk [21] omputed the saliency by taking symmetric surrounding pixels into consideration and averaging the color differences between pixels within each region. Goferman *et al*. [22] applied multi-scale patches and calculated both color differences and locations between different patches. Zhai and Shah [23] constructed spatial and temporal saliency maps by using a spatiotemporal attention model. Based on local image contrast, Ma and Zhang [24] determined the salient regions by fuzzy growing which extracts regions or objects of interest when forming the saliency map. Recently, Wang *et al*. [25] proposed a biological inspired approached and derived visual saliency based on site entropy rate for saliency detection. Nevertheless, finding visual saliency in images or video frames would provide promising results and infer the region of the foreground objects. However, since real-world videos might encounter low contrast or insufficient lighting, etc. problems, one might not be able to obtain desirable visual saliency maps for identifying candidate foreground objects. As a result, one cannot simply apply visual saliency methods for segmenting foreground objects in real world videos.

Our Contributions: In this paper, we aim 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 as [13], [14] did, 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 VOE (i.e., visual saliency, shape, foreground/background color models, and spatial/temporal energy terms). From our quantitative and qualitative experiments, we verify that our VOE performance exhibits spatial consistency and temporal continuity, and our method is shown to outperform state-of the- art unsupervised VOE approaches. It is worth noting that, our proposed VOE 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 MODELLING AND EXTRACTION

Most existing unsupervised VOE 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 VOE 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. 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 [26] for segmentation, and the resulting image segments (superpixels) are applied to perform saliency detection. The use of Turbopixels allows us to produce edgepreserving superpixels with similar sizes, which would achieve improved visual saliency results as verified later. For the *k*th superpixel *rk* , we calculate its saliency score *S(rk)* as follows:

$$\begin{aligned} S(r_k) &= \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i) D_r(r_k, r_i) \\ &\approx \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2) D_r(r_k, r_i) \qquad (1) \end{aligned}$$

where *Ds* is the Euclidean distance between the centroid of *rk* and that of its surrounding superpixels *ri* , while *σs* controls the width of the kernel. The parameter *ω(ri)* is the weight of the neighbor superpixel *ri* , which is proportional to the number of pixels in *ri* . The last term *Dr (rk , ri)* measures the color difference between *rk* and *ri* , which is also in terms of Euclidean distance. 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 *ˆS (i)* for each pixel *i* as follows:

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

where *S(i)* is the original saliency score derived by (1), and dist*(i)* measures the nearest Euclidian distance to the salient point set. We note that distmax 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.

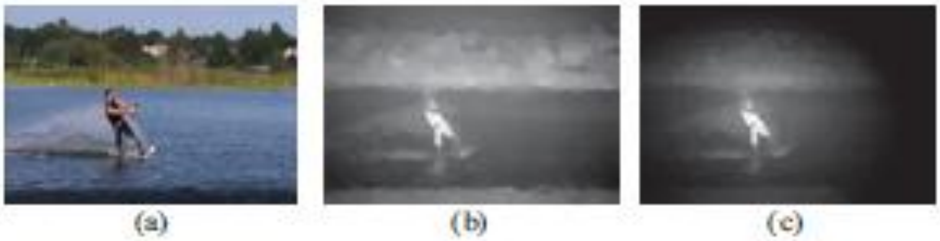


Fig. 2. Example of visual saliency calculation.

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

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 VOE 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 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 *qt* at frame *t* is determined by

$$q_t = \widehat{q}_{t,\ t-1} \bigcap \widehat{q}_{t,\ t+1} \qquad (3)$$

Where$^{\wedge}$ *q* denotes the pixel pair detected by forward or backward 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). Consider Fig. 1, 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 important and supplementary information for identifying foreground objects.

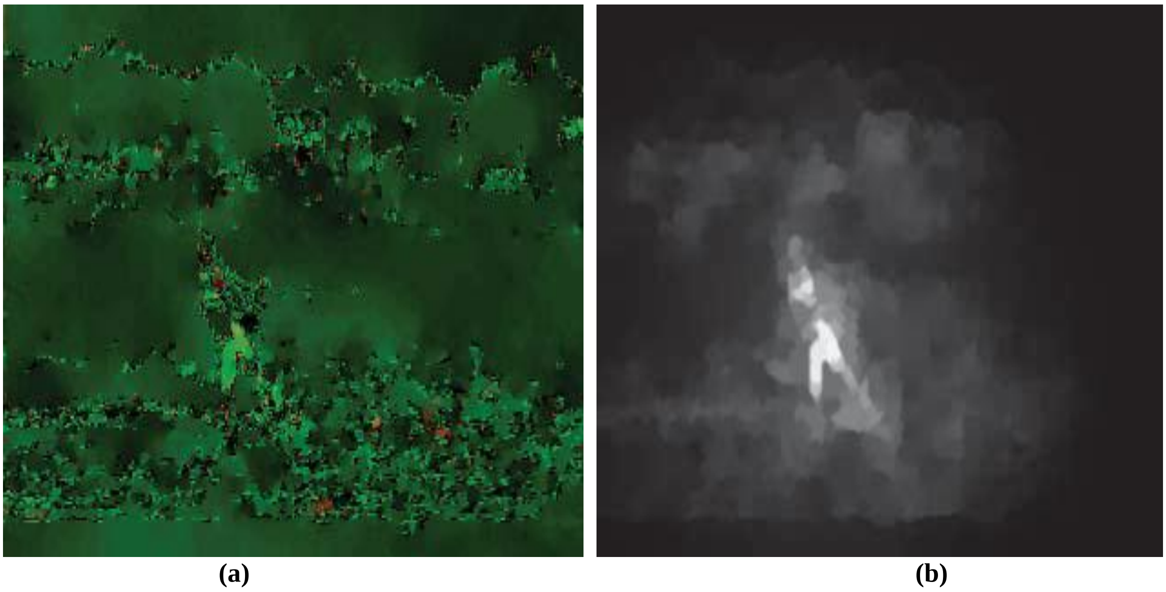


Fig. 3. Motion saliency calculated for Fig. 2.

Fig. 1 illustrates the overview of our proposed VOE framework.

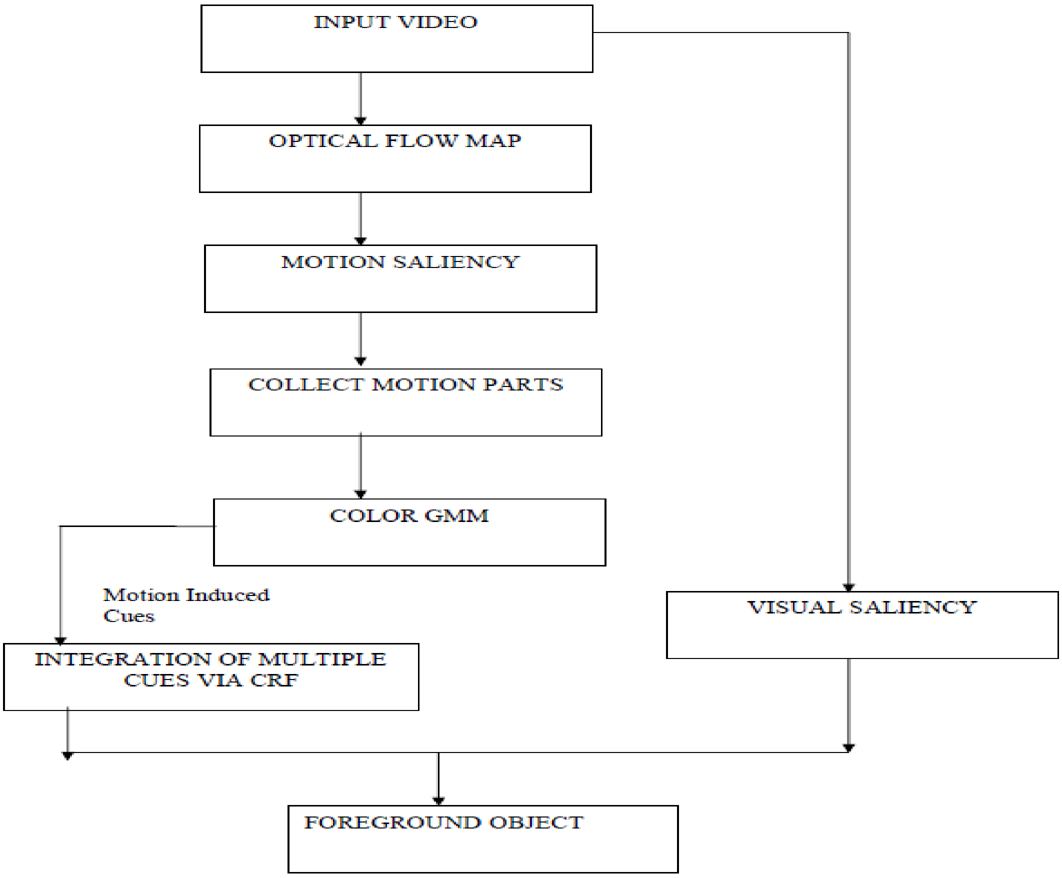


Fig 1. Block Diagram of Proposed System

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 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 foreground object[5][6][13][14], we advance part-based shape information induced by motion cues for characterizing the foreground object. 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. 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 {**h***n*, *n* = 1, 2, . . . , *N*} in a *p*-dimensional space, we construct an over-complete dictionary **D***p* ?*K* which includes *K* basis vectors, and we determine the corresponding sparse coefficient *αn* of each HOG descriptor. Therefore, the sparse coding problem can be formulated as

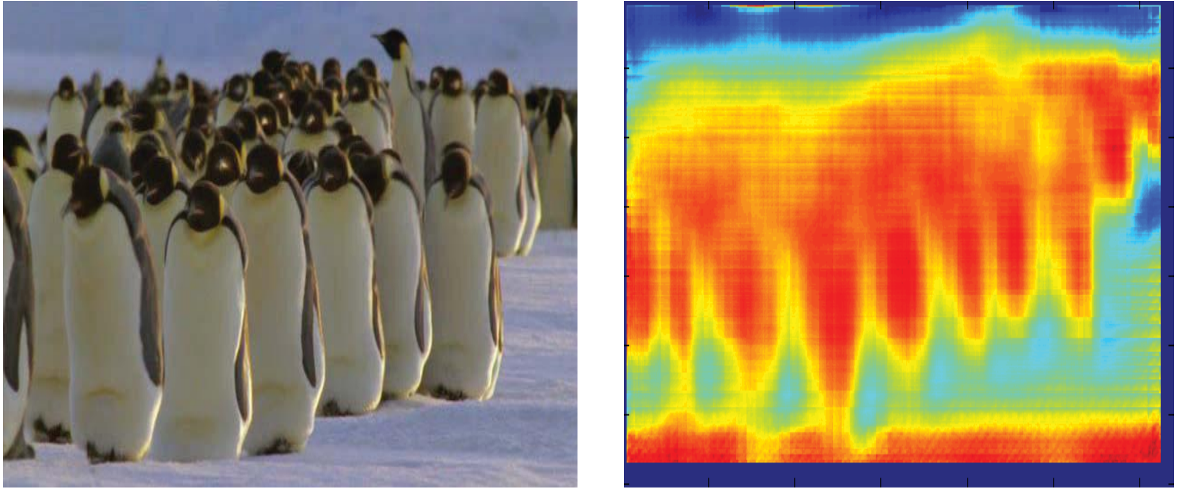
$$\min_{\mathbf{D},\boldsymbol{\alpha}} \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}||\mathbf{h}_n - \mathbf{D}\boldsymbol{\alpha}_n||_2^2 + \lambda||\boldsymbol{\alpha}_n||_1 \tag{4}$$

where *λ* balances the sparsity of *αn* and the *l*2-norm reconstruction error. 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 *αn* to obtain the map *Mk*. As a 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.

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 α, 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*^i^S *t* , which is calculated as follows:

$$\hat{X}_t^S = \sum_{n \in I_t} \sum_{k=1}^{K} (\alpha_{n,k} \cdot M_k) \qquad (5)$$

where α_n,*k* is the weight for the *n*th patch using the *k*th codeword. Fig. 4 shows an example of the reconstruction of a video frame using the motion-induced shape information of the foreground object.



(a) **(b)**

Fig 4. Shape likelihood reconstructed by sparse shape representation.

(a) Original frame. (b) Shape likelihood.

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 obtained from the previous step, and we threshold this likelihood by 0.5 to determine the candidate foreground (FS_{shape}) and background ($BS_{\text{shape)}}$ 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 expectation-maximization (EM) algorithm. Finally, we integrate both foreground and background color models with visual saliency and shape likelihood into a unified framework for VOE.

IV. CONDITION RANDOM FIELD FOR VOE

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 *zi* , while the hidden node *Fi* indicates its corresponding label (i.e. foreground or background). In this framework, the label *Fi* is calculated by the observation *zi* , while the spatial coherence between this output and neighboring observations *z j* and labels *Fj* are simultaneously taken into consideration. Therefore, predicting the label of an observation node is equivalent to maximizing the following posterior probability function

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

where *ψi* is the unary term which infers the likelihood of *Fi* with observation *zi* . *ψi, j* is the pairwise term describing the relationship between neighboring pixels *zi* and *z j*, and that between their predicted output labels *Fi* and *Fj* . 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

$$\begin{aligned} E &= -\log(p) \\ &= \sum_{i\in I}(\psi_i) + \sum_{\substack{i\in I \\ j\in\text{Neighbor}}}(\psi_{i,j}) \\ &= E_{\text{unary}} + E_{\text{pairwise}}. \end{aligned} \quad (7)$$

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

$$E^S = -w^s \log(\widehat{X}_t^S). \quad (8)$$

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

$$E^C = w^c(E^{C\mathcal{F}} - E^{C\mathcal{B}}). \quad (9)$$

Note that the foreground and background color GMM models $G^C f$ and $G^C b$ are utilized to derive the associated energy terms $E^C F$ and $E^C B$, which are calculated as

$$\begin{cases} E^{C\mathcal{F}} = -\log\left(\sum_{i\in I} G^{C_f}(i)\right) \\ E^{C\mathcal{B}} = -\log\left(\sum_{i\in I} G^{C_b}(i)\right). \end{cases}$$

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:

$$E^V = -w^v \log(\widehat{S}_t). \quad (10)$$

We note that in the above equations, parameters *ws* , *wc*, and *wv* 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 VOE. It is also worth noting that, Liu and Gleicher[13] only considers the construction of foreground color models for VOE. As verified by[14], it can be concluded that the disregard of background color models would limit the performance of VOE, since the only use of foreground color model might not be sufficient for distinguishing between foreground and background regions. In the proposed VOE framework, we now utilize multiple types of visual and motion salient features for VOE, and our experiments will confirm the effectiveness and robustness of our approach on a variety of real-world videos.

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 VOE. Since 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 VOE: When applying a pixel-level prediction process for VOE (like ours and some prior VOE methods do), the spatial structure of the extracted foreground region is typically not considered during the VOE 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 pair wise term in our CRF framework. The introduced pairwise term *Ei, j* is defined as

$$E_{i,j} = \sum_{\substack{i \in I \\ j \in \text{Neighbor}}} |F_i - F_j| \times \left(\lambda_1 + \lambda_2 \left(\exp\left(-\frac{\|z_i - z_j\|}{\beta}\right)\right)\right). \qquad (11)$$

Note that *β* is set as the averaged pixel color difference of all pairs of neighboring pixels. In (11), *λ*1 is a data-independent Ising prior to smoothen the predicted labels, and *λ*2 is to relax the tendency of smoothness if color observations *zi* and *z j* form an edge (i.e. when _*zi* − *z j* _ is large). This pair wise 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 VOE 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 ˆ$F_{\text{foreground}}$ and ˆ$F_{\text{background}}$) 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

$$\begin{cases} FS_{t+1} = FS_{\text{shape}}(t+1) \bigcup FS_{\text{shape}}(t) \bigcup \hat{F}_{\text{foreground}}(t) \\ BS_{t+1} = BS_{\text{shape}}(t+1) \bigcup BS_{\text{shape}}(t) \bigcup \hat{F}_{\text{background}}(t) \end{cases} \qquad (12)$$

where ˆ$F_{\text{foreground}}(t)$ indicates the pixels at frame *t* to be predicted as foreground, and $FS_{\text{shape}}(t)$ is the set of pixels whose shape likelihood is above 0.5 as described in Section III.B3. Similar remarks apply for ˆ$F_{\text{background(t)}}$ and $BS_{\text{shape(t)}}$.

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

$$\begin{aligned} E &= E_{\text{unary}} + E_{\text{pairwise}} \\ &= \left(E^{\mathcal{S}} + E^{\mathcal{CF}} - E^{\mathcal{CB}} + E^{\mathcal{V}}\right) + E_{i,j} \\ &= E^{\mathcal{S}} + E^{\mathcal{C}} + E^{\mathcal{V}} + E_{i,j}. \end{aligned} \qquad (13)$$

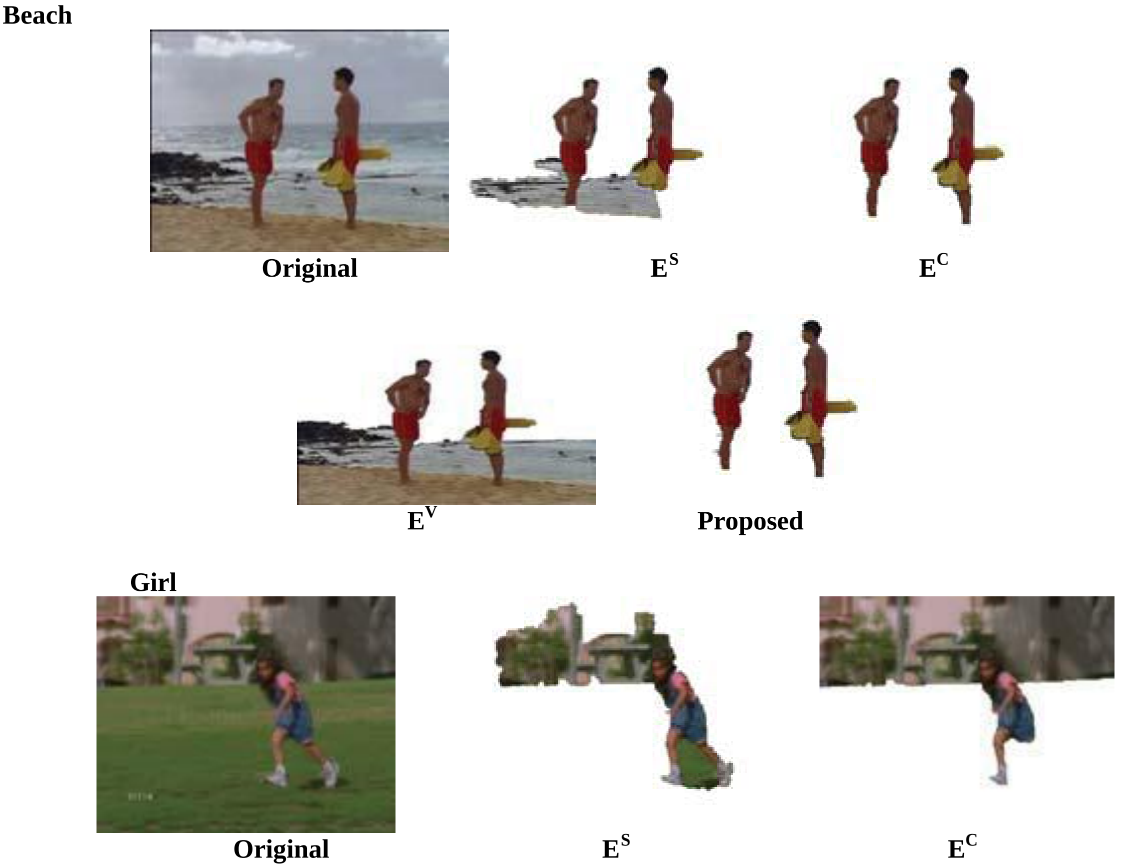
To minimize, one can apply graph-based energy minimization techniques such as max-flow/min-cut algorithms. When the optimization process is complete, the labeling function output *F* would indicate the class label (foreground or background) of each observed pixel at each frame, and thus the VOE problem 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 VOE, 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 VOE methods. Both qualitative and quantitative results will be presented to support the effectiveness and robustness of our proposed method.

A.Data Sets and Parameters: We consider eight different video sequences from three different datasets ([33]–[35]) in our experiments. Five out of the eight videos (Girl, Parachute, Penguin, Fox, Redbird) are selected from [33] and [34]. To quantitative evaluate the VOE performance, we use the ground truth provided with the original data (i.e., label information at the pixel level) except for the Penguin sequence. This is because that the ground truth information for the Penguin sequence (provided by [33]) is designed for object tracking in a weakly supervised setting, in which only one penguin is manually annotated by the original user at each frame. As mentioned in [36]), this might not be preferable since all the penguins should be considered as foreground objects. Therefore, we manually label the ground truth for that sequence. We also note that, videos in [35] do not contain any foreground or background information, and thus we also manually label their ground-truth information. It is worth noting that, both Penguin and Beach sequences are used to demonstrate that our proposed method is able to handle videos with multiple object instances (i.e., one type of foreground objects but multiple instances are presented). To learn the CRF model, we set $\lambda 1 : \lambda 2 \approx 1:5$ for the pairwise term. As for different energy unary terms, we have two sets of parameter: $wv = 2ws = 2wc = 0.5$ and $wv = ws = wc = 0.33$ for weighting visual saliency, shape, and color, respectively. We select the better results for our evaluation. To construct the foreground and background color models with GMM, we consider the number of Gaussian mixtures as 10 for both cases.

B. Integration of Multiple Motion-Induced Features for VOE: We first verify the effectiveness of fusing multiple types of features selected in our proposed framework. As shown in Fig. 1 and discussed in Section III, we consider visual saliency together with two motion-induced cues (i.e., shape and color) in a unified CRF model for predicting the label information of each pixel. To confirm that it is necessary to combine the features considered, Fig. 5 shows example video frames of three videos (Beach, Girl, Penguin) and their VOE results using single or multiple types of features (i.e., shape, color, and visual saliency). We note that, for the VOE results shown in Fig. 5 using a single type of feature, both pair wise and temporal terms are enforced for the corresponding CRF models. In other words, the only difference between those results and that of ours is the use of one or multiple unary terms describing the associated features.



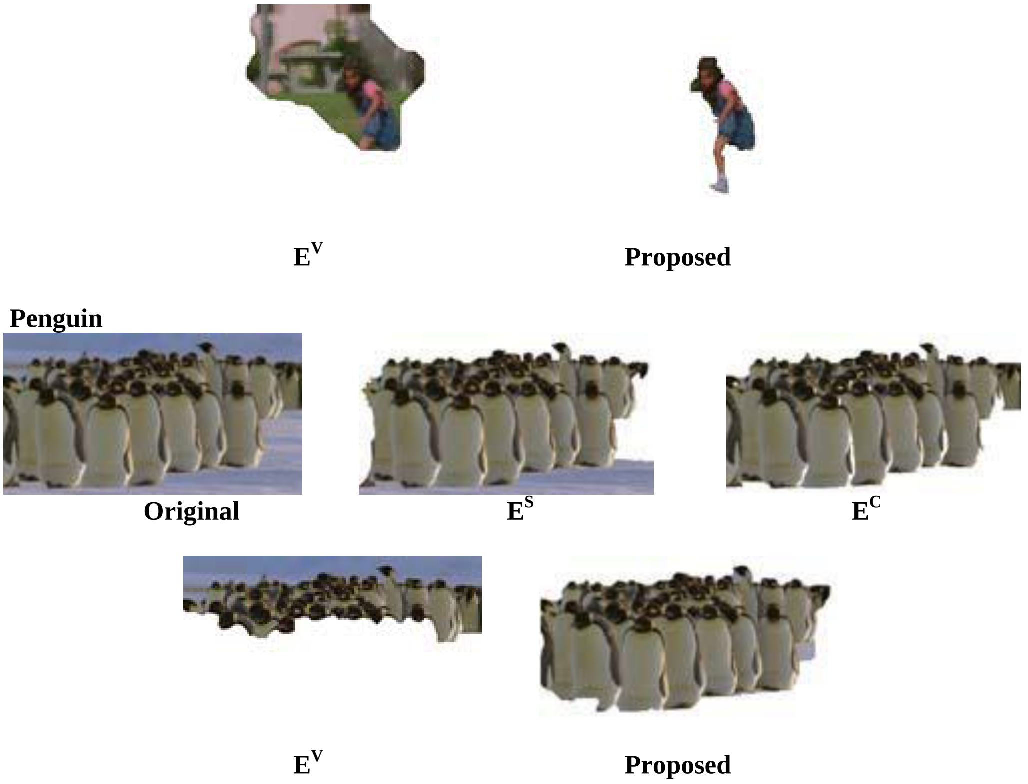
Beach

Original **E^S** **E^C**

E^V **Proposed**

Girl

Original **E^S** **E^C**



E^V **Proposed**

Penguin

Original E^S E^C

E^V **Proposed**

Fig. 5. VOE results using different feature cues (the CRF pair wise term is considered for all cases for fair comparisons).

VOE results using only shape, color, and visual saliency are shown in Fig. 8(a)–(c), respectively, while those produced by our approach are shown in Fig. 5(Proposed). For the Beach video (first row in Fig. 5), since background motion due to sea waves is easily detected by optical flow, plus the high visual contrast between the seashore and foreground objects, the corresponding features are not sufficient to discriminate between the foreground and background regions. Although the motion-induced foreground and background color cues share a portion of the sea (background), our definition of color energy term in (9) is able to disregard the associated

Common Gaussian components. As a result, only the use of color cues could produce satisfactory results. For the Girl video shown in the second row of Fig. 5, both foreground and background exhibit remarkable motion, while the visual contrast between them is not significant. As a result, the use of any single type of feature was not able to achieve proper segmentation results. Finally, for the Penguin video shown in the last row in Fig. 8, the use of visual saliency was not able to identify the body parts of the penguins, while shape and color cues extracted foreground objects with missing parts. Nevertheless, our proposed framework integrating all three types of features can be observed to achieve the most satisfactory VOE results for all three videos, as shown in the last column of Fig. 5.

C. Comparisons With Saliency-Based Approaches: We now compare our method with state-of-the-art visual saliency detection approaches [20]–[23], [27]. In particular, we consider CA (context-aware) proposed by Goferman *et al*. [22], LD (learning to detect) proposed by Liu *et al*. [20], ST (spatio-temporal cues) of Zhai and Shah [23], MSSS (maximum symmetric surround saliency) of Achanta and Süsstrunk [21], HC (histogram-based contrast) and RC (region-based contrast) proposed by Cheng *et al*. [27]. From the visual saliency results shown in Fig. 6, it can be observed that our approach was able to extract visual salient regions, even for videos with low visual contrast (e.g., Girl and Penguin). Later we will verify the use of our derived visual saliency along with motion-induced cues would produce promising VOE results.

COMPARISONS OF MAXIMUM F-MEASURE SCORES FOR DIFFERENT VISUAL SALIENCY DETECTION APPROACHES

Methods	CA	LD	ST	MSSS	HC	RC	PROPOSED METHOD
F-measure	0.9100%	0.8667%	0.6222%	0.7839%	0.7032%	0.8067%	0.8617%

Table I

In order to quantitatively compare the above results, we subsample the number of video frames for each sequence by a factor of 10 and perform quantitative evaluation. While one could use precision-recall curves to evaluate the performance of each method, the goal of this work is to utilize the retrieved visual saliency information for VOE. Therefore, we choose to provide the maximum F-measure scores (i.e., 2 · *(*Precision · Recall*/*Precision + Recall*)*) produced by different methods, as listed in Table I.

From the results shown in Table I, we see that our approach did not produce the highest F-measure scores in terms of visual saliency detection, since both CA [22] and LD [20] performed slightly better than ours. As pointed out in [27], both CA and LD tend to produce higher saliency values along object edges due to the use of local contrast information. However, if the scale of the foreground object is large, high visual saliency along object edges will not be able to provide sufficient visual cues for VOE. In [27], RC has been shown to exhibit better capabilities than CA/LD on benchmark datasets for visual saliency detection. Moreover, both CA and LD are computationally more expensive than RC and ours due to the use of multi-scale patches for feature extraction/selection. Based on the above quantitative and qualitative evaluations, the use of our proposed visual saliency detection algorithm for VOE can be verified.

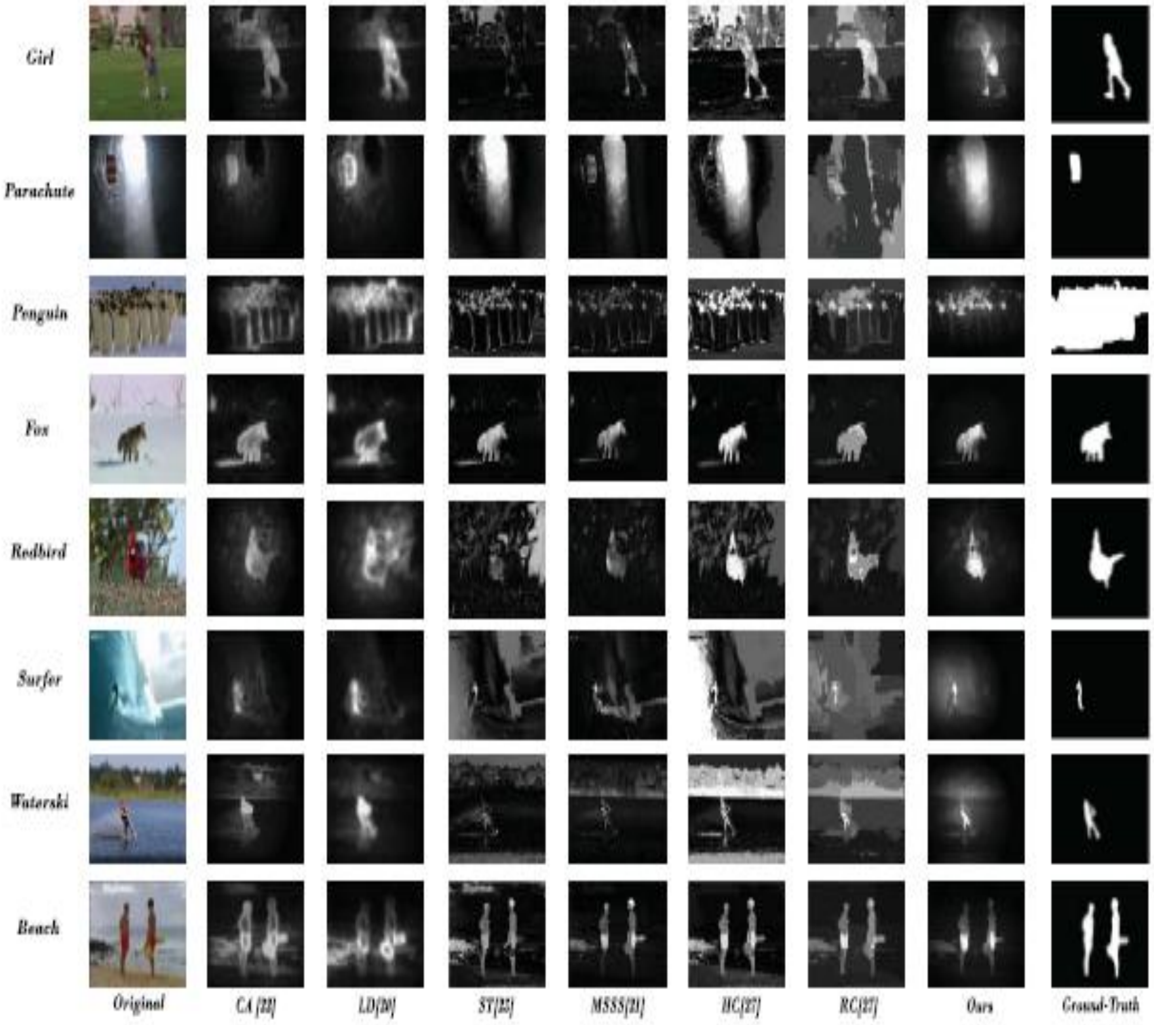


Fig. 6. Selected video frames and their visual saliency results produced by different methods.

D. Comparison With Unsupervised VOE Methods: Since our proposed VOE method is able to automatically extract foreground objects of interests without any prior knowledge or the need to collect training data in advance, we compare our results with those produced by three state-of-the-art unsupervised VOE approaches. We first consider the approach of proposed in [13], which also applies CRF to combine color and locality features for VOE. However, no background color model and temporal consistency is considered in their proposed framework. Since our saliency detection stage is inspired by [27], it is necessary for us to consider the approach of [27], followed by performing saliency cut (i.e., an iterative GrabCut technique [27]) to segment the detected visually salient regions as foreground objects. We also compare our method to a recently proposed unsupervised VOE approach of [36]. Based on the image segmentation and object ranking results of [37], the approach of [36] aims at automatically discovering the key image segments across video frames as foreground objects using multiple appearance and motion cues. The code for [27] and [36] is available at the websites of the authors.

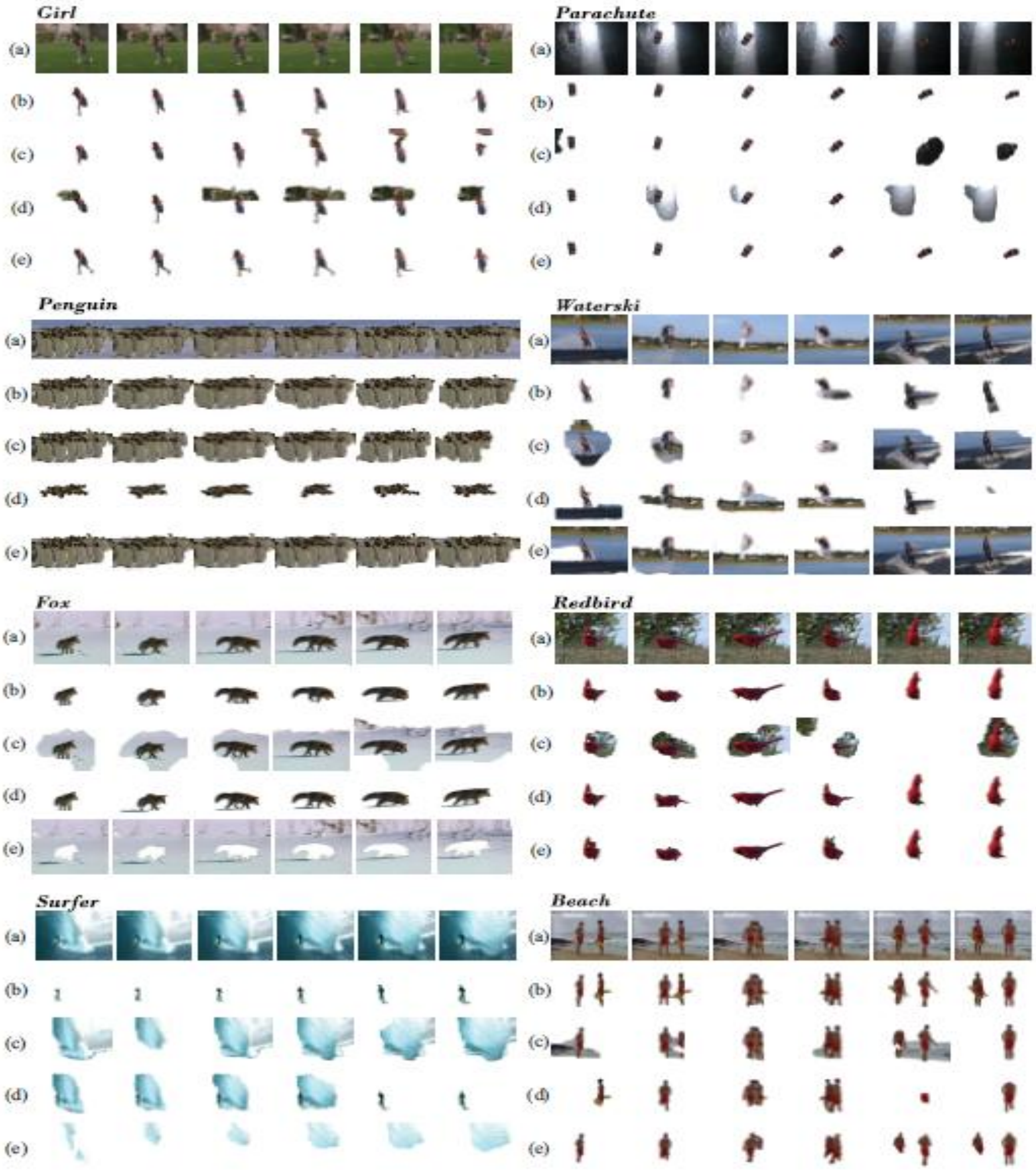


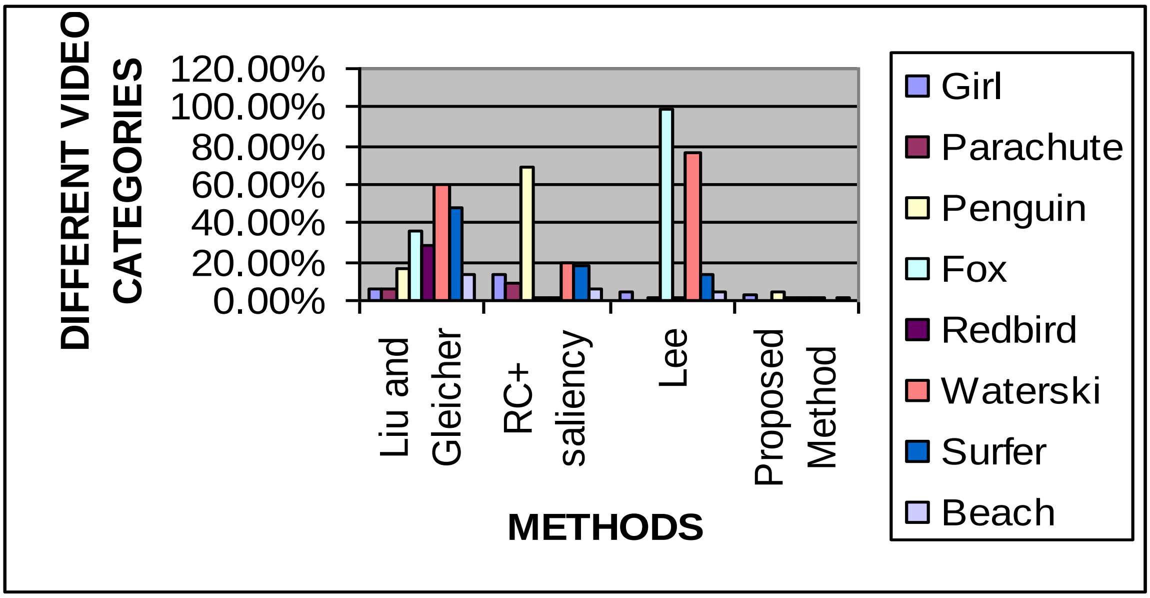
Fig. 7. Example VOE results of different unsupervised approaches. (a) Original video frames. (b) Our method. (c) Liu and Gleicher [13]. (d) RC+saliency cut [27]. (e) Lee *et al.* [36].

To quantitatively evaluate the VOE performance, we consider the use of mis-segmentation rates€*(S)* = |**XOR***(S, GT))*|*/F* · *P*, where *S* is the VOE output image, *GT* is the ground-truth, *F* is the total number of video frames, and *P* is the total number of pixels in each frame. Table II lists the mis-segmentation rates of different videos for all approaches considered. From Table II and Fig. 10, it can be seen that we achieved significantly better or comparable VOE results on most of the video sequences. We also verify that our proposed method is able to handle videos captured by freely moving camera (e.g., Girl), or with complex background motion (e.g., Waterski and Surfer). We also produce satisfactory results on videos with low visual contrast (e.g., Parachute), and those with articulated foreground objects presented (e.g., Beach).

TABLE II

Methods	**Girl**	**Parachute**	**Penguin**	**Fox**	**Redbird**	**Waterski**	**Surfer**	**Beach**	**Avg**
Liu and Gleicher	6.31%	5.36%	17.03%	35.32%	28.97%	59.33%	47.5%	14.21%	**26.75%**
RC+ saliency cut	13.42%	9.05%	68.28%	1.98%	1.73%	18.86%	17.82%	6.64%	**17.22%**
Lee	3.83%	0.13%	1.66%	99.01%	1.19%	75.90%	13.36%	5.14%	**25.03%**
Proposed Method	**2.30%**	**0.15%**	**5.01%**	**2.22%**	**1.68%**	**2.24%**	**0.38%**	**1.59%**	**1.95%**

Comparisons of mis-segmentation rates of different unsupervised voe methods.



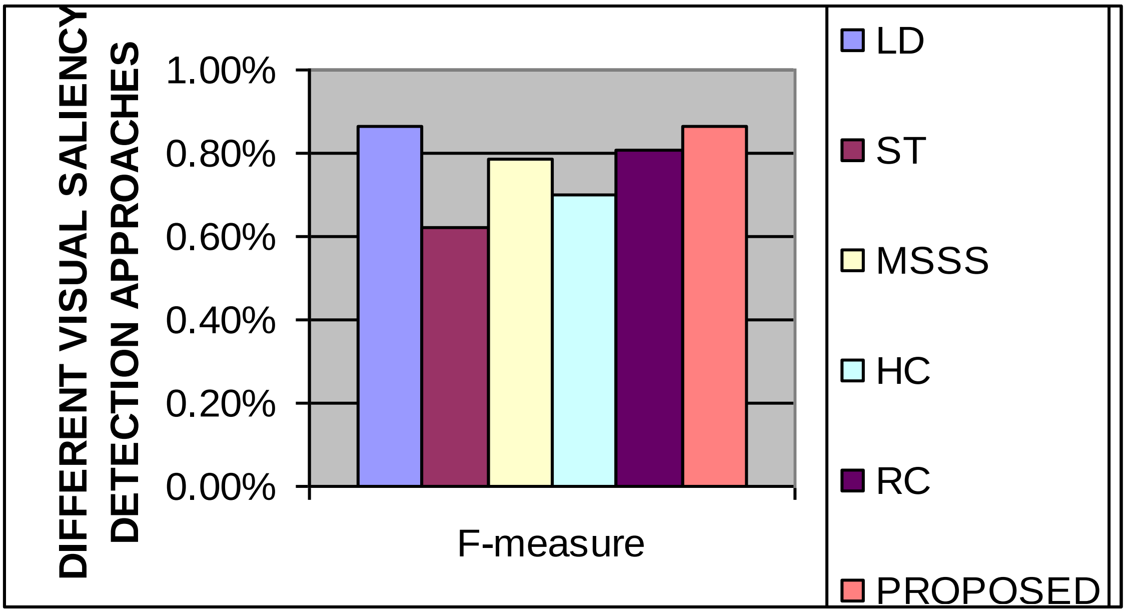
Comparisons of mis-segmentation rates of different unsupervised voe methods.

We note that the method of [13] constructs the foreground color model from video frames with dominant foreground motion detection results, and thus only prefers scenarios where the foreground object exhibits significant motion. Since no background color model is considered in [13], plus the background clutter might contribute to motion salient regions, the lack of discrimination ability between candidate foreground and background regions makes the method of [13] difficult to achieve satisfactory VOE results. As for the visual-saliency based method of [27], it would fail to detect the foreground object which is not visually salient within a video frame. We observe that the method of [36] tends to treat foreground as one single object and thus restricts the generalization for cases like Beach. This is because the use of objectness for ranking their image segmentation results for VOE. We note that this approach had very high mis-segmentation rates for the sequence Fox, since it detected the background region as the foreground; as for the sequence Waterski, the VOE results of [36] were not as good as those reported in [36] even we direct applied their release code. When comparing the averaged mis-segmentation rates in Table II, we also list the result without using these two sequences. Besides presenting quantitative VOE results, we also provide qualitative results and comparisons in Fig. 7, and it can be seen that our approach generally produced satisfactory results. For the video like Waterski which contains visual and motion salient regions for both the foreground object (i.e., water-skier) and background clutter (e.g., surf), and it will be very difficult for unsupervised VOE methods to properly detect the foreground regions even multiple types of visual and motion induced features are considered. Discrimination between such challenging foreground and background regions might require one to observe both visual and motion cues over a longer period. Or, if the video is with sufficient resolution, one can consider to utilize trajectory information of the extracted local interest points for determining the candidate foreground regions. In such cases, one can expect improved VOE results.

We finally comment on the computation time of our proposed method. When applying our approach for a video frame with 320 240 pixels (implemented by MATLAB), it takes about 5 s, 1 min, and 20 s for computing optical flow, visual/motion saliency, and deriving the shape likelihood, respectively. About another 1 s is required for inducing the foreground/background color GMM models and predicting pixel labels using CRF. While it is possible to accelerate the implementation by C/C++ for most of the above procedures, calculation of optical flow is still computationally expensive even using GPU. Since the goal of this paper is to automatically extract the foreground objects without using training data or user interaction, real-time processing will be among future research directions. All unsupervised VOE approaches considered in this paper (including ours) are performed offline.

Maximum F-Measure Scores For Different Visual Saliency Detection Approaches

Methods	LD	ST	MSSS	HC	RC	PROPOSED METHOD
F-measure	0.8667%	0.6222%	0.7839%	0.7032%	0.8067%	0.8617%



Maximum F-Measure Scores For Different Visual Saliency Detection Approaches

VI. CONCLUSION

In this paper, we proposed an automatic VOE 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 VOE. Compared with state-of-the-art unsupervised VOE 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.

REFERENCES

[1]. 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.

[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]. F.-C. Cheng, S.-C. Huang, and S.-J. Ruan, “Advanced background subtraction approach using Laplacian distribution model,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2010, pp. 754–759.

[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 singleconcept 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. San Mateo, 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.