INFORMATICA, 2019, Vol. 30, No. 1, 53–72

DOI: http://dx.doi.org/10.15388/Informatica.2019.197

Video Saliency Detection Using Motion Distinctiveness and Uniform Contrast Measure

Rahma KALBOUSSI*, Mehrez ABDELLAOUI, Ali DOUIK
Networked Object Control and Communication Systems Laboratory,
University of Sousse, Tunisia
e-mail: rahma.kalboussi@gmail.com

Received: May 2018; accepted: October 2018

Abstract. Saliency detection has been deeply studied in the last few years and the number of the designed computational models is increasing. Starting from the assumption that spatial and temporal information of an input video frame can provide better saliency results than using each information alone, we propose a spatio-temporal saliency model for detecting salient objects in videos. First, spatial saliency is measured at patch-level by fusing local contrasts with spatial priors to label each patch as a foreground or a background one. Then, the newly proposed motion distinctiveness feature and gradient flow field measure are used to obtain the temporal saliency maps. Finally, spatial and temporal saliency maps are fused together into one final saliency map.

On the challenging SegTrack v2 and Fukuchi benchmark datasets we significantly outperform the state-of-the-art methods.

Key words: saliency detection, motion estimation, object of interest, optical flow, contrast measure.

1. Introduction

The human visual system is able to select the visually most important regions in its visual field. Such cognitive process allows humans to interpret complex scenes in a short and real time lapse with no need to training. Visual saliency detection is originally a problem to predict where the observer may fixate (Borji *et al.*, 2015). Then, it has been extended to detect the object that attracts his gaze.

While visual saliency is highly related to human visual perception and processing, it is studied by different researches in various fields including neuro-biology (Mannan *et al.*, 2009), computer vision (Borji *et al.*, 2015) and cognitive psychology (Wolf, 2004). And it was used in different vision applications like object of interest detection (Donoser *et al.*, 2009), object recognition (Gu *et al.*, 2015), image compression (Christopoulos *et al.*, 2000), image editing based on content aware (Zhang *et al.*, 2009; Cheng *et al.*, 2010), image retrieval (Chen *et al.*, 2009), etc.

Classic segmentation problems aim to partition the input into coherent regions while salient object detection approaches aim to segment the object of interest from its surroundings.

*Corresponding author.

However, detecting the salient object automatically, accurately and efficiently is very desired if we consider the high and immediate ability to grant computational resources for image processing, extract the objects features, isolate it from the background and produce the final salient object. In recent years, image saliency detection has achieved good results and the number of computational models is quite big compared to video saliency which is a quite new topic and a promising research field.

Image saliency detection covers only the spatial domain, while video saliency includes spatial and temporal information which is incorporated by the video motion information.

Actually, exploiting and using the spatial and temporal information into a video saliency framework has become a research trend in the field of video saliency detection.

The saliency of a given input is the most visible content that is able to define the human attention, called saliency map. Saliency map computation is a usually bottom-up process issued from a surprising or distinctive visual stimuli and is often assigned to brutal change in image features such as edges, boundaries, colour or gradient (Borji and Itti, 2013). The first visual saliency models were devoted to image saliency and can be grouped into two groups, namely, local and global saliency approaches. Local approaches measure rarity of a region over its neighbourhoods (Itti *et al.*, 1998; Harel *et al.*, 2006). In contrast, global approaches are based on the rarity and uniqueness of image regions with respect to the whole scene (Cheng *et al.*, 2015; Kim *et al.*, 2014). Mao *et al.* (2015) propose a saliency method inspired from the human visual system that combines local and global saliency features with high-level features (prior-knowledge, object detection) and structure saliency to highlight the object that attracts human gaze. For each saliency map, features are extracted using Adaptive-Subspace Self-Organizing Map for image retrieval. Gu *et al.* (2015) present an application of the visual saliency detection which aims to recognize the object of interest. Since the biological visual system naturally tends to focus on the region that contains the most informative object, saliency detection is used as a robust object detector. Then, features of the region that contains the object of interest are extracted using the dense Scale Invariant Feature Transform. A linear support vector machine classifier is used to define the object class. Here we only review the main papers on video saliency, for an excellent review of saliency methods in still images we refer to Borji *et al.* (2014, 2015). While saliency detection in still images has been intensively treated, spatiotemporal saliency detection is a new problem. Motion cues are a crucial foreground indicator in a video saliency detection framework; however, some background motions can blur the location of a salient object.

Only few methods adress the video saliency problem (Itti and Baldi, 2005; Zhong *et al.*, 2013; Gao *et al.*, 2008; Wang *et al.*, 2015; Mauthner *et al.*, 2015) and most of them make use of an image saliency method and simply add a motion feature as a saliency clue. Therefore, in this paper we propose a simple and effective framework that detects salient objects in videos based on spatio-temporal saliency estimation. First, for a robust saliency estimation, we use the change of contrast to indicate the main object locations. To do so, we propose a uniform contrast measure which makes use of traditional contrast features (local contrast and contrast consistency) and our novel contrast cue named spatial consistency (see Section 4.2). Spatial saliency is designed as a growing process by propagating the

influence of the proposed local uniform contrast measure in the foreground-background patch assignment.

Then, for more accuracy, temporal saliency estimation is derived where we use the inter-frame temporal coherence incorporated into our motion distinctiveness feature (Section 5.1) and the intra-frame motion information presented as our four-sided motion estimator (Section 5.2). Finally, spatial and temporal saliency maps are fused into one final saliency map. The main steps of our proposed approach are introduced in Algorithm 1.

For the evaluation we use two standard benchmark data sets for video saliency, namely, the SegTrack v2 dataset proposed by Li *et al.* (2013) and Fukuchi dataset of Fukuchi *et al.* (2009).

The rest of the paper is organized as follows. In Section 2, we discuss related works. In Section 3, we present an overview of our proposed model. In Section 4, we detail our uniform contrast measure for spatial saliency estimation. In Section 5, we introduce our temporal motion estimation. Then, the final saliency map fusion is presented in Section 6. Experiments and results will be discussed in Section 7. Finally, conclusions will be provided in Section 8.

2. Related Work

Video saliency detection aims to identify the object that catches our attention from video sequences.To the best of our knowledge, the number of methods designed to address this problem is reduced. When an observer watches a video, he does not have enough time to examine the whole scene, so his gaze is always directed towards the moving object. For this reason, motion is the most important cue for detecting salient objects in videos which makes a deep exploration of the inter-frame information more crucial than ever.

Recently, different spatio-temporal saliency models have been proposed using different methods and theories such as the theory of information, control theory, the frequency domain analysis, machine learning and low rank decomposition.

Information theory based spatio-temporal saliency models use the video frames self information, the conditional entropy and the various formulations of the incremental coding length as saliency indicators.

Spatio-temporal saliency models based on the control theory represent first the video sequence with the linear systems state space model, then exploit the controllability or the observability of the linear system to discriminate the salient object and the background motion and produce the exact saliency measure (Gopalakrishnan *et al.*, 2012).

Spatio-temporal saliency methods based on Frequency domain analysis generate the master saliency map using the Quatrain Fourrier Transform (QFT) phase spectrum over the feature space which contains the luminance, two chrominance components and the frame difference, and the fourrier transform amplitude spectral over the time slices at vertical and horizontal directions (Cui *et al.*, 2009).

Machine learning methods use training data to build models and testing data to predict saliency map of an input video frame. Machine learning methods like probabilistic learning, support vector regression with or without Gaussian kernels are widely used (Rudoy *et al.*, 2013).

The low rank decomposition methods decompose the matrix of the temporal slices into a low rank matrix that characterizes the background and a sparse matrix for salient objects (Xue *et al.*, 2012).

In the literature, since there is a gap between the spatial and temporal domains, a lot of spatio-temporal duels measure the spatial and temporal saliency apart, then combine them using either a linear or a non-linear fusion schemes to provide the final spatio-temporal saliency map. Gao *et al.* (2008) proposed a spatio-temporal model issued from an image saliency model by adding a motion channel to characterize the temporal feature. Also, Mahadevan and Vasconcelos (2010) used an image saliency model to model spatio-temporal saliency using dynamic texture. Rahtu *et al.*'s (2010) saliency model is suitable for videos and still images. It combines a conditional random field model with saliency measures formulated using local features and statistical framework. Using the centre-surround and colour orientation histograms as spatial cues and temporal gradient differences as temporal cues, Kim *et al.* (2011) proposed spatio-temporal saliency models.

In Luo and Tian (2012), using the temporal consistency and the entropy gain, a video saliecy detection model is proposed.

Wang *et al.* (2015) proposed a spatio-temporal framework where first they determine the spatial edges of the input frame and the optical flow (used to highlight the dynamic object), then mix both spatial and temporal information to produce the exact salient object. Saliency scores are assigned using the geodesic distance. Later, the Gestalt principle of figure-ground segregation for appearance and motion cues is used by Mauthner *et al.* (2015) to predict video saliency. A spatio-temporal saliency framework using colour dissimilarity, motion difference, objectness measure, and boundary score feature is proposed by Singh *et al.* (2015) to determine each saliency score of every superpixel in the input frame.

Shanableh (2016) proposed a video saliency method that uses intra- and inter-frame distances for the saliency maps computation. Hamel *et al.* (2016) integrated the colour information into the bottom-up saliency detection model. An effective attention model (Annum *et al.*, 2018) based on texture smoothing and contrast enhancement applications is introduced for improving saliency maps in complex scenes. Bhattacharya *et al.* (2017) investigated the video decomposition model to extract the motion salient information from a video. Imamoglu *et al.* (2017) developed a multi-model saliency detection fusing salient features through both top-down and bottom-up salient cues. Given consistency of spatio-temporal saliency maps, video saliency research is still an emerging hard issue to be more investigated.

Although the aforementioned approaches process the input video in a frame by frame basis, they ignore that a prefect saliency map should be spatio-temporally coherent. It is obvious that video saliency detection is a challenging research problem to further be investigated.

3. Proposed Model

In this section, we propose an overview of our spatio-temporal saliency framework.

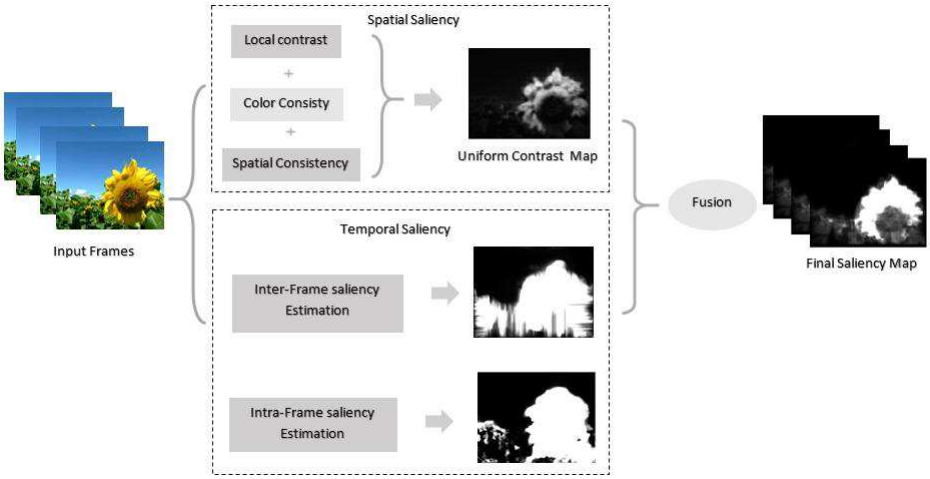

Fig. 1. Framework of the proposed method from left to right: in-out video frames, spatial and temporal saliency maps, final saliency map.

Unlike state-of-the-art methods, this work produces an accurate spatio-temporal saliency maps where the object of interest is perfectly highlighted and segmented from the background. Our framework has three main steps: spatial saliency estimation, temporal saliency estimation and final spatio-temporal map generation.

Our model takes a $n \times m \times t$ video frame F and produces a saliency map S, where each pixel x has a saliency score $S(x) = l$ and the higher this score is, the more salient this pixel is. Spatial saliency map is measured at patch level by the newly proposed uniform contrast measure which combines spatial priors with traditional contrast measure. Patches with higher uniform contrast measure are considered as salient foreground. Then, we measure the temporal saliency using our inter-frame and intra-frame motion estimators: Along a video sequence, for each frame, pixels with distinctive motion attract human gaze. Inter-frame motion estimation is performed using our motion distinctiveness measure which will highlight the object of interest that has a distinctive activity when moving from one frame to another. Intra-frame motion information is measured using our four-sided motion estimator where for each frame, we compute the magnitude gradient, then a gradient flow field is derived from the cumulative sum of the magnitude gradient through four sides of the frame. Motion information alone is insufficient to identify the object of interest in case of complex scenarios like a moving object with small optical flow or dynamic background. The use of our spatial saliency map have improved the results (see Fig. 1).

4. Spatial Saliency Estimation

4.1. *Contrast Measure*

Salient region is usually distinctive from the rest of the scene. Previous works on saliency detection have proved that a change in contrast is the main cue to highlight the object of

interest (Cheng *et al.*, 2015). Given a frame divided into patches, local contrast is defined as the brutal change of colour independently of the spatial distance between considered patches. While salient patches are generally spatially grouped, spatial distance is considered for a good local contrast representation. In this context, surrounding contrast cue is presented by Goferman *et al.* (2012) which assumes that not only colour distinctiveness is necessary for saliency detection but also the surrounded patches characteristics. To do so, local contrast distinctiveness for each patch P_i is defined as follows

$$LC(p_i) = \sum_{p_j \forall j} \frac{d_c(p_i, p_j)}{1 + \alpha . d_p(p_i, p_j)}, \tag{1}$$

where α is a control parameter of the colour by spatial distance rate, $d_c(p_i, p_j)$ is the Euclidean distance between p_i and p_j in the CIE L∗a∗b colour space and $d_p(p_i, p_j)$ is the Euclidean distance between p_i and p_j positions.

4.2. *Uniform Contrast Measure*

Local Contrast *LC* can be considered as one of the strongest object boundaries detection in colour images. This detector is not adequate for saliency estimation, which needs to detect the whole object, because it only highlights the object's of interest boundaries. This is due to the uniformity of the patches' colour characteristics inside an object (Yeh *et al.*, 2014). Therefore, we use a local-contrast weighted sum over a q neighbouring patches that have a similar colour as the current one. Contrast uniformity may depend also on the spatial distance which separate the current patch from its neighbouring one where our uniform contrast measure is defined as follows

$$UC(p_i) = \sum_{p_j \in N_q} LC(p_j) \times CC(p_i, p_j) \times SC(p_i, p_j). \tag{2}$$

$CC(p_i, p_j)$ measures the colour consistency between the current patch p_i and patch p_j and is defined as follows

$$CC(p_i, p_j) = \frac{\exp(-d_c(p_i, p_j)}{\sum_{p_h \in N_q} \exp(-d_c(p_i, p_h)}, \tag{3}$$

p_i and p_j are given patches, to figure out if those two patches belong to the same object or not, we compute a feature which compares the colour distance of p_i and p_j over the sum of colour distances between the current patch p_i and its immediate neighbours (p_h is a patch which belongs to the set of neighbouring patches of p_i). As more p_i and p_j have similar colour contrast, *CC* will assign them high scores indicating that they belong to foreground. This feature was mainly proposed by Yeh *et al.* (2014) and we noticed that in case of similar colour contrast between the object and the background, contrast consistency *CC* will fail to highlight the exact object of interest. For that, we propose a

spatial consistency feature $SC(p_i, p_j)$ which measures the spatial consistency between the current patch p_i and patch p_j and can be defined as follows

$$SC(p_i, p_j) = \frac{\exp(-d_p(p_i, p_j))}{\sum_{p_r \in N_p} \exp(-d_c(p_i, p_r))}. \tag{4}$$

The spatial consistency SC computes the Euclidean distance between two patches p_i and p_j over the sum of colour distance between the patch p_i and N_p the set of neighbouring patches of p_j. Spatial consistency will serve to rectify the problem of object/background colour similarity.

Experiments showed that the bigger is the increase of the number of neighbours q, the more the patches inside the object are highlighted. But a large number of q neighbours will cause a higher foreground/background smoothness. After several tests, we fix $q = 8$ and $r = 8$ in our experiments.

4.3. *Static Saliency Map Generation*

In general, patches with higher contrast values arouse the attention. Thus, the proposed uniform contrast measure in equation (2) will be used to select foreground and background patches. The first thing to do is to sort the UC values in the ascending order, then patches are ranked according to their UC degree, where patches with high UC degree are marked as foreground patches and patches with lower UC degree are marked as background patches. More precisely, the P_f are the first 10% patches and the P_b are the last 70% patches.

Given foreground and background patches sets (F, B), we can define foreground and background probabilities of a given patch as a superimposed mixture distribution

$$Pr(P|F) = \frac{UC(P)}{|F|} \sum_{Y \in F} \exp\left(-\frac{d_c(P, Y)}{\sigma_c}\right) \exp\left(-\frac{d_p(P, Y)}{\sigma_p}\right) \tag{5}$$

and

$$Pr(P|B) = \frac{1 - UC(P)}{|B|} \sum_{X \in B} \exp\left(-\frac{d_c(P, X)}{\sigma_c}\right) \exp\left(-\frac{d_p(P, X)}{\sigma_p}\right). \tag{6}$$

Foreground and background probabilities of a given patch P_i depend on the distance in the space and colour domains regarding the other patches of the whole frame and on the uniform contrast measure. The final static saliency map will be refined according to the following equation

$$S(P) = \frac{Pr(P|F)}{Pr(P|F) + Pr(P|B)}. \tag{7}$$

5. Video Saliency Estimation

5.1. *Motion Distinctiveness*

In this paper we define a salient region as a region that has a distinctive motion compared to the previous frame. Therefore, we define a new metric to quantify the motion distinctiveness at pixel level.

Here the concept of motion distinctiveness is defined as a region that has a low motion commonality compared to the previous.

The pixel-level motion vectors mv are calculated for each frame using the optical flow estimation method proposed by Brox and Malik (2011). Motion vector provides the information about the motion activity of objects in the current frame. A uniform motion vector at frame-level when moving from frame f_t to frame f_{t+1} shows that there is no new motion activity in the frame, while a fluctuating motion denotes that it contains a new or distinctive moving object. In general, the newly appearing moving object catches attention.

Therefore, for every video frame, we exploit correlation measure between current and previous frames. The sum of squared difference (SSD) is used to measure the similarity between a pair of frames in different previous works (Shi and Malik, 2000). However, after testing different similarity measures, we have found that Pearson correlation measure is more adequate to compute the motion distinctiveness.

Unlike the Euclidean distance score which is scaled to vary between 0 and 1, Pearson correlation measure $\rho_i(f_t, f_{t-1})$ is scaled between 1 and -1 given by equation (8)

$$\rho_i(f_t, f_{t-1}) = \frac{cov_{mv_t, mv_{t-1}}}{\sigma_{mv_t} \times \sigma_{mv_{t-1}}}, \tag{8}$$

where mv_t is the motion vector at frame f at instant t, and $cov_{mv_t, mv_{t-1}}$ is given by

$$cov_{mv_t, mv_{t-1}} = \frac{\sum_{i=1}^{N} (mv_t - \overline{mv_t})(mv_{t-1} - \overline{mv_{t-1}})}{N-1}, \tag{9}$$

and σ_{p_i} is defined as

$$\sigma_{mv_t} = \sqrt{\frac{\sum_{i=1}^{N} (mv_t - \overline{mv_t})^2}{N-1}}. \tag{10}$$

The Pearson measure indicates how two variables are correlated and is varied from -1 to 1, where a value of 1 indicates that both patches are similar and a score of -1 indicates that the two patches are not correlated and are totally distinctive. Since we are interested in measuring the motion distinctiveness score, we propose a new metric defined as follows

$$M_d^t = \frac{1}{\alpha} \exp\left(- \frac{\rho_i(f_t, f_{t-1}) - 1}{2} \right), \tag{11}$$

where α is a parameter equal to 0.5. M_d is the motion distinctiveness measured at frame-level and is used as an inter-frame saliency indicator.

5.2. *Four-Sided Motion Estimation*

While contrast measure is a good saliency indicator in still images, it can not be discriminated in a complex scene with a high-textured background. In the last section, we introduced a new measure to quantify the inter-frame distinctive motion. To ensure the motion consistency, we have found that integrating intra-frame motion information into the same framework can be more effective.

To compute the optical flow, we use the large displacement motion estimation algorithm (Brox and Malik, 2011) Given a video frame f_i, let v_i be its optical flow field. We propose a temporal gradient field which uses an exponential function to highlight the optical flow gradient magnitude $\|\nabla v_i\|$ and eliminates noise

$$M_i = 1 - \exp\big(-\lambda\|\nabla v_i\|\big), \tag{12}$$

λ is used to scale the exponential function and is set to the value 1.

The temporal gradient magnitude reveals the boundaries of the moving object. Logically, a video frame is crossed by many flows where some of them start from the right side to the left side (or inversely), and some others from top side to down side (or inversely). In general, when a flow crosses a frame, its value increases with the value of the corresponding temporal magnitude gradient field.

Given those two assumptions, we define a four-sided gradient flow field estimator. Let a frame f_i be $L \times H$, we first define left-to-right gradient as the cumulative set of pixels in the same row starting from the left direction as

$$G^l = \sum_{j=1}^{H}\sum_{i=1}^{L} M_q(i, j). \tag{13}$$

Then, we define right-to-left gradient as the cumulative set of pixels in the row starting from the right direction to the left as

$$G^r = \sum_{j=1}^{L}\sum_{i=1}^{H} M_q(i, y + 1 - j), \tag{14}$$

G^r and G^l estimate the gradient flow in the horizontal direction, it will be useful to estimate the gradient flow in the vertical direction, we define top to down gradient flow as

$$G^t = \sum_{i=1}^{H}\sum_{j=1}^{L} M_q(i, j). \tag{15}$$

And we define down to top gradient flow as

$$G^d = \sum_{i=1}^{L}\sum_{j=1}^{H} M_q(x + 1 - i, j). \tag{16}$$

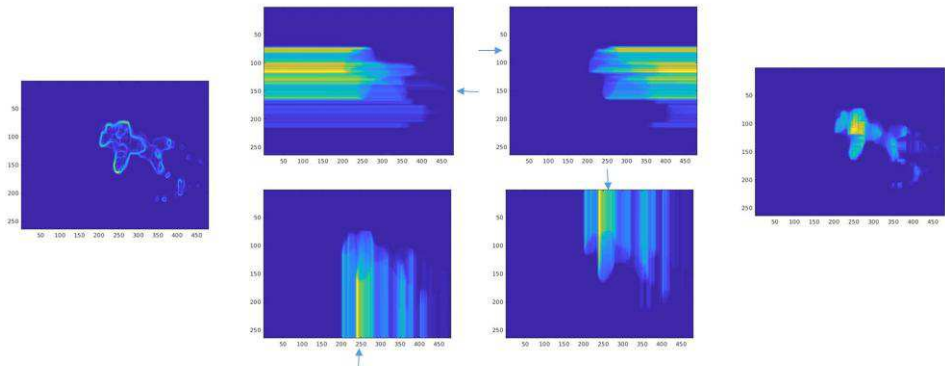

Fig. 2. Four-sided motion estimation, from left to right: temporal gradient magnitude, four-sided estimation, estimated motion map.

Given the aforementioned four-sided gradient flow estimation, and starting from the assumption that a flow field increases when it crosses a video frame, we define a gradient flow field by minimizing the overlap between the four-sided gradient flow as follows

$$G = \min\left(G^l, G^r, G^t, G^d\right). \tag{17}$$

This method draws the exact appearance of the salient object. Figure 2 shows that our flow field can perfectly estimate the salient regions.

6. Saliency Maps Fusion

The final saliency map is the fusion of the static and dynamic saliency maps. The combination is performed to modify static saliency maps with the corresponding dynamic saliency value.

According to previous works on video saliency (Goferman *et al.*, 2012), locations that are distant from the region of attention are less attractive than those which are around. Which means that pixels that are closer to the object of interest get higher saliency scores than further ones.

Hence, the saliency at location $X = (x, y)$ can be defined as

$$SM(x, y) = \left(S_m(x, y)(1 - d(X, C)\right), \tag{18}$$

where $d(x, y)$ is the Euclidean distance between $X = (x, y)$ and the centre $C = (x_c, y_c)$, $S_m(x, y)$ is the saliency values at location (x, y) and is given by

$$S_m(x, y) = N\left(S(x, y)\right) \times \exp\left(M_d(x, y) \times G(x, y)\right) * I_{k*k}, \tag{19}$$

the exponential function is used to widen the contrast of the dynamic saliency weights and $N(S(x, y))$ is a normalization operation used to normalize the values of $S(x, y)$ to

the range of [0, 1]. To minimize noise caused by camera motion we use a 2D Gaussian low-pass filter I_{k*k}, k is the kernel value equal to 5.

All the aforementioned actions that helped to achieve better results are organized in Algorithm 1 as follows

Algorithm 1 Video saliency detection.

Input: A video Frame

1: Separate the input frame into patches $P_1, P_2, P_3, \ldots, P_N$
2: Compute for each patch the Uniform Contrast measure using Eq. (2)
3: Compute Foreground/Background probabilities using Eq. (5) and Eq. (6)
4: Compute final static map using Eq. (7)
5: Compute motion distinctiveness measure using Eq. (11)
6: Compute the four-sided motion estimator using Eq. (17)
7: Generate the final saliency map using Eq. (19)

Output: A saliency map

7. Experiments

Our method detects automatically salient objects in video sequences. In this section, we compare our spatio-temporal saliency framework against state-of-the-art methods on the Segtrack v2 (Li *et al.*, 2013) and Fukuchi (Fukuchi *et al.*, 2009) datasets. In our proposed method, we utilize the spatial and temporal information of the input frame at pixel and frame levels to decide saliency probability of each pixel. Spatial information includes contrast cues, while temporal information makes use of motion distinctiveness and magnitude gradient flow field. The fusion of spatial and temporal saliency maps leads to a saliency map which highlights region of interest and segments salient object from the background.

7.1. *Datasets*

We evaluate our approach on two benchmark datasets that are used by most of the state-of-the-art video saliency methods.

Fukuchi dataset (Fukuchi *et al.*, 2009) contains 10 video sequences with a total of 768 frames with one dynamic object per video. The ground truth consists of the segmented images. The dynamic objects are from different classes including horse, flower, sky-man, snow cat, snow fox, bird, etc.

SegTrack v2 dataset (Li *et al.*, 2013) contains 14 sequences with a total of 1066 frames. Videos can contain more than one dynamic object. Salient objects include objects with challenging deformable shapes such as birds, a frog, cars, a soldier, etc.

7.2. *Evaluation Metrics*

The performance of our method is evaluated on two extensively used metrics including Precision-recall **PR** curves, Receiver operating characteristic **ROC** curves **F-measure** and **AUC**.

In the saliency detection field, precision is defined as the salient pixels that are rightly detected and is given by (20)

$$\text{precision} = \frac{\sum_{x,y} S(x, y)G(x, y)}{\sum S(x, y)}. \tag{20}$$

Recall is the percentage of the detected salient pixels and is given by (21)

$$\text{recall} = \frac{\sum_{x,y} S(x, y)G(x, y)}{\sum G(x, y)}, \tag{21}$$

where $S(x, y)$ is the saliency degree of pixel $p(x, y)$ in the obtained saliency map, and $G(x, y)$ is the saliency degree of the pixel $p(x, y)$ in the ground truth. The precision-recall curves are build by binarizing the saliency maps of each method using a fixed threshold. They are computed by varying the threshold from 0 to 255.

PR curves offer a reliable comparison of how good the saliency maps are and how well they highlight salient regions in a video frame. The **F-measure** is defined as:

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}, \tag{22}$$

where we use $\beta^2 = 0.3$ following (Li *et al.*, 2014).

The **ROC** curve plots the true positive rate against the false positive rate. A perfect approach has a 0 value for the false positive rate and 100 per cent value for the true positive rate which indicates that predictions are identical to the ground truth.

While ROC curves evaluate the performance of a model as two-dimensional representation, the **AUC** elaborates this information into a single measure. As the name signifies, it is calculated as the area under the ROC curve. A perfect model will score an AUC of 1, while random prediction will score an AUC of around 0.5.

7.3. *Implementation*

The implementation of the proposed Algorithm 1 can be divided into two main steps, namely, static map generation and dynamic map generation. To generate the static map, we divide the input video into frames, where each frame is treated as an independent image. Then each video frame is divided into non-overlapping square patches (patch width equal to 2). For each patch, the second step of Algorithm 1 is computed using local contrast, contrast consistency and spatial consistency features. Local contrast aims to compute the contrast change between a pair of patches, which highlight the object's boundaries. Contrast consistency measures the contrast weight between two patches regarding

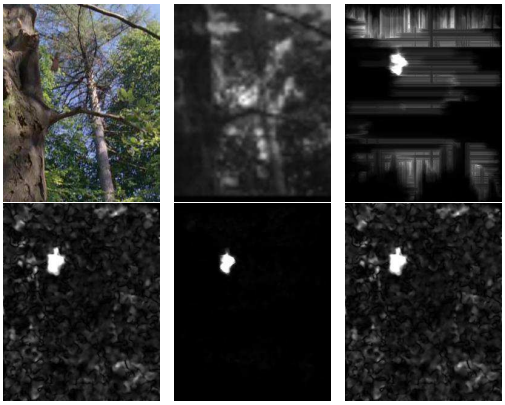

Fig. 3. Spatio-temporal saliency map. From left to right: Input frame, spatial map, four-sided motion map, motion distinctiveness map, saliency map and ground-truth.

the neighbouring ones which will emphasize the whole object. Local contrast and contrast consistency will emphasize the object and its border. Spatial consistency measure is based on the assumption that distant patches highly belong to different objects. Unified contrast measure will be used to label each patch as a foreground or background patches where foreground patches belong to the salient object and background patches belong to background which are then used to compute foreground/background probabilities. The final static map is then computed (Algorithm 1, fourth step).

The second part of our proposed method starts at step 5 of the proposed Algorithm 1 which consists on computing the temporal saliency degree of each frame. To do so, two saliency measures are proposed: the first is the motion distinctiveness measure which is proposed under the assumption that suspicious motion attracts attention (see Section 5.1), and the four-sided motion estimator which is used to compute motion consistency between each pair of frames (see Section 5.2). Temporal saliency measures and static saliency map are fused together to generate the final saliency map.

Figure 3 details the exact resultant map at each step. First the spatial map is generated, we notice that the colour of the falling bird is quite similar to the colour of the trees limbs which enables the uniform contrast measure to detect the salient object. So, we use the four-sided motion estimator and motion distinctiveness measure (see Section 5.2 and Section 5.1) to extract the salient object (falling bird). We notice that in challenging cases where the contrast and spatial consistencies can not segment the object of interest from the background, temporal cues are essential to highlight the salient object.

7.4. *Results*

We compare our video saliency approach to seven state-of-the-art methods, namely, CBS (Jiang *et al.*, 2011), GB (Harel *et al.*, 2006), GVS (Wang *et al.*, 2015), ITTI (Itti and Baldi, 2005), RR (Mancas *et al.*, 2011) and RT (Rahtu *et al.*, 2010).

On both Segtrack v2 and Fukuchi datasets we clearly outperform the other methods in terms of F-measure and AUC. The precision-recall curves in Fig. 4 provide similar con-

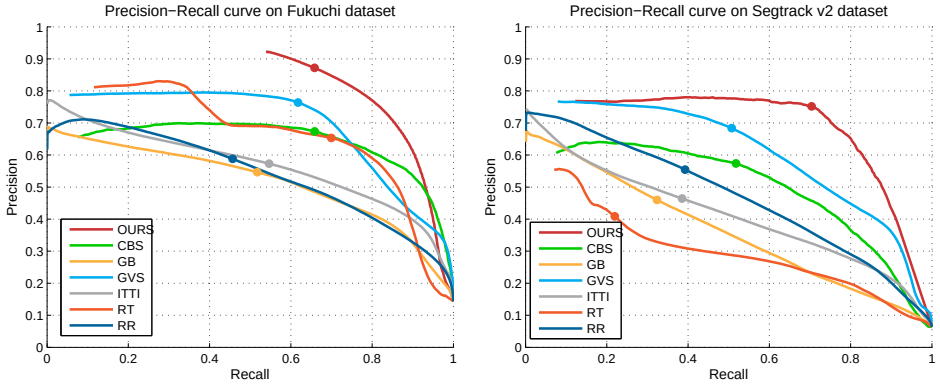

Fig. 4. PR curves on Fukuchi and Segtrack v2 datasets.

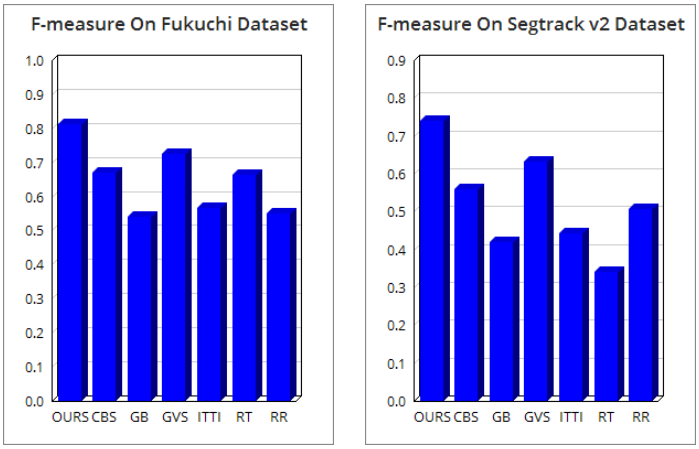

Fig. 5. F-measure values on Fukuchi and Segtrack v2 datasets.

clusions where our method obtains best results compared to the state-of-the-art methods for most recall values.

Recall values of **RR** (Mancas *et al.*, 2011) and **GVS** (Wang *et al.*, 2015) are very small when we vary the threshold to 255 and even decrease it to 0 in case of **ITTI** (Itti and Baldi, 2005), **RT** (Rahtu *et al.*, 2010) and **GB** (Harel *et al.*, 2006) since the output saliency maps do not respond to the salient object detection. For the Segtrack v2 and Fukuchi datasets, the minimum value of recall does not decrease to zero which means that in the worst case and with the most complex background, our method detects the region of interest with good response values. Moreover, our saliency method attains the best precision rate, which denotes that our detected saliency maps are more responsive to regions of interest. The obtained F-score results are 0.739 on Segtrack v2, and 0.829 on Fukuchi (see Fig. 5).

ROC curves are presented in Fig. 6. For low false positive rate our method obtains much higher true positive rates. The area under ROC curves are also reported in Fig. 7 where we reach best values on both datasets.

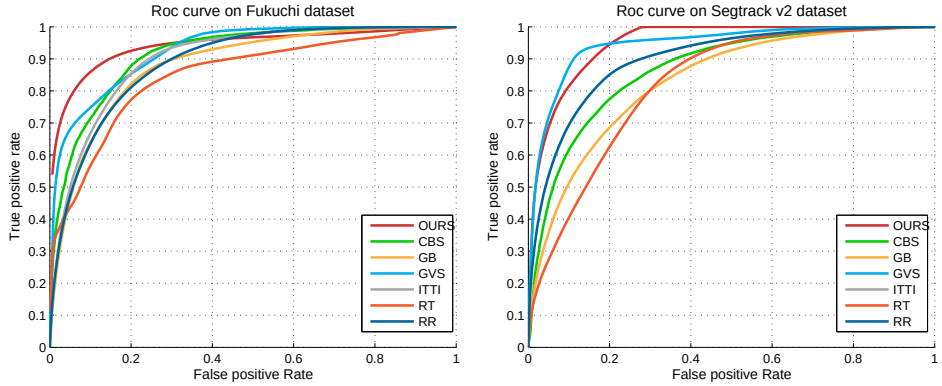

Fig. 6. ROC curves on Fukuchi and Segtrack v2 dataset.

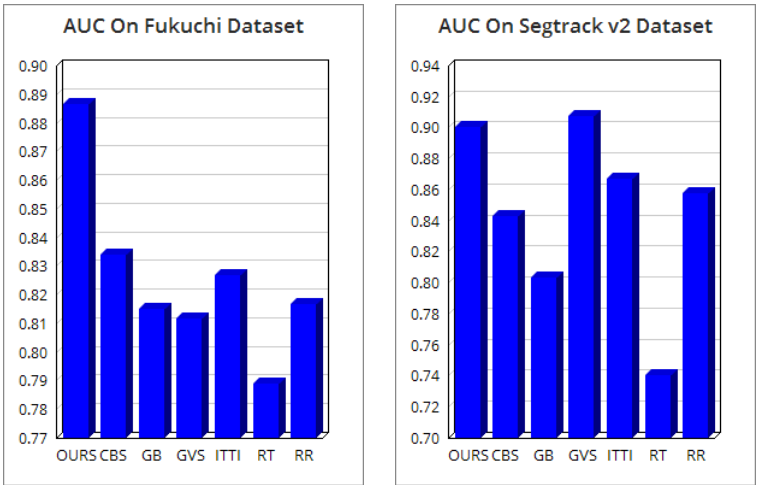

Fig. 7. AUC values on Fukuchi and Segtrack v2 dataset.

We added qualitative comparison on different challenging cases in Fig. 8 and in each situation, our method outperforms other methods. Saliency maps provided by **GB** (Harel *et al.*, 2006) and **ITTI** (Itti and Baldi, 2005) and do not show the exact location of the salient object because of lack of motion information, especially with complex backgrounds. **RT** (Rahtu *et al.*, 2010) is quite good, the salient object is correctly detected but the background gets high saliency probability. While optical flow is one of the most used techniques to detect moving objects, it can not be a good saliency estimator. The performance of video saliency detector **RR** (Mancas *et al.*, 2011) based on optical flow, assign low saliency probability to static pixels which belong to salient object (see third and sixth rows). In most cases, **CBS** (Jiang *et al.*, 2011) and **GVS** (Wang *et al.*, 2015) are able to locate the salient object even in complex situations where foreground-background colours are similar (see eighth rows) since their motion information is very informative. Results

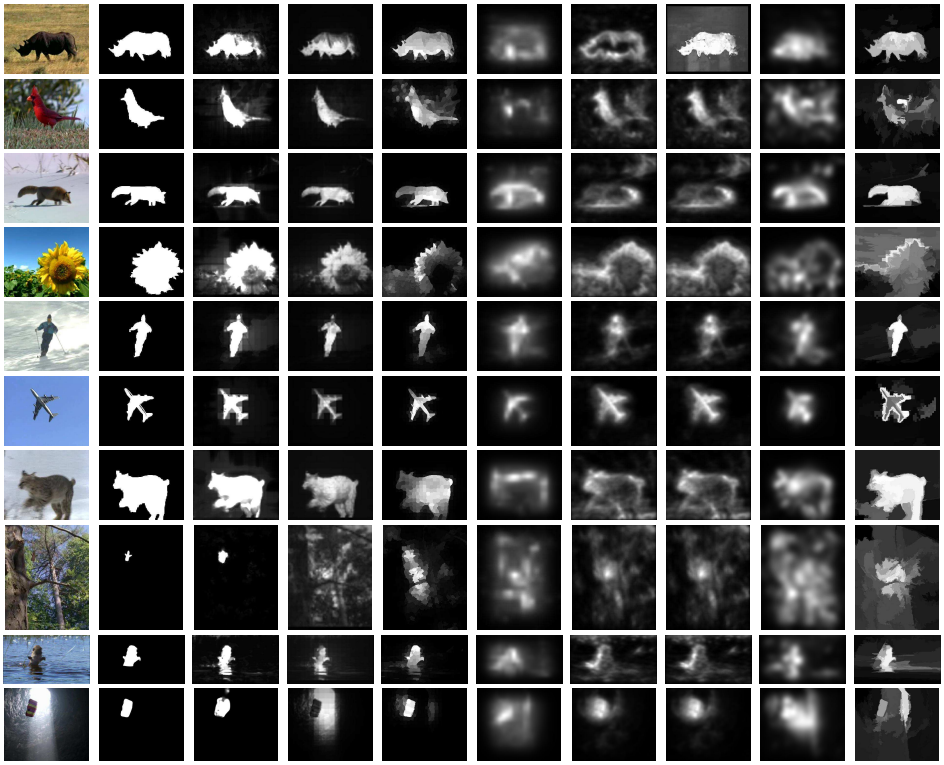

Fig. 8. Visual comparison of saliency maps generated from 6 different methods, including our method using the spatio-temporal features, our method using only spatial features, GVS (Wang *et al.*, 2015), GB (Harel *et al.*, 2006), RR (Mancas *et al.*, 2011), RT (Rahtu *et al.*, 2010), ITTI (Itti and Baldi, 2005) and CBS (Jiang *et al.*, 2011)

of a moving object with higher speed and a static camera are shown in the third row, and produce a good saliency map. In case of an object with high speed and a moving camera (fifth and sixth rows), our proposed motion feature highlights only the moving object. Based on the aforementioned analysis, two main conclusions can be drawn. First, to detect a salient object in videos, it is essential to examine motion information. Second, developing a method that depends only on motion information is not an excellent idea. Combining spatial and temporal information into a video saliency framework leads to the best results.

We performed an extra test, where we use only uniform contrast measure to detect saliency (see second column of Fig. 8). We notice that when there is a change of contrast between the background and the object, the salient object can be correctly detected and the role of motion feature is to set up the regions of the object that made a remarkable movement (e.g. the flower and the wolf). In case of colour similarity between the salient object and the background, the use of static information will fail to point out the object of interest, and the use of motion features will accomplish the mission (e.g the falling bird and the parachute). In this work, we use an inter-frame and intra-frame motion estimation to reinforce temporal saliency detection. Inter-frame motion estimation is performed

using our motion distinctiveness measure which will highlight the object of interest that has a distinctive activity when moving from one frame to another. Intra-frame motion information is measured using our four-sided motion estimator where for each frame, we compute the magnitude gradient, then a gradient flow field is derived from the cumulative sum of the magnitude gradient through four sides of the frame. Our motion features are able to face challenging situations like slow motion, noise caused by optical flow.

8. Conclusion

In this paper, we propose a spatio-temporal framework for video saliency detection. First we derive a spatial saliency map at patch level using local contrast and a uniform contrast measure to highlight the change in contrast of the object of interest by integrating colour and spatial priors. Temporal information is derived from the motion distinctiveness measure and gradient flow field estimator. Spatial and temporal saliency maps are fused into one master saliency maps. In the experiments, we show that the motion features and unified contrast feature greatly improve results. Furthermore, we show that our framework obtains good results for video saliency, compared to the results of the state-of-the-art methods on the Segtrack v2 and Fukuchi datasets.

References

Annum, R., Riaz, M.M., Ghafoor, A. (2018). Saliency detection using contrast enhancement and texture smoothing operations. *Signal, Image and Video Processing*, 12(3), 505–511.

Bhattacharya, S., Venkatsh, K., Gupta, S. (2017). Background estimation and motion saliency detection using total variation-based video decomposition. *Signal, Image and Video Processing*, 11(1), 113–121.

Brox, T., Malik, J. (2011). Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3), 500–513.

Borji, A., Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 185–207.

Borji, A., Cheng, M.-M., Jiang, H., Li, J. (2014). Salient object detection: a survey. arXiv preprint, arXiv:1411.5878.

Borji, A., Cheng, M.-M., Jiang, H., Li, J. (2015). Salient object detection: a benchmark. *IEEE Transactions on Image Processing*, 25(12), 5706–5722.

Chen, T., Cheng, M.-M., Tan, P., Shamir, A., Hu, S.-M. (2009). Sketch2photo: internet image montage. *ACM Transactions on Graphics (TOG)*, 124–129.

Cheng, M.-M., Zhang, F.-L., Mitra, N.J., Huang, X., Hu, S.-M. (2010). Repfinder: finding approximately repeated scene elements for image editing. *ACM Transactions on Graphics (TOG)*, 29(4), 84.

Cheng, M.-M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.-M.G. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 569–582.

Christopoulos, C., Skodras, A., Ebrahimi, T. (2000). The JPEG2000 still image coding system: an overview. *IEEE Transactions on Consumer Electronics*, 46(4), 1103–1127.

Cui, X., Liu, Q., Metaxas, D. (2009). Temporal spectral residual: fast motion saliency detection. In: *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 617–620.

Donoser, M., Urschler, M., Hirzer, M., Bischof, H. (2009). Saliency driven total variation segmentation. In: *IEEE International Conference on Computer Vision*, pp. 817–824.

Fukuchi, K., Miyazato, K., Kimura, A., Takagi, S., Yamato, J. (2009). Saliency-based video segmentation with graph cuts and sequentially updated priors. In: *IEEE International Conference on Multimedia and Expo*, pp. 638–641.

Gao, D., Mahadevan, V., Vasconcelos, N. (2008). The discriminant center-surround hypothesis for bottom-up saliency. *Advances in Neural Information Processing Systems*, 497–504.

Goferman, S., Zelnik-Manor, L., Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10), 1915–1926.

Gopalakrishnan, V., Rajan, D., Hu, Y. (2012). A linear dynamical system framework for salient motion detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 683–692.

Gu, G., Zhu, J., Liu, Z., Zhao, Y. (2015). Visual saliency detection based object recognition. *Journal of Information Hiding and Multimedia Signal Processing*, 6, 1250–1263.

Hamel, S., Guyader, N., Pellerin, D., Houzet, D. (2016). Contribution of color in saliency model for videos. *Signal, Image and Video Processing*, 10(3), 423–429.

Harel, J., Koch, C., Perona, Po. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systes*, 545–5524.

Imamoglu, N., Shimoda, W., Zhang, C., Fang, Y., Kanezaki, A., Yanai, K., Nishida, Y. (2017). An integration of bottom-up and top-down salient cues on RGB-d data: saliency from objectness versus non-objectness. *Signal, Image and Video Processing*, 12(2), 1–8.

Itti, L., Baldi, P. (2005). A principled approach to detecting surprising events in video. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 631–637.

Itti, L., Koch, C., Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S. (2011). Automatic salient object segmentation based on context and shape prior. In: *British Machine Vision Conference*, pp. 2083–2090.

Kim, W., Jung, C., Kim, C. (2011). Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4), 446–456.

Kim, J., Han, D., Tai, Y.-W., Kim, J. (2014). Salient region detection via high-dimensional color transform. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890.

Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M. (2013). Video segmentation by tracking many figure-ground segments. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2192–2199.

Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L. (2014). The secrets of salient object segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287.

Luo, Y., Tian, Q. (2012). Spatio-temporal enhanced sparse feature selection for video saliency estimation. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–38.

Mahadevan, V., Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1), 171–177.

Mancas, M., Riche, N., Leroy, J., Gosselin, B. (2011). Abnormal motion selection in crowds using bottom-up saliency. In: *IEEE International Conference on Image Processing*, pp. 229–232.

Mannan, S K., Kennard, C., Husain, M. (2009). The role of visual salience in directing eye movements in visual object agnosia. *Current Biology*, R247–R248.

Mao, Y., Guo, B.-L., Yan, Y., Sun, W. (2015). Multiple structure based saliency detection and its application in image retrieval. *Journal of Information Hiding and Multimedia Signal Processing*, 6(4), 771–782.

Mauthner, T., Possegger, H., Waltner, G., Bischof, H. (2015). Encoding based saliency detection for videos and images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2494–2502.

Rahtu, E., Kannala, J., Salo, M., Heikkilä, J. (2010). Segmenting salient objects from images and videos. In: *European Conference on Computer Vision*, pp. 366–379.

Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L. (2013). Learning video saliency from human gaze using candidate selection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1147–1154.

Shanableh, T. (2016). Saliency detection in mpeg and hevc video using intra-frame and inter-frame distances. *Signal, Image and Video Processing*, 10(4), 703–709.

Shi, J., Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.

Singh, A., Chu, Chee-Hung, H., Pratt, M. (2015). Learning to predict video saliency using temporal superpixels. In: *International Conference on Pattern Recognition Applications and Methods*, pp. 201–209.

Wang, W., Shen, J., Porikli, F. (2015). Saliency-aware geodesic video object segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3402.

Wolf, J.M. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6), 1–7.

Xue, Y., Guo, X., Cao, X. (2012). Motion saliency detection using low-rank and sparse decomposition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1485–1488.

Yeh, H.-H., Liu, K.-H., Chen, C.-S. (2014). Salient object detection via local saliency estimation and global homogeneity refinement. *Pattern Recognition*, 47(4), 1740–1750.

Zhang, G.-X., Cheng, M.-M., Hu, S.-M., Martin, R.R. (2009). A shape-preserving approach to image resizing. In: *Computer Graphics Forum*, pp. 1897–1906.

Zhong, S.-H., Liu, Y., Ren, F., Zhang, J., Ren, T. (2013). Video saliency detection via dynamic consistent spatio-temporal attention modelling. In: *National Conference of the American Association for Artificial Intelligence*, pp. 1063–1069.

R. Kalboussi received her bachelor and engineer degree from the Higher Institute Higher Institute of Computer Science and Communication Techniques. She is currently a PhD candidate in computer sciences. She is also a researcher at the Networked Objects, Control and Communication Systems (NOCCS) Research Laboratory, National School of Engineering of Sousse, Tunisia. Her research interests include image and video processing, pattern recognition, computer vision and machine learning.

M. Abdellaoui received his engineer, MS and PhD degrees from the National School of Engineering of Monastir, Tunisia, in 2003, 2005 and 2012, respectively. He is currently an assistant professor in Signal and Image Processing at the High Institute of Applied Technologies, University of Kairouan. He is also a researcher at the Networked Objects, Control and Communication Systems (NOCCS) Research Laboratory, National School of Engineering of Sousse, Tunisia. His research interests include image and video processing, computer vision and machine learning.

A. Douik received his MSEE degree in 1990 and PhD degree in 1996, both from ENSET, University of Tunis. He is currently a full professor in signal and image processing at the National Engineering School of Sousse. He is also a researcher at the Networked Objects, Control and Communication Systems (NOCCS) Research Laboratory, National School of Engineering of Sousse, Tunisia. His research interests include image and video processing, machine learning and control.