Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18126-x

Hybrid time-spatial video saliency detection method to enhance human action recognition systems

Abdorreza Alavi Gharahbagh[1] · Vahid Hajihashemi[1] · Marta Campos Ferreira[1] · J.J.M. Machado[2] · João Manuel R.S. Tavares[2]

Received: 20 July 2022 / Revised: 27 December 2023 / Accepted: 1 January 2024

Abstract
Since digital media has become increasingly popular, video processing has expanded in recent years. Video processing systems require high levels of processing, which is one of the challenges in this field. Various approaches, such as hardware upgrades, algorithmic optimizations, and removing unnecessary information, have been suggested to solve this problem. This study proposes a video saliency map based method that identifies the critical parts of the video and improves the system's overall performance. Using an image registration algorithm, the proposed method first removes the camera's motion. Subsequently, each video frame's color, edge, and gradient information are used to obtain a spatial saliency map. Combining spatial saliency with motion information derived from optical flow and color-based segmentation can produce a saliency map containing both motion and spatial data. A nonlinear function is suggested to properly combine the temporal and spatial saliency maps, which was optimized using a multi-objective genetic algorithm. The proposed saliency map method was added as a preprocessing step in several Human Action Recognition (HAR) systems based on deep learning, and its performance was evaluated. Furthermore, the proposed method was compared with similar methods based on saliency maps, and the superiority of the proposed method was confirmed. The results show that the proposed method can improve HAR efficiency by up to 6.5% relative to HAR methods with no preprocessing step and 3.9% compared to the HAR method containing a temporal saliency map.

Keywords Video processing · Optical flow · Genetic algorithm · Time saliency · Spatial saliency · Deep learning

1 Introduction

Visual saliency detection identifies regions of an image that are more significant and facilitates image and video processing algorithms. Video Saliency Detection (VSD) is a visual detection category that identifies important regions and objects in videos. Usually, video frames have

✉ João Manuel R.S. Tavares
tavares@fe.up.pt

Extended author information available on the last page of the article

Published online: 14 February 2024

lower resolution and more noise than typical images and should be continuously processed; therefore, video processing encounters more challenges than image processing. Human eye is focused more on moving regions of videos when recognizing relevant features; therefore, motion detection is a critical component of VSD [1]. Processing the regions of interest (ROIs) of a video detected by a saliency map or Video Salient Object Detection (VSOD) can significantly reduce the computational cost and increase the efficiency of video processing algorithms, such as HAR systems.

A review of video processing algorithms revealed that most of them use random regions or frame selections throughout the input video frames [2–4], which decreases their efficiency. Particularly, if only ROIs of video frame sequences are used as the input, video processing systems can be better trained. Therefore, this article proposes a hybrid time-spatial video saliency detection method to enhance the accuracy of HAR systems. The existing VSOD methods can be categorized into Supervised, Unsupervised, and Semi-supervised methods [5]. Some studies have divided VSODs into Deep Learning (DL) methods and other methods [6]. DL-based methods, which are mainly supervised methods, have been widely developed in recent years but have shortcomings, such as the long-time learning process, the need for a large amount of training data, and powerful hardware for implementation. Furthermore, general DL systems must be retrained for each new application to achieve maximum efficiency. Wang et al. [7] presented a saliency detection method based on multiple instance learning, which extracts low-, mid-, and high-level features and evaluates the possibility of each region being salient. Li et al. [8] suggested a DL-based approach called Flow Guided Recurrent Neural Encoder for VSOD. Sun et al. [9] proposed a Step-gained Fully Convolutional Network (SGF) model for VSOD, which uses temporal and spatial information simultaneously. Lee et al. [10] presented a Saliency Map Prediction method based on a Convolutional Neural Network (CNN) for 360-degree video streaming, and Li et al. [11] proposed a Deep Neural Network (DNN) for VSOD that employs spatial and optical flow features. Yan et al. [12] introduced a Semi-Supervised Convolutional Neural Network for VSOD. Fan et al. [13] analyzed different VSOD methods on datasets commonly used in this field and compared their results. They also proposed an improved in-depth learning method for VSOD.

Yang et al. [14] suggested superpixels, a Fully Convolutional Network (FCN), and a random walk method on a graph for image salient object detection. Liu et al. [15] proposed an approach based on motion saliency and mid-level patches for action recognition, which uses a threshold-based motion region segmentation algorithm to extract motion saliency maps. Gu et al. [16] proposed a Pyramid Constrained Self-Attention Network for VSOD. Kousik et al. [17] combined a CNN with a Recurrent Neural Network (RNN) to improve salient object detection accuracy. Ji et al. [18] reviewed CNN-based encoder-decoder networks used in salient object detection. Zong et al. [19] proposed a motion saliency-based multi-stream Multiplier ResNets (MSM-ResNets) structure for action recognition, which can also extract motion saliency maps. Ji et al. [20] proposed a DNN structure for VSOD, including an encoder-decoder backbone network, self-and cross-attention modules, and a multiscale feature fusion operation called CASNet.

Zhang et al. [21] proposed a VSOD termed the dynamic context-sensitive filtering network, which includes convolutional layers, optical flow and long-term memory (LSTM). Wang et al. [22] suggested a hybrid feature-aligned network for semantic salient object detection in images, which mitigates the disturbance of a complex background. Liu et al. [23] proposed a dual-branch network based on multitask learning, which extracts the localization information of salient objects from scene classification and then employs dynamic guided learning to enhance saliency detection. The authors suggested another salient object detection method using a universal super-resolution-assisted learning approach [24]. Alavigharahbagh et al.

[25] proposed a time saliency map based on motion features, which takes into account the disturbance of the camera movement in the used motion features. As an alternative to VSOD methods, multiscale object detection methods such as the one proposed by Liu et al. [26] can also extract salient objects from image frames.

DL networks with different layers typology, combinations and time-memory specifications have been proposed for DL-based VSOD systems. The efficiency of the used DL network depends significantly on the application, mainly of the type of the input video. The most used layers in DL-based VSOD systems are convolutional layers, LSTM, and rectified linear units (ReLU). The shortcomings of DL methods also exist in supervised non DL-based VSOD systems. However, non DL-based systems require fewer samples for training and not so powerful hardware. In this category of research, one can find the work of Vijayan and Ramasundaram [27], which combined the PSO framework with the saliency map technique for more accurate object detection, and the work of Huang and McKenna [28], which considered a combination of superpixels and optical flow for VSOD. Generally, non DL-based VSOD systems, which have attracted much less research attention, demonstrate a lower level of performance than DL-based systems.

Unsupervised and semi-supervised systems are less efficient than supervised systems in real-world applications, and their advantages are their high generalizability, independence from training samples, and ability to work under different conditions. In all unsupervised VSOD methods, the combination of information in video frames, that is, the spatial information, and information between image frames, i.e., the motion, is used to identify the ROIs of the input video. The use of object recognition algorithms as part of or in combination with VSOD is usual in applications such as the ones that can be found in transportation and security. Mahapatra [29] showed that the human visual system is more sensitive to moving objects than static objects in a scene.

Therefore, motion is an essential part of the video saliency map. Lee et al. [30] performed motion analysis based on a video dynamic visual saliency map model using frames' optical flow and spatial characteristics. Jeong et al. employed a combination of motion, color, edge, and fuzzy logic to create video saliency maps [31]. Cui et al. [32] relied on the temporal spectral residual to separate moving objects from the static scene background and considered spectral-temporal features a fast descriptor. Woo et al. [33] used general and local features for image obstacle categorization which can be combined with motion characteristics and applied to videos. Belardinelli et al. [34] exploited spatiotemporal filtering to extract motion saliency maps. Morita [35] used only motion information to detect saliency maps in videos. Then, the author modified his method by combining color and motion information to increase the efficiency of the obtained saliency maps [36]. Hu et al. [37] examined three ways to obtain motion saliency maps in videos: optical flow, motion contrast, and spatiotemporal filtering.

Mejía-Ocaña et al. [38] removed the motion of the used camera from the acquired video and estimated the motion vector in different parts of the video hierarchically. The result was then reprocessed to eliminate possible errors in the motion vector calculation. Gkamas and Nikou [39] tried to increase the optical flow accuracy using superpixels. Li et al. [40] applied visual and motion saliency information for object extraction in videos. Their method considered the spatiotemporal consistency of structural information, color information, and optical flow. Chang and Wang [41] used superpixels to increase the optical flow performance in low-resolution or noisy videos. Huang et al. removed scene motion from image frames and used spatial, entropy, and temporal features to detect moving objects in videos and video saliency maps [42]. Dong et al. [43] proposed a HAR method that combines optical flow and superpixel residual motion characteristics to detect actions. Giosan et al. [44] used a combination of superpixel, optical flow, and image depth information for obstacle segmentation in videos.

Dellaert and Roberts [45] used optical flow templates to tag motion superpixels in video frames. Xu et al. [46] applied the discrete cosine transform to create an initial saliency map and then modified it using the global motion estimation method. Srivatsa and Babu [47] proposed a method based on foreground connectivity in superpixels for salient object detection. Donn'e et al. [48] introduced a dynamic optical flow calculation method using superpixels that increased the optical flow calculation speed in high-quality videos. Li et al. [49] used superpixels, color information, and spatial features for VSOD, and Hu et al. [50] proposed an enhanced optical flow algorithm using superpixels. Guo et al. [51] also applied a combination of superpixels and optical flow to extract the initial saliency map and corrected it using cross-frame cellular automata. Tu et al. [52] extracted motion objects from the input video, removed the noise, and obtained the final saliency map using motion continuity.

Hu et al. [53] proposed a video segmentation method using motion information and spatial properties of video frames, which used optical flow and edge information to create a neighborhood graph. Ling et al. [54] corrected camera motion using a feedback-based robust video stabilization method. Video stabilization enhances the efficiency of motion features such as of the ones based on optical flow. Wang et al. [55] extracted motion features with kernelized correlation filters using superpixels, color, and spatial features for visual object tracking. Chen et al. [56] employed a combination of principal component pursuit and motion saliency to distinguish the scene foreground in video images. Temporal and spectral residuals were considered to calculate the motion saliency.

Maczyta et al. [57] relied on optical flow for motion saliency map estimation. Zhu et al. [58] proposed an optical flow estimation method using light field superpixels. Kim et al. [59] considered a set of superpixels for object tracking, assuming significant motion in the videos. Ngo et al. [60] applied a dynamic mode decomposition for motion saliency detection and a difference-of-Gaussian filter in the frequency domain to improve the detection. Qiu et al. [61] considered superpixels to enhance motion detection efficiency. Tian et al. [62] used Lucas-Kanade optical flow and Kalman filter to model the regular movement in videos. They also used a template update scheme to ensure timely adaptation to feature changes. In a concise analysis of the aforementioned studies, one can perceive the following:

- Motion is the most common feature in unsupervised methods, along with spatial, spectral, color, and edge features;
- The principle of motion continuity is essential when extracting motion saliency maps;
- As motion features, the optical flow and frame difference are the mostly used;
- Superpixels and image segmentation have also been used for motion detection;
- The final output is commonly achieved after some postprocessing accuracy improvement steps.

According to the above descriptions, mainly as to the advantages of the unsupervised systems, this study aimed to propose an unsupervised method for VSOD. The method is independent of the objects and uses only temporal and spatial features, including based on color, motion, contrast, frequency, and edge information. A bank of filters is used to extract the spatial saliency map. Additionally, different features are used to detect and remove the motion of the scene to improve the VSOD performance. The combination of the used features considers both spatial and motion information. In summary, the innovations achieved with the current study are as follows:

- Several spatial saliency maps containing color, frequency, edge, and frequency information are used to obtain a comprehensive expression of the ROIs in each image frame;
- Extraction of motion information using optical flow and segmented images;
- Removal of the scene and camera motion to enhance the accuracy;

- Integration of spatial and temporal saliency maps using a nonlinear function.

The remainder of this article is organized as follows. Section 2 presents the flowchart of the proposed method and an explanation of its steps. The results of the proposed method and their comparison with related methods are presented in Section 3, and Section 4 presents the conclusions and perspectives of future work.

2 Proposed method

The flowchart of the proposed method is given in Fig. 1. Its first step consists of retrieving several image frames from the input videos. The main blocks of the method are related to image registration, spatial and temporal saliency maps, and the final saliency map. Three extra blocks, namely region selection, frame cropping, and HAR, were used to evaluate the performance of the proposed method.

2.1 Image registration

In video processing and VSOD, it is vital to separate the motion of the sued camera from that of the objects of interest. The motion of the camera can effectively affect motion detection algorithms and, therefore, interfere with detecting moving objects. Many algorithms have been proposed for image registration, and one of the most common and effective is Speeded Up Robust Features (SURF), which is similar to the scale-invariant feature transform (SIFT) but is faster and less complex in computing [63].

SURF uses a Gaussian filter, Hessian matrix, blob detection, and feature matching to select the key points in an image and detect the rotation and shift in these points. Due to the robustness of the SURF algorithm against rotation and translation, various optimizations have been performed, mainly focusing on execution speed. Due to the aforementioned advantages,

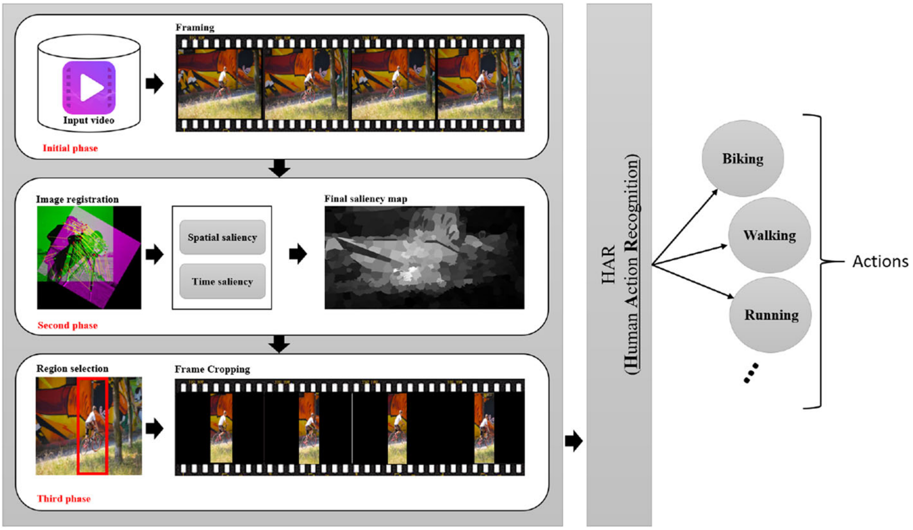


Fig. 1 Flowchart of the proposed method

the SURF operator was implemented in the first step of the proposed method to detect the motion of the scene. Supposing the scene moves because of the motion of the used camera, the motion detection algorithm can distinguish between the motion of the scene and of the moving objects after the image registration step, and the regions of the image frames that are not overlapped can be discarded for further processing. Therefore, it is assumed that the camera moves with the objects of interest and so eliminating these regions does not lead to any performance loss in saliency detection. Figure 2 shows the effect of this step in two image frames of a video containing the motion of a scene. Note that the left border of frame t and the right border of frame $t + 1$, indicated in the frames by black rectangles, were cropped after the registration step.

2.2 Spatial and time saliency maps

2.2.1 Spatial saliency

In this step, the spatial saliency map in the image frames, regardless of their sequence, is extracted. For this purpose, a modified version of salient region detection via high dimensional color transform [64] is used. The features employed in the proposed method are classified into spatial, color, texture, and shape features, Table 1.

The modifications in the proposed method compared to [64] include: first, the proposed method uses parallel processing to improve speed. Second, Simple Linear Iterative Clustering (SLIC) is used to find superpixels in the original method, which is replaced in the proposed method by the SLIC0 algorithm [65] to refine compactness adaptively. In SLIC, the user chooses the compactness parameter, which is constant for all superpixels. If the video has

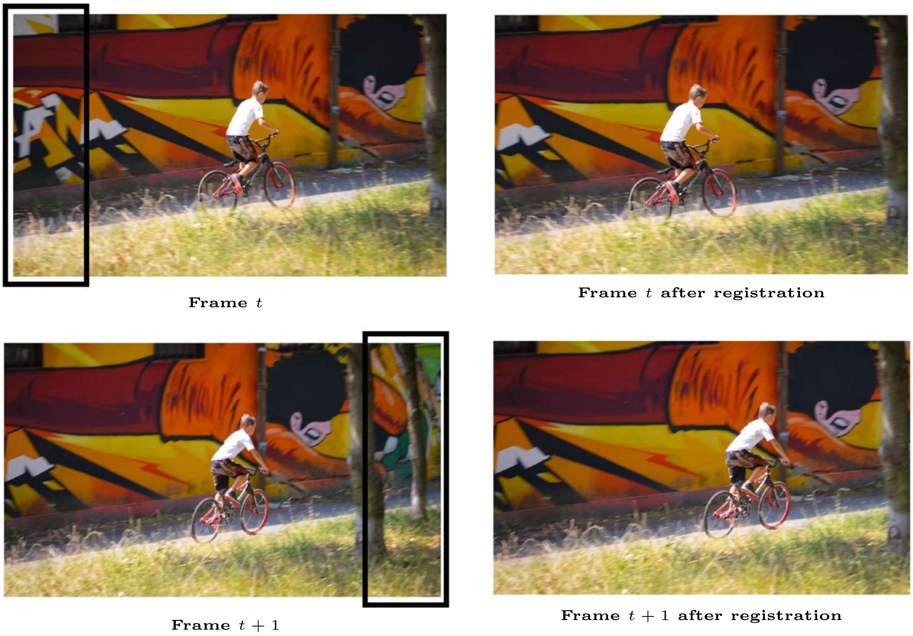
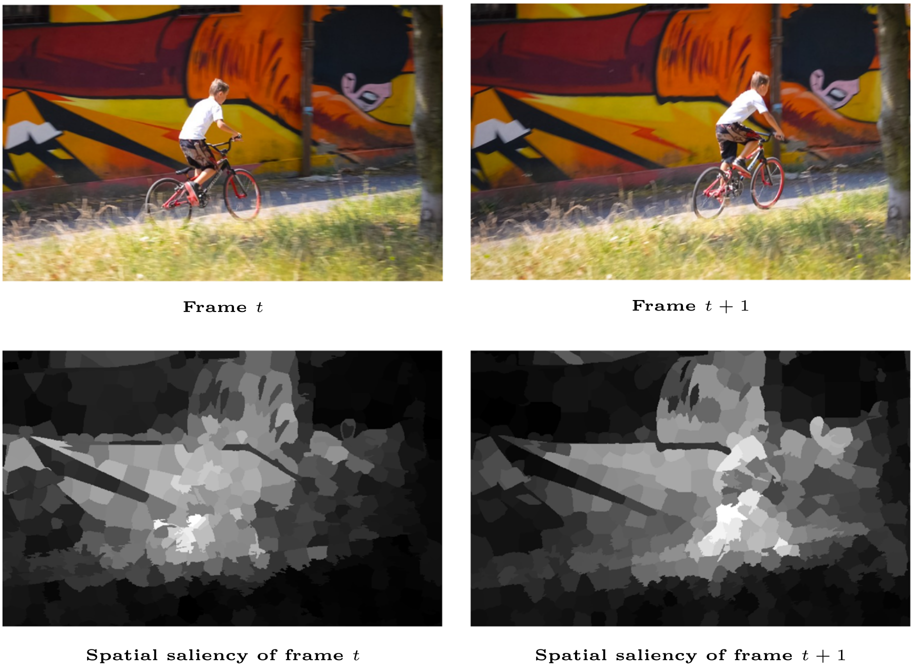

Frame t
Frame t after registration
Frame $t + 1$
Frame $t + 1$ after registration

Fig. 2 Two consecutive image frames before and after the image registration step

Table 1 Features and their length used in the spatial saliency detection step of the proposed method

Feature	Length
Spatial features	
Average normalized x coordinates	1
Average normalized y coordinates	1
Color features	
Average RGB values	3
Average CIELab values	3
Average HSV values	3
RGB histogram	1
CIELab histogram	1
Hue histogram	1
Saturation histogram	1
Global contrast of color features	9
Local contrast of color features	9
Element distribution of color features	9
Texture and shape features	
Area of superpixel	1
Histogram of gradient (HOG)	36
Singular value feature	1


Frame t

Frame $t + 1$

Spatial saliency of frame t

Spatial saliency of frame $t + 1$

Fig. 3 Spatial saliency of the two consecutive video frames of Fig. 2 obtained by the proposed method

different smooth properties, The output of the SLIC algorithm includes irregular superpixels in non-smooth regions. SLIC0 is an adaptive method that generates regular-shaped superpixels in the smooth and non-smooth regions. The third is the number of Histogram of Oriented Gradient (HOG) bins, which is 31 in the original method [64], but in the proposed method is 36. Figure 3 shows the spatial saliency of the two iamge frames in Fig. 2 after the image registration step. In this case, the cyclist and bicycle are among the highlighted regions on the saliency map as ROIs, which demonstrates the ability of the suggested spatial saliency algorithm to detect ROIs.

2.2.2 Time saliency

This study uses optical flow to extract temporal changes in a video [66, 67]. Optical flow extracts the motion pattern in a visual scene that results from the movement of scene objects between an observer and a camera. Some advantages of optical flow are different optical flow extraction methods with distinct hypotheses and optimized algorithms in terms of speed and relative robustness to noise. Its relatively high sensitivity to the motion of the scene and used camera and, in most algorithms, the extraction of the border of moving objects and not the whole objects are among the disadvantages of optical flow [68]. In a two-dimensional (2D) grey-scale video, if the motion in x and y directions is Δx and Δy, respectively, at Δt time interval, and, assuming that the motion between two consecutive image frames is small, this change can be approximated as [69]:

$$\begin{aligned} &I(x+\Delta x, y+\Delta y, t+\Delta t) = \\ &\qquad I(x,y,t) + \tfrac{\partial I}{\partial x}\Delta x + \tfrac{\partial I}{\partial y}\Delta y + \tfrac{\partial I}{\partial t}\Delta t + higher - order \quad terms, \end{aligned} \tag{1}$$

where I is the brightness, x and y the coordinates in the first frame, Δx and Δy the changes in the horizontal and vertical directions, i.e., in the motion in the following frame, Δt the time of motion, and ∂ the differential operator, respectively. Since after Δt the brightness of an object is almost the same, (2) can be written as:

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t), \tag{2}$$

then, by discarding the higher-order terms, one has:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0, \tag{3}$$

and, by dividing (3) by Δt and performing some simplifications, it is possible to obtain [69]:

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0, \tag{4}$$

therefore, (4) can be rewritten as [69]:

$$I_x V_x + I_y V_y = -I_t, \qquad or \qquad \nabla I \cdot \mathbf{V} = -I_t, \tag{5}$$

where $V_x and V_y$ are the vector of motion, i.e., the optical flow, and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y} and \frac{\partial I}{\partial t}$ derivatives or gradients of the image in the x and y, directions and time [69], respectively. Since the number of unknowns in (5) is more than that of considered equations, different optical flow methods simplify and balance the number of equations and unknowns by adding a constraint or hypothesis. After some simulations, the Horn–Schunck [70] method was selected from the current optical flow algorithms, which assumes that the pattern of brightness changes across the video is smooth and uses an optimal iterative method to solve (5). Figure 4 shows

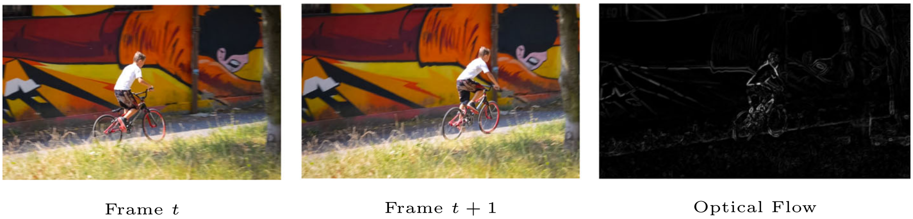


Fig. 4 Output obtained by the Horn–Schunck method for the two consecutive video frames of Fig. 2

the output of the Horn–Schunck method for two consecutive video frames. As can be seen in Fig. 4, the Horn–Schunck method [71] only extracts the boundaries of the moving object, therefore, to obtain the entire moving object, optical flow is combined with color based image segmentation. Thus, the input image is first simplified according to the color based image segmentation method proposed by Bensaci et al. [72] using quantization based on variance [73]. In this quantization, the color variance in the segmented objects is minimized. The dithering technique reduces the quantization error in the neighboring pixels after quantization. Several dithering methods [74] were used in the proposed method. After dithering, the quantized image is corrected using J-image calculations [75].

This correction minimizes the variance in the quantized blocks using the growing region. Figure 5 shows the segmented image frames. Finally, a temporal saliency map is made of the detected segments and their optical flow energies (Fig. 6).

2.3 Final saliency map

The most important part of the proposed method is the combination of temporal and spatial saliency maps. Various research suggested different combinations and schemes. Some researchers used multiplication and a weighted linear function of temporal and spatial saliency maps to obtain the final saliency [76–78]. In most VSOD research, the weight of the temporal saliency is assumed to be greater than that of the spatial saliency [78, 79]. Here, a different scheme is suggested. First, a nonlinear function containing seven expressions combines the time and saliency maps. The expressions include the multiplicative and weighted summation, power of two, and exponential term of temporal and spatial saliency maps. The suggested function that combines temporal and spatial saliency maps is:

$$VSM = x_1 S_s + x_2 T_s + x_3 S_s^2 + x_4 T_s^2 + x_5 S_s T_s + x_6 e^{S_s} + x_7 e^{T_s}, \tag{6}$$

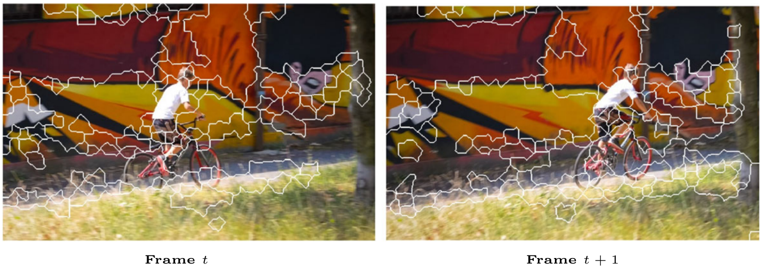
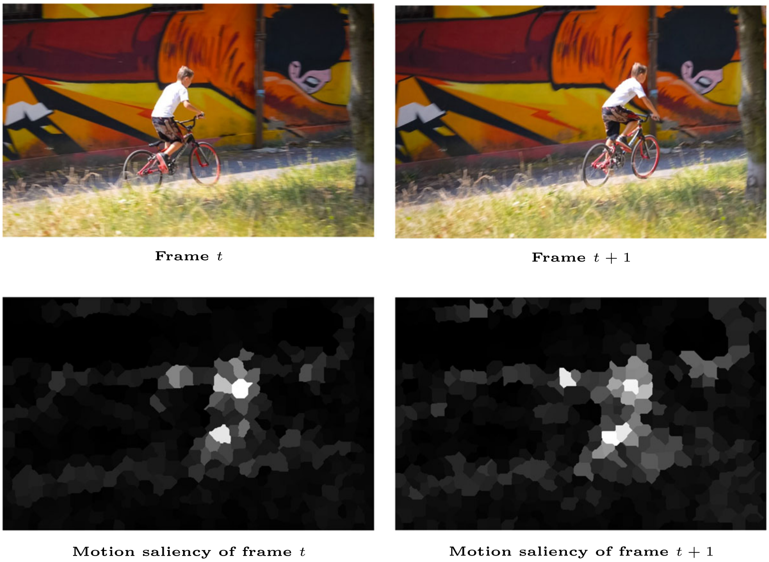


Fig. 5 Output of the time saliency step of the proposed method for the two consecutive frames of Fig. 2



Frame t

Frame $t+1$

Motion saliency of frame t

Motion saliency of frame $t+1$

Fig. 6 Output of the time saliency step of the proposed method for the two consecutive frames of Fig. 2

where x_is are coefficients, VSM the video saliency map, S_s the spatial saliency and T_s the temporal saliency map. The coefficients of (6), which model the weights of the temporal and spatial saliency parts, were obtained using a multi-objective optimization scheme to satisfy two objective functions, which were designed to maximize the efficiency of the saliency map. The objective function 1 is the ratio of the total energy of the 10^{st} decile of the necessary image pixels to the 1st decile of the unimportant image pixels, which must be maximized (Algorithm 1). In other words, the weakest VSM values assigned to necessary image pixels must be greater than the highest VSM values assigned to non-important image pixels. The second objective function is the percentage of unimportant pixels in the input image, higher than 1^{st} decile of essential pixels (Algorithm 2), which must be minimized.

```
Algorithm 1 Objective function 1.
Input: Spatial and Time Saliency maps ($S_s$ and $T_s$) of all image frames. Annotations:
1: Compute the value of VSM based on (6).
2: Separate image pixels belonging to the motion area from other pixels using Annotations, where $VSM_{sm}$
   and $VSM_{ss}$ are pixels that belong to moving and static image regions, respectively.
3: Compute the summation of the first decile of $VSM_{sm}$: $D1_M$.
4: Compute the summation of the last decile of $VSM_{ss}$: $D10_S$.
5: The value of Objective Function 1 is: $f_1 = \frac{D1_M}{D10_S}$.
Output: $f_1$
```

3 Results

The efficiency of the proposed method was evaluated in two steps. In the first step, unknown parameters of the method and the coefficients of (6) were determined by implementing the

Algorithm 2 Objective function 2.

```
Input: Spatial and Time Saliency maps ($S_s$ and $T_s$) of all image frames.
Annotations:
1: Compute the value of VSM based on (6).
2: Separate image pixels belonging to the motion area from other pixels using Annotations, where $VSM_{sm}$ and $VSM_{ss}$ are pixels that belong to moving and static image regions, respectively.
3: Compute the summation of the first decile of $VSM_{sm}$: $D1_M$.
4: The value of Objective Function 2 is defined as the percentage of $VSM_{ss}$ pixels higher than $D1$: $f_2 = \frac{Number\ of\ VSM_{sm} > D1}{Total\ VSM_{sm}\ pixels}$.
Output: $f_2$
```

proposed VSOD. The effectiveness of the method in increasing the accuracy of HAR systems was addressed in the second step. In this case, the HAR system response was analyzed before and after integrating the proposed method. All simulations were performed on MATLAB R2020b and PyCharm IDE Professional Edition 2020, with a personal computer Intel Core i7-9700 CPU (8 cores, 12M Cache, up to 4.70 GHz) with 16GB of RAM and a GeForce RTX 2070 SUPER 8GB GPU.

3.1 Datasets

Four datasets were used in this study. The first two datasets were used to determine the unknown parameters of (6), and the remaining two were used to investigate the effect of adding the proposed method to current HAR systems. Thus, to determine the parameters of the proposed VSOD method and obtain the unknown coefficients of (6), DAVIS 2016 and 2017 datasets[1] were used [75, 87].

DAVIS 2016 and DAVIS 2017 datasets contain a collection of 50 and 150 video frames with their moving object masks, respectively. All videos are in color, and each video's number of frames and image quality differ. After identifying the unknown parameters of (6), the proposed method was added to the input of current HAR systems as a preprocessing step, and the performance of each HAR system with and without the proposed method was evaluated. Two datasets, UCF101[2] and HMDB51[3], were used in this step to provide different videos and a separate file containing the human actions performed in the videos. The UCF101 dataset contains 13,320 videos of 101 human actions, and the HMDB51 dataset contains 6766 videos of 51 actions.

3.2 Proposed method's parameters

The parameters of the proposed method are used in its second phase (Fig. 1), which includes the image registration step, and the building of the spatial saliency map, temporal saliency map, and final saliency map. A set of different features and feature-matching algorithms were examined for the image registration step, as shown in Table 2. Among the available features, the best response was obtained using the SURF algorithm. This algorithm is fast and can be used in online implementations due to its feature extraction speed. Regarding the matching key points, the rotation was assumed to be 0 (zero) because of the type of the studied videos,

[1] https://davischallenge.org/

[2] https://www.crcv.ucf.edu/data/UCF101.php

[3] https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

and only shift and scale transforms were considered. However, any error in this step can significantly reduce the efficiency of the temporal saliency map. The algorithms studied in the optical flow calculation include the Farneback [88], Horn-Schunck, Lucas-Kanade [89], and Lucas-Kanade derivative of the Gaussian [90] algorithms. The Horn-Shank algorithm was selected based on its robustness against errors in the image registration step. In the color image segmentation step [72], the scale was set to 2, μ_j was assumed equal to 0.4, and σ_j was equal to 0.03. A Genetic Algorithm (GA) was used to calculate the coefficients of (6).

In the optimization step, All coefficients were assumed to be between 0 (zero) and 5, and the initial population was set to 200. Two different optimizations were performed. The first was done with objective function 1 (Section 2.3), and the second was performed by combining objective functions 1 and 2 (Section 2.3) with equal weight. In the second optimization, the difference between the two objective functions was used as the objective function due to the different goals of the objective functions, i.e., one should be maximized, and the other should be minimized, as:

$$\mathrm{F}_{\mathrm{cost}} = 1 + \text{objective function } 2 - \text{objective function } 1. \tag{7}$$

Figure 7 shows the changes in (7) when minimized by the GA and the changes in objective functions 1 and 2 during the iterative process. As can be seen in Fig. 7, the changes in objective function 1 are much higher, and its effect on output is more import. This could be due to the large area of the non-relevant region of the video frames, which makes the changes of the objective function 2 less important. The function obtained in the first optimization process was:

$$VSM = 0.9e^{S_s} + 5e^{T_s}, \tag{8}$$

and the function obtained in the second optimization was:

$$VSM = 0.02S_s + 0.01T_s + 0.8e^{S_s} + 4.9e^{T_s}. \tag{9}$$

Equations (8) and (9) output is then used as the final saliency maps.

3.3 Comparison with other methods

By adding the proposed method to some current HAR methods, they can be trained using the salient objects of video frames detected by the proposed method. The results of these methods were compared with those obtained when trained using random or static region selection. Five different HAR methods were used for comparison: four use arbitrary region and frame

Table 2 Features studied in the image registration step of the proposed method

Algorithm	Type of detection	Output object
Accelerated Segment Test (FAST) [80]	Corners	Corner points
Harris–Stephens [81]	Corners	Corner points
KAZE [82]	Points	KAZE points
Minimum eigenvalue [83]	Corners	Corner points
MSER [84]	Regions	MSER regions
ORB keypoints [85]	Points	ORB points
Scale-invariant feature transform (SIFT) [86]	Points	SIFT points
Speeded-Up Robust Features (SURF) [63]	Points	SURF points

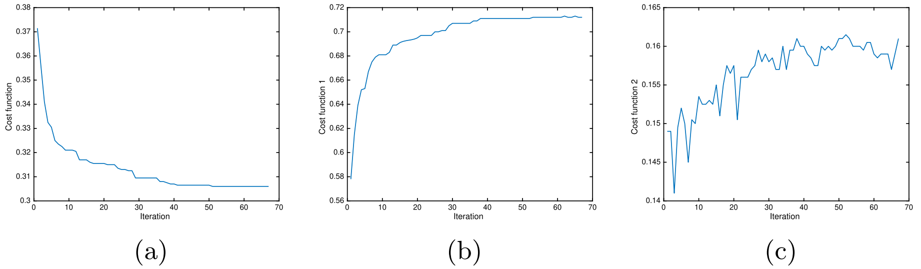


Fig. 7 Behaviour of a) (7), b) objective function 1, and c) objective function 2 during the optimization process

selection methods [91–94], the other uses a temporal saliency map for frames and region selection [95]. All studied HAR models were trained under similar conditions on UCF101 and HMDB51 datasets and then evaluated. In all methods, the number of video frames used for training was 20, and the dimension of the selected frames was 111 × 111. The frames were cropped to 111 × 111 according to the saliency map. Data augmentation algorithms were also used to generate more data for training. The remaining conditions were similar to those suggested in [95]. Table 3 presents the results when the proposed method was added to each studied HAR system using (8) and (9) and applied to the UCF101 and HMDB51 datasets. According to the results, the accuracy of the studied HAR methods improved when the proposed method was added to them and the input frames were selected based on VSOD. The performance improvement of (9) obtained from the summation of the two objective functions was better than that of (8), which was optimized only based on objective function 1.

After identifying (9) as the best choice for producing VSOD, the result of the proposed method was compared to that obtained by the method proposed in [95], which has been used to improve the HAR input. According to Table 4, the proposed method yields better results than the method of [95], which shows that the proposed VSOD is better than the one based on the saliency criterion proposed in that method. The increase in the accuracy achieved by the studied HAR systems with the proposed method is more evident in the case of the HMDB51 dataset. Two reasons can be mentioned for the lower performance improvement achieved with the UCF101 dataset relative to the HMDB51 dataset. First, HAR methods show high performance on the UCF101 dataset, which is difficult to improve. Second, the ease of action detection in the UCF101 dataset decreases the effectiveness of the proposed method. The complexity of the HMDB51 dataset led to a more improvement in performance by adding a preprocessing step. Table 5 shows the 4% improvement achieved by the method suggested in [95] relative to the original method and shows that the proposed method is approximately 1.5 times better than the one proposed in [95].

The total percentage improvement achieved by the proposed method in all studied HAR methods was equal to 97.7%, which is approximately 42.4% more than that the one achieved by the method proposed in [95], which led to an improvement of 55.3%, confirming the superior efficiency of the proposed method. The main reason for the better performance of the proposed method is its spatial-tempora saliency map. In [95], the saliency map uses only a simple gradient operator; however, in the proposed method, a combination of features is used to produce the saliency map. Finally, the proposed method can be added to the current HAR systems that do not select the salient regions in their input to improve their accuracy.

Table 3 Accuracy obtained by the studied current HAR systems with and without (Baseline) the integration of the proposed method through (8) and (9) (best values in bold)

Method		UCF101			HMDB 51		
		Baseline	Equation 8	Equation 9	Baseline	Equation 8	Equation 9
Two-stream I3D [91]	Train	93.5	95.8	**95.9**	72.1	75.9	**78.6**
	Test	92.5	93.1	**93.6**	65.2	67.0	**69.7**
Motion guided network [92]	Train	96.4	97.1	**97.3**	72.3	75.9	**78.5**
	Test	94.1	95.2	**96.0**	68	69.9	**72.4**
Spatiotemporal network [93]	Train	94.7	97.8	**98.1**	67.3	69.0	**73.7**
	Test	93.8	95.0	**95.3**	66.4	66.9	**68.9**
Correlation net [94]	Train	96	97.2	**98.1**	73	73.1	**78.6**
	Test	92.8	94.9	**95.7**	68.1	70.2	**71.3**

Table 4 Comparison of the original HAR method (Baseline) and the proposed method, along with the one suggested in [95], in terms of efficiency (best values in bold)

Method		UCF101			HMDB 51		
		Baseline	[95]	Proposed	Baseline	[95]	Proposed
Two-stream I3D [91]	Train	93.5	95.6	**95.9**	72.1	76.1	**78.6**
	Test	92.5	93.1	**93.6**	65.2	68.5	**69.7**
Motion guided network [92]	Train	96.4	97.2	**97.3**	72.3	77.4	**78.5**
	Test	94.1	95.1	**96.0**	68	70.3	**72.4**
Spatiotemporal network [93]	Train	94.7	97.4	**98.1**	67.3	69.8	**73.7**
	Test	93.8	94.7	**95.3**	66.4	67.4	**68.9**
Correlation net [94]	Train	96	98	**98.1**	73	74.7	**78.6**
	Test	92.8	95.2	**95.7**	68.1	70.7	**71.3**
Model-agnostic multi-domain learning [96]	Train	96.03	96.49	**96.76**	74.59	75.88	**76.96**
	Test	94.94	95.01	**95.35**	74.62	75.01	**76.27**
TCLR [97]	Train	87.92	88.22	**88.4**	59.84	60.26	**62.03**
	Test	86.83	87.14	**87.53**	59.87	60.65	**62.15**
HAR-depth [98]	Train	92.56	92.98	**93.35**	69.61	71.68	**71.84**
	Test	92.6	92.62	**92.96**	68.5	68.96	**71.36**

Table 5 Percentage improvement of the proposed method, along with the one suggested in [95], compared to the original HAR method (Baseline) in terms of efficiency (best values in bold)

Method		UCF101		HMDB 51	
		[95]	Proposed	[95]	Proposed
Two-stream I3D [91]	Train	2.25	**2.52**	5.55	**9.01**
	Test	0.65	**1.18**	5.06	**6.89**
Motion guided network [92]	Train	0.83	**0.90**	7.05	**8.53**
	Test	1.06	**1.98**	3.38	**6.47**
Spatiotemporal network [93]	Train	2.85	**3.61**	3.71	**9.48**
	Test	0.96	**1.65**	1.51	**3.76**
Correlation net [94]	Train	2.08	**2.20**	2.33	**7.73**
	Test	2.59	**3.11**	3.82	**4.74**
Model-agnostic multi-domain learning [96]	Train	0.48	**0.76**	1.73	**3.18**
	Test	0.07	**0.43**	0.52	**2.21**
TCLR [97]	Train	0.34	**0.55**	0.7	**3.66**
	Test	0.36	**0.81**	1.30	**3.81**
HAR-depth [98]	Train	0.45	**0.85**	2.97	**3.20**
	Test	0.02	**0.39**	0.67	**4.18**

4 Conclusion

This study proposed a VSOD method based on a time-spatial saliency map. A weighted nonlinear combination of two saliency maps is used in the proposed method. A different scheme is used to obtain each saliency map. With an image registration step, the motion of the used camera is detected in the proposed method, and the non-overlapping regions of each of the two consecutive video frames are removed. Spatial, color, texture, and shape based features are used to extract the spatial saliency map. An optical flow algorithm, which is one of the most common approaches used in motion detection, is employed to extract the temporal saliency map after the image registration step. A color-based image segmentation method is used because the selected optical flow method only detects the boundaries of the moving objects in the video frame. Subsequently, the optical flow energy of each detected segment is considered as the temporal saliency of that segment. To maximize the efficiency of the proposed method, which combines temporal and spatial saliency maps, a function with seven degrees of freedom was implemented and optimized on two labeled datasets using two different objective functions. In summary, the main contributions of the proposed method are as follows: first, the use of several spatial saliency maps containing spatial, color, edge, and frequency based features to obtain a comprehensive expression of the ROIs of each image frame and the integration of spatial and temporal saliency maps by a nonlinear function, and second, the removal of the motion of the scene and camera to enhance the detection accuracy. The final VSOD method was added to four different current HAR system as a preprocessing step. The performance of each HAR system was evaluated before and after using the proposed method. The results showed that the proposed method increased the accuracy of the studied HAR systems. Future research can focus on speeding up the preprocessing step or suggesting a saliency map for grey-scale videos. Another future work could be to compare the impact of state-of-the-art VSOD methods on the accuracy of various HAR systems and to study saliency maps in videos with multiple motions.

Acknowledgements This article is a result of the projects Safe Cities—"Inovação para Construir Cidades Seguras" (reference POCI-01-0247-FEDER-041435) and Sensitive Industry, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement. The second author would like to thank "Fundação para a Ciência e Tecnologia (FCT)", in Portugal, for his PhD grant with reference 2021.08660.BD.

Funding Open access funding provided by FCT|FCCN (b-on).

Data Availability The used datasets are available in publicly available repositories, and the correspondent URLs are provided as footnotes in Section 3.1.

Compliance with ethical standards

Conflicts of interest The authors declare no conflict of interest.

References

1. Walther D (2006) Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics, California Institute of Technology
2. Hajihashemi V, Pakizeh E (2016) Human activity recognition in videos based on a two levels k-means and hierarchical codebooks. Int J Mechatron, Electr Comput Technol
3. Song X, Lan C, Zeng W, Xing J, Sun X, Yang J (2019) Temporal-spatial mapping for action recognition. IEEE Trans Circuits Syst Video Technol 30(3):748–759
4. Deshpnande A, Warhade KK (2021) An improved model for human activity recognition by integrated feature approach and optimized SVM. In: 2021 International conference on emerging smart computing and informatics (ESCI). IEEE, pp 571–576
5. Cong R, Lei J, Fu H, Cheng MM, Lin W, Huang Q (2018) Review of visual saliency detection with comprehensive information. IEEE Trans Circuits Syst Video Technol 29(10):2941–2959
6. Gupta AK, Seal A, Prasad M, Khanna P (2020) Salient object detection techniques in computer vision–a survey. Entropy 22(10):1174
7. Wang Q, Yuan Y, Yan P, Li X (2013) Saliency detection by multiple-instance learning. IEEE Trans Cybern 43(2):660–672
8. Li G, Xie Y, Wei T, Wang K, Lin L (2018) Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3243–3252
9. Sun M, Zhou Z, Hu Q, Wang Z, Jiang J (2018) SG-FCN: a motion and memory-based deep learning model for video saliency detection. IEEE Trans Cybern 49(8):2900–2911
10. Lee S, Jang D, Jeong J, Ryu ES (2019) "Motion-constrained tile set based 360-degree video streaming using saliency map prediction. In: Proceedings of the 29th ACM workshop on network and operating systems support for digital audio and video, pp 20–24
11. Li H, Chen G, Li G, Yu Y (2019) Motion guided attention for video salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7274–7283
12. Yan P, Li G, Xie Y, Li Z, Wang C, Chen T, Lin L (2019) Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7284–7293

13. Fan DP, Wang W, Cheng MM, Shen J (2019) Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8554–8564
14. Yang J, Fang X, Zhang L, Lu H, Wei G (2020) Salient object detection via double random walks with dual restarts. Image Vis Comput 93:103822
15. Liu F, Zhao L, Cheng X, Dai Q, Shi X, Qiao J (2020) Fine-grained action recognition by motion saliency and mid-level patches. Appl Sci 10(8):2811
16. Gu Y, Wang L, Wang Z, Liu Y, Cheng MM, Lu SP (2020) Pyramid constrained self-attention network for fast video salient object detection. Proceedings of the AAAI conference on artificial intelligence 34:10869–10876
17. Ji Y, Zhang H, Zhang Z, Liu M (2021) CNN-based encoder-decoder networks for salient object detection: a comprehensive review and recent advances. Inf Sci 546:835–857
18. Kousik N, Natarajan Y, Raja RA, Kallam S, Patan R, Gandomi AH (2021) Improved salient object detection using hybrid convolution recurrent neural network. Expert Syst Appl 166:114064
19. Zong M, Wang R, Chen X, Chen Z, Gong Y (2021) Motion saliency based multi-stream multiplier resnets for action recognition. Image Vis Comput 107:104108
20. Ji Y, Zhang H, Jie Z, Ma L, Wu QJ (2020) CASNet: a cross-attention Siamese network for video salient object detection. IEEE Trans Neural Networks Learn Syst 32(6):2676–2690
21. Zhang M, Liu J, Wang Y, Piao Y, Yao S, Ji W, Li J, Lu H, Luo Z (2021) Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1553–1563
22. Wang Q, Liu Y, Xiong Z, Yuan Y (2022) Hybrid feature aligned network for salient object detection in optical remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–15
23. Liu Y, Xiong Z, Yuan Y, Wang Q (2023) Transcending pixels: boosting saliency detection via scene understanding from aerial imagery. IEEE Trans Geosci Remote Sens
24. Liu Y, Xiong Z, Yuan Y, Wang Q (2023) Distilling knowledge from super resolution for efficient remote sensing salient object detection. IEEE Trans Geosci Remote Sens
25. Alavigharahbagh A, Hajihashemi V, Machado JJ, Tavares JM (2023) Deep learning approach for human action recognition using a time saliency map based on motion features considering camera movement and shot in video image sequences. Information 14(11):616
26. Liu Y, Li Q, Yuan Y, Du Q, Wang Q (2021) ABNet: adaptive balanced network for multiscale object detection in remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–14
27. Vijayan M, Ramasundaram M (2019) A fast DGPSO-motion saliency map based moving object detection. Multimed Tools Appl 78(6):7055–7075
28. Huang T, McKenna S (2018) Sequential recognition of manipulation actions using discriminative superpixel group mining. In: 2018 25th IEEE International conference on image processing (ICIP). IEEE, pp 579–583
29. Mahapatra D, Winkler S, Yen SC (2008) Motion saliency outweighs other low-level features while watching videos. In: Human vision and electronic imaging XIII, vol 6806. SPIE, pp 246–255
30. Lee I, Ban SW, Fukushima K, Lee M (2006) Selective motion analysis based on dynamic visual saliency map model. In: International conference on artificial intelligence and soft computing. Springer, pp 814–822
31. Jeong S, Ban SW, Lee M (2008) Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. Neural Netw 21(10):1420–1430
32. 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
33. Woo JW, Lim YC, Lee M (2009) Obstacle categorization based on hybridizing global and local features. In: International conference on neural information processing. Springer, pp 1–10
34. Kim S, Kim M (2014) Improvement of saliency map using motion information. In: Proceedings of the Korean society of broadcast engineers conference. The Korean Institute of Broadcast and Media Engineers, pp 259–260
35. Morita S (2008) Generating saliency map related to motion based on self-organized feature extracting. In: International conference on neural information processing. Springer, pp 784–791
36. Morita S (2009) Generating self-organized saliency map based on color and motion. In: International conference on neural information processing. Springer, pp 28–37
37. Hu J, Pitsianis N, Sun X Motion saliency map generations for video data analysis: spatio-temporalsignatures in the array operations
38. Mejía-Ocaña AB, De Frutos-López M, Sanz-Rodríguez S, del Ama-Esteban Ó, Peláez-Moreno C, Díaz-de María F (2011) Low-complexity motion-based saliency map estimation for perceptual video coding. IEEE

39. Gkamas T, Nikou C (2011) Guiding optical flow estimation using superpixels. In: 2011 17th International Conference on Digital Signal Processing (DSP). IEEE, pp 1–6
40. Li WT, Chang HS, Lien KC, Chang HT, Wang YC (2013) Exploring visual and motion saliency for automatic video object extraction. IEEE Trans Image Process 22(7):2600–2610
41. Chang HS, Wang YC (2013) Superpixel-based large displacement optical flow. In: 2013 IEEE international conference on image processing, pp 3835–3839
42. Huang CR, Chang YJ, Yang ZX, Lin YY (2014) Video saliency map detection by dominant camera motion removal. IEEE Trans Circuits Syst Video Technol 24(8):1336–1349
43. Dong X, Tsoi AC, Lo SL (2014) Superpixel appearance and motion descriptors for action recognition. In: 2014 International joint conference on neural networks (IJCNN). IEEE, pp 1173–1178
44. Giosan I, Nedevschi S (2014) Superpixel-based obstacle segmentation from dense stereo urban traffic scenarios using intensity, depth and optical flow information. In: 17th International IEEE conference on intelligent transportation systems (ITSC). IEEE, pp 1662–1668
45. Roberts R, Dellaert F (2014) Direct superpixel labeling for mobile robot navigation using learned general optical flow templates. In: 2014 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 1032–1037
46. Xu J, Tu Q, Li C, Gao R, Men A (2015) Video saliency map detection based on global motion estimation. In: 2015 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, pp 1–6
47. Srivatsa RS, Babu RV (2015) Salient object detection via objectness measure. In: 2015 IEEE international conference on image processing (ICIP). IEEE, pp 4481–4485
48. Donné S, Aelterman J, Goossens B, Philips W (2015) Fast and robust variational optical flow for high-resolution images using slic superpixels. In: International conference on advanced concepts for intelligent vision systems. Springer, pp 205–216
49. Li J, Liu Z, Zhang X, Le Meur O, Shen L (2015) Spatiotemporal saliency detection based on superpixel-level trajectory. Signal Process Image Commun 38:100–114
50. Hu Y, Song R, Li Y, Rao P, Wang Y (2016) Highly accurate optical flow estimation on superpixel tree. Image Vis Comput 52:167–177
51. Guo J, Ren T, Huang L, Liu X, Cheng MM, Wu G (2017) Video salient object detection via cross-frame cellular automata. In: 2017 IEEE international conference on multimedia and expo (ICME). IEEE, pp 325–330
52. Tu Z, Guo Z, Xie W, Yan M, Veltkamp RC, Li B, Yuan J (2017) Fusing disparate object signatures for salient object detection in video. Pattern Recognit 72:285–299
53. Hu YT, Huang JB, Schwing AG (2018) Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: Proceedings of the European conference on computer vision (ECCV), pp 786–802
54. Ling Q, Deng S, Li F, Huang Q, Li X (2016) A feedback-based robust video stabilization method for traffic videos. IEEE Trans Circuits Syst Video Technol 28(3):561–572
55. Wang J, Liu W, Xing W, Zhang S (2018) Visual object tracking with multi-scale superpixels and color-feature guided kernelized correlation filters. Signal Process Image Commun 63:44–62
56. Chen R, Tong Y, Yang J, Wu M (2019) Video foreground detection algorithm based on fast principal component pursuit and motion saliency. Comput Intell Neurosci 2019
57. Maczyta L, Bouthemy P, Le Meur O (2019) Unsupervised motion saliency map estimation based on optical flow inpainting. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 4469–4473
58. Zhu H, Sun X, Zhang Q, Wang Q, Robles-Kelly A, Li H, You S (2019) Full view optical flow estimation leveraged from light field superpixel. IEEE Trans Comput Imaging 6:12–23
59. Kim C, Song D, Kim CS, Park SK (2019) Object tracking under large motion: combining coarse-to-fine search with superpixels. Inf Sci 480:194–210
60. Ngo TT, Nguyen V, Pham XQ, Hossain MA, Huh EN (2020) Motion saliency detection for surveillance systems using streaming dynamic mode decomposition. Symmetry 12(9):1397
61. Qiu G, Wang Y, Wei Y (2020) An algorithm for the hole filling of motion foreground based on superpixel segmentation. In: 2020 International conference on communications, information system and computer engineering (CISCE). IEEE, pp 450–453
62. Tian H, Cai W, Ding W, Liang P, Yu J, Huang Q (2023) Long-term liver lesion tracking in contrast-enhanced ultrasound videos via a siamese network with temporal motion attention. Front Physiol 14
63. Bay H, Tuytelaars T, Van Gool L (2006) “SURF: speeded up robust features. In: European conference on computer vision. Springer, pp 404–417
64. Kim J, Han D, Tai YW, 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

65. Nan B, Mu Z (2014) Slic0-based superpixel segmentation method with texture fusion. Chin J Sci Instrum 35(3):527–534
66. Hetherington R (1952) The perception of the visual world. by James J. Gibson. USA: Houghton mifflin company, 1950 (George Allen & Unwin, Ltd., London). price 35s. J Mental Sci 98(413):717–717
67. Gibson JJ, Carmichael L (1966) The senses considered as perceptual systems, vol 2. Houghton Mifflin, Boston
68. Barron JL, Fleet DJ, Beauchemin SS (1994) Performance of optical flow techniques. Int J Comput Vis 12(1):43–77
69. Bronshtein IN, Semendyayev KA (2013) Handbook of mathematics. Springer
70. Horn BK, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203
71. Brox T (2020) Optical flow: traditional approaches. In: Computer vision: a reference guide, pp 1–5
72. Bensaci R, Khaldi B, Aiadi O, Benchabana A (2021) Deep convolutional neural network with KNN regression for automatic image annotation. Appl Sci 11(21):10176
73. Wan S, Prusinkiewicz P, Wong S (1990) Variance-based color image quantization for frame buffer display. Color Res Appl 15(1):52–58
74. Floyd RW (1976) An adaptive algorithm for spatial gray-scale. Proceedings of the Society for Information Display 17:75–77
75. Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A, Van Gool L (2017) The 2017 Davis challenge on video object segmentation. arXiv:1704.00675
76. Chen J, Li Z, Jin Y, Ren D, Ling H (2021) Video saliency prediction via spatio-temporal reasoning. Neurocomputing 462:59–68
77. Chen C, Wang G, Peng C, Fang Y, Zhang D, Qin H (2021) Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. IEEE Trans Image Process 30:3995–4007
78. Huang X, Zhang YJ (2021) Fast video saliency detection via maximally stable region motion and object repeatability. IEEE Trans Multimedia
79. Shang J, Liu Y, Zhou H, Wang M (2021) Moving object properties-based video saliency detection. J Electron Imaging 30(2):023005
80. Rosten E, Drummond T (2005) Fusing points and lines for high performance tracking. In: 10th IEEE international conference on computer vision (ICCV'05) vol 1, vol 2. IEEE, pp 1508–1515
81. Harris C, Stephens M et al (1988) A combined corner and edge detector. In: Alvey vision conference, vol 15. Citeseer, pp 10–5244
82. Alcantarilla PF, Bartoli A, Davison AJ (2012) KAZE features. In: European conference on computer vision. Springer, pp 214–227
83. Shi J et al (1994) Good features to track. In: 1994 Proceedings of IEEE conference on computer vision and pattern recognition. IEEE, pp 593–600
84. Nistér D, Stewénius H (2008) Linear time maximally stable extremal regions. In: European conference on computer vision. Springer, pp 183–196
85. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision. IEEE, pp 2564–2571
86. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
87. Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 724–732
88. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. Springer, pp 363–370
89. Lucas BD, Kanade T et al (1981) An iterative image registration technique with an application to stereo vision, vol 81
90. Baker S, Matthews I (2004) Lucas-Kanade 20 years on: a unifying framework. Int J Comput Vis 56(3):221–255
91. Carreira J, Zisserman A (2017) Quo vadis, action recognition, a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
92. Zheng Z, An G, Ruan Q (2020) Motion guided feature-augmented network for action recognition. In: 2020 15th IEEE international conference on signal processing (ICSP), vol 1. IEEE, pp 391–394
93. Chen E, Bai X, Gao L, Tinega HC, Ding Y (2019) A spatiotemporal heterogeneous two-stream network for action recognition. IEEE Access 7:57267–57275
94. Yudistira N, Kurita T (2020) Correlation Net: spatiotemporal multimodal deep learning for action recognition. Signal Process Image Commun 82:115731
95. Gharahbagh AA, Hajihashemi V, Ferreira MC, Machado JJ, Tavares JMR (2022) Best frame selection to enhance training step efficiency in video-based human action recognition. Appl Sci 12(4):1830

96. Omi K, Kimata J, Tamaki T (2022) Model-agnostic multi-domain learning with domain-specific adapters for action recognition. IEICE Trans Inf Syst 105(12):2119–2126
97. Dave I, Gupta R, Rizve MN, Shah M (2022) TCLR: temporal contrastive learning for video representation. Comput Vis Image Understand 219:103406
98. Sahoo SP, Ari S, Mahapatra K, Mohanty SP (2020) HAR-depth: a novel framework for human action recognition using sequential learning and depth estimated history images. IEEE Trans Emerg Top Comput Intell 5(5):813–825

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Abdorreza Alavi Gharahbagh[1] · Vahid Hajihashemi[1] · Marta Campos Ferreira[1] · J.J.M. Machado[2] · João Manuel R.S. Tavares[2]

Abdorreza Alavi Gharahbagh
abalavi.gh@gmail.com

Vahid Hajihashemi
Hajihashemi.vahid@ieee.org

Marta Campos Ferreira
mferreira@fe.up.pt

J.J.M. Machado
jjmm@fe.up.pt

[1] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, Porto 4200-465, Portugal

[2] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, Porto 4200-465, Portugal