Cogn Comput (2016) 8:629–647
DOI 10.1007/s12559-016-9387-7

A New Spatio-Temporal Saliency-Based Video Object Segmentation

Zhengzheng Tu[1] · Andrew Abel[2] · Lei Zhang[1] · Bin Luo[1] · Amir Hussain[1,2]

Received: 17 September 2015 / Accepted: 19 February 2016 / Published online: 8 March 2016

Abstract Humans and animals are able to segment visual scenes by having the natural cognitive ability to quickly identify salient objects in both static and dynamic scenes. In this paper, we present a new spatio-temporal-based approach to video object segmentation that considers both motion- and image-based saliency to produce a weighted approach which can segment both static and dynamic objects. We perform fast optical flow and then calculate the motion saliency based on this temporal information, detecting the presence of global motion and adjusting the initial optical flow results accordingly. This is then fused with a region-based contrast image saliency method, with both techniques weighted. Finally, our joint weighted saliency map is used as part of a foreground–background labelling approach to produce the final segmented video files. Good results in a wide range of environments are presented, showing that our spatio-temporal system is more robust and consistent than a number of other state-of-the-art approaches.

✉ Bin Luo
luobinahu@163.com; luobin@ahu.edu.cn

Zhengzheng Tu
zhengzhengahu@163.com

Andrew Abel
aka@cs.stir.ac.uk

Lei Zhang
lzhang15@foxmail.com

Amir Hussain
ahu@cs.stir.ac.uk

[1] School of Computer Science and Technology, Anhui University, Hefei 230601, China

[2] Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, Scotland, UK

Keywords Video object segmentation · Global motion · Spatio-temporal saliency · Foreground–background labelling

Introduction

Video object segmentation is a very important subject of interest. The overall research domain of video analysis can be divided into a number of individual research focuses, including moving object detection [1], saliency detection [2], and video object segmentation [3], but in this research we are most interested in video object segmentation. The purpose of this is to segment an image into regions of coherent groupings that can be used to separate the foreground of an object from the background. To improve the accuracy of this, both appearance and motion information can be exploited [3–8]. In order to segment the objects which are moving in a discontinuous manner, [5] proposed a novel approach using dense point trajectories, which automatically distinguishes foreground objects from background based on trajectory clustering results. The method proposed by Wang et al. [6] incorporates saliency as a prior for object segmentation by computing robust geodesic measurement, and combines spatio-temporal saliency maps and appearance models with dynamic location models into an energy minimisation framework to attain both spatially and temporally coherent object segmentation. Other recent research [7, 8], focuses on the issue of object co-segmentation. Taking advantage of intra-frame saliency, inter-frame consistency, and across-video correspondence, [7] proposes a spatio-temporal scale-invariant feature transform (SIFT) flow descriptor to capture the relationship between foreground objects, and establishes an object discovery energy function to segment out the common foreground objects without imposing any prior constraints regarding their appearance and motion patterns. Using global and local energy optimisation, Dong et al. [8] propose a novel interactive co-segmentation method, which outperforms other state-of-the-art unsupervised or interactive co-segmentation methods.

Of further interest is how we as humans segment visual scenes, by having the natural cognitive ability to quickly be able to identify salient objects in both static and dynamic scenes. This is easy for us to do, but much more challenging for machines to be able to do in an unsupervised manner. By salient objects, we refer to those objects that have been defined as being the most visually noticeable foreground objects in a scene [4]. The visual selection method of humans is of great interest to many researchers in fields as diverse as psychology [9], computer vision [10], and neurobiology [11]. We are capable of identifying salient objects, quickly, but these salient regions do not always correspond simply to the area with the most motion, or the greatest contrast, intensity, or colour but are more complex than this. Machine-based approaches do not always match up well to human results with regard to saliency detection.

There are a variety of approaches for identifying saliency in a scene, and inspired by theories about human vision processing [12], these can be divided into slower top-down approaches, which consider high-level, task-dependent, scene processing [13, 14], and bottom-up approaches [2], which are considered to be task independent, are data-driven, and focus more on detecting saliency from image low-level features, such as intensity, location, texture, and many kinds of contrast. Top-down attention is considered to be task-driven intentional behaviour, and so top-down approaches require prior constraints, often supervised, and so in comparison with bottom-up approaches, top-down methods tend to be slower. There are a wide range of saliency detection approach applications, including image representation [15], video analysis [16], and cognitive robotics [17]. We discuss this in more depth in the background section.

The classic bottom-up saliency model [2] considers local contrast and central-surround image feature differences to define image saliency. However, as humans are not only influenced by static features, but also by motion features [18], video saliency approaches have also been developed [19, 20] to detect salient areas from the image series of video. Generally, they use not only static image saliency features but also temporal motion information and are consequently described as being spatio-temporal based.

In this paper, building on the foundations of preliminary moving object detection research by the authors [21], we present a new spatio-temporal-based approach to video object segmentation. By this, we mean that rather than consider purely motion saliency or image saliency as in other research work, we make use of both approaches in combination to produce an unsupervised weighted integrated approach, which can be used to identify foreground objects, and so segment a video by automatically extracting the relevant objects. Our approach firstly identifies motion information between successive frames by performing fast optical flow [22]. Rather than simply using the optical flow results as motion saliency, as is common, we present a novel approach to identify global motion and produce more accurate saliency. This is then fused in a weighted fusion with a region-based contrast (RC) [4] saliency method, which considers only spatial information within a single image frame. This produces an overall weighted combined grey-level saliency map, and thresholding is then applied to produce a binary mask. Finally, our weighed saliency map is used as part of a foreground–background labelling approach [3] to produce our final segmented video files. The key novel components here are the detection of global motion and calculation of the motion saliency, our use of the spatio-temporal saliency map as an overall feature to improve the final labelling result, detecting both static and dynamic objects, and the overall powerful, robust, and quick-fused state-of-the-art approach.

The results show that as well as being fast, and with good results in a wide range of environments, as we have employed novel global motion detection fused with image-based saliency, our spatio-temporal system is more robust than other approaches that make use of a single technique, delivering consistent results. We evaluate our approach with videos containing single and multiple objects, plain and dynamic backgrounds, moving and static camera position, and from multiple datasets, to demonstrate robustness in a wide range of scenarios. Finally, we show with detailed analysis that using both the contrast- and centre-based spatial techniques, fused with the temporal motion-based saliency, provides the basis of a good overall system with much potential for more general application.

The remainder of this paper is divided as follows. “Background Work” section introduces the background to this research, discussing relevant state-of-the-art and historic methods, which is followed by our fused motion and image saliency approach, fully described in “Proposed Fast Object Detection Approach” section. We present the results of thorough testing in “Experimental Results” section and then provide some additional discussion and analysis in “Discussion” section. Finally, “Summary and Future Work” section concludes the paper.

Background Work

Moving Object Detection

One key topic in video analysis is object detection, a topic which has been researched for many years. While there is some similarity to video object segmentation (which is discussed separately below), there are some significant differences. Object segmentation research focuses on the task of separating foreground objects from background in a video when the foreground objects include both moving and static objects, which presents its own challenges, whereas the aim of moving object detection is to identify objects that are distinct from the background, but change rapidly over time.

One classic approach used in this domain is background modelling and subtraction [1, 23, 24]. This approach functions by constructing a background model consisting of a probability density function of the intensity of each individual pixel for a scene, and this background model is then compared with each frame in a video sequence in order to identify significant differences between the background model and the frame. These differences can then be identified as areas where motion is present. In addition to this, the background model also needs to be gradually updated throughout the detection process to take account of potential scene changes over time. One powerful example of this approach is the ViBe algorithm [25], which is a pixel-based background subtraction technique that chooses background samples to build background models.

In addition to background modelling and subtraction, another very important and widely used detection approach is optical flow. This technique has a long history, starting from pioneering research by Horn and Schunck (H–S) [26] and Lucas and Kanade (L–K) [27] in 1981. This technique attempts to describe object motion by analysing brightness patterns between two consecutive video frames. As this approach is independent of features like colours and texture, it is not influenced by moving objects that are similar to the background. In addition to the pioneering work described above, there are also many new and enhanced approaches, such as [28–30], which enhance the accuracy of estimated motion fields. In our work, we adopt a fast optical flow method [22], which can handle large displacement motions. Previous experimental results [22] show that this method is significantly faster than other state-of-the-art methods while still producing good results, especially with scenes that contain a significant degree of motion.

There are other additional approaches to detecting moving objects; for example, a recent trend is to use saliency detection methods and models [31, 32], which has potential for aiding the locating and detection of objects.

Saliency Detection

As discussed in “Introduction” section, a salient object can be defined as being the most visually noticeable foreground object in a scene, and the classical visual saliency model for a single image (i.e. not considering video sequences) was proposed by Itti et al. [2]. Here, they considered local contrast and use biologically motivated centre-surround image feature differences to define the image saliency.

Following this initial research, many recent approaches have been proposed to detect image saliency in a broad range of applications. This includes image representation [15], video analysis [16], object-of-interest image segmentation [33], object recognition [34], adaptive compression of images [35], and image retrieval [36]. In addition to this, there has been much recent research into saliency detection, with a number of models developed, including graph-based visual saliency(GBVS) [37], which was reported [37] to produce better results with regard to human vision fixation than those reported by Itti et al. Zhang et al. proposed the SUN model [38], a Bayesian framework for saliency detection. Given the visual features observed, this approach assumes that visual saliency is the probability of a target at every location and this probability is estimated by computing feature response maps on an image set of natural scenes. Also, Cheng et al. [4] proposed a Region Contrast (RC)-based salient object detection algorithm, which simultaneously evaluates global contrast differences and spatial weighted coherence cores, and can produce full-resolution, high-quality saliency maps(RC model). In addition to these key works, a wide range of other methods have been proposed [39–42].

Although the centre-surround-based approaches first proposed by Itti et al. [2] were proven to be effective, some methods utilise additional foreground and background information for helping detect saliency. An example of this is provided by Xie et al. [39], who identify that the foreground location is another important prior for modelling salient regions, so employ the convex hull of interesting points to estimate foreground location. Wei et al. [40] suggested that background priors are equally important for saliency detection. Considering a natural image can theoretically be decomposed into a distinctive salient foreground and a homogeneous background, other research presented in [41 42], utilises fresh compressive sensing techniques such as low-rank and sparse matrix decomposition methods to detect salient image features.

Video Object Segmentation

In video object segmentation, appearance and motion cues are generally employed in order to provide the most accurate video segmentation [3, 4, 43]. One state-of-the-art approach is SaliencyCut [4], an enhancement of GrabCut [44] (which considers edge and texture information to minimise an energy function, and operates by allowing a user to drag a rectangle around an area to extract an object). SaliencyCut was developed by Cheng et al. [4] and uses the

RC model discussed previously to generate an initial saliency map. This map can then be used as the initial location selection to perform accurate segmentation with GrabCut. Papazoglou et al. [3] propose an approach that produces a rough estimate of the motion cues, which are then combined with an appearance model based on the initial foreground estimate to perform object segmentation. However, the results showed that while this approach was very fast, performance was not ideal when attempting to segment static or very small objects. Another approach was developed by Li et al. [43], who presented a saliency-based video object extraction framework. A conditional random field (CRF) [45] is then applied to effectively integrate the associated features (i.e. shape, visual saliency, foreground/background colour models, and spatial/temporal energy terms) [43]. The CRF [45] is a powerful technique to estimate the structural information (e.g. class label) of a set of variables with the associated observations. For video foreground object segmentation, the CRF was applied in this case to predict the label of each observed pixel in an image, by solving a CRF optimisation problem. However, the downside of this approach is that it is time-consuming.

In addition to these approaches, there are also additional techniques that have been proposed. Zhang et al. [46] proposed a novel layered directed acyclic graph (DAG)-based framework to extract primary object segments in videos in the 'object proposal' domain. The extracted primary object regions are then used to build object models for optimised video segmentation. Li et al. [47] proposed Sub-Optimal Low-rank Decomposition (SOLD) and, based on this, developed an effective and efficient framework that automatically segments streaming videos in an unsupervised and interactive way.

To evaluate the performance of our approach, in "Experimental Results" section, we compare the results produced by our approach to a number of the approaches reviewed above. This includes [3, 4, 43, 46]. In addition to this, we also compare our approach with ViBe [25], which is a classic moving object detection method, and RC [4], which is an image saliency-based object segmentation algorithm.

Main Contributions

The main contributions of this paper are firstly that rather than simply taking the optical flow intensity as the motion saliency, as is commonly the case, we present a novel approach to identify global motion and produce what is therefore arguably a much more accurate and relevant saliency map. Secondly, as one of the inputs of the final labelling step, fusing the spatial saliency map and motion saliency map into an overall feature map can segment both static and dynamic objects. This improves the object segmentation results in comparison with using methods that utilise only a single saliency map, either spatial alone or motion alone.

Proposed Fast Object Detection Approach

Approach Overview

As the overall aim of our system is to present a fast and accurate object detection approach for video files, there are a number of important stages. These are shown in the system diagram given in Fig. 1.

Initially, the input video file is divided into individual frames and is then processed in two ways. Firstly, we perform fast optical flow [22] and then calculate motion saliency using the results, taking account of global motion. This provides us with an initial saliency map. However, to improve results, we also fuse this with an image-only saliency calculation approach, using RC saliency [4].

After calculating both the RC saliency and adjusted motion saliency for each individual image frame, the two resulting image masks are then weighted and fused to produce a single overall saliency map for each frame in a video. This is then thresholded to output a binary map. The final step of our process is to then use the binary image map and then label parts of each image frame as being either

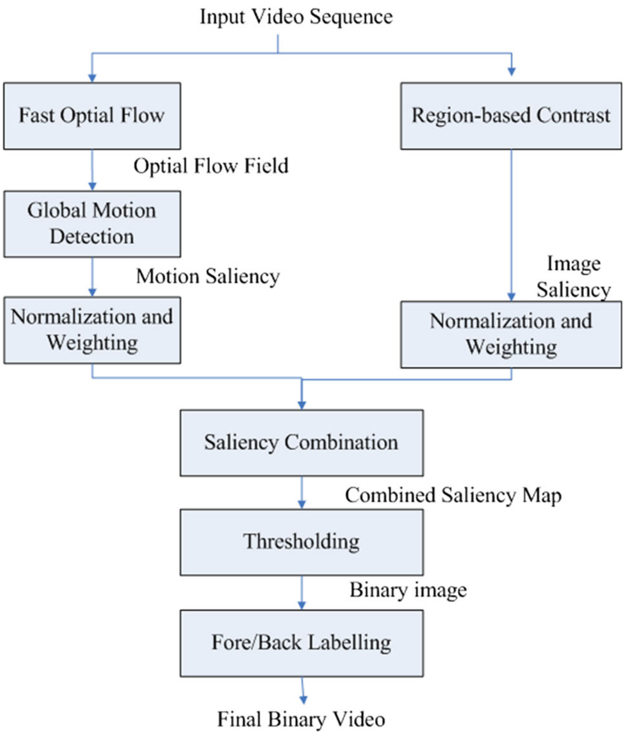


Fig. 1 Overall diagram for our approach

foreground or background. For this, we use a state-of-the-art labelling approach proposed by Papazoglou and Ferrari [3], to produce our final salient object regions. An example of the result of each stage is shown in Fig. 2.

Figure 2 shows two example frames: a car (left) and a fox (right). The top row shows the original images, followed by the initial optical flow intensity results in the second row. However, it can be seen that in this case, the optical flow is darker on the object of interest, indicating that more motion is present in the background. As global motion is therefore determined to be present, the image frame is recalculated with our approach, resulting in an improved region of interest in (c), which is then used as the motion saliency input. On the fourth row of this figure, (d) shows the image saliency (RC) result. (c) and (d) are then weighted and combined to produce the fused saliency maps in (e). Finally, after foreground–background labelling has been applied, the final output of our approach is shown in the bottom row of figure (f).

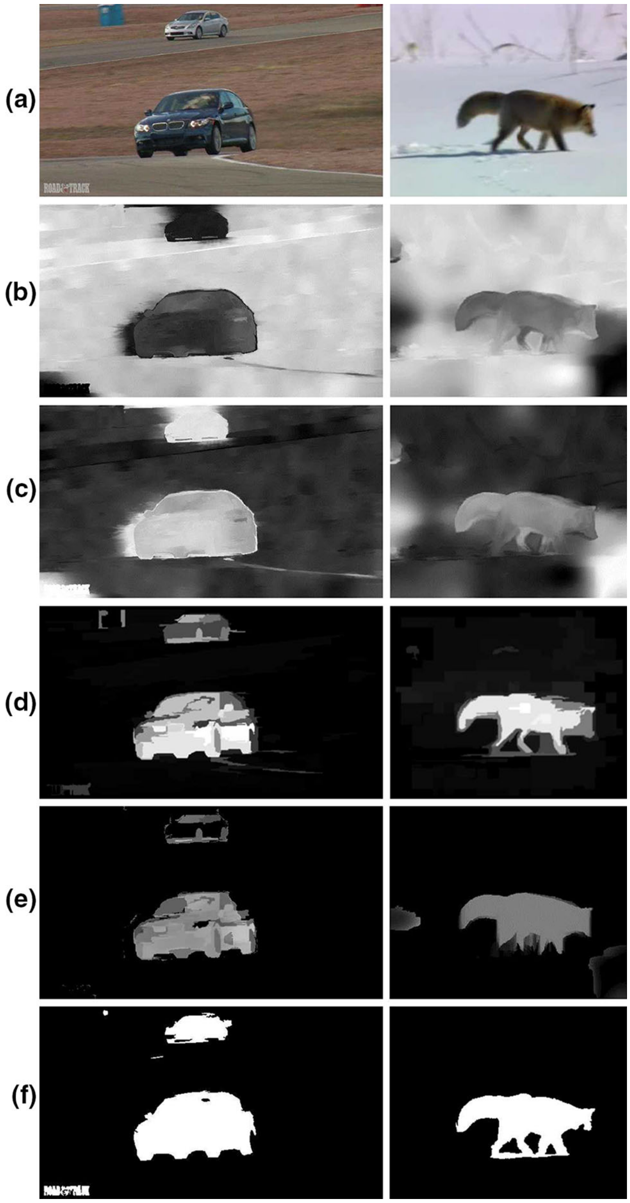


Fig. 2 Example of each stage of our technique, showing example frames for car and fox. **a** The original images, **b** the initial optical flow results, **c** the affect of our motion map adjustment, **d** the image-only saliency, **e** the fused saliency map, using both image and motion saliency, and **f** the final system output, using foreground–background labelling

Fast Optical Flow

The initial step of our method is to calculate the optical flow between successive pairs of frames. Since the first optical flow approaches devised by Horn–Schunck (the global model) [26] and Lucas–Kanade (the local model) [27], there have been many different versions of optical flow. In this work, we wish to make use of a state-of-the-art fast optical flow method in order to calculate the flow as quickly as possible. We therefore make use of Fast Edge-Preserving PatchMatch for Large Displacement Optical Flow, as proposed by Bao et al. [22].

In their work, Bao et al. [22] identify that one issue that traditional optical flow has is the handling of larger motions (such as those where there is camera movement). They devise a local method (i.e. one that does not involve optimisation across the whole image) for reasons of speed, with the use of an approximate nearest neighbour field (NNF) in order to estimate large displacement. A NNF is defined as a correspondence field that identifies the closest image patch (i.e. groups of pixels) pairings between two images, with the benefit of there being no limitation on the distance, which is therefore identified as being a good way of handling large displacement motion. The 'approximate' part of the description is because while NNF is described as being very computationally expensive to calculate exactly, algorithms exist which make use of approximate approaches. [48, 49].

The approach of Bao et al. [22] follows a traditional local correspondence searching framework [50], consisting of four steps, firstly, matching cost computation, then cost aggregation, correspondence selection, and finally, refinement. They propose a modified version of PatchMatch [48] that aims to preserve edges in a much better manner. They also developed an approximation algorithm to control pixel selection (i.e. ignoring dissimilar pixels). More details can be found at [48]. In addition to this, they also reduce the computational complexity by downsampling the initial images to a lower resolution (generally, they downsize twice) and performing NNF on the lower resolution. This can be used to provide a coarse NNF on higher-resolution images to reduce computational time. They also perform forward–backward consistency checking between consecutive NNF results and smooth small outliers.

We use this approach to perform optical flow on our input image sequence I, taken from the video sequence consisting of T frames, and generate our optical flow vectors $\bar{U}_t$ (for vertical optical flow motion) and U_t (for horizontal), for pairwise image frames I_t, I_{t+1}.

Global Motion Detection and Motion Saliency

One recent development in the field has been the adoption of the classical optical flow algorithm to generate motion saliency. [51, 52]. Here, the theory is that image regions or objects with strong motion must therefore have temporal saliency. Existing work exploits this to define the magnitude of the temporal saliency map as being equal to the intensity of optical flows. This map is then fused with other spatial saliency maps to construct final maps of overall saliency. However, while it is true that optical flow vectors are capable of providing a good description of motion field intensity, we do not feel that stronger motion necessarily always represents temporal saliency, and so in our approach, we take account of global motion.

This is because in some complex scenes, such as those with dynamic backgrounds, a lot of redundant motion exists in the background, which is picked up by the optical flow. For example, in a natural scene, there could be tree branches moving in the wind, water flow, and other redundant motion that is not part of the foreground, but are tracked by optical flow. Using the conventional approach described above can identify these redundant motions as being salient. In addition, there is the issue of camera movement or shaking that may result in the appearance of stronger background motion. However, it is our opinion, as also argued by Pylyshyn [53], that human eyes always focus on foreground objects [54], even those that have less strong motion than some of the redundant background information.

The motion intensity Θ of an optical flow vector at location (m, n) can be calculated as:

$$\Theta(m,n) = \sqrt{U(m,n)^2 + \bar{U}(m,n)^2} \tag{1}$$

where U and $\bar{U}$ represent the horizontal and vertical motion components of optical flow, calculated during the optical flow process. The issue is to determine whether global motion is present in an image frame and therefore process the frame differently depending on this. We define global motion as being scenarios where the camera is moving, shaking, or zooming, following the definition provided by Dufaux and Konrad [55]. This means that because the camera is being moved, by definition, all objects in the camera will also appear to be moving. In these scenarios, the motion intensity is not appropriate for use to as temporal saliency. An example of this is in Fig. 2b. Here, the car object (left hand image) optical flow intensity is smaller than areas of the surrounding background, which would affect the saliency results if used in this format.

We have therefore created a new approach to identify if global motion is present in a frame and therefore adjust the image accordingly. Using the motion intensity calculated from Eq. 1 for each pixel, we identify the maximum motion intensity of an individual frame t from all pixels $\Theta_\beta = \max(\Theta)$ and define a threshold $\Theta_{\bar{\beta}}$ as being half of this value:

$$\Theta_{\bar{\beta}} = 0.5 \times \max(\Theta) \tag{2}$$

That is, $\Theta_{\bar{\beta}} = \omega_1 \times \Theta_\beta$, (here $\omega_1 = 0.5$).

Given the value calculated by Eq. (2), for each frame, we can then determine whether global motion is present. For this, if more than 50 % ($\omega_2 = 50\,\%$) of optical flow vectors in the tth frame have a motion intensity greater than $\Theta_{\bar{\beta}}$, then it is decided that global motion exists. If global motion is not present, then the motion intensity calculated in Eq. (1) is used unchanged to determine motion saliency. However, if global motion is determined to be present, then for all pixels in a frame, motion saliency is calculated as:

$$\Theta(m,n) = \Theta_\beta - \sqrt{U(m,n)^2 + \bar{U}(m,n)^2} \tag{3}$$

The difference between the original optical flow results in Fig. 2b and the global motion calculated results is shown in Fig. 2c, showing here that moving objects, particularly in the car, are made much more prominent. After using either Eq. (3) or (1), the overall motion saliency, Θ_t, for each tth image in the video sequence is then normalised and can then be fused with RC saliency.

Region-Based Contrast

In addition to considering motion with our global motion approach, we fuse it with a spatial saliency detection technique, designed to work for a single image, rather than across multiple video frames. We hypothesise that this approach will identify different details from motion saliency and thus add flexibility. RC is a state-of-the-art approach for salient region detection, developed by Cheng et al. [4]. They argue that humans tend to pay more attention to high-contrast regions of their vision field, and in particular, high-contrast regions surrounding the salient object are of importance. They therefore developed a bottom-up data-driven saliency extraction technique.

Since it can be very computationally expensive with a bottom-up approach to introduce spatial relationships for each pixel, the RC approach first segments an image into regions, using a graph-based region segmentation [56]. It then computes a colour histogram for each region. In a true colour image, there may be a very large number of possible colours for each pixel (256^3), and so to boost computational efficiency and speed of calculation further, they quantise each colour channel to use only 12 values for each red–green–blue (RGB) channel, resulting in only 12^3 possible colours. By then choosing only the most frequently used colours that cover 95 % of pixels in an image, the number of colours can be reduced to approximately 85, with the remaining pixels being replaced by the closest matching colours in the histogram.

For a region $c_{\hat{j}}$ of an image I_t, containing $\hat{J}$ regions, colour contrast and spatial information can be used to calculate the saliency of a region $\hat{\Theta}\left(c_{\hat{j}}\right)$ as follows:

$$\hat{\Theta}\left(c_{\hat{j}}\right) = w_s(c_{\hat{j}}) \sum_{c_{\hat{k}} \neq c_{\hat{j}}} e^{\frac{D_s(c_{\hat{k}}, c_{\hat{j}})}{\phi^2}} w_c(c_{\hat{j}}) D_c(c_{\hat{k}}, c_{\hat{j}}) \tag{4}$$

Here, firstly, Cheng et al. [4] measure the $\hat{j}$th region in comparison with all other regions present in the image, and for each alternative region $c_{\hat{k}}$, they also add two weights. Firstly, the spatial prior weight, $w_s(c_{\hat{j}})$, is described as being similar to centre bias [57] and is calculated by Cheng et al. [4] as being $-\exp(-9D_j^2)$, with D_j being the average distance between the centre of the image and the pixels in the jth region. This essentially helps to ensure a higher value if the region is close to the centre of an image, and a lower value if the region is closer to the image border. In addition to this, they also make use of region weighting, $w_c(c_{\hat{j}})$, which weights the region by the number of pixels, in order to emphasise the contrast in larger regions.

To take account of spatial information, in addition to the spatial prior weighting, they also use the spatial distance between regions, $D_s(c_{\hat{k}}, c_{\hat{j}})$, and a spatial distance weight, ϕ, which affects the effect of spatial information, with a larger weight reducing the effect (meaning that more distant regions contribute more). In this work, we use the same weighting as in [4], with $\phi^2 = 0.4$.

Finally, they also consider the colour distance between regions, as represented by $D_c(c_{\hat{k}}, c_{\hat{j}})$. The colour distance between two regions is calculated by:

$$D_c(c_1, c_2) = \sum_{\hat{i}=1}^{c_{n1}} \sum_{\bar{i}=1}^{c_{n2}} f(\mu_{1,\hat{i}}) f(\mu_{2,\bar{i}}) D(\mu_{1,\hat{i}}, \mu_{2,\bar{i}}) \tag{5}$$

Here, in Eq. 5, $f(\mu_{1,\hat{i}})$ represents the probability of the $\hat{i}$th colour $\mu_{\hat{j},\hat{i}}$ in the $\hat{j}$th region. $D(\mu_{1,\hat{i}}, \mu_{2,\bar{i}})$ is the Euclidean colour distance between colour $\mu_{1,\hat{i}}$ and $\mu_{2,\bar{i}}$ in the $L \times a \times b$ colour space for perceptual accuracy, as defined by the following:

$$D(\mu_{1,\hat{i}}, \mu_{2,\bar{i}}) = ||\mu_{1,\hat{i}}, \mu_{2,\bar{i}}|| \tag{6}$$

Finally, this is improved by using colour space smoothing. As discussed in [4], some similar colours, when quantised, as mentioned previously, may by quantised to different values and introduce artefacts The saliency of each colour can therefore be replaced by the weighted average of the saliency of different colours.

Overall, this is used as it is state of the art, is reported to be fast, and delivers good results [4]. As part of the comprehensive analysis of results, we compare our new approach to SaliencyCut (also known as RCC), which segments objects based on RC saliency. For each image frame I_t, an overall RC saliency map $\hat{\Theta}_t$ is produced. This is normalised and can be fused with the motion saliency, calculated as discussed in "Global Motion Detection and Motion Saliency" section, to produce an overall saliency.

Saliency Fusion and Thresholding

In "Global Motion Detection and Motion Saliency" section, we calculated the motion saliency Θ_t for each tth frame of a video sequence of T frames, and in "region-based contrast" section, we also calculated an equivalent RC saliency map $\hat{\Theta}_t$. Both of these were normalised to ensure that all pixels had an image value ranging between 0 and 255. As shown in Fig. 1 in "region-based ontrast" section, these are then fused to produce a single binary image. Figure 2c, d shows examples of motion saliency and (the equivalent RC saliency result. These are then fused to produce (e).

For each tth frame in T, the normalised motion saliency Θ_t and the RC saliency $\hat{\Theta}_t$ are weighted and fused to produce the overall saliency map $\hat{\theta}_t$ as follows:

$$\hat{\theta}_t = \omega_{\Theta} \Theta_t + \omega_{\hat{\Theta}} \hat{\Theta}_t \tag{7}$$

where ω_{Θ} and $\omega_{\hat{\Theta}}$ represent the weights applied to the motion and RC saliency, respectively. Here, we set both values to 0.5 to provide equal weighting to both.

For each image frame I_t, an overall saliency map $\hat{\theta}_t$ is produced, see Fig. 1e. This is a fusion of methods, which to the best of our knowledge, along with our global motion calculation, has not been previously reported.

Finally, to produce the final saliency binary map θ_t, we apply a threshold to each pixel of the image, which, after preliminary investigation, we set to be equal to 90. Here, any pixels of $\hat{\theta}_t < 90 = 0$, and all other values $\hat{\theta}_t \geq 90 = 1$. This produces a binary mask θ which can subsequently be used for foreground and background labelling.

Foreground and Background Labelling

For the final labelling of each frame of the video into foreground and background models, we make use of an approach reported by Papazoglou and Ferrari [3]. In this approach, they make use of an energy function and several models, with the aim of identifying clearly which parts of each video frame are foreground, and which parts are background. As the results identified in [3] are promising, we use this approach as part of our research. However, one key difference with our overall framework is that while in [3] they use inside-outside maps, we use the fused mask described above.

A full description of this technique can be found in [3], and this section presents a brief summary. Firstly, superpixels are calculated for all video frames, using the simple linear iterative clustering (SLIC) superpixel method [58]. This means that in each frame, similar adjacent pixels are fused into larger groupings in order to reduce computational complexity and allow for faster calculation. Each superpixel P^t then has a label l, which determines whether it is foreground or background, i.e. $l_j^t \in \{0, 1\}$. The aim is therefore to label all superpixels in all frames of a video, which can be represented as $L = \left\{ l_j^t \right\}_{t,j}$.

Following [3], we use an energy function to evaluate $L = \left\{ l_j^t \right\}_{t,j}$, as shown in (8). The desired output labelling L is that which minimises this.

$$E(L) = \sum_{t,j} G_j^t\left(l_j^t\right) + v_1 \sum_{t,j} Q_j^t\left(l_j^t\right) + v_2 \sum_{(j,k,t)\in\Gamma_p} \bar{M}_{jk}^t\left(l_j^t, l_k^t\right) + v_3 \sum_{(j,k,t)\in\Gamma_t} \hat{M}_{jk}^t\left(l_j^t, l_k^{t+1}\right) \tag{8}$$

We can see that the energy function shown in Eq. (8) consists of four parts. Firstly, there is an appearance model G, which determines how likely a superpixel is to be part of the foreground based on appearance models. The second part of the function is a location model Q, which builds a model using all super pixels in all frames of the video. The final parts of the function are a pair of contrast-modulated Potts Potentials, as used in [44, 59, 60], which are used to calculate both spatial ($\bar{M}$) and temporal ($\hat{M}$) smoothness.

Finally, each part of the function is weighted using scalars v, which are defined manually. Equation (8) is minimised exactly with graph cuts, as [3] describe E as being a binary pairwise energy function with submodular pairwise potentials. At each stage, the output segmentation is used to re-estimate the appearance models, ultimately producing the final labelled output mask as shown in Fig. 2f. We very briefly summarise the individual components below, a full description can be found in Papazoglou and Ferrari [3].

Appearance Model G^t

The appearance model is a unary potential (i.e. based on only this value), which determines how likely a superpixel is to be part of the foreground based on appearance models. The appearance model G^t consists of two GMM models, which consider the average Red/Green/Blue (RGB) colour of superpixels. The colour of individual pixels in each superpixel is averaged, to produce one single average RGB value per superpixel, which is used in the GMM models. Appearance models are estimated for each frame (Q^t), but their estimation integrates information throughout the video, meaning that each individual model can be seen as one single dynamic model. Using this model as part of the overall energy function is considered to produce more accurate segmentation than simply using motion, with models that operate across the entire video, transfer information between frames, and correct motion tracking errors, for example, in frames where only part of an object is moving.

In our implementation, we automatically estimate parameters based on the mask from the previous saliency detection steps. To estimate the models, at each tth frame, we use all superpixels in the video. The foreground and background models are calculated in a similar manner, but weighted by their closeness to the tth frame. Using the estimated models, the potential is then calculated as being the log probability of superpixel p_j^t having label l_j^t, with the appropriate model.

Location Model Q^t

The location model Q is created using the super pixels in all frames of the video, with one model being initialised at the first frame and then being updated each frame, while another model is initialised in the final frame and is propagated backwards. This is used as considering only appearance runs the risk of poor results in image regions where foreground and background are of similar colour (as the appearance model described above uses RGB colours). In this paper, we adapt the work of [3]. Our approach makes use of the output generated by the saliency fusion as a location prior. To ensure maximum accuracy (and compensate for examples of inaccurate optical flow or thresholding), the location priors can be used over time to build a more accurate location model Q^t with temporal propagation.

The location prior Q_j^t at superpixel p_j^t is defined as being the percentage of the pixels of p_j^t that are considered to be part of the object as defined by the saliency, defined as $Q_j^t := n_j^{t'}$. The model is propagated forwards, starting at frame 1, and advancing to the end of the video. In addition, an analogous backwards propagation takes place, starting at the final frame and finishing at the first video frame. At each step, the location prior for a superpixel p_k^{t+1} is updated, using a summation of all superpixels in frame t. The forward and backwards steps are run independently of each other, and then the final location prior for the tth frame Q^t is calculated to be the normalised sum of both forwards and backwards steps.

Spatial and Temporal Smoothing

In addition to the location and appearance models, we also have the pairwise potentials, representing the temporal ($\hat{M}$) and spatial ($\bar{M}$) smoothness. These are standard contrast-modulated Potts potentials, see [44, 59, 60].

Firstly, the temporal smoothness ($\hat{M}$) is defined as:

$$\hat{M}^t_{jk}\left(l^t_j, l^{t+1}_k\right) = \eta\left(p^t_j, p^{t+1}_k\right)\left[l^t_j, \neq l^{t+1}_k\right] \exp\left(-\tau col\left(p^t_j, p^{t+1}_k\right)^2\right) \tag{9}$$

In Eq. 9, as discussed previously, η represents the percentage of connected pixels within two superpixels and is invariant of the speed of motion. In addition to this, *col* is defined as the average difference in RGB colour between two superpixels, and a parameter τ is used for control. Here, we consider superpixels to be temporally connected (i.e. connected between adjacent temporal frames) if at least one individual pixel of p^t_j moves into superpixel p^{t+1}_k.

The spatial smoothness $\bar{M}$ is similar, with the key difference being that rather than considering superpixels in different frames, they consider only superpixels that are adjacent to each other and in the same frame (i.e. spatially connected), as defined by the edge set Γ_p. Here *dis* represents the Euclidean distance between the centres of the two relevant superpixels.

Experimental Results

Datasets Used

To fully evaluate the research presented in this paper, we make use of several datasets. The first is the widely used Fukuchi [61, 62] dataset. The videos include different types of objects, including a plane, red bird, fox, skiing, horse, cat, rhinoceros, yellow bird, and sunflower. These videos have both static and dynamic backgrounds and have standard ground truths available, therefore ensuring that a good comparison can be made. Example frames can be seen in (a) and (b) of Fig. 3.

In addition, to fully demonstrate the effectiveness of our proposed approach, we also created a customised additional dataset. This dataset comprises ten videos taken from two different sources, six from the YouTube-Objects dataset [63], as used by other research [3], and four surveillance videos recorded by the authors. All videos required the creation of custom ground truths, as the YouTube-Objects dataset only provides bounding boxes around the objects of interest. The videos from the YouTube-Objects dataset include videos for a car, plane, football match, and others, and have a wide range of foreground and background motion. An example frame can be seen in (c) of Fig. 3.

Neither the Fukuchi and YouTube-Objects datasets contain any examples of surveillance videos, which are common scenes, so we therefore added some custom recorded surveillance videos. While these surveillance videos have mostly static backgrounds, they contain different object motions, they are recorded in different environments (indoors and outdoors), some show walking people, and some show standing people and talking. An example frame can be seen in (d) of Fig. 3.

Finally, for additional verification, the SegTrack v2 dataset [64] is another widely used dataset. However, this dataset only provides a separate annotation of every moving object as the ground truth, so for each video, several ground-truth sequences are available, as this dataset was developed to evaluate the tracking of multiple segments per video. Therefore, we test our method on ten good quality videos, and to provide a single ground truth, we incorporate all separate annotations for each frame into a single combined frame. The ten videos are 'bird_of_paradise', 'birdfall', 'cheetah', 'drift', 'frog', 'girl', 'hummingbird', 'monkey', 'penguin', and 'worm'. Most of the videos have multiple objects (hence the original cases of multiple ground-truth examples) and complex foreground and background motion. We test our proposed method using these videos.

Overall, the datasets we selected take account of having a number of objects, objects moving at different speeds, close up and more distant objects, different background motion from things like shaking leaves, and static and moving camera, which allows for a robust, comprehensive, and objective evaluation.

Experiment Setup and Evaluation

We have selected some popular object segmentation methods [3, 4, 25, 43, 46] for comparison. These are state-of-the-art methods, as discussed in "Background Work" section. These methods include three recently developed video object segmentation algorithms with good results, Zhang et al. [46], Papazoglou and Ferrari (PF) [3], and Li et al. [43]. In addition to this, we also compare our approach with ViBe [25], which is a classic moving object detection method, and image saliency alone, and SaliencyCut (RCC) [4].

Fig. 3 Example video frames from our chosen dataset, showing frames from **a** red bird from the Fukuchi dataset, **b** skiing from the Fukuchi dataset, **c** football from the YouTube-Objects dataset, and **d** surveillance 4 from our surveillance dataset

In our experiments, we have to define a number of parameters. Firstly, for the foreground and background labelling, the weighted scalars v, as used in Eq. (8), are defined as $v_1 = 5$, $v_2 = 5000$, and $v_3 = 4000$. In addition to this, for our saliency fusion and thresholding in (7), we define $\omega_{\Theta} = 0.5$ and $\omega_{\hat{\Theta}} = 0.5$, and the threshold for thresholding the fused saliency map is set as being 90, based on preliminary trials.

There were some key thresholds that needed to be defined. In (2), we defined a threshold value with ω_1 set to 0.5, and also defined ω_2 as being 50 %. These values were chosen based on experimentation performed using the Fukuchi dataset. Figure 4 shows the mis-segmentation rate when $\omega_1 = 0.1$–0.9 on the Fukuchi dataset.

It can be seen from Fig. 4 that a range of values for ω_1 delivers similar results (with values ranging from 0.5 to 0.9 being acceptable), and so we chose to use 0.5. Figure 5 also shows the mis-segmentation rates for the Fukuchi dataset when $\omega_2 = 0.1$–0.9.

Figure 5 shows that a wide range of values are acceptable, and so we chose a threshold of 50 %.It can clearly be seen from both figures that when $\omega_1 = 0.5$ and $\omega_2 = 50\,\%$, the separate mis-segmentation rates are smaller; therefore, the use of these thresholds in this research was considered to be justified.

An additional parameter was defined in the saliency fusion step in Eq. (7). Here, we chose to assign an equal weight of 0.5 to ω_{Θ} (weight of motion saliency) and $\omega_{\hat{\Theta}}$ (weight of RC saliency). The justification for this choice of weighting is shown in Fig. 6. Here, the mis-segmentation rates are shown for the combined Fukuchi and the Youtube/surveillance datasets when $\omega_{\Theta} = 0.1$–0.9 (this corresponds to $\omega_{\hat{\Theta}} = 0.9$–$0.1$). We can see from Fig. 6 that assigning equal weighting to ω_{Θ} and $\omega_{\hat{\Theta}}$ produces the lowest mis-segmentation rate.

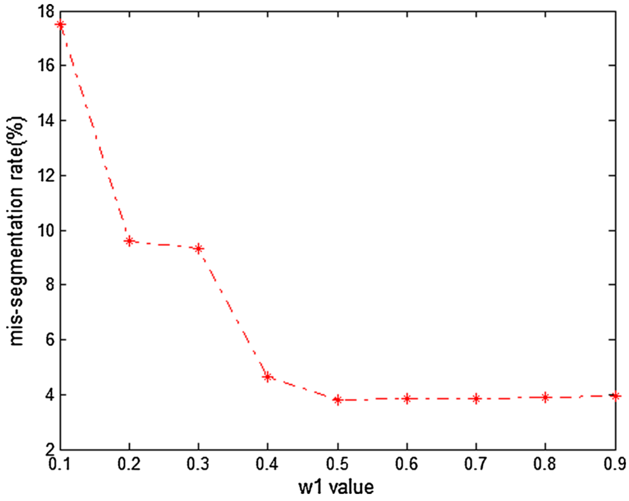


Fig. 4 Mis-segmentation rates when $\omega_1 = 0.1$–0.9 on the Fukuchi dataset

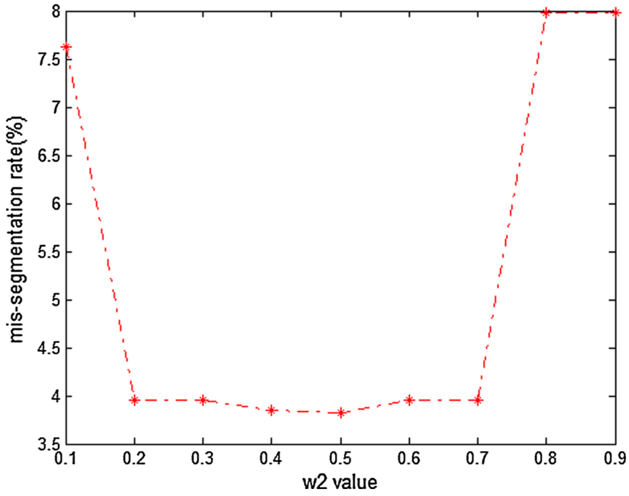


Fig. 5 Mis-segmentation rates when $\omega_2 = 0.1$–0.9 on the Fukuchi dataset

In terms of evaluation, we quantitatively evaluate video object segmentation performance by calculating the mis-segmentation rate as used in [43]. Given a binary segmentation result image, and a matching ground truth, we can calculate the mis-segmentation rate:

$$\Upsilon(S) = \sum_{t=1}^{T} (XOR(S_t, \hbar_t))/T \cdot \hat{p} \qquad (10)$$

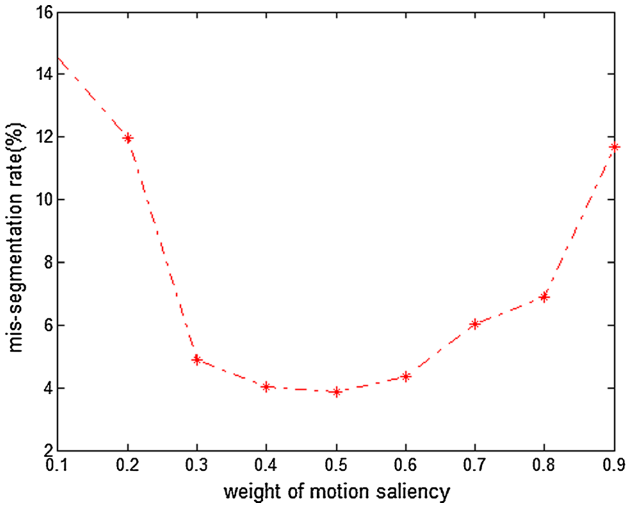


Fig. 6 Averge mis-segmentation rates when $\omega_\Theta = 0.1$–0.9 on two datasets

In (10), T represents the number of t frames, $\hbar_t$ the ground truth pixel map for each tth frame, S_t the video object segmentation output images in comparison with each ground truth value, where $\hbar$ and S are both binary labellings, $\hbar \in \{0, 1\}$, $S \in \{0, 1\}$, with 0 denoting background, and 1 denoting object. Finally, $\hat{p}$ represents the number of pixels in each frame.

In addition to this, we also evaluate using Recall, Precision, and the F-measure (F). Recall and Precision are used to assess the output of a foreground detection algorithm when given a series of ground truth segmentation maps. To do this, for each pixel in a video sequence, the number of true positives (TP), the number of correctly detected foreground pixels, is calculated, along with the false positives (FP), the number of background pixels incorrectly classified as foreground, and the number of false negatives (FN), which accounts for the number of foreground pixels incorrectly classified as background. The bigger the values of Recall and Precision, the better. These measures are then calculated as follows,:

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{11}$$

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{12}$$

The Recall and Precision values given in Eqs. (11) and (12), respectively, can then be used to calculate the F-measure. In the results given in this paper, we give the average F-measure score F, as calculated by:

$$F = (2 \times \mathrm{Precision} \times \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall}) \tag{13}$$

Quantitative Evaluation Results Comparison

Table 1 shows the mis-segmentation rate result in comparison with other methods for videos from the Fukuchi dataset. Firstly, looking at the performance on individual videos, although our approach is only the absolute top performing method on one skiing video, it is the best overall method for this dataset due to having the lowest mean mis-segmentation rate, meaning that the results are consistently reliable and of high quality.

We can see that the ViBe method, the older classic approach, consistently has a higher mis-segmentation rate, which is to be expected, because this approach is more suited to extracting a moving object, leading to large fluctuations in performance between different videos. In addition to this, the approach proposed by Li et al. [43] is an improvement on the method of Zhang et al. [46], but the average result is worse than our method, due to producing poor results in scenarios where the camera is moving or zooming. This approach can locate the main object, but always misses some objects, meaning that the segmented object contour is not always very accurate.

The state-of-the-art PS approach [3] is of interest because they use the same foreground–background labelling approach as we adopt in this paper, and the RCC approach [4] uses the RC technique that we adopt as part of our overall system, and so is also of relevance. Firstly, considering PS, the results show that although it delivers some better individual results, the overall mis-segmentation rate is much higher than our method.

Finally, considering the RCC approach, again, the overall score is very similar to our method, with only a 6.37 % improvement identified with our method, suggesting very similar results. Overall, our method is comparable to two other state-of-the-art approaches. However, to compare performance in more objective and challenging environments, as discussed previously, we also used an additional dataset, comprising of surveillance footage and videos from the YouTube-Objects database, and the mis-segmentation results for this dataset are shown in Table 2.

Firstly, we can see that our own results are consistently good with this dataset, producing the lowest mis-segmentation rate on five of the videos (cars, plane, surveillance 1 and surveillance 2, and motorcycle). We can see that different methods are more suited to different videos, and so it is of interest to consider the average overall performance, which will be discussed later.

Of the other results, firstly, we can see that the ViBe approach consistently has a higher mis-segmentation rate, although the average rate shows an improvement over the first dataset, which is similar to the results of Zhang et al. This is because ViBe has many holes inside the objects and also detects a lot of dynamic background. The results of Li et al. [43] are more inconsistent, because in this dataset, it loses some objects.

Considering the PS approach, there is no obvious outlier video, but the overall results show that our approach

Table 1 Segmentation accuracy comparison using the widely used Fukuchi dataset, giving mis-segmentation rates

	Our method (%)	Zhang et al. [46] (%)	PS [3] (%)	RCC [4] (%)	ViBe [25] (%)	Li et al. [43] (%)
Plane	2.03	1.90	1.68	4.82	6.39	5.98
Red bird	2.20	11.60	11.28	1.75	11.17	1.68
Fox	1.09	1.26	0.88	0.53	12.87	2.22
Skiing 1	2.13	1.77	4.33	11.92	32.04	6.43
Skiing 2	1.95	5.37	3.32	3.88	24.88	4.01
Horse	3.30	3.23	3.35	7.58	23.57	12.42
Cat	5.99	1.60	1.19	1.90	14.85	8.88
Yellow bird	6.12	32.35	61.29	3.73	31.82	15.87
Rhinoceros	1.27	1.20	0.84	0.73	22.37	7.68
Sunflower	12.14	24.69	6.63	3.93	28.43	5.41
Avg. mis-seg. rate	3.82	8.50	9.48	4.08	20.84	7.06
Improvement mis-seg. rate	–	55.06	59.70	6.37	81.67	45.89

generally produces a better performance, as shown by the average results. Of more interest is the difference in performance of the RCC method. The RCC result is not as good as for the first dataset, because this dataset contains a greater range of scenes and objects, and presents larger object segmentation challenges. It was found that RC does not always perform well with multiple object segmentation, resulting in poor performance for RCC, whereas our method was more consistent, leading to the best overall mean result.

Finally, we combine the results in Tables 1 and 2 to present a mean segmentation accuracy comparison between our approach and other relevant approaches. This is shown in Table 3.

Table 3 confirms that our method shows a consistent improvement over the other approaches used for comparison, for datasets with a wide range of motions, both single and multiple objects, and diverse real scenes.

In addition to considering the mis-segmentation rate, we also consider the Recall, Precision, and F-measure. We calculate these as described above and present the overall mean results for all methods discussed in this paper.

Table 4 shows that we produce a much higher (and better) Precision and Recall values than other approaches, leading to a higher F-measure. The poor performance of the older methods (ViBe and Li et al.) is prominent here, reinforcing our mis-segmentation results.

As a final comparison, the results given in this paper attribute the good performance of our method to our weighted fusion of both motion and image saliency. To demonstrate this further, we consider the effects if we only use one saliency technique (i.e. set the weight of one of the two components to zero). We also evaluate the effect of global motion detection to determine its utility. The results are shown in Table 5.

Table 2 Segmentation accuracy comparison our own combined dataset, giving mis-segmentation rates

	Our method (%)	Zhang et al. [46] (%)	PS [3] (%)	RCC [4] (%)	ViBe [25] (%)	Li et al. [43] (%)
Cars	2.01	2.62	2.77	2.73	10.69	13.12
Football game	6.58	4.57	12.50	15.34	20.13	4.74
Plane	0.82	0.85	10.54	10.03	5.2	66.78
Surveillance 1	1.72	4.42	4.64	4.47	9.72	3.41
Surveillance 2	3.04	3.14	3.06	3.71	6.86	4.58
Horse	12.89	14.23	13.62	5.58	15.34	12.65
Motorcycle	2.43	3.19	3.92	2.70	14.72	18.96
Cow	1.09	1.22	1.3	0.84	15.11	23.45
Surveillance 3	1.32	4.54	0.84	10.56	7.74	1.14
Surveillance 4	1.10	4.85	4.3	8.24	0.82	1.26
Avg. mis-seg. rate	3.92	4.57	5.36	6.42	10.63	15.01
Improvement mis-seg. rate	–	14.22	26.87	38.94	63.12	73.88

Table 5 confirms that not only does our saliency fusion-based approach deliver improved results over other methods, as discussed previously, but that using only one or the other of the image and motion saliency techniques produces worse results, thus confirming our hypothesis that the motion and the image saliency are both contributing to the overall performance, proving the strength of our new technique. Finally, the ‘global motion’ step has also been shown to play a role in the good results found with our proposed method. If the datasets contain more videos with camera motion, the improvement with global motion detection will be more obvious, as this is where global motion is of most use.

We also performed further experiments on the SegTrack v2 [64] subdataset and compared our method with a number of other methods. The average mis-segmentation rate, F-measure, Precision, and Recall of our proposed method and other methods are given in Table 6. It can be shown that our method still outperforms all others, verifying the results found with the previously used datasets.

Segmentation Visual Examples

To illustrate our conclusions above, we provide some examples of the results discussed in the previous section. Firstly, Fig. 7 shows an example of frames from two videos from the Fukuchi dataset.

Figure 7a shows the original image frames from two videos. Firstly, the 14th, 24th, and 34th frames of “skiing 1”, and also the 53rd, 63rd, and 73rd frames of “horse”, the ground truth is shown in (b). Fig. 7c shows the results of the method used by Zhang et al. [46], (d) the PS approach [3], (e) Li et al. [43], ViBe [25] is shown in (f), RCC [4], i.e. image saliency alone, is displayed in (g), our motion saliency-based object segmentation is shown in (h), and (i) gives the final result of our spatio-temporal saliency-based segmentation.

In addition to verifying the results reported in this paper, the comparison of (g), (h), and (i) is of particular interest, as (i) is created from the fusion of (g), and (h), and thus combines the strengths of these approaches to improve

Table 3 Segmentation accuracy comparison combining both datasets

	Our method (%)	Zhang et al. [46] (%)	PS [3] (%)	RCC [4] (%)	ViBe [25] (%)	Li et al. [43] (%)
Avg. mis-seg. rate	3.87	6.54	7.42	5.25	15.74	11.04
Improvement mis-seg. rate	–	40.83	47.84	26.29	75.41	64.95

Table 4 Precision, Recall, and F-measure for our approach compared to other approaches, results give the average of all 20 videos used in this paper

	Our method	Zhang et al. [46]	PS [3]	RCC [4]	ViBe [25]	Li et al. [43]
Average Precision	0.7787	0.7751	0.6800	0.6656	0.3680	0.4830
Average Recall	0.8178	0.6794	0.6554	0.7247	0.3679	0.6900
Average F-measure	0.7631	0.6808	0.6163	0.6484	0.3406	0.4982

Table 5 Average mis-segmentation rate, F-measure, Precision, and Recall comparisons based on motion saliency only, image saliency only, no ‘global motion’ step, and saliency fusion results shown on two datasets

	Motion saliency only	Image saliency only	No ‘global motion’ step in the proposed method	Proposed spatio-temporal saliency approach
Average mis-segmentation	11.72 %	15.07 %	4.03 %	3.87 %
Average Precision	0.3543	0.2736	0.7843	0.7787
Average Recall	0.4059	0.2430	0.8011	0.8178
Average F-measure	0.3277	0.1487	0.7569	0.7631

Table 6 Segmentation accuracy, Precision, Recall, and F-measure for our approach compared to other approaches on the video from SegTrack V2 dataset

	Our method	Zhang et al. [41]	PS [3]	RCC [4]	ViBe [20]	Li et al. [38]
Mis-seg.rate	6.17 %	12.17 %	7.34 %	8.72 %	15.26 %	16.41 %
Precision	0.6448	0.6834	0.7067	0.6160	0.3588	0.1216
Recall	0.6369	0.5907	0.6285	0.6003	0.3953	0.2318
F-measure	0.6408	0.5651	0.6329	0.5423	0.3064	0.1046

results where needed. This can also be seen in the examples shown in Fig. 8, which gives two further examples. One from the YouTube-objects dataset, and frames from one of our surveillance videos (surveillance 1). Similarly to Figs. 7, 8 shows another two examples, this time from the second dataset. We can see that our method produces much better results for multiple objects with different motion than the other compared methods, especially can segment out the static and the moving objects.

Runtime Comparison

Finally, we also consider the runtime of all of the methods. The mean values for all videos are given in Table 7. All results were calculated on a Windows 7 64 bit Operating system running MATLAB 2012, with a Intel i7-4790K 4.0 GHz CPU and 32 GB of RAM.

Firstly, considering the individual components of our method, optical flow computation takes approximately 0.08 s/frame, visual saliency computation takes 0.0016 s/frame, and superpixel calculation requires 0.125 s/frame. Our approach combines all of these stages as part of its process, and our overall runtime takes 0.53 s per frame (s/frame).

Excluding optical flow and over-segmentation, the PS method takes 0.4 sec/frame. For the purposes of comparison, if we adjust PS to use the same optical flow and over-segmentation algorithm as our approach, PS takes approximately 0.6 sec/frame. The ViBe approach is much quicker and takes 0.00314 sec/frame, which is effectively almost real time. However, it should be noted that the previously discussed accuracy results show that the speed comes at the cost of accuracy. The approaches of Zhang et al. and Li et al. are both extremely time-consuming, resulting in a performance of over 80 and 60 s per frame, respectively, and also require a lot of memory.

The closest result to our approach is RCC, which is quicker to calculate than ours, as it deals only with image saliency and not the additional motion saliency. Overall, our approach has been shown to be quick to calculate, and the good results reported in previous sections justify the fractionally slower calculation time.

Discussion

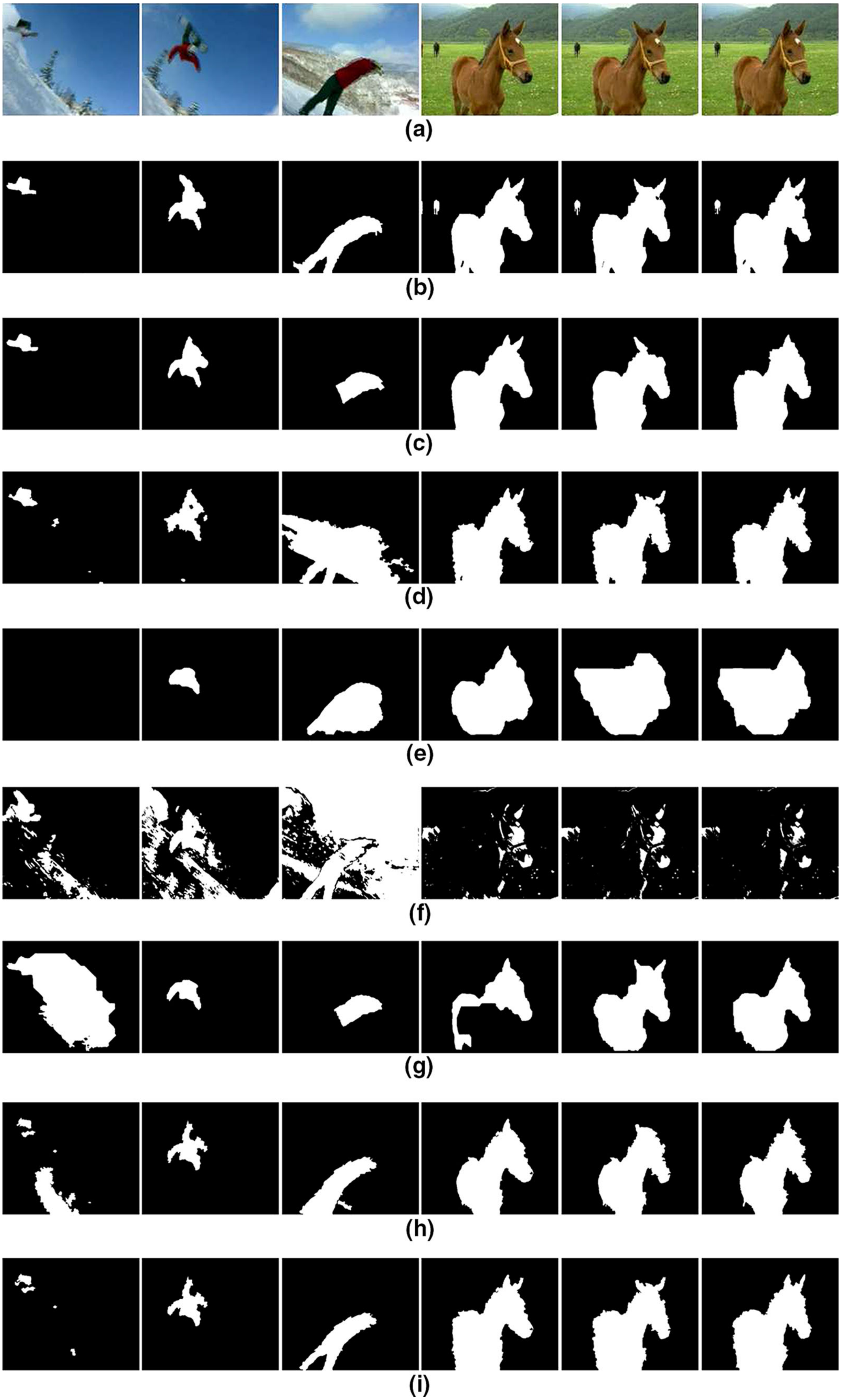
The results in the previous section show that our system has delivered better performance, with high overall F-measure, and a very low mis-segmentation rate. It is also fast, being quicker than all except two methods, ViBe and the image saliency only RCC approach. While ViBe delivers much poorer results, RCC is much closer and also faster. This is because our approach fuses more than simply one saliency approach, resulting in additional computational time per frame. However, the additional time is still very low, resulting an approach that is still fast.

Fig. 7 Example frames from two videos from Fukuchi dataset, ▶ **a** original image (14th, 24th, and 34th frames of "skiing 1", and 53rd, 63rd, and 73rd frames of "horse"), **b** ground truth, **c** Zhang et al. [46], **d** PS [3], **e** Li et al. [43], **f** ViBe [25], **g** RCC [4], **h** motion saliency-based object segmentation, and **i** final output of our method

As discussed previously, the results confirm that our fused weighted saliency method shows a consistent improvement over the other approaches used for comparison. Of most relevance are the approaches that share some techniques with our method. The RCC approach has good results with the first dataset, but the second dataset and overall results show that the image-only saliency approach has higher overall mis-segmentation rates than our approach, which fuses RC with a weighted motion saliency (fusing RC and motion cues). This demonstrates the merits of using our approach rather than simple RC saliency alone. With static saliency cues, our approach can detect static objects that many methods do not successfully track. While the PS method uses the same foreground–background labelling approach as we use, their approach does not detect these static objects very well. Using global motion, our generated motion maps are more accurate than optical flow alone, especially in scenarios with camera motion present. With our weighted fusion of the motion map with visual saliency, our approach can detect more objects. In addition to this, the use of foreground–background labelling is required as when the spatio-temporal saliency is not fully accurate, foreground–background labelling can remove noise, in particular temporal related noise, meaning that consistent foreground motion is kept, in addition to good static image saliency.

However, while the individual video results show that our fused performance was generally consistent, we found noticeably worse results with the football and horse videos, which the other methods also have issues with. This shows one disadvantage of our method, in that while it combines the benefits of both saliency approaches, if one individual technique performs particularly poorly (motion or image), then this will affect our overall results. This is because our method will naturally take its cues from the image saliency and the motion map. For example, RCC also produces poor results for the football video, and although the addition of the motion map results in a lower mis-segmentation than RC-only, there is still a higher mis-segmentation rate than for other videos.

While this is a disadvantage, the results in Table 5 show that using both approaches in our weighted method is an improvement over using a single approach alone, confirming the overall benefits of our approach.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

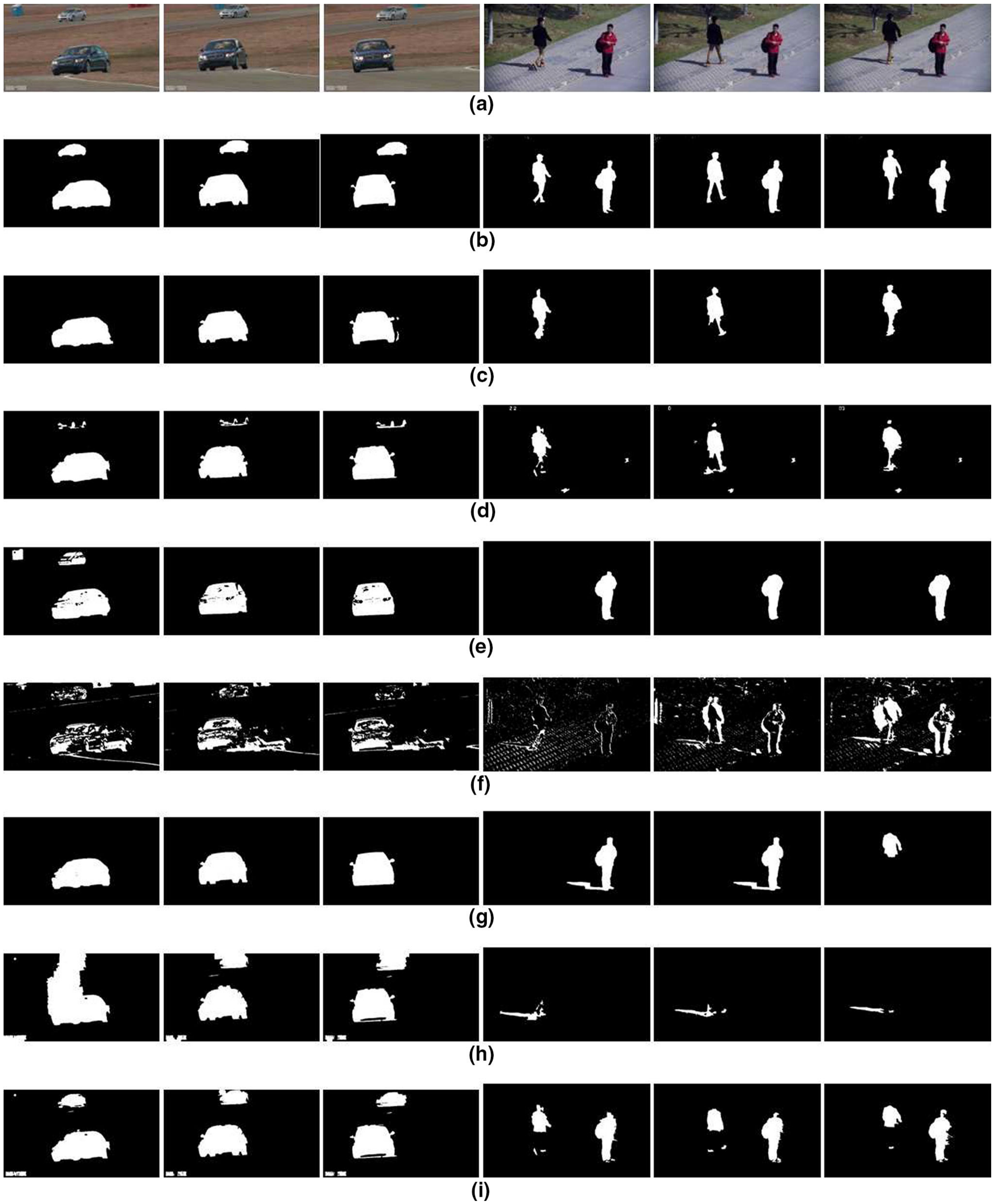


Fig. 8 Example frames from one video from YouTube-Objects and one from our surveillance dataset, **a** original image (9th, 19th, and 29th frames of "cars", and 2nd, 12th, and 22nd frames of "surveillance 1"), **b** ground truth, **c** Zhang et al. [46], **d** PS [3], **e** Li et al. [43], **f** ViBe [25], **g** RCC [4], **h** motion saliency-based object segmentation, and **i** final output of our method

Table 7 Average runtime comparison of all methods, given in seconds per frame

	Our method	Zhang et al. [46]	PS [3]	RCC [4]	ViBe [25]	Li et al. [43]
Runtime (sec/frame)	0.53	>80	0.6	0.004	0.00314	>60

It should also be noted that while we have compared with a number of datasets and with a number of state-of-the-art and classic techniques, there is much scope for future comparison with a range of other techniques. In addition, while we used two widely used datasets, augmented with some custom recordings, there is potential to use further video datasets in order to confirm our results.

Likewise, there are a small number of examples where RCC produces better results on individual videos, which is to be expected, but these better results still result in a poor overall average over all videos, because of very poor performance on other videos, which our approach overcomes. Overall, although there are some examples where results are sub-optimal (as was found with all approaches compared), the benefits of using our approach are clear.

Summary and Future Work

In this paper, we presented a new spatio-temporal saliency-based video object segmentation approach that considers saliency as a bottom-up problem. To do this, we fused motion saliency, calculated by developing an approach to check for global motion, and modify the initial optical flow results if required, with image-only saliency, in a weighted fusion to create a novel design of a fused saliency map. This can be used as the input into a foreground–background labelling approach. The results presented showed that in comparison with a number of both classic and state-of-the-art approaches, our method delivered consistently quick results, as well as having a lower overall mis-segmentation rate and higher Precision, Recall, and F-measure than all other approaches. As part of our thorough evaluation, we considered two datasets, and both overall and individual video performance. This showed the good performance of our method and was confirmed by considering the effect of altering the weights to only consider using one technique, so rather than weighting both image and motion saliency equally, we consider both techniques individually. The results show that the fused approach is better than either individual approach, thus confirming the merits of our approach. Finally, the system is shown to be very quick, faster than a number of other approaches. There is potential to apply our successfully tested object segmentation to other domains, for example multimodal speech filtering [65]. In Abel and Hussain [65], visual information (specifically, lip region images extracted from videos) is used as part of a Wiener filtering-based signal processing, and therefore, the use of our approach to improve the region-of-interest extraction is feasible and worth investigating as part of a cognitively inspired speech filtering system.

Although our method shows a good ability for segmenting dynamic and static objects, we still find its disadvantages from the visual results. For example, some redundant motions are detected in the background for the ‘skiing’ video, or some foreground objects are not so complete, as shown in the results for the ‘surveillance’ video. Therefore, the further improvement that could be made to our research is to focus on improving the motion saliency results, as even with our global motion recalculation, there is much further room for refinement.

As a general situation, there are a smaller number of objects in a frame to be tested. If there are a large number of moving objects in the scene, for example, vehicles on crowded highway, pedestrians in a shopping centre, the global motion proposed here may be not the right motion to maintain. For this issue, in our future work, we will add a step ‘foreground motion judgement’ after ‘global motion detection’ to detect the right foreground objects.

Our results can be improved further besides the above, by developing the weighting system. One possible option is to implement some knowledge about the video; however, as discussed earlier in this paper, the human inspired approach to saliency is generally to consider it as mainly a bottom-up problem, and adding information about the video is both challenging and makes it into more of a top-down problem. Top-down approaches tend to be more time-consuming, which affects practical functionality. A balance must therefore be found between integration of additional information and speed.

Acknowledgments This work is supported in part by the National High Technology Research and Development Program (863 Program) of China (2014AA015104). This research is also supported by the Natural Science Foundation of Anhui Higher Education Institution of China (KJ2015A110), the “Sino-UK” Higher Education Research Partnership for PhD studies” joint-project (2013–2015) funded by the British Council China and the China Scholarship Council (CSC), the Natural Science Foundation of Anhui Higher Education Institution of China (No. KJ2014A015), and the Natural Science Foundation of China (No. 61472002). This research is also partially supported by the Engineering and Physical Sciences Research Council (No. EP/M026981/1) and a Royal Society of Edinburgh Travel Grant (ABEL/NNS/INT).

Compliance with Ethical Standards

Conflict of Interest Zhengzheng Tu, Andrew Abel, Lei Zhang, Bin Luo, and Amir Hussain declare that they have no conflict of interest.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

Human and Animal Rights This article does not contain any studies with human or animal subjects performed by the any of the authors.

References

1. Edelstein A, Rabbat M. Background subtraction for online calibration of baseline rss in rf sensing networks. IEEE Trans Mobile Comput. 2013;12(12):2386–98.
2. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell. 1998;11:1254–9.
3. Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video. In: 2013 IEEE international conference on computer vision (ICCV). IEEE; 2013. p. 1777–1784.
4. Cheng M, Mitra NJ, Huang X, Torr PHS, Hu S. Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell. 2015;37(3):569–82.
5. Chen L, Shen J, Wang W, Ni B. Video object segmentation via dense trajectories. IEEE Trans Multimed. 2015;17(12):2225–34.
6. Wang W, Shen J, Porikli F. Saliency-aware geodesic video object segmentation. In: Proceedings of IEEE CVPR; 2015.
7. Wang W, Shen J, Li X, Porikli F. Robust video object cosegmentation. IEEE Trans Image Process. 2015;24(10):3137–48.
8. Dong X, Shen J, Shao L, Yang M-H. Interactive cosegmentation using global and local energy optimization. IEEE Trans Image Process. 2015;24(11):3966–77.
9. Wolfe JM, Horowitz TS. What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci. 2004;5(6):495–501.
10. Cheng M-M, Warrell J, Lin W-Y, Zheng S, Vineet V, Crook N. Efficient salient region detection with soft image abstraction. In: 2013 IEEE international conference on computer vision (ICCV). IEEE; 2013. p. 1529–1536.
11. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Annu Rev Neurosci. 1995;18(1):193–222.
12. Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer; 1987. p. 115–141.
13. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum H-Y. Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell. 2011;33(2):353–67.
14. Yang J, Yang M-H. Top-down visual saliency via joint crf and dictionary learning. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2012. p. 2296–2303.
15. Fang Y, Chen Z, Lin W, Lin C-W. Saliency detection in the compressed domain for adaptive image retargeting. IEEE Trans Image Process. 2012;21(9):3888–901.
16. Li H, Ngan KN. Saliency model-based face segmentation and tracking in head-and-shoulder video sequencesd. J Vis Commun Image Represent. 2008;19(5):320–33.
17. Siagian C, Itti L. Biologically inspired mobile robot vision localization. IEEE Trans Robot. 2009;25(4):861–73.
18. Born RT, Groh JM, Zhao R, Lukasewycz SJ. Segregation of object and background motion in visual area mt: effects of microstimulation on eye movements. Neuron. 2000;26(3):725–34.
19. Seo HJ, Milanfar P. Static and space–time visual saliency detection by self-resemblance. J Vis. 2009;9(12):15.
20. Wang W, Shen J, Shao L. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans Image Process. 2015;24(11):4185–96.
21. Tu Z, Zheng A, Yang E, Luo B, Hussain A. A biologically inspired vision-based approach for detecting multiple moving objects in complex outdoor scenes. Cogn Comput. 2015;7:539–51.
22. Bao L, Yang Q, Jin H. Fast edge-preserving patchmatch for large displacement optical flow. IEEE Trans Image Process. 2014;23(12):4996–5006.
23. Heikkilä M, Pietikäinen M. A texture-based method for modeling the background and detecting moving objects. IEEE Trans Pattern Anal Mach Intell. 2006;28(4):657–62.
24. Zivkovic Z. Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International conference on pattern recognition, 2004. ICPR 2004, volume 2. IEEE; 2004. p. 28–31.
25. Barnich O, Van Droogenbroeck M. Vibe: a universal background subtraction algorithm for video sequences. IEEE Trans Image Process. 2011;20(6):1709–24.
26. Horn BK, Schunck BG. Determining optical flow. In: 1981 Technical symposium east. International Society for Optics and Photonics; 1981. p. 319–331.
27. Lucas BD, Kanade T, et al. An iterative image registration technique with an application to stereo vision. IJCAI. 1981;81:674–9.
28. Sun D, Roth S, Black MJ. Secrets of optical flow estimation and their principles. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2010. pp. 2432–2439.
29. Wedel A, Cremers D, Pock T, Bischof H. Structure- and motion-adaptive regularization for high accuracy optic flow. In: 2009 IEEE 12th international conference on computer vision. IEEE; 2009. p. 1663–1668.
30. Wedel A, Pock T, Zach C, Bischof H, Cremers D. An improved algorithm for tv-l 1 optical flow. In: Statistical and geometrical approaches to visual motion analysis. Springer; 2009. p. 23–45.
31. Yubing T, Cheikh FA, Guraya FFE, Konik H, Trémeau A. A spatiotemporal saliency model for video surveillance. Cogn Comput. 2011;3(1):241–63.
32. De Croon GCHE, Postma EO, van den Herik HJ. Adaptive gaze control for object detection. Cogn Comput. 2011;3(1):264–78.
33. Donoser M, Urschler M, Hirzer M, Bischof H. Saliency driven total variation segmentation. In: 2009 IEEE 12th international conference on computer vision. IEEE; 2009. p. 817–824.
34. Rutishauser U, Walther D, Koch C, Perona P. Is bottom-up attention useful for object recognition? In: Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on, volume 2. IEEE; 2004. p. II–37.
35. Christopoulos C, Skodras A, Ebrahimi T. The jpeg2000 still image coding system: an overview. IEEE Trans Consum Electron. 2000;46(4):1103–27.
36. Gao Y, Wang M, Zha Z-J, Shen J, Li X, Xindong W. Visual-textual joint relevance learning for tag-based social image search. IEEE Trans Image Process. 2013;22(1):363–76.
37. Harel J, Koch C, Perona P. Graph-based visual saliency. In: Advances in neural information processing systems; 2006. p. 545–552.
38. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. Sun: a bayesian framework for saliency using natural statistics. J Vis. 2008;8(7):32.

39. Xie Y, Huchuan L, Yang M-H. Bayesian saliency via low and mid level cues. IEEE Trans Image Process. 2013;22(5):1689–98.
40. Wei Y, Wen F, Zhu W, Sun J. Geodesic saliency using background priors. In: Computer vision—ECCV 2012. Springer; 2012. p. 29–42.
41. Shen X, Wu Y. A unified approach to salient object detection via low rank matrix recovery. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2012. p. 853–860.
42. Lang C, Liu G, Jian Y, Yan S. Saliency detection by multitask sparsity pursuit. IEEE Trans Image Process. 2012;21(3):1327–38.
43. Li W-T, Chang H-S, Lien K-C, Chang H-T, Wang YF. Exploring visual and motion saliency for automatic video object extraction. IEEE Trans Image Process. 2013;22(7):2600–10.
44. Rother C, Kolmogorov V, Blake A. Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans Graph (TOG). 2004;23(3):309–14.
45. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data; 2001. p. 282–289.
46. Zhang D, Javed O, Shah M. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: 2013 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2013. p. 628–635.
47. Li C, Lin L, Zuo W, Yan S, Tang J. Sold: sub-optimal low-rank decomposition for efficient video segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 5519–5527.
48. Barnes C, Shechtman E, Finkelstein A, Goldman D. Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Trans Graph (TOG). 2009;28(3):24.
49. He K, Sun J. Computing nearest-neighbor fields via propagation-assisted kd-trees. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2012. p. 111–118.
50. Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vis. 2002;47(1–3):7–42.
51. Loy CC, Xiang T, Gong S. Salient motion detection in crowded scenes. In: 2012 5th International symposium on communications control and signal processing (ISCCSP). IEEE; 2012. p. 1–4.
52. Mathe S, Sminchisescu C. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In: Computer vision–ECCV 2012. Springer; 2012. p. 842–856.
53. Pylyshyn ZW. Visual indexes, preconceptual objects, and situated vision. Cognition. 2001;80(1):127–58.
54. Scholl BJ. Objects and attention: the state of the art. Cognition. 2001;80(1):1–46.
55. Dufaux F, Konrad J. Efficient, robust, and fast global motion estimation for video coding. IEEE Trans Image Process. 2000;9(3):497–501.
56. Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. Int J Comput Vis. 2004;59(2):167–81.
57. Jiang H, Wang J, Yuan Z, Liu T, Zheng N, Li S. Automatic salient object segmentation based on context and shape prior. BMVC. 2011;6:9.
58. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–82.
59. Wang T, Collomosse J. Probabilistic motion diffusion of labeling priors for coherent video segmentation. IEEE Trans Multimed. 2012;14(2):389–400.
60. Lee YJ, Kim J, Grauman K. Key-segments for video object segmentation. In: 2011 IEEE international conference on computer vision (ICCV). IEEE; 2011. p. 1995–2002.
61. Fukuchi K, Miyazato K, Kimura A, Takagi S, Yamato J. Saliency-based video segmentation with graph cuts and sequentially updated priors. In: IEEE international conference on multimedia and expo, 2009 (ICME 2009). IEEE; 2009. p. 638–641.
62. Singh A, Chu C-HH, Pratt M. Learning to predict video saliency using temporal superpixels. In: Pattern Recognition Applications and Methods, 4th International Conference on; 2015.
63. Prest A, Leistner C, Civera J, Schmid C, Ferrari V. Learning object class detectors from weakly annotated video. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2012. p. 3282–3289.
64. Li F, Kim T, Humayun A, Tsai D, Rehg JM. Video segmentation by tracking many figure-ground segments. In: 2013 IEEE international conference on computer vision (ICCV). IEEE; 2013. p. 2192–2199.
65. Abel A, Hussain A. Novel two-stage audiovisual speech filtering in noisy environments. Cogn Comput. 2014;6(2):200–17.