

Unsupervised motion saliency map estimation based on optical flow inpainting

Léo Maczyta, Patrick Bouthemy, Olivier Le Meur

▶ **To cite this version:**

Léo Maczyta, Patrick Bouthemy, Olivier Le Meur. Unsupervised motion saliency map estimation based on optical flow inpainting. ICIP 2019 - IEEE International Conference on Image Processing, Sep 2019, Taipei, Taiwan. pp.4469-4473, 10.1109/ICIP.2019.8803542 . hal-02065112v2

HAL Id: hal-02065112

https://inria.hal.science/hal-02065112v2

Submitted on 4 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IEEE copyright notice

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
International Conference on Image Processing (ICIP) 2019.
DOI: 10.1109/ICIP.2019.8803542

UNSUPERVISED MOTION SALIENCY MAP ESTIMATION BASED ON OPTICAL FLOW INPAINTING

L. Maczyta, P. Bouthemy*

Inria, Centre Rennes - Bretagne Atlantique, France

O. Le Meur

Univ Rennes, CNRS, IRISA Rennes, France

ABSTRACT

The paper addresses the problem of motion saliency in videos, that is, identifying regions that undergo motion departing from its context. We propose a new unsupervised paradigm to compute motion saliency maps. The key ingredient is the flow inpainting stage. Candidate regions are determined from the optical flow boundaries. The residual flow in these regions is given by the difference between the optical flow and the flow inpainted from the surrounding areas. It provides the cue for motion saliency. The method is flexible and general by relying on motion information only. Experimental results on the DAVIS 2016 benchmark demonstrate that the method compares favourably with state-of-the-art video saliency methods.

Index Terms—
Motion saliency, optical flow inpainting, video analysis

1. INTRODUCTION

Motion saliency map estimation corresponds to the task of estimating saliency induced by motion. More specifically, regions whose motion departs from the surrounding motion should be considered as dynamically salient. Estimating motion saliency can be useful for a number of applications, such as navigation of mobile robots or autonomous vehicles, alert raising for video-surveillance, or attention triggering for video analysis. In contrast to video saliency approaches, we estimate motion saliency based on motion information only. Indeed, we do not resort to any appearance cues to make the method as general as possible. Furthermore, the method does not require any supervised (or unsupervised) learning stage. Our main contribution will consist in introducing flow inpainting to address in an original way the motion saliency problem.

Video saliency has been first developed as an extension of image saliency, with the objective to extract salient objects in videos. Considering video means that temporal information becomes available, and that motion is usable as an additional saliency cue. In [1], Wang et al. rely on intra-frame boundary information and contrast, as well as motion to predict video saliency. In [2], Le and Sugimoto propose a center-surround framework with a hierarchical segmentation model. In [3], Karimi et al. exploit spatio-temporal cues and represent videos as spatio-temporal graphs with the objective of minimizing a global function.

Apparent motion in each frame is strongly influenced by camera motion. While some approaches directly combine spatial and temporal information without first cancelling the camera motion such as in [4, 5, 6], other methods explicitly compensate the camera motion such as in [7, 8].

Recently, deep learning methods have been explored to estimate saliency in videos. In [9], Wang et al. propose a CNN exploiting explicitly the spatial and temporal dimensions, yet without computing any optical flow. In [10], Le and Sugimoto resort to spatio-temporal deep features to predict dynamic saliency in videos. They extend conditional random fields (CRF) to the temporal domain, and they make use of a multi-scale segmentation strategy. In [11], Wang et al. introduce saliency information as a prior for the task of video object segmentation (VOS), by using spatial edges and temporal motion boundaries as features.

The methods presented above are mostly directed toward the problem of video saliency, that is, extracting foreground objects departing from their context due to their appearance and motion. We are more specifically concerned with the problem of motion saliency, which is more general in some way, and highlights motion discrepancy only. Configurations of interest may arise due to motion only, as in crowd anomaly detection [12] (a person moving differently than the surrounding crowd, or similarly an animal in a flock or a herd, a car in the traffic, a cell in a tissue). In addition, appearance can be of very limited use or even helpless for some types of imagery, like thermal video or fluorescence cell microscopy.

The rest of the paper is organised as follows. Section 2 presents our method for motion saliency estimation. Section 3 reports comparative results with state-of-the-art methods for video saliency. Section 4 contains concluding comments.

2. MOTION SALIENCY ESTIMATION

As stated in the introduction, we estimate motion saliency maps in video sequences only from optical flow cues. We ex-

*This work was supported in part by the DGA and the Région Bretagne as co-funding of Léo Maczyta's PhD thesis

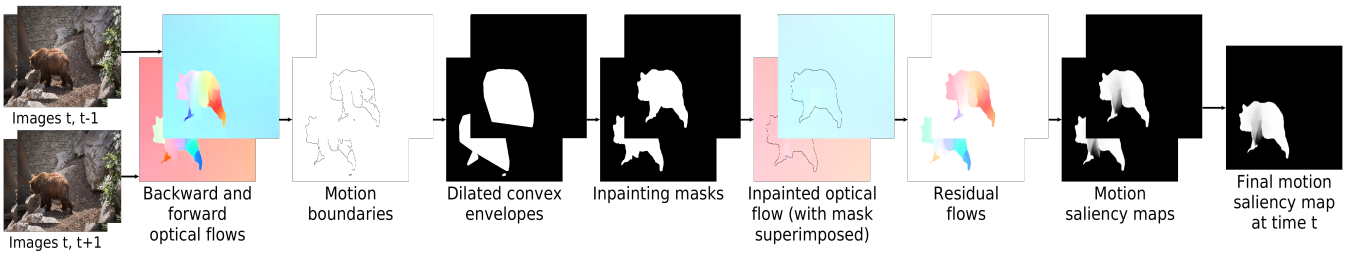


Fig. 1. Overall framework of our method for motion saliency map estimation with the two backward and forward streams.

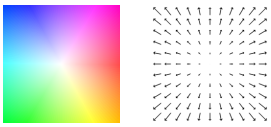


Fig. 2. Colour code (left) for the corresponding optical flow field (right).

pect that the optical flow field will be distinguishable enough in salient regions. We have to compare the flow field in a given area, likely to be a salient moving element, with the flow field that would have been induced in the very same area with the surrounding motion. The former can be computed by any optical flow method. The latter is not directly available, since it is not observed. Yet, it can be predicted by a flow inpainting method. This is precisely the originality of our motion saliency approach. Our method is then two-fold. We extract candidate salient regions and compare the inpainted flow to the original optical flow in these regions. A discrepancy between the two flows is interpreted as an indicator of motion saliency. In addition, we combine a backward and forward processing. Our overall framework is illustrated in Figure 1.

2.1. Extraction of inpainting masks

First, we have to extract the masks of the regions to inpaint. We will rely on the optical flow field computed over the image, and more precisely on its discontinuities. Indeed, the silhouette of any salient moving element should correspond to motion boundaries, since its motion should differ from the surrounding motion. The surrounding motion will be generally given by the background motion, also referred to as global motion in the sequel.

For the motion boundary extraction, one could directly apply a threshold on the norm of the gradient of the velocity vectors. This is however likely to produce noisy contours. Instead, we choose to rely on the classical contour extraction method proposed by Canny [13]. For this, we convert the optical flow to its HSV representation which is commonly used for visualisation, the hue representing the direction of motion and the saturation its magnitude (see Fig.2).

Then, we build region masks from these possibly fragmented contours as illustrated in Figure 1. Contours are first organised into connected parts. For each connected part, the convex envelope is computed and dilated with a 5x5 kernel. Each final region mask is given by the corresponding union of overlapping dilated convex envelopes. By construction, region masks tend to be larger than actual salient areas. Nevertheless, this is desirable for inpainting, since inpainting must start from global motion information only. Yet, a too rough mask can decrease the accuracy of motion inpainting, especially for the salient areas which are non convex (see Fig.1). The masks are then refined by applying the GrabCut algorithm [14] on the HSV representation of the optical flow. To avoid small localisation errors which would include salient pixels in the inpainting mask, a dilatation with a 5x5 kernel is again applied to the resulting mask.

2.2. Optical flow inpainting

We dispose of a set $\mathcal{R}$ of inpainting masks in the image domain Ω. The issue now is to perform the flow inpainting in these masks from the surrounding motion. We have investigated three inpainting techniques to achieve it: two PDE-based methods [15, 16] and a parametric method. Since the background motion to inpaint is globally smooth, a diffusion-based approach for inpainting is well-suited.

We apply the image inpainting method based on fast marching [15] as done in [17] for video completion, which is a different goal than ours. We similarly extend the Navier-Stokes based image inpainting method of [16] to flow inpainting. We adopt the floating point representation of the velocity vectors $\{\boldsymbol{\omega}(p), p \in \Omega\}$ with $\boldsymbol{\omega}(p) \in \mathbb{R}^2$. The two components of the flow vectors are inpainted separately. Finally, we developed a parametric alternative. We assume that the surrounding motion, i.e., the background motion, can be approximated by a single affine motion model. The latter is estimated by the robust multiresolution method Motion2D [18]. The inpainting flow is then simply given by the flow issued from the estimated affine motion model over the masks. The three variants are respectively named MSI-fm, MSI-ns and MSI-pm (MSI stands for Motion Saliency Inpainting).

2.3. Motion saliency map computation

The residual flow, i.e., the difference between the optical flow and the inpainted flow, is then computed over the masks $r, r \in \mathcal{R}$. The motion saliency map g, normalised within $[0, 1]$, is derived from the residual flow as follows:

$$\forall p \in \mathcal{R}, \quad g(p) = 1 - \exp(-\lambda ||\boldsymbol{\omega}_{inp}(p) - \boldsymbol{\omega}(p)||_2), \quad (1)$$

where $\boldsymbol{\omega}$ is the optical flow, $\boldsymbol{\omega}_{inp}$ the inpainted flow, $g(p) = 0$ for $p \notin \mathcal{R}$, and λ modulates the saliency score. Function g expresses that non-zero residual motion highlights salient moving elements. Parameter λ allows us to establish a trade-off between robustness to noise and ability to highlight small but still salient motions.

Let us note that, if we were interested in an explicit motion segmentation, that is, producing binary maps, we would just need to set λ to a high value. Indeed, by applying a threshold τ to $g(p)$, we can deduce from (1) that p will be segmented if:

$$||\boldsymbol{\omega}_{inp}(p) - \boldsymbol{\omega}(p)||_2 \geqslant -\frac{ln(1-\tau)}{\lambda}. \quad (2)$$

With τ arbitrarily set to $\frac{1}{2}$ (the middle value of $[0, 1]$), the decision depends only on λ. Pixels with residual flow magnitude greater than $\frac{ln(2)}{\lambda}$ will be segmented. This shows that our method is flexible, since we can shift from the motion saliency problem to the video segmentation problem just by tuning parameter λ.

Finally, we propose to further leverage the temporal dimension to reduce the number of false positive, in particular close to motion boundaries. To do this, we introduce a bidirectional processing (see Fig. 1). The whole workflow is applied twice in parallel, backward and forward, that is, to the image pair $I(t), I(t-1)$ and the image pair $I(t), I(t+1)$. This yields two motion saliency maps, which we combine by taking their pixel-wise minimum. The reported experimental results for our main method and the NM method introduced in Section 3.3, will include this bidirectional processing.

3. EXPERIMENTAL RESULTS

3.1. Experimental setting

For the computation of the optical flow, we employ FlowNet 2.0 [19]. This algorithm can run almost in real time and estimates sharp motion boundaries. This is important for the successful extraction of inpainting masks.

For all the experiments, the parameters are set as follows. The Canny edge detector is applied to the image smoothed with a Gaussian filter of standard deviation $\sigma = 5$. The two thresholds for the Canny edge detector are set to 20 and 60 respectively. For the inpainting algorithm, a radius of 5 pixels around the region to inpaint is used. Finally, the parameter λ for the computation of the saliency map has been set to $\frac{3}{2}$.

There is no available benchmark dedicated to motion saliency. Therefore, we choose the DAVIS 2016 dataset for the evaluation of our method. This dataset has been initially introduced in [20] for the video object segmentation (VOS) task. It has also been recently used to evaluate methods estimating saliency maps in videos, as in [9, 10]. For the VOS task, the object to segment is a foreground salient object of the video, which has a distinctive motion compared to the rest of the scene. It makes this dataset exploitable for motion saliency estimation, although appearance plays a role.

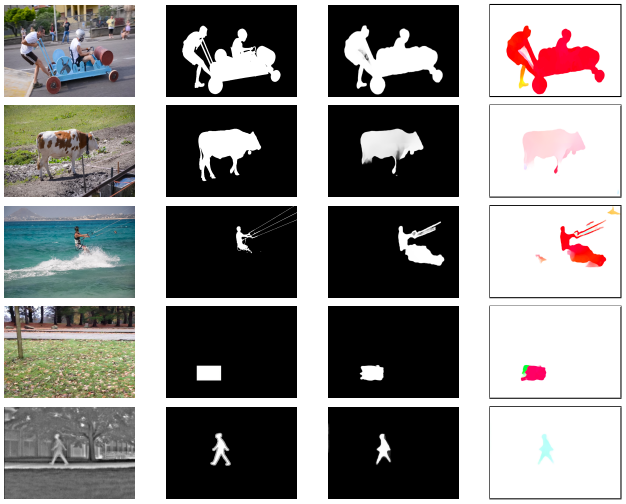


Fig. 3. From left to right: one image from the video, binary ground truth, motion saliency maps predicted by our method MSI-ns, and the estimated forward residual flow (displayed with the motion colour code of Fig. 2).

3.2. Qualitative evaluation

First, we present a visual evaluation of our method MSI-ns, which turns out to be the best of the three variants as reported in Table 1. Fig. 3 displays the output of our method for frames of the videos *soapbox, cows* and *kite-surf* of the DAVIS 2016 dataset and for two other types of videos. In the fourth example (*lawn* video), a rectangular region in the lawn was artificially moved in the image as indicated by the ground truth. It provides us with an example where the only discriminative information is supplied by the undergone motion. The fifth image comes from the *park* video of the changedetection.net dataset [21]. It was acquired with a thermal camera, providing us with an example where appearance is of limited help.

Both computed motion saliency maps and residual flows are shown in Fig. 3. Indeed, the residual flow, although an intermediate step in our method, is meaningful on its own. It provides valuable additional information about the direction and magnitude of salient motions in the scene. It could be viewed as an augmented saliency map.

For the *soapbox* example, the salient element with clearly distinctive motion has been almost perfectly extracted. The *cows* example exhibits an interesting behaviour. The cow is globally moving, except for its legs which are intermittently static. This illustrates the difference between the video object segmentation task, for which the whole cow should be segmented, and the motion saliency estimation task, for which the elements of interest are elements with distinctive motion. Our method consistently does not involve the two legs in the saliency map.

In the *kite-surf* example, the sea foam has a non rigid but strong motion, and consequently, it is likely to belong to the

Method	STCRF [10]	MSI-ns	MSI-pm	MSI-fm	VSFCN [9]	RST [2]	LGFOGR [1]	SAG [11]	NM
MAE ↓	**0.033**	<u>0.043</u>	0.044	0.045	0.055	0.077	0.102	0.103	0.453
F-Adap ↑	**0.803**	<u>0.735</u>	0.724	0.716	0.698	0.627	0.537	0.494	0.367
F-Max ↑	**0.816**	<u>0.751</u>	0.750	0.747	0.745	0.645	0.601	0.548	0.612
Appearance	*Yes*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Motion	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Supervised	*Yes*	*No*	*No*	*No*	*Yes*	*No*	*No*	*No*	*No*

Table 1. Comparison with state-of-the-art methods for saliency map estimation on the test set of DAVIS 2016. In bold, the best performance; underlined, the second best. We also indicate whether the method relies on appearance information, on motion information, and whether it is supervised.

salient moving region, whereas for the VOS task, the kite-surfer is the only foreground object to segment as defined in the ground truth.

In the *lawn* example, the square region is easy to detect when seeing the video, but is much harder to localize in a single frozen image. Our method based on optical flow is able to recover the salient moving region. Finally, in the *park* example involving an IR video with less pronounced appearance, our method also yields a correct motion saliency map.

3.3. Quantitative comparison

We introduce a naive method (named NM) to motion saliency estimation to better assess the contribution of the main components of our method. It merely consists in first computing the dominant (or global) motion in the image. To this end, we estimate an affine motion model with the robust multi-resolution algorithm Motion2D [18]. No inpainting masks are extracted. The residual flow contributing to the motion saliency map is now the difference, over the whole image, between the computed optical flow and the estimated parametric dominant flow. As reported in Table 1, we observe that the method NM yields poor performance. It demonstrates the importance of the flow inpainting approach for motion saliency.

Table 1 also collects comparative results of our three variants, MSI-ns, MSI-pm and MSI-fm, with state-of-the-art methods for saliency map estimation in videos: LGFOGR [1], SAG [11], RST [2], STCRF [10] and VSFCN [9]. Results of these methods are those reported in [10], except for [9], for which we used saliency maps provided by the authors to compute the metrics.

We carried out the experimental evaluation on the test set of DAVIS 2016, which contains 20 videos. The quantitative evaluation on the DAVIS 2016 dataset is useful, but may generate a (small) bias. The available ground truth on DAVIS 2016 may not fully fit the requirements of the motion saliency task as illustrated in Fig. 3 and commented in Section 3.2, since it is object-oriented and binary.

For the evaluation, we use the Mean Average Error (MAE), F-Adap and F-Max metrics, that we compute the same way as in [10]. The MAE is a pixel-wise evaluation of the saliency map g compared to the binary ground truth. F-Adap and F-Max are based on the weighted F-Measure, in which the weight β^2 is set to 0.3 following [10]:

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (3)$$

F-Adap involves an adaptive threshold on each saliency map based on the mean and standard deviation of each map, while F-Max is the maximum of the F-Measure for thresholds varying in [0,255].

Our method MSI-ns obtains consistently satisfactory results, as it ranks second for the three metrics. The two other variants, MSI-pm and MSI-fm, respectively rank third and fourth, but follow MSI-ns by a small margin. Let us recall that we obtain our results without any learning on saliency and any appearance cues in contrast to [10], which performs the best. Our parametric and diffusion-based flow inpainting methods have close performance on the DAVIS 2016 dataset. However, the latter should be more easily generalisable, since the surrounding motion cannot be always approximated by a single parametric motion model.

Regarding the computation time, the MSI-ns method takes 10.3s to estimate the motion saliency map for a 854x480 frame on a 2.9 GHz processor. Our code is written in Python and can be further optimised. Notably, the forward and backward streams of the workflow could be parallelised.

4. CONCLUSION

We proposed a new paradigm to estimate motion saliency maps in video sequences based on optical flow inpainting. It yields valued saliency maps to highlight the presence of motion saliency in videos. We tested our method on the DAVIS 2016 dataset, and we obtained state-of-the-art results, while using only motion information and introducing no learning stage. This makes our method of general applicability. Additionally, the computed residual flow on its own provides augmented information on motion saliency, which could be further exploited. Our current method relies on three successive frames. Future work will aim to further leverage the temporal dimension by exploiting longer-term dependencies.

5. REFERENCES

[1] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, Nov 2015.

[2] T.-N. Le and A. Sugimoto, "Contrast based hierarchical spatial-temporal saliency for video," in *Image and Video Technology*, 2016, pp. 734–748.

[3] A. H. Karimi, M. J. Shafiee, C. Scharfenberger, I. Ben-Daya, S. Haider, N. Talukdar, D. A. Clausi, and A. Wong, "Spatio-temporal saliency detection using abstracted fully-connected graphical models," in *ICIP*, 2016.

[4] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, Sept 2014.

[5] W. Kim and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 646–659, April 2014.

[6] D. Mahapatra, S. O. Gilani, and M. K. Saini, "Coherency based spatio-temporal saliency detection for video object segmentation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 454–462, June 2014.

[7] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[8] C. R. Huang, Y. J. Chang, Z. X. Yang, and Y. Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1336–1349, Aug 2014.

[9] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, Jan 2018.

[10] T. Le and A. Sugimoto, "Video salient object detection using spatiotemporal deep features," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5002–5015, Oct 2018.

[11] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, 2015.

[12] J.-M. Pérez-Rúa, A. Basset, and P. Bouthemy, "Detection and localization of anomalous motion in video sequences from local histograms of labeled affine flows," *Frontiers in ICT, Computer Image Analysis*, 2017.

[13] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, Nov 1986.

[14] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut - interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH)*, pp. 309–314, August 2004.

[15] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics Tools*, vol. 9, pp. 23–34, Jan 2004.

[16] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *CVPR*, 2001.

[17] M. Strobel, J. Diebold, and D. Cremers, "Flow and color inpainting for video completion," in *CVPR*, 2014.

[18] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348 – 365, 1995.

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.

[20] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[21] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *CVPRW*, 2012.