Eurographics Workshop on Intelligent Cinematography and Editing (2015)
W. Bares, M. Christie, and R. Ronfard (Editors)

Efficient Salient Foreground Detection for Images and Video using Fiedler Vectors

Federico Perazzi[1,2] Olga Sorkine-Hornung[2] Alexander Sorkine-Hornung[1]

[1]Disney Research [2]ETH Zurich

Abstract

Automatic detection of salient image regions is a useful tool with applications in intelligent camera control, virtual cinematography, video summarization and editing, evaluation of viewer preferences, and many others. This paper presents an effective method for detecting potentially salient foreground regions. Salient regions are identified by eigenvalue analysis of a graph Laplacian that is defined over the color similarity of image superpixels, under the assumption that the majority of pixels on the image boundary show non-salient background. In contrast to previous methods based on graph-cuts or graph partitioning, our method provides continuously-valued saliency estimates with complementary properties to recently proposed color contrast-based approaches. Moreover, exploiting discriminative properties of the Fiedler vector, we devise an SVM-based classifier that allows us to determine whether an image contains any salient objects at all, a problem that has been largely neglected in previous works. We also describe how the per-frame saliency detection can be extended to improve its spatiotemporal coherence when computed on video sequences. Extensive evaluation on several datasets demonstrates and validates the state-of-the-art performance of the proposed method.

1. Introduction

The ability to identify visually important content (see Figure 1) is a fundamental problem in image and video processing. In particular, applications related to automated cinematography and video editing benefit considerably from methods for detecting salient image elements. For instance, algorithms for intelligent camera control require knowledge of which actions are potentially interesting to the viewer [ZLZZ08, AWC10]. Similarly, saliency information allows automatic video editing techniques to consider viewer attention [DDN08, GL08, JSSH15], or to better adapt to output display constraints [WHSL11, SSH11]. But even basic video processing techniques such as stabilization [GKE11], summarization [EZP*13], segmentation [LCL*13], and compression [ZPS*13] provide improved results when the salient foreground is known.

Estimating saliency is a multidisciplinary problem. Initially founded on psychological and neuroscientific studies [KU85, Ros99], computational saliency detection has become a highly active research area in computer vision [IKN98, HZ07, AHES09, GZMT10]. The literature shows a considerable heterogeneity in the characterization of saliency, as the utility of different definitions is largely application driven, ranging from the detection of eye fixation points [IKN98] to accurate object segmentation masks for video processing applications [FMK*09]. Our method is in the line of works that automatically identify and accu-

Figure 1: *Saliency results on video. Top left: input video frame. Top right: superpixel segmentation. Bottom left: our per-frame saliency. Bottom right: final, temporally coherent saliency. Note how spurious, salient background elements on the car and sky are removed.*

DOI: 10.2312/wiced.20151074

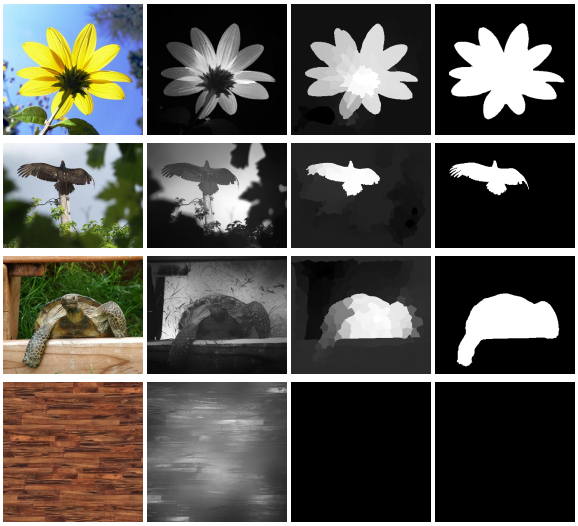

Figure 2: *Saliency computation based on concepts such as global contrast and color uniqueness is less suitable for multi-colored objects (first row), multiple unique colors (second row), and cases where fore- and background are rather similar (third row). The second column shows corresponding saliency maps computed with a representative method based on various contrast measures [PKPH12]. Our approach successfully handles such challenging cases and produces results closer to ground truth. Moreover, a unique feature of our approach is its ability to identify whether an image contains any salient object at all (bottom row).*

rately segment a salient foreground object from a less salient background [AEWS08, CLL11].

Motivated by studies on the importance of contrast-based stimuli, various measures of color contrast have been investigated and successfully employed for computational saliency detection in images; see, e.g., [CZM*11, PKPH12, MTZM13] for an overview of recent works. While those methods have been shown to produce good results in general, natural images often violate some of their fundamental assumptions, which can lead to wrong saliency estimates (see Figure 2 for examples). Therefore, various alternative approaches based, e.g., on object detection, shape analysis and other *objectness* measures have been proposed [LSZ*07, ADF10, LZLX11] in order to complement the set of available tools for saliency detection. Computational complexity is a further challenge that hinders the straightforward application of many image saliency algorithms to video.

In this paper we propose a highly effective method for computational saliency estimation, which has complementary properties to the above mentioned techniques and shows similar or better performance on standard benchmarks compared to current state-of-the-art. Our algorithm is based on the basic assumption that most of the image boundaries are covered by non-salient background. Background color priors and local color similarities are encoded in a graph structure defined over a superpixel segmentation of images or video frames. We then show that the Fiedler vector of the corresponding graph Laplacian is a very effective and robust way to compute saliency masks. In particular, our formulation allows us to determine whether a salient object is present at all by training a support vector machine using properties of the Fiedler vector. This problem has largely been neglected in previous works, which always highlight some regions as salient, even if an image contains background only. In addition, differently from previous approaches that use various heuristics or graph-cut segmentation to binarize saliency maps, the entries of the Fiedler vector yield both a continuous estimate and a content-adaptive binary partition. Finally, we describe an extension of the per-frame saliency detection to video, which improves the spatiotemporal coherence of the results.

Despite its computational simplicity, we show in our examples and evaluation that our method compares favorably to the recent methods and efficiently handles various image and video types that are challenging for previous approaches. To demonstrate the complementary nature of our method, we also show that the performance can be further increased when combining our approach with a recent color contrast-based technique.

2. Related work

The pre-attentive human visual system is driven by bottom-up, low-level stimuli such as color, contrast, orientation of edges, disparity and sudden movements [KU85]. Depending on the nature of their features, methods that model bottom-up visual saliency can be categorized into biologically inspired or computationally based approaches. Biologically inspired methods [IB05, IKN98, HKP06, KB01] are generally built upon the visual attention architecture proposed by Koch and Ullman [KU85], in which biologically motivated features are first selected, then processed with center-surround operations such as Difference of Gaussians that extract local gradients, and finally linearly combined to form the saliency map. Recently, methods such as [JEDT09, Bor12] investigated the combination of biologically motivated features based on low-level stimuli with top-down memory-dependent information such as face or object detectors. Those methods show improved performance but at the same time they establish a dependency with the image context that might not always be desirable. Biologically inspired methods aim to determine eye fixations, i.e. a set of points or blobs in the image that are likely to attract the viewer's eye attention. As a result, saliency maps are often blurry and highlight sparse local features, making their usage in computer vision applications such retargeting or image segmentation impracticable [CZM*11].

In contrast, computational methods are often inspired by biological principles but strongly focus on their practical usage in computer vision and graphics. Central to those application is the ability to determine salient objects, instead of eye fixation points. Hence, an important aspect to consider is the ability to segment and assign a uniform saliency value to the entire salient object [CLL11,AEWS08], preserving edges and producing a pixel-level accurate saliency map. Furthermore, it has been noted in [PKPH12, WWZS12] that there is great variability of results in methods implementing the same visual cues with different computational models. For this reason, the choice of the computational model becomes as important as the visual cues it is based upon.

Perceptual research studies indicate that color-based contrast is a fundamental cue to determine bottom-up visual attention [PLN02, EKK03]. As a result, many interpretations of color contrast have been proposed in recent years. Among those one can distinguish between methods that perform the analysis in the frequency domain [HZ07, GMZ08] and methods operating directly in color space. Achanta et al. [AHES09] interpret saliency as the dissimilarity of a pixel from the mean image color. Different variants of patch-based methods measure the dissimilarity of square image patches [GZMT10, WKI*11, DWM*11]. The exhaustive nearest-neighbor search performed by those methods becomes impractical even for mid-resolution images, such that downsampling or dimensionality reduction is often necessary. To reduce the computational complexity, other methods measure contrast using image histogram bins [CZM*11] or abstract the image into superpixels and upsample it to pixel-level using bilateral filters [PKPH12]. Most contrast-based methods rely on the notion of color uniqueness and distribution [LSZ*07, PKPH12]. Intuitively, distinct colors concentrated in a small region should belong to a salient object. Recently, these concepts have been extended by Margolin et al. [MTZM13] to also integrate pattern distinctiveness and high-level cues.

While such contrast-based methods have proven to be very effective, their basic assumptions do not always hold, as illustrated in Figure 2. Therefore, recent research has also focused on additional visual cues. With a basic motiviation similar to our work, Wei et al. [WWZS12] note that image boundaries are most likely to be part of the background and introduce a measure of saliency based on the color-based geodesic distance between interior image regions and boundaries. Their method produces good results in high-recall areas, but it may suffer from non-smooth backgrounds, producing noisy saliency maps. As we demonstrate in Section 4, our approach effectively resolves such issues and produces more globally coherent saliency maps. User-defined background pixels are typically also employed for graph-cut based binary image-segmentation [YS04]. Alexe et al. [ADF10] combine multiple *objectness* cues such as color contrast, edge density and multiscale saliency into a Bayesian framework to determine the existence of an object

(a) (b) (c) (d)

Figure 3: *Graph augmentation with background prior. (a) Source image, (b) saliency map computed without our graph augmentation, (c) saliency map using our method, and (d) ground truth. The boundary prior and our graph augmentation are key to separating the background from the salient foreground object.*

withing the image. In [LZLX11], the discriminative power of geometric cues such as concavity and symmetry is taken into consideration. Their research, motivated by experimental tests on humans, suggests that the convex side of a curved boundary tends to be figural, leading to successful salient object detection. Yang et al. [YZL*13] employs graph-based manifold ranking to detect salient foreground pixels. The recent method of Liu et al. [LCLS14] exploits PDEs to describe the evolution of visual attention in saliency diffusion. In contrast to those methods, our algorithm has the benefit of guaranteeing to identify a single connected foreground object rather than individual salient (super-) pixels, and furthermore is able to distinguish between images with and without salient foreground objects.

3. Algorithm

We first explain the basic algorithm on a per-image (per-frame) basis, and then extend the approach to video sequences.

3.1. Image representation

As a first step our algorithm decomposes an input image into superpixels, as they provide an effective and perceptually meaningful level of abstraction, and remove unnecessary detail such as small scale non-salient structures and noise [PKPH12]. To segment the image into superpixels we use a variant of [ASS*12] as proposed in [PKPH12], which is fast and preserves discontinuities such as edges well.

The superpixel-decomposition of the image induces an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the vertices $\mathcal{V}$ correspond to superpixels and the edges $\mathcal{E}$ represent an adjacency relationship between the superpixels. Similarly to segmentation algorithms such as [SM97], we model only local relationships, i.e. $(i, j) \in \mathcal{E}$ only if the superpixels corresponding to the nodes v_i, v_j share contiguous pixels in the image. We assign each node v_i the mean Lab color of the superpixel it belongs to, denoted as c_i. The Lab color space is chosen be-

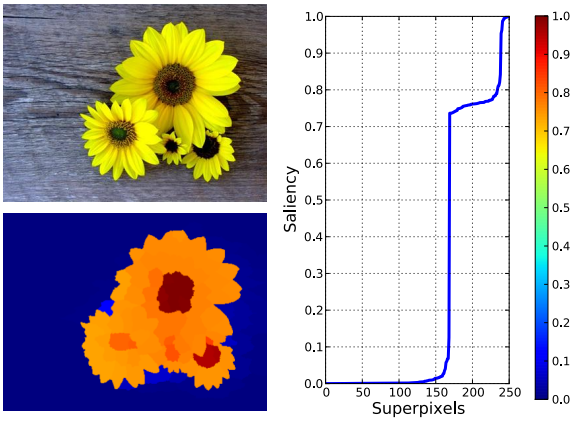

Figure 4: *Saliency computation using the Fiedler vector. In our approach the input image on the top left is represented by a graph structure that encodes color similarities between superpixels and a background color prior computed from the image boundary. The Fiedler vector of the graph Laplacian results in a continuously-valued saliency estimate for every superpixel, illustrated by the saliency map on the bottom left and the plot.*

cause its Euclidean metric mimics the human color perception. Each edge (i, j) is assigned a positive weight $w_{i,j}$ that measures the color similarity between superpixels v_i and v_j, higher values corresponding to higher similarity:

$$w_{i,j} = \frac{1}{\|c_i - c_j\|^2 + \varepsilon} \tag{1}$$

where ε is a small constant (to avoid infinite weights).

Performing a straightforward partitioning of the graph, e.g., using *RatioCut* [HK92], to separate the superpixels into potential fore- and background regions is not sufficient to obtain a reliable salient object estimate. A quantitative evaluation on the MSRA dataset showed that the performance is substantially below state-of-art methods. See also Figure 3(b) for a representative saliency result.

We therefore incorporate a simple prior assuming that the majority of boundary superpixels belongs to non-salient background, motivated by recent studies in gaze prediction which indicate that humans have a tendency to focus attention on the center of an image. This is also reflected in various photographic rules and utilized in saliency estimation techniques such as [WWZS12]. This prior is integrated by augmenting the graph with a background node b and a set of edges $\mathcal{U}$ connecting b to the nodes forming the image boundaries, i.e., to those superpixels that are in immediate contact with the image border.

The augmented graph is hence $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ with $\mathcal{V}_a = \mathcal{V} \cup \{b\}$ and $\mathcal{E}_a = \mathcal{E} \cup \mathcal{U}$. The edge weights in $\mathcal{U}$ model the confidence of a node in being part of the background. We use the Euclidean distance to the mean boundary color, which

Figure 5: *Robustness to salient objects being part of the image boundaries. From left to right: source, our saliency map S_{cont}, ground truth saliency.*

was demonstrated to be effective in [AHES09]: we assign the mean boundary color to b and compute the weights of the edges in $\mathcal{U}$ with Eq. (1). With this formulation, most of the edges in $\mathcal{U}$ are likely to be attached to background superpixels and carry high weight, while few edges (if any) are attached to salient regions and have low weights.

3.2. Saliency estimation

Denote $n = |\mathcal{V}_a|$. We compute an eigendecomposition of the weighted graph Laplacian matrix $L \in \mathbb{R}^{n \times n}$ of $\mathcal{G}_a$:

$$L_{i,j} = \begin{cases} -w_{i,j} & i \neq j,\ (i,j) \in \mathcal{E}_a \\ \sum_{(i,k)\in\mathcal{E}} w_{i,k} & i = j \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

The eigenvector f corresponding to the second smallest eigenvalue, also known as the *Fiedler* vector, represents an optimal *soft* segmentation of $\mathcal{G}_a$ according to a relaxed, continuously valued *RatioCut* objective [vL07] by minimizing

$$\min_f \sum_{i,j\in\mathcal{E}} w_{i,j}(f_i - f_j)^2. \tag{3}$$

The entries of this vector can be interpreted as a one-dimensional (linear) embedding of $\mathcal{G}_a$, where vertices are closer to each other if they are connected by large weights.

We found that this property of the Fiedler vector f provides a meaningful, continuously valued saliency score (see Figure 4 for an illustration). We can derive either a saliency score $S_{\text{cont}} \in [0,1]^n$ or a binary partition $S_{\text{bin}} \in \{0,1\}^n$. Both measures are based on the sign of the entries of the Fiedler vector. Entries having the same sign as the entry f_b corresponding to the background node b will be less salient than those having the opposite sign. Hence we define the continuously-valued saliency score S_{cont} as:

$$S_{\text{cont}} = -\text{sign}(f_b) \cdot f \tag{4}$$

This sign-corrected S_{cont} is then scaled to the range $[0,1]$, possibly with pre-cropping of the value range such that the resulting mean saliency is at least 0.1 [PKPH12].

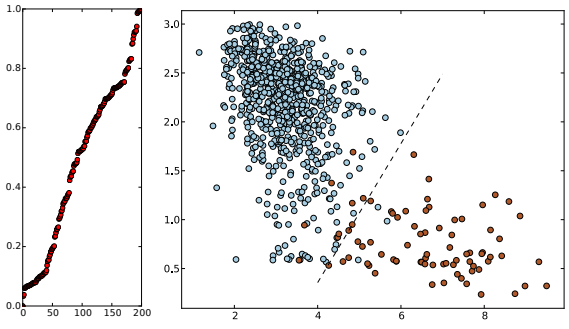

Figure 6: *Left: example of a plot of a Fiedler vector for a non-salient image. Right: 2D plot of features used for distinguishing images with (in blue) and without (in red) salient foreground. The x-axis represents scaled eigenvalues corresponding to respective Fiedler vectors, while the y-axis is the residual of a linear fit to respective Fiedler vectors. The dotted line shows the decision boundary of the trained SVM.*

A binary partition is obtained by discretizing the entries of the Fiedler vector f. This operation can be performed based on the sign of f_b, such that entries having opposite sign to f_b are defined as salient:

$$S_{\text{bin}}(i) = 1 \text{ if } f_i \cdot f_b \leq 0, \;\; S_{\text{bin}}(i) = 0 \text{ otherwise} \qquad (5)$$

To subdivide the graph into more than two partitions, i.e., to identify multiple individual salient objects, the entries of f can be interpreted as points in $\mathbb{R}$ and partitioned by a clustering algorithm such as k-means.

The above approach is quite robust even in challenging cases where the salient object is actually part of the image boundary. As long as the majority of the superpixels is part of the background, the graph partitioning correctly distinguishes between salient and non-salient areas (see Figure 5).

3.3. Detection of non-salient images

Existing methods for saliency estimation are generally unable to identify whether an image actually contains a salient foreground object or not, simply because they are designed to find the most salient pixels or image regions (see Figure 2). We discovered that the properties of the Fiedler vector allow us to devise a simple yet reliable test whether a salient foreground region is present in an image.

On the one hand, the eigenvalue associated with the Fiedler vector indicates how strongly connected the graph is [vL07]. For images without any sufficiently separable foreground object, we can hence generally expect a larger eigenvalue than for images containing one or more foreground objects. On the other hand, a 2D plot of the sorted elements of the Fiedler vector reveals an S-shaped structure in the presence of foreground objects (Figure 4), while non-salient images such as stochastic textures feature a more linear structure (Figure 6). Hence, as a second indicator, we perform a linear regression on the sorted elements of the Fiedler vector and measure the residual error.

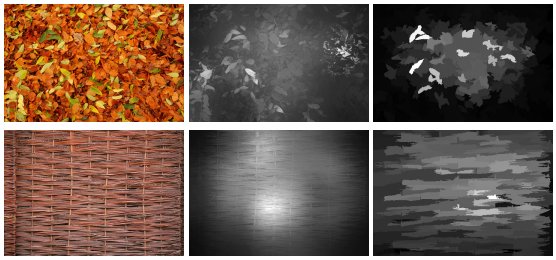

Figure 7: *Examples of texture images used for training an SVM-based classifier to distinguish between images with and without salient foreground, and the corresponding saliency maps produced by existing methods. Since these methods have been designed to always identify the most salient regions in an image, they cannot distinguish between images with and without salient foreground. Our SVM-based classifier identifies such images as non-salient.*

The combination of these two features results in two well separable sets of 2D points and provides strong discriminative power to decide whether a salient object is actually present in an image or not. We trained a linear SVM classifier [CST10], using images of simple textures as non-salient examples [VS14] (see Figure 7) and the images from the benchmark datasets below as salient examples, and evaluated the classification performance using 10-fold cross-validation. The SVM classifier consistently achieved a correct classification performance of 96% for distinguishing between images with and without a salient foreground, while other methods for saliency estimation always detect at least some salient regions (see Figure 7).

3.4. Extension to video

Thanks to the computational efficiency of our algorithm, it can be easily used to process entire videos. In our experiments (see supplemental video) we found the algorithm to be sufficiently temporally stable, such that computation on a per-frame basis can be sufficient for many applications. Several options exist to enforce explicit temporal coherence if required. One idea would be to explicitly extend the dimensionality of the underlying graph structure and the corresponding Laplacian to the temporal dimension. This, however, could be problematic in terms of memory requirements for longer video sequences, and limits interaction of pixels to their direct neighbors only.

We instead propose to utilize an approach for efficient high dimensional filtering [KK11] as a temporal regularizer of the per-frame saliency results. The individual saliency maps represent the unary terms in a fully connected CRF, while the pair-wise connections are based on a 6D feature

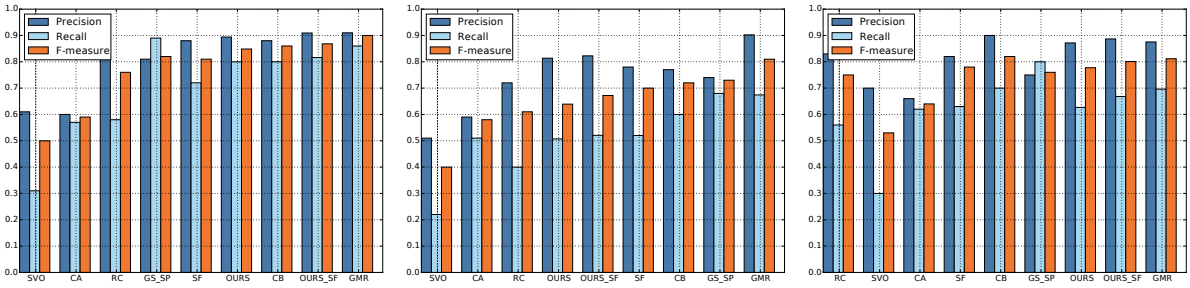

Figure 8: *Precision, recall, and F-measure for* ***MSRA*** *[CZM*11, AHES09] (left),* ***SED1*** *[AGBB12] (middle), and* ***ImgSal*** *[LLA*13] (right).*

space including color and the spatiotemporal pixel positions, analogous to previous works using a similar approach for regularizing object segmentation [LKG11] or depth [YG14]. The CRF converges after only a few iterations to a temporally coherent saliency estimate and can be implemented efficiently on a GPU [KK11], adding only negligible computational overhead to the overall saliency computation pipeline.

4. Results and evaluation

In order to estimate the quality of our computed saliency maps with respect to previous works, we evaluate the per-image saliency maps on three different standard datasets, each of them with manually labeled ground-truth saliency: the **MSRA** [CZM*11, AHES09] dataset with 1000 images, **SED1** [AGBB12] with 100 images, and **ImgSal** [LLA*13] consisting of 50 images.

In accordance with [BSI12] we compare our results to the top four performing methods: context-aware saliency (CA [GZMT10]), global-contrast (RC [CZM*11]), context-prior (CB [JWY*11]), and generic-objectness (SVO [CLCL11]). Furthermore we include comparisons with more recently developed state-of-the-art approaches that showed the best overall performance on these benchmarks: saliency filters (SF [PKPH12]), geodesic saliency (GSSP [WWZS12]), and graph manifold ranking (GMR [YZL*13]). We also combine our method with the contrast-based *saliency filters* SF [PKPH12] approach by simple averaging of the saliency maps (denoted by OURS+SF) in order to demonstrate the potentially complementary nature of our approach to contrast-based techniques.

Similar to evaluations in previous works we binarize the saliency maps using the image dependent adaptive threshold proposed by Achanta et al. [AHES09], defined as twice the mean saliency of a given saliency map S:

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x,y), \tag{6}$$

where W and H are the width and height of the saliency map S, respectively.

In order to provide a unique score that weighs the trade-off between precision and recall we also compute a combination of the two, the *F-measure*, which is defined as:

$$F_\beta = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}. \tag{7}$$

Following the experimental setups of previous approaches that emphasize the importance of precision over recall, we set $\beta^2 = 0.3$.

Quantitative evaluation. The performance evaluation on the three datasets in Figure 8 shows that our algorithm is comparable to current state-of-the-art approaches. The qualitative comparison in Figure 9 confirms the quantitative analysis, showing that our saliency maps are very close to ground truth and often less cluttered than previous approaches, with salient objects uniformly highlighted and well segmented from the background.

The evaluation also shows the complementary properties of our algorithm to contrast-based methods such as SF [PKPH12]. Despite both methods alone performing already very well, even a rather simple combination (OURS+SF) readily produces a performance gain. This can be explained by our method addressing limitations of contrast-based techniques such as difficulties caused by multicolored objects, multiple unique colors and strong color similarities between background and foreground (shown in Figure 2 and Figure 9).

Binary heuristics. There are at least three possible ways to binarize a continuously valued saliency map S_{cont}: our sign-based split described in Eq. (5), k-means clustering, or the adaptive threshold proposed in [AHES09]. We have discovered that in terms of their F-measure they all yield comparable performance on these datasets. We found that k-means provides slightly better precision at the expense of recall, and sign-based split has slightly better recall at the expense of accuracy compared to adaptive thresholding.

Parameters. The only relevant parameter in our algorithm is the quality of the superpixel segmentation. In our experiments we found regularly shaped, compact superpixels to work best. The size or number of superpixels solely influ-

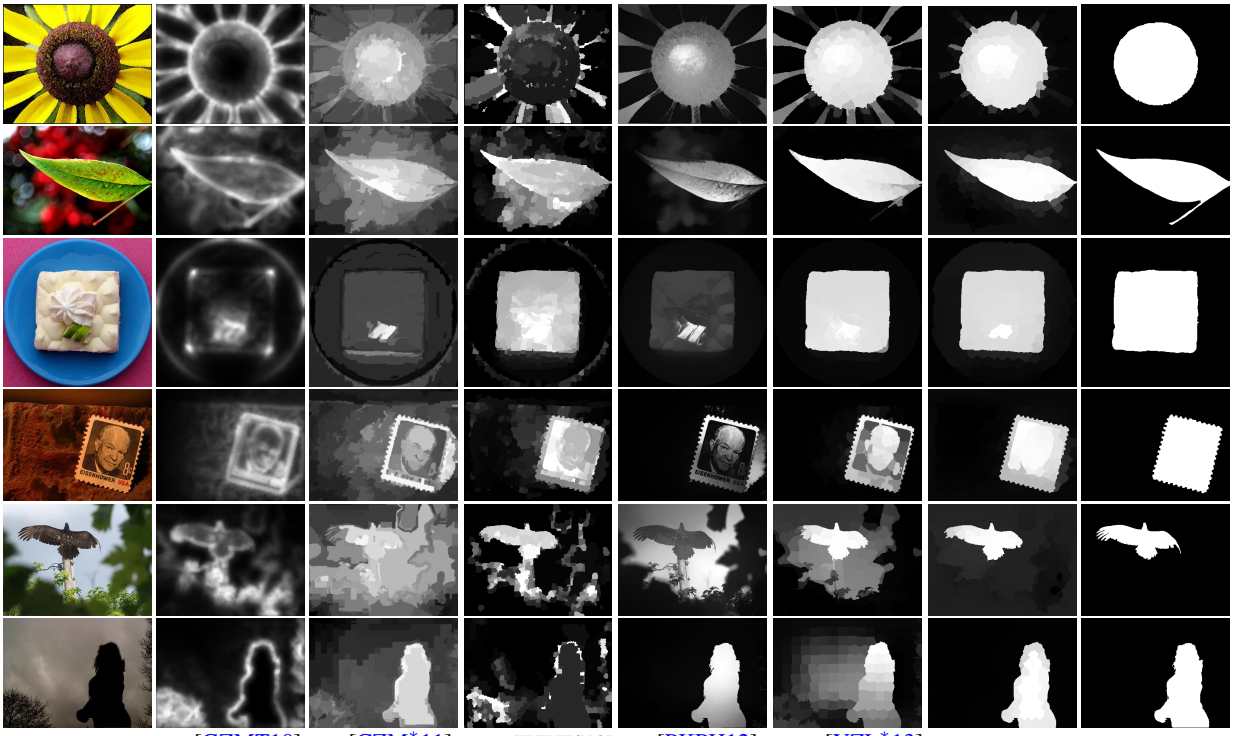

Figure 9: *Qualitative comparison of the results of our algorithm* (OURS) *with ground truth* (G) *and several other state-of-the art approaches. Our method consistently produces a foreground-background separation close to ground truth. From left to right: source image* (SRC)*, context-aware saliency* (CA *[GZMT10]*)*, global-contrast* (RC *[CZM*11]*)*, *geodesic saliency* (GSSP *[WWZS12]*)*, saliency filters* (SF *[PKPH12]*) *and graph manifold ranking* (GMR *[YZL*13]*).

ences the preservation of small image details. We employed the superpixel implementation proposed in [PKPH12] and simply used 250 superpixels for all our results. If desired, the upsampling proposed in that paper could be employed to recover fine salient structures that were removed during the superpixel segmentation.

Running time. The average running time on images from the above datasets, measured on an Intel Core i7-3820QM 2.7 GHz with 16 GB of RAM, is approximately 0.1 seconds, with most of the time required for the superpixel computation and the eigendecomposition of the Laplacian matrix. Even though our current prototype implementation in Python is not optimized for speed, its performance is already comparable to other fast state-of-the-art methods outlined in [CZM*11]. The additional regularization step depends on the respective shot length. For example, regularizing a shot of 250 frames takes about 0.5 seconds.

Video results. Due to the lack of established ground truth data for video saliency, we demonstrate the video performance and saliency quality of our method by computing saliency on a short movie [Dis14] with various types of challenging real world scenes. Please see Figure 10 for example frames and the supplemental video for longer video sequences. Note how inconsistencies between per-frame saliency maps are effectively removed.

Limitations. In its current formulation our method is particularly effective for the detection of single salient objects. For example, in Figure 3 our algorithm correctly detects the salient object with the strongest separation from the background, but fails to detect the remaining pieces. Multiple salient objects (Figure 11) could be retrieved by repeatedly segmenting the salient region [LZLX11] or using a sliding window approach [FWT*11]. For the latter, the shape of the Fiedler vector might well serve as an additional indicator of the number of salient objects within the window.

5. Conclusions and future work

We presented a method that combines the assumption of image boundaries covered mostly by background with soft graph segmentation using the Fiedler vector, yielding a continuously-valued solution to salient foreground detection and segmentation. Our approach exhibits a comparable performance to the state-of-the-art on several benchmarks. We furthermore showed that it is possible to train an SVM-based classifier on properties of the Fiedler vector to distinguish

Figure 10: *More saliency results on video. Top left: input video frame. Top right: superpixel segmentation. Bottom left: our per-frame saliency. Bottom right: final, temporally coherent saliency. Please also refer to the supplemental video for the full sequences.*

between images with and without an actual salient foreground, a unique feature of our approach. We described a simple and efficient extension to temporally regularize per-frame saliency maps, which makes the method applicable to video. The method is straightforward to implement and practically parameter-free, making it a good candidate for practical applications and further development.

As discussed in the limitations, an interesting direction for future work is the detection of multiple salient objects. Moreover, as the integration of our method with the contrast-based saliency filters shows, already simple combinations of different computational saliency methods with complementary properties can lead to improved accuracy of the computed saliency maps. In particular, in the context of automated cinematography and intelligent video editing, actors are usually the most important elements in a video. Our

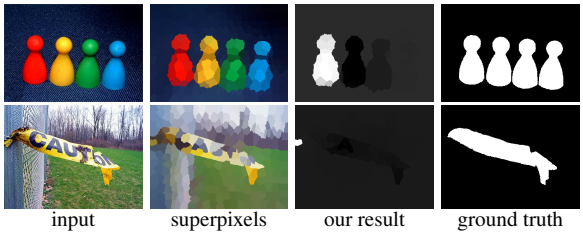

Figure 11: *Failure cases. In the case of multiple disconnected objects our current algorithm correctly detects only the most salient one. Non-salient objects with distinctive colors cause the method to fail in some instances.*

method is negligent with respect to such higher-level knowledge. A principled approach for automatically combining low-level saliency detection with techniques for actor detection and tracking (e.g. [GR13]), possibly involving machine learning, could be a promising direction for future research.

References

[ADF10] ALEXE B., DESELAERS T., FERRARI V.: What is an object? In *CVPR* (2010). 2, 3

[AEWS08] ACHANTA R., ESTRADA F. J., WILS P., SÜSSTRUNK S.: Salient region detection and segmentation. In *ICVS* (2008), pp. 66–75. 2, 3

[AGBB12] ALPERT S., GALUN M., BRANDT A., BASRI R.: Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE PAMI 34*, 2 (2012), 315–327. 6

[AHES09] ACHANTA R., HEMAMI S. S., ESTRADA F. J., SÜSSTRUNK S.: Frequency-tuned salient region detection. In *CVPR* (2009). 1, 3, 4, 6

[ASS*12] ACHANTA R., SHAJI A., SMITH K., LUCCHI A., FUA P., SÜSSTRUNK S.: Slic superpixels compared to state-of-the-art superpixel methods. *IEEE PAMI 34*, 11 (2012), 2274–2282. 3

[AWC10] ASSA J., WOLF L., COHEN-OR D.: The virtual director: a correlation-based online viewing of human motion. *Comput. Graph. Forum 29*, 2 (2010), 595–604. 1

[Bor12] BORJI A.: Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR* (2012). 2

[BSI12] BORJI A., SIHITE D. N., ITTI L.: Salient object detection: A benchmark. In *ECCV (2)* (2012), pp. 414–429. 6

[CLCL11] CHANG K.-Y., LIU T.-L., CHEN H.-T., LAI S.-H.: Fusing generic objectness and visual saliency for salient object detection. In *ICCV* (2011), pp. 914–921. 6

[CLL11] CHANG K.-Y., LIU T.-L., LAI S.-H.: From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR* (2011). 2, 3

[CST10] CRISTIANINI N., SHAWE-TAYLOR J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2010. 5

[CZM*11] CHENG M.-M., ZHANG G.-X., MITRA N. J., HUANG X., HU S.-M.: Global contrast based salient region detection. In *CVPR* (2011). 2, 3, 6, 7

[DDN08] DESELAERS T., DREUW P., NEY H.: Pan, zoom, scan - time-coherent, trained automatic video cropping. In *CVPR* (2008). 1

[Dis14] DISNEY RESEARCH: Lucid Dreams of Gabriel. http://www.disneyresearch.com/luciddreamsofgabriel/, 2014. 7

[DWM*11] DUAN L., WU C., MIAO J., QING L., FU Y.: Visual saliency detection by spatially weighted dissimilarity. In *CVPR* (2011). 3

[EKK03] EINHAUSER W., KONIG P., KONIG P.: Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience 17*, 5 (2003). 3

[EZP*13] EVANGELOPOULOS G., ZLATINTSI A., POTAMIANOS A., MARAGOS P., RAPANTZIKOS K., SKOUMAS G., AVRITHIS Y. S.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimedia 15*, 7 (2013). 1

[FMK*09] FUKUCHI K., MIYAZATO K., KIMURA A., TAKAGI S., YAMATO J.: Saliency-based video segmentation with graph cuts and sequentially updated priors. In *ICME* (2009). 1

[FWT*11] FENG J., WEI Y., TAO L., ZHANG C., SUN J.: Salient object detection by composition. In *ICCV* (2011). 7

[GKE11] GRUNDMANN M., KWATRA V., ESSA I. A.: Auto-directed video stabilization with robust L1 optimal camera paths. In *Proc. CVPR* (2011). 1

[GL08] GLEICHER M., LIU F.: Re-cinematography: Improving the camerawork of casual video. *TOMCCAP 5*, 1 (2008). 1

[GMZ08] GUO C., MA Q., ZHANG L.: Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *CVPR* (2008). 3

[GR13] GANDHI V., RONFARD R.: Detecting and Naming Actors in Movies using Generative Appearance Models. In *CVPR* (2013). 8

[GZMT10] GOFERMAN S., ZELNIK-MANOR L., TAL A.: Context-aware saliency detection. In *CVPR* (2010). 1, 3, 6, 7

[HK92] HAGEN L. W., KAHNG A. B.: New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. on CAD of Integrated Circuits and Systems 11*, 9 (1992), 1074–1085. 4

[HKP06] HAREL J., KOCH C., PERONA P.: Graph-based visual saliency. In *NIPS* (2006), pp. 545–552. 2

[HZ07] HOU X., ZHANG L.: Saliency detection: A spectral residual approach. In *CVPR* (2007). 1, 3

[IB05] ITTI L., BALDI P.: Bayesian surprise attracts human attention. In *NIPS* (2005). 2

[IKN98] ITTI L., KOCH C., NIEBUR E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI 20*, 11 (1998), 1254–1259. 1, 2

[JEDT09] JUDD T., EHINGER K. A., DURAND F., TORRALBA A.: Learning to predict where humans look. In *ICCV* (2009), pp. 2106–2113. 2

[JSSH15] JAIN E., SHEIKH Y., SHAMIR A., HODGINS J.: Gaze-driven video re-editing. *ACM Trans. Graph.* (2015). 1

[JWY*11] JIANG H., WANG J., YUAN Z., LIU T., ZHENG N.: Automatic salient object segmentation based on context and shape prior. In *BMVC* (2011), pp. 1–12. 6

[KB01] KADIR T., BRADY M.: Saliency, scale and image description. *IJCV 45*, 2 (2001), 83–105. 2

[KK11] KRÄHENBÜHL P., KOLTUN V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS* (2011), pp. 109–117. 5, 6

[KU85] KOCH C., ULLMAN S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology 4* (1985), 219–227. 1, 2

[LCL*13] LI W., CHANG H., LIEN K., CHANG H., WANG Y. F.: Exploring visual and motion saliency for automatic video object extraction. *IEEE TIP 22*, 7 (2013), 2600–2610. 1

[LCLS14] LIU R., CAO J., LIN Z., SHAN S.: Adaptive partial differential equation learning for visual saliency detection. In *Proc. CVPR* (2014). 3

[LKG11] LEE Y. J., KIM J., GRAUMAN K.: Key-segments for video object segmentation. In *ICCV* (2011), pp. 1995–2002. 6

[LLA*13] LI J., LEVINE M. D., AN X., XU X., HE H.: Visual saliency based on scale-space analysis in the frequency domain. *IEEE PAMI 35*, 4 (2013), 996–1010. 6

[LSZ*07] LIU T., SUN J., ZHENG N., TANG X., SHUM H.-Y.: Learning to detect a salient object. In *CVPR* (2007). 2, 3

[LZLX11] LU Y., ZHANG W., LU H., XUE X.: Salient object detection using concavity context. In *ICCV* (2011), pp. 233–240. 2, 3, 7

[MTZM13] MARGOLIN R., TAL A., ZELNIK-MANOR L.: What makes a patch distinct? In *CVPR* (2013). 2, 3

[PKPH12] PERAZZI F., KRÄHENBÜHL P., PRITCH Y., HORNUNG A.: Saliency filters: Contrast based filtering for salient region detection. In *CVPR* (2012). 2, 3, 4, 5, 6, 7

[PLN02] PARKHURST D., LAW K., NIEBUR E.: Modeling the role of salience in the allocation of overt visual attention. *Vision Research 42*, 1 (2002), 107–123. 3

[Ros99] ROSENHOLTZ R.: A simple saliency model predicts a number of motion popout phenomena. *Vision Research 39* (1999), 3157–3163. 1

[SM97] SHI J., MALIK J.: Normalized cuts and image segmentation. In *CVPR* (1997). 3

[SSH11] SHAMIR A., SORKINE O., HORNUNG A.: Modern approaches to media retargeting. *SIGGRAPH ASIA Courses* (2011). 1

[vL07] VON LUXBURG U.: A tutorial on spectral clustering. *Statistics and Computing 17*, 4 (2007), 395–416. 4, 5

[VS14] VIJFWINKEL M., STARAK W.: CG Textures. http://www.cgtextures.com/, 2014. 5

[WHSL11] WANG Y., HSIAO J., SORKINE O., LEE T.: Scalable and coherent video resizing with per-frame optimization. *ACM Trans. Graph. 30*, 4 (2011), 88. 1

[WKI*11] WANG M., KONRAD J., ISHWAR P., JING K., ROWLEY H. A.: Image saliency: From intrinsic to extrinsic context. In *CVPR* (2011). 3

[WWZS12] WEI Y., WEN F., ZHU W., SUN J.: Geodesic saliency using background priors. In *ECCV (3)* (2012), pp. 29–42. 3, 4, 6, 7

[YG14] YU F., GALLUP D.: 3D reconstruction from accidental motion. In *CVPR* (2014). 6

[YS04] YU S. X., SHI J.: Segmentation given partial grouping constraints. *IEEE PAMI 26*, 2 (2004). 3

[YZL*13] YANG C., ZHANG L., LU H., RUAN X., YANG M.-H.: Saliency detection via graph-based manifold ranking. In *CVPR* (2013). 3, 5, 6, 7

[ZLZZ08] ZHANG C., LIU Z., ZHANG Z., ZHAO Q.: Semantic saliency driven camera control for personal remote collaboration. In *MMSP* (2008). 1

[ZPS*13] ZUND F., PRITCH Y., SORKINE-HORNUNG A., MANGOLD S., GROSS T. R.: Content-aware compression using saliency-driven image retargeting. In *ICIP* (2013). 1