PISA: Pixelwise Image Saliency by Aggregating Complementary Appearance Contrast Measures with Spatial Priors

Keyang Shi[1], Keze Wang[1], Jiangbo Lu[2], Liang Lin[1*]
[1]Sun Yat-Sen University, Guangzhou, China
[2]Advanced Digital Sciences Center, Singapore

Abstract

Driven by recent vision and graphics applications such as image segmentation and object recognition, assigning pixel-accurate saliency values to uniformly highlight foreground objects becomes increasingly critical. More often, such fine-grained saliency detection is also desired to have a fast runtime. Motivated by these, we propose a generic and fast computational framework called PISA – Pixelwise Image Saliency Aggregating complementary saliency cues based on color and structure contrasts with spatial priors holistically. Overcoming the limitations of previous methods often using homogeneous superpixel-based and color contrast-only treatment, our PISA approach directly performs saliency modeling for each individual pixel and makes use of densely overlapping, feature-adaptive observations for saliency measure computation. We further impose a spatial prior term on each of the two contrast measures, which constrains pixels rendered salient to be compact and also centered in image domain. By fusing complementary contrast measures in such a pixelwise adaptive manner, the detection effectiveness is significantly boosted. Without requiring reliable region segmentation or post-relaxation, PISA exploits an efficient edge-aware image representation and filtering technique and produces spatially coherent yet detail-preserving saliency maps. Extensive experiments on three public datasets demonstrate PISA's superior detection accuracy and competitive runtime speed over the state-of-the-arts approaches.

1. Introduction

Saliency detection in natural images is an important task useful for many computer vision applications. Given an input image, the general objective is to automatically detect salient objects and assign consistently high saliency values to them, while the background part should take on zero values ideally. Though quite challenging, being able to separate salient objects from the background is a very useful tool for many computer vision and graphics applications such as object recognition [22], content-aware image retargeting [23], and image classification [20]. Driven by these recent applications, saliency detection has also evolved to aim at assigning pixel-accurate saliency values to uniformly highlight foreground objects, going far beyond its early goal of mimicing human eye fixation. More often, such fine-grained saliency detection is also desired to have a fast runtime. This paper focuses on addressing these challenges increasingly pressed by recent application requirements.

Without any user intervention, inferring (pixel-accurate) saliency assignment for diversified natural images is a highly ill-posed problem, because of the lack of a rigorous definition of saliency itself. To tackle this problem, a myriad of computational models [21, 11, 7, 24, 8, 15, 4] have been proposed using various principles or priors ranging from high-level biological vision [12] to low-level image properties [10, 8]. Focusing on bottom-up, low-level saliency computational models in this paper, we can classify most of the previous methods into two basic classes depending on the way the saliency cues are defined: *contrast priors* and *background priors* [24]. Contrast priors have been widely adopted in many previous methods to model the appearance contrast between foreground objects and the background. Various appearance contrast measures can be computed either in a local neighborhood of a pixel or patch [11, 15] or from an entire image context globally [5, 1]. Typical limitations of the existing methods based on contrast priors include attenuated object interior e.g. Fig. 1(e) and ambiguous saliency detection for images with rich structures in foreground or/and background e.g. Fig. 1(f-h). Complementing the prime role of contrast priors in this research topic, background priors [24] have been proposed recently to exploit two interesting priors about backgrounds – boundary and connectivity priors. This approach

[*]Corresponding author is Liang Lin (linliang@ieee.org). This work was supported by the 863 Program of China (no.2012AA011504), the National Natural Science Foundation of China (no. 61173082), and the Special Project on the Integration of Industry, Education and Research of Guangdong Province (no. 2012B091100148). This study was also supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

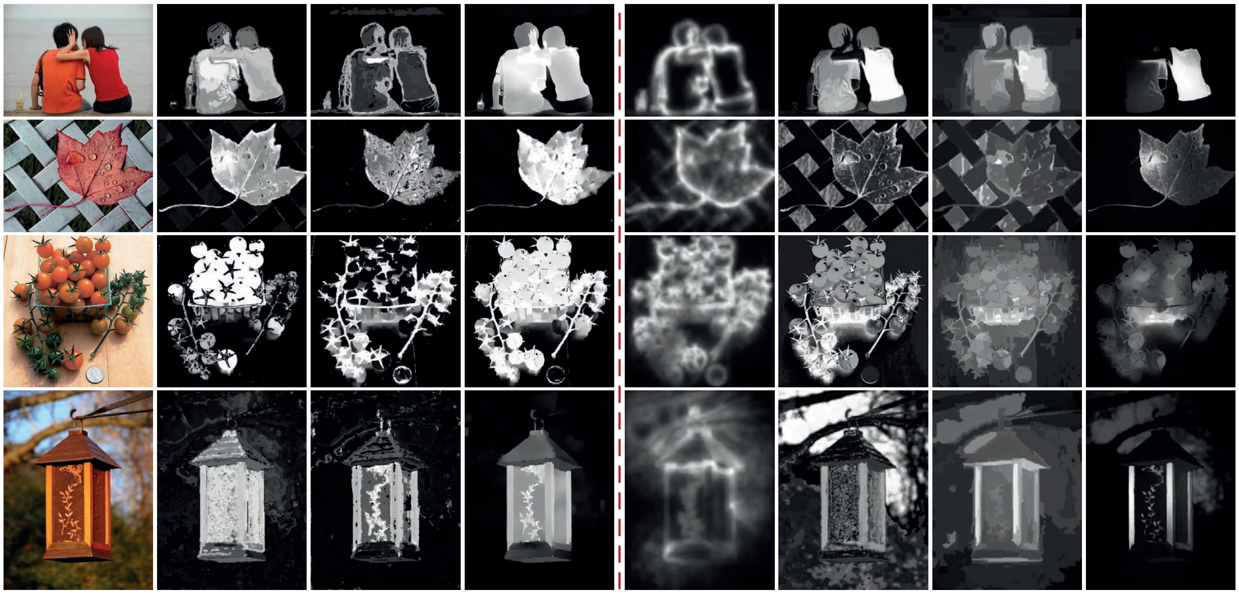


Figure 1. Saliency map comparisons on (a) four example images detected by (d) our PISA method and (e-h) a few representative contrast prior based methods modeling only the color contrast [5, 19, 7]. (b/c) Raw saliency detection result using the color/structure contrast measure alone in the proposed PISA framework.

demonstrates the detection effectivenesses from a new perspective. However, this method fails when objects touch the image boundary to quite some extent, or when connectivity assumptions are invalid in the presence of complex backgrounds or textured scenes. For instance, the maple leave case in Fig. 1 poses a challenge for this method [24].

Inspired by the insights and lessons from the significant amount of previous work, we target studying this challenging saliency detection problem in a more holistic manner. We also keep computational efficiency as one of important desiderata. More specifically, this work is primarily motivated by three key principles or priors supported by psychological evidence and observations of natural images:

Complementary appearance contrast in a global context. Though the color contrast is a popular saliency cue used dominantly in many methods [5, 19, 14], other influential factors do exist, which make certain pixels or regions outstanding. For instance, they can have unique appearance features in edge/texture patterns [11].

Attention cue-adaptive receptive field and region-based non-parametric feature modeling. It is known from perceptual research [6] that different local receptive fields are associated with different kinds of visual stimuli, so local analysis regions where saliency cues are extracted should be adapted to match specific image attributes. In addition, using a non-parametric distribution to summarize the extracted features tends to be more robust than relying on just a few quantities computed for a pixel or region [7, 15].

Spatial priors and edge-preserving spatial coherence. Previous works have used the spatial variance to further modulate saliency values computed from a single visual attribute (e.g. color [19, 5]). This spatial prior can also be generalized to consider the spatial distribution of different saliency cues, including also other useful location priors such as the center prior [15]. Another observation is that pixel-accurate saliency maps are often spatially coherent with the discontinuities well aligned to image edges.

Based on these principles, we propose a generic and fast computational framework called PISA – Pixelwise Image Saliency Aggregating complementary saliency cues based on color and structure contrasts with spatial priors holistically. Overcoming the limitations of previous methods, PISA advances in following aspects: (*i*) Instead of using homogeneous superpixel-based and color contrast-only treatment, PISA directly performs saliency modeling for each individual pixel on two complementary measures (color and structure contrast) and makes use of densely overlapping, feature-adaptive ovservations for saliency measure computation. (*ii*) We further impose a spatial prior term on each of the two contrast measures, which constrains pixels rendered salient to be compact and also centered in image domain. By fusing complementary contrast measures in such a pixelwise adaptive manner, the detection effectiveness is significantly boosted. (*iii*) Without requiring reliable region segmentation and then post-relaxation for pixelwise saliency assignment, PISA exploits an efficient edge-aware image representation and filtering technique [16] to produce spatially coherent yet edge-preserving saliency maps. Fig. 1 shows a few motivating examples that highlight the advantage of our PISA method, compared with some leading methods [5, 19, 7].

To balance the accuracy-efficiency trade-off, we also propose a faster version called F-PISA. It first performs saliency computation for a feature-driven, subsampled image grid, and then uses an adaptive upsampling scheme with the color image as the guidance signal to recover a full-resolution saliency map. Compared to segmentation-based saliency methods [19], our F-PISA method reduces the computational complexity similarly by considering a coarse image grid, while having the advantage of utilizing image structural information for saliency reasoning over [19]. Our extensive experiments on three public datasets demonstrate the superior detection accuracy and competitive runtime speed of our approach over the state-of-the-arts.

2. Overview of the PISA Framework

As motivated in Sect. 1, we propose PISA in this paper as a computational framework for effective and efficient pixel-accurate saliency detection, aggregating complementary saliency cues based on color and structure contrasts with spatial priors holistically. In the framework, a saliency measure representing the structure contrast is proposed in addition to the well exploited color-based measure. These two measures complement each other in detecting saliency cues from different perspectives, and are combined together to give the initial saliency value. More formally, given an image I, we compute the initial saliency value $\tilde{\mathcal{S}}(p)$ for each pixel p by aggregating the two contrast measures $\{U^c(p), U^g(p)\}$ with spatial priors $\{D^c(p), D^g(p)\}$, giving a general PISA framework as:

$$\tilde{\mathcal{S}}(p) = U^c(p) \cdot D^c(p) + U^g(p) \cdot D^g(p) \ . \qquad (1)$$

Four terms are computed for pixel p in (1), which are:

Appearance contrast terms $\{U^c(p), U^g(p)\}$. They are evaluated based on the general contrast prior principle that rare or infrequent visual features in a global image context give rise to high salient values. $U^c(p)$ denotes the rarity of pixel p with respect to the entire image in the color feature space (Sect. 3.1). $U^g(p)$ computes the uniqueness of pixel p in the orientation-magnitude (OM) feature space for all the pixels (Sect. 3.2). Rather than describing the features for pixel p by a single or just a few quantities, we use non-parametric histogram distributions to capture and represent both the color and OM features within an appropriate pixelwise adaptive neighborhood around p.

Spatial prior terms $\{D^c(p), D^g(p)\}$. They are evaluated based on the generally valid spatial prior that salient pixels tend to have a compact spatial distribution or small spatial variance in image domain, while background can distribute quite widely over the entire image. Therefore, a pixel p should not be rendered salient, if its visually similar peers have a high spatial variance. It is also often useful to integrate the center prior in this saliency reweighting process. We use $D^c(p)$ and $D^g(p)$ to denote such an integrative spatial reweighting term imposed on the color and structure contrast measure contrast, respectively (Sect. 3.3).

By fusing the two complementary saliency cues in such a pixelwise adaptive manner, the saliency detection effectiveness is significantly boosted. Though the initial saliency estimation map $\tilde{\mathcal{S}}$ is already good for some applications, it is not pixel-accurate and still exhibits many spurious noises or unsmooth saliency values even within a small neighborhood. We hence employ an efficient edge-aware image filtering [16] to smooth out $\tilde{\mathcal{S}}$ to generate a filtered output $\mathcal{S}$, which is spatially coherent and with the saliency discontinuities aligned to the guidance color image edges (Sect. 3.4).

In fact, the aforementioned four terms and their aggregation as in (1) present only one specific implementation of our PISA framework, other kinds of saliency cues or priors can be integrated as well. In this paper, we instead continue introducing a faster method to evaluate $\tilde{\mathcal{S}}$, which is called F-PISA (Sect. 3.5). F-PISA generates saliency maps at the detection accuracy close to that achieved by PISA, but it brings an over 18-times speedup over PISA.

3. PISA Algorithm

3.1. Color-Based Contrast Term

Directly computing pixelwise color contrast in a global image context is computationally expensive, as its complexity is $O(N^2)$ with N being the number of pixels in I. Recently, Cheng *et al.* [5] proposed an effective and efficient color-based contrast measure HC. They assume that if the spatial correlation is not accounted for, pixels with the similar appearance should be assigned the same saliency values. However, their strategy of defining the contrast on the color value of a single pixel individually is sensitive to noise, and it is not extensible for measuring additional attribute. In this work, we compute the color contrast based on non-parametric color distributions extracted from a locally adaptive homogeneous region [16]. As pixels within the adaptive region share similar appearances with the central pixel, they provide a more robust color-based measure. Different from those methods [19, 5] that count on image oversegmentation to delineate compact and boundary-preserving regions, our method allows topologically more flexible region construction for each pixel. Moreover, taking segments as the atomic units for saliency evaluation does not lend itself to easy integration of other appearance contrast measures.

For each pixel p, we first construct a shape-adaptive observation region Ω_p efficiently using the CLMF method [16] (see Fig. 2). A color histogram $\mathbf{h}^c(p)$ for pixel p is then built from the pixels $q \in \Omega_p$ covered in the localized homogeneous region. Using $\mathbf{h}^c(p)$ rather than I_p is more consistent with the psychological evidence on human eyes' receptive field on homogeneous regions. Using the *Lab* color space, we quantize each color channel uniformly into 12 bins, so the color histogram $\mathbf{h}^c(p)$ is a 36-d descriptor (see Fig. 2).

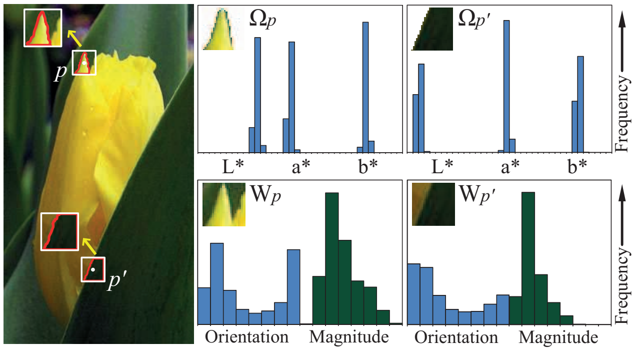


Figure 2. The color descriptor is extracted from the pixel-adaptive region $\Omega_{p/p'}$ (top) and the orientation-magnitude (OM) descriptor captures the structures within a local window $W_{p/p'}$ (bottom).

Next, we cluster pixels that share similar color histograms together using *kmeans*. The whole color feature space for the input image I is quantized into K_c clusters, indexed by $\{\phi_1, \ldots, \phi_{K_c}\}$. As a result, we use the rarity of color clusters as the proxy to evaluate the rarity or contrast measure for pixels. Let ϕ_i denote the cluster that pixel p, or more precisely $\mathbf{h}^c(p)$, is assigned to. We estimate the color-based contrast measure $U^c(p)$ for pixel p as,

$$U^c(p) = U^c(\mathbf{h}^c(p)) = \sum_{j=1}^{K_c} \omega_j \|\mathbf{h}^c(\phi_i), \mathbf{h}^c(\phi_j)\| \, . \qquad (2)$$

ω_j uses the number of pixels belonging to the cluster ϕ_j as a weight to emphasize the color contrast to bigger clusters. $\mathbf{h}^c(\phi_i)$ is the average color histogram of the cluster ϕ_i.

Feature space quantization may cause undesirable artifacts. Similar color histograms can sometimes be quantized into different clusters. To tackle this problem, we have applied two schemes. First, we slightly modify *kmeans* in its distance calculation when clustering. In addition to the L2 distance between the two histograms, we add the color dissimilarity between the center pixels into the distance measurement. Second, we adopt a linearly-varying smoothing scheme [5] to refine the quantization-based saliency measurement. The saliency value of each cluster is replaced by the weighted average of the saliency values of visually similar clusters. Larger weights are assigned to those clusters which share more similar color features. Such a refinement smooths the saliency assignment to each pixel.

The cluster number K_c of the color feature space is adaptively decided with regard to image content. Specifically, we choose the most frequently occurring color features by ensuring they cover 95% of histogram distributions of all pixels in the input image I. K_c typically takes values in the range of 10 to 256. This scheme is similar to that used in [5], and it reduces the computational complexity from $O(N^2)$ to $O(N \cdot K_c) + O(K_c^2)$, where the second term corresponds to the complexity of *kmeans*.

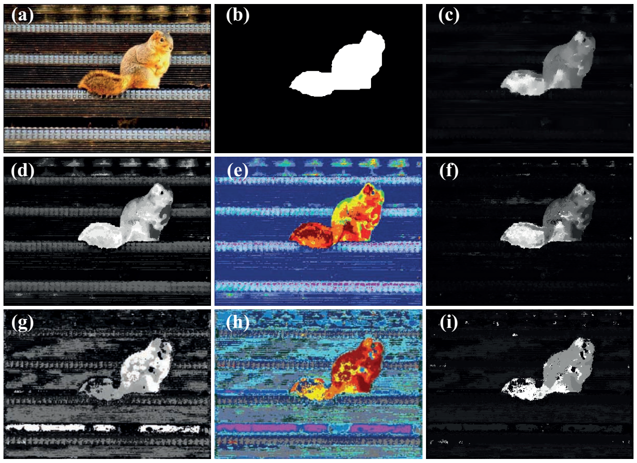


Figure 3. Main workflow of PISA. (a) Input image [1]. (b) Ground-truth map. (c) PISA result $\mathcal{S}$. (d) Color contrast measure U^c. (e) Cluster assignment in the color feature space. (f) Spatial prior-modulated color measure $U^c \cdot D^c$. (g) Structure contrast measure U^g. (h) Cluster assignment in the OM feature space. (i) Spatial prior-modulated structure contrast measure $U^g \cdot D^g$.

3.2. Structure-Based Contrast Term

As motivated in Fig. 1(second and third rows), using color information only is not adequate to discriminatively describe and detect salient objects or parts of them from the background. Even in the event that the color uniqueness measurement gives a good saliency value to foreground objects, other complementary contrast measures can still be helpful in reinforcing the saliency assignment e.g. Fig. 1(fourth row). Based on the PISA framework, we propose a structure-based descriptor to complement the color descriptor here. The proposed structure descriptor models the image gradient distribution for pixel p by a histogram $\mathbf{h}^g(p)$ in a rectangular region W_p. $\mathbf{h}^g(p)$ measures the occurrence frequency of a concatenated vector consisting of the gradient orientation component and magnitude component. Similarly, we quantize both components into 8 bins, and call the resulting feature space the OM space. It is clear that a point in such a OM space is of 16-d (see Fig. 2). In this paper, we fix the local window W_p to 9×9, which is comparable to Ω_p used for the color histogram extraction. As will be shown later, we find that our OM structure descriptor, though simple, is more effective and reliable than other gradient features e.g., Gabor [17] and LBP [9] in the image saliency detection task.

Similar to the color contrast measure, *kmeans* is utilized to partition the OM feature space into K_g clusters indexed by $\{\varphi_1, \ldots, \varphi_{K_g}\}$. The structure contrast measure for pixel p is equivalent to measuring φ_i that p is grouped to, as,

$$U^g(p) = U^g(\mathbf{h}^g(p)) = \sum_{j=1}^{K_g} \omega_j \|\mathbf{h}^g(\varphi_i), \mathbf{h}^g(\varphi_j)\| \, , \qquad (3)$$

ω_j is the weight stressing the contrast against bigger clus-

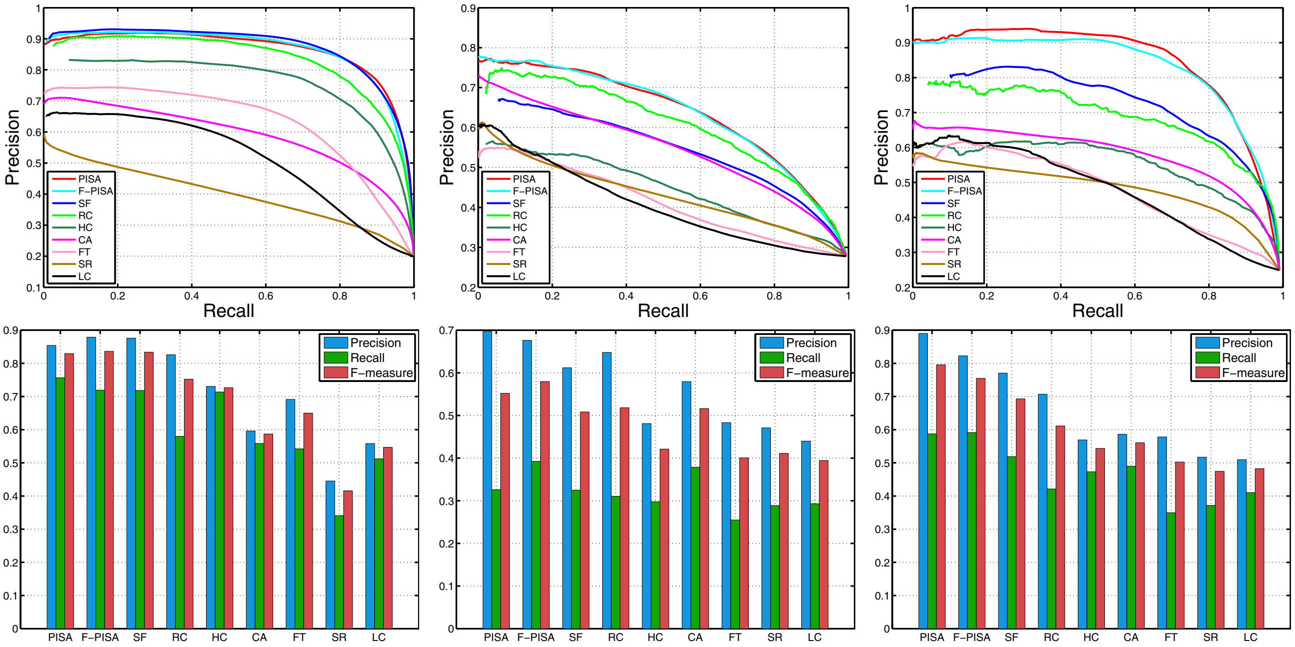


Figure 4. Precision-recall curves (top) and precision-recall bars with F-measure (bottom) for comparing previous works with the proposed PISA and F-PISA methods on the three datasets from left to right: ASD [1], SOD [18] and SED1 [2], respectively. PISA performs consistently better than the other methods. The visual comparisons of our methods and previous works are shown in Fig. 5.

ters. $\mathbf{h}^g(\varphi_i)$ is the average OM histogram of the cluster φ_i.

U^g can suffer from the influence of side effects caused by the brute-force feature space quantization process. Again, we use the local smoothing scheme to alleviate these artifacts. The cluster number K_g is determined by representing the most frequent OM vectors and accounting for at least 95% pixels. We observe K_g typically varies from 10 to 40.

3.3. Spatial Priors

Motivated by recent works [3, 19, 7, 15], we impose a spatial prior term on each of the two contrast measures $\{U^c(p), U^g(p)\}$, constraining pixels rendered salient to be compact and centered in image domain based on intra-cluster distance which is more compelling than the use of simpler center-surround structures. For each pixel p, we evaluate the initial spatial prior term $\tilde{D}^{c/g}(p)$ based on the cluster ϕ_i/φ_i that contains p from two aspects: 1) compactness of salient objects defined by the intra-cluster spatial variance, and 2) preference to the image center. Combining these two criteria, we compute $\tilde{D}^{c/g}(p)$ as follows,

$$\tilde{D}^{c/g}(p) = 1/n_i \sum_{l=1}^{n_i} (\|\mathbf{x}_l, \mu_i\| + \lambda \cdot \|\mathbf{x}_l, \mathbf{c}\|) \ . \qquad (4)$$

n_i is the number of pixels which are contained in the same color (or OM) cluster ϕ_i (or φ_i) with p. The mean spatial position of the cluster ϕ_i/φ_i is defined by $\mu_i = \frac{1}{n_i}\sum_{l=1}^{n_i} \mathbf{x}_l$. $\mathbf{c}$ is the image center position. We use a user-specified parameter λ to control the relative weight of the center prior.

Since clusters exhibiting higher spatial variance or farther from the image center are quite unlikely to be salient, we compute the final spatial prior term $D^{c/g}(p)$ for pixel p using a threshold $\mathcal{T}$ as,

$$D^{c/g}(p) = \begin{cases} \exp(-\kappa \cdot \tilde{D}^{c/g}(p)) & \tilde{D}^{c/g}(p) \le \mathcal{T} \\ 0 & \text{otherwise} \ . \end{cases} \qquad (5)$$

κ controls the fall-off rate of the exponential function.

By now we have all the four terms necessary for computing $\tilde{\mathcal{S}}(p)$ in (1) defined. Fig. 3 illustrates these dense maps visually and their respective effects to saliency assignment.

3.4. Saliency Coherence

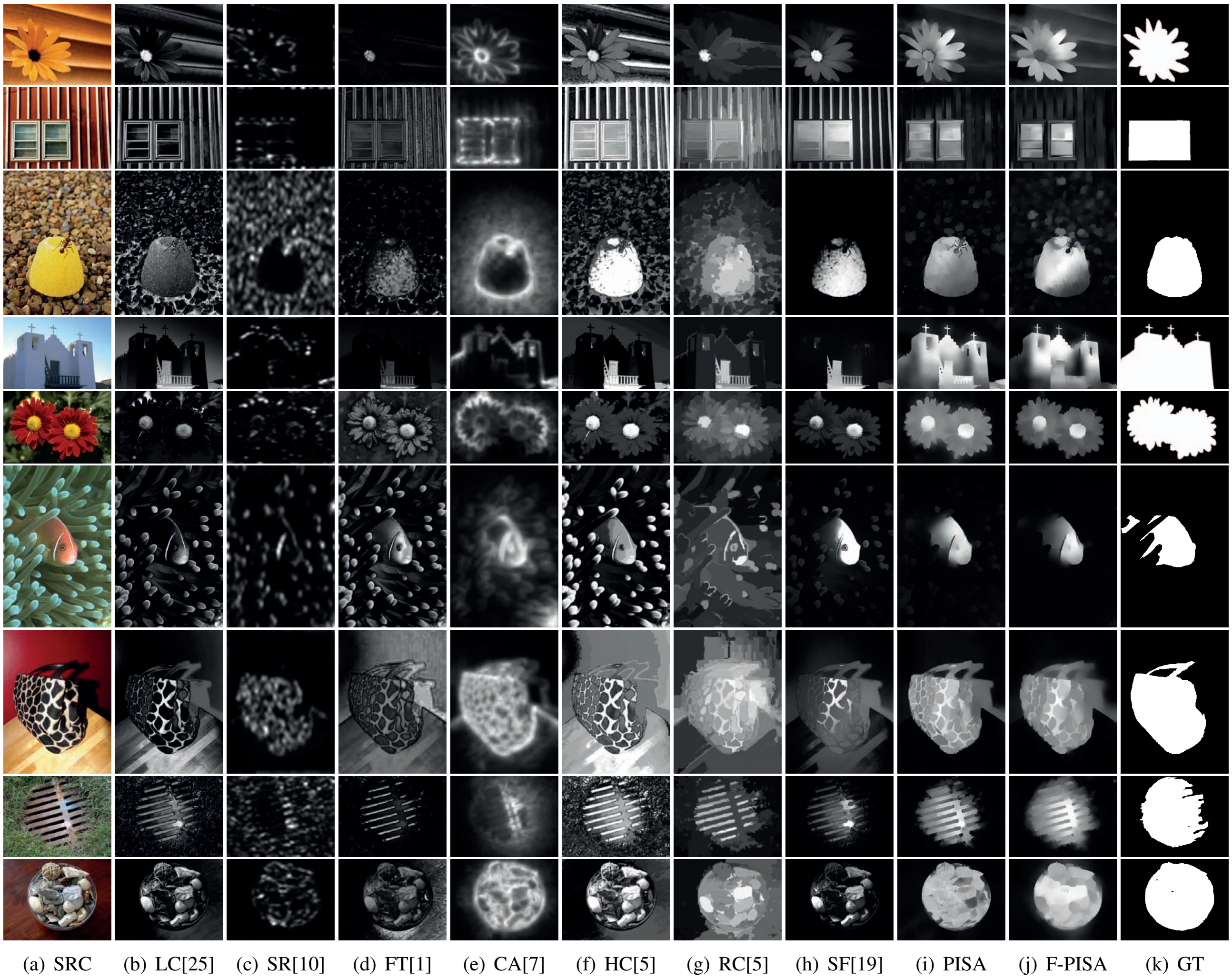
Based on (1), an initial saliency estimation map $\tilde{\mathcal{S}}$ is generated. Though good for certain applications, this initial saliency map does not consider the spatial coherence in its evaluation, resulting in spurious noises and non-uniform saliency assignment even for pixels close to each other. We therefore employ the efficient CLMF filtering technique [16] here to smooth out $\tilde{\mathcal{S}}$ and produce a spatially coherent yet discontinuity-preserving saliency map $\mathcal{S}$. In fact, the same cross-based data structure already computed when evaluating $U^c(p)$ can be reused here. This refinement step takes the following form:

$$\mathcal{S}(p) = \sum_{q\in\Omega_p} \omega_{pq}\tilde{\mathcal{S}}(q) \ . \qquad (6)$$

Ω_p is the shape-adaptive support region defined for pixel p in Sect. 3.1. ω_{pq} is the normalized support weight [16].

3.5. F-PISA: Fast Implementation

Salient object detection is often applied as a preprocessing technique for subsequent applications. To op-



(a) SRC (b) LC[25] (c) SR[10] (d) FT[1] (e) CA[7] (f) HC[5] (g) RC[5] (h) SF[19] (i) PISA (j) F-PISA (k) GT

Figure 5. Visual comparisons between existing methods and our PISA and F-PISA methods on all the three datasets: ASD [1] (top three rows), SOD [18] (middle three rows) and SED1 [2] (bottom three rows).

timize accuracy-complexity trade-off, we present a faster version F-PISA. Instead of processing the full image grid, we perform a gradient-driven subsampling of the input image I, so the saliency computation in (1) is only applied to this set of selected pixels. More specifically, for a given image I, we pick the pixel with the largest gradient magnitude from a 3×3 rectangular patch on the regular image grid to form a sparse image I^l. The two proposed contrast saliency measures are then computed for I^l, giving a sparse saliency map $\tilde{\mathcal{S}}^l$. To obtain a full-sized saliency map $\mathcal{S}$, we propagate the saliency values among the pixels within the same cross support region [16], as they share the similar appearance. This propagation scheme resembles the principle of joint bilateral upsampling [13], using a high-resolution color image I as a guidance to upsample a sparsely-valued solution map $\tilde{\mathcal{S}}^l$. It can produce a smoothly varying dense saliency map $\mathcal{S}$ without blurring the edges of salient objects. Thus given a pixel $p \in I$, its saliency value is obtained as,

$$\mathcal{S}(p) = \frac{1}{m}\sum_{i=1}^{m} \alpha_{pq_i}\tilde{\mathcal{S}}^l(q_i) \ , \qquad (7)$$

where $q_i \in I^l$ and its cross support region Ω_{q_i} contains p, namely $p \in \Omega_{q_i}$. m is the total number of such q_i pixels. $\alpha_{pq_i} = \exp(-\frac{\|\mathbf{x}_p, \mathbf{x}_{q_i}\|}{\sigma})$, which gives higher weights to the support pixels with a shorter spatial distance to pixel p.

4. Experiments

4.1. Evaluation on Benchmarks

We evaluate the proposed algorithm for saliency detection on three public datasets which have been used as standard benchmarks in [3]. The ASD dataset [1] contains 1,000 images, which has been widely used by recent methods [5, 19, 1]. The SOD dataset [18] is more challenging, including complex objects and scenes, and we obtain the groundtruth for this dataset from the authors of the work [24]. The SED1 dataset [2] is exploited recently, and we consider a pixel salient if it is annotated as salient by all subjects. For all the three datasets, we use the following

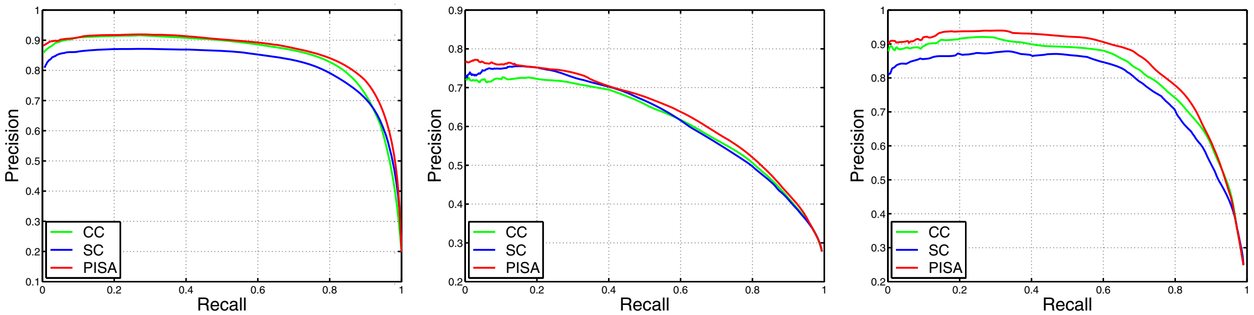


Figure 6. Extensive study for different saliency measures (CC, SC) in our method. The experiments are executed on all the three datasets, from left to right: ASD [1], SOD [18] and SED1 [2]. We observe the advantage of aggregating the two complementary contrast measures.

CA [7]	HC [5]	RC [5]	SF [19]	PISA	F-PISA
42.9	0.011	0.115	0.125	0.650	0.035

Table 1. Comparisons of the average runtime (seconds per image) on the ASD [1] dataset. CA [7] uses the Matlab implementation, while the rest are implemented in C++.

parameter settings $\{\lambda, \kappa, \sigma\} = \{1.0, 0.01, 0.17\}$. The only exception is $\mathcal{T}$, we set $\mathcal{T} = 25, 40, 30$ for ASD, SOD and SED1 respectively, as the spatial distributions of the foreground objects in the three datasets are different. We also set $\tau = 30, L = 4$ for the adopted CLMF technique.

We compare our methods on all datasets with several state-of-the-art works: Spatial-temporal Cues (LC [25]), Spectral Residual saliency (SR [10]), Frequency-Tuned saliency (FT [1]), Context-Aware saliency (CA [7]), Histogram-based Contrast (HC [5]), Region-based Contrast (RC [5]) and Saliency Filter (SF [19]). Results of LC, SR, FT, HC, RC are generated by using the codes provided by [5], and we adopt the public implementations from the original authors for CA and SF.

We use (P)precision-(R)recall curves and $F_{0.3}$ metric to evaluate the detection performance similar as [1, 5, 19]. The results of PR curves and precision, recall and F-measure are shown in the first and second row of Fig. 4 respectively. Based on the results, our PISA method achieves state-of-the-art accuracy on all the three datasets, demonstrating the advantages consistently. Fig. 5 shows the visual comparisons between our methods against other competing approaches. Our methods consistently perform better than the others on a variety of challenging images.

The average runtimes of our approaches and competing methods on the ASD [1] dataset are reported in Table 1. The experiments are carried out on an Intel Core i5 3.0GHz with 2GB RAM. Our fast implementation F-PISA significantly improves the efficiency (i.e. 18 times faster than PISA), while keeping good detection effectiveness (see Fig. 4, 5).

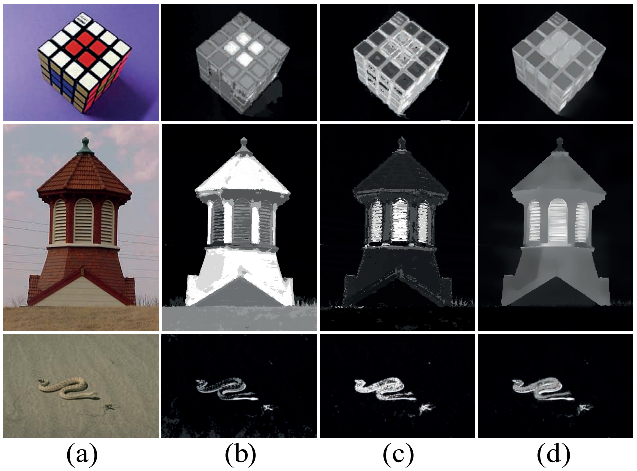


Figure 7. Visual comparisons. (a) Input image. (b) Spatial prior-modulated color measure. (c) Spatial prior-modulated structure measure. (d) PISA result.

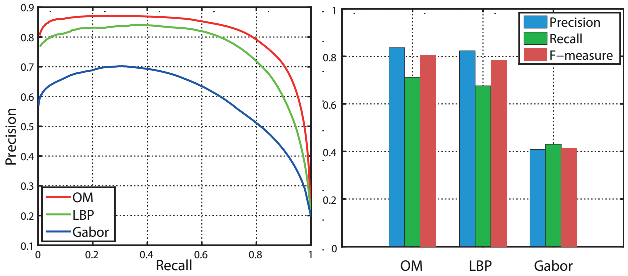


Figure 8. Empirical study on two common structure features Gabor and LBP for replacing our OM feature in the PISA framework. Our OM descriptor performs better on the ASD [1] dataset.

4.2. Component Analysis

We further analyze the effectiveness of the two complementary measures, i.e. color-based contrast (CC) and structure-based contrast (SC). The quantitative results in Fig. 6 demonstrate the requisite of aggregating the two measures. We can observe that the aggregated saliency detection achieves superior performance, as CC and SC capture saliency from different aspects, verified by the visual results in Fig. 7. It is worth noting that we obtain favorable results on the images in the second and third rows in Fig. 1 and the fourth row in Fig. 5, which are exhibited in [24] and [5] as failure cases. They serve as good evidences to advocate our choice in fusing complementary saliency cues.

We also explored other commonly used features Gabor [17] and LBP [9] to substitute OM for capturing structure information. For all the features, we choose their best

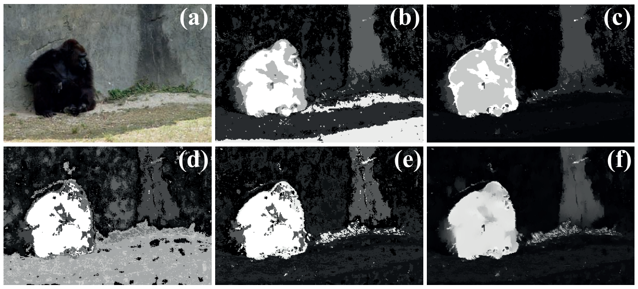


Figure 9. An example case challenging PISA. (a) Input image [1]. (b) Color contrast measure. (c) Spatial prior-modulated color measure. (d) Structure contrast measure. (e) Spatial prior-modulated structure measure. (f) PISA result.

results for comparison by tuning their quantizations. The dimensions for Gabor and LBP features are 72 and 256, respectively. The PR-curves of the experiments evaluated on the ASD dataset [1] are shown in Fig. 8. The OM descriptor outperforms the others. Meanwhile, under the proposed computational model, our OM descripor also shows higher efficiency than Gabor and LBP due to its small dimension.

4.3. Limitations

In Fig. 9, we present an unsatisfying result generated by PISA. As our approach uses the spatial priors, it has problems when such priors are invalid. For example, if the center prior does not hold, the background regions located near the image center cannot be effectively suppressed in saliency evaluation (see Fig. 9 (e)). By adjusting the contribution of this prior through tuning λ, we can alleviate the influence of this prior. Another limitation stems from the additive form of our formulation. For any background regions that have been assigned high saliency values from either of the contrast cues after the modulation of the spatial priors, they remain salient in the final saliency map (see Fig. 9 (f)). This problem could be tackled by incorporating high-level knowledge to adjust the confidence of two measures in the formulation.

5. Conclusion

We have presented a generic framework for pixelwise saliency detection via aggregating two complementary appearance contrast measures (color and structure) with spatial priors. We extensively evaluate our methods on three public datasets by comparing with previous works. Experimental results demonstrate the advantages of the proposed PISA methods in detection accuracy consistency and speed. For future work, we plan to incorporate high-level knowledge, which could be beneficial to handle more challenging cases and investigate other kinds of saliency cues or priors to be embedded into the PISA framework.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*, 2007.

[3] A. Borji, D. Sihite, and L. Itti. Salient object detection: a benchmark. In *ECCV*, 2012.

[4] N. Bruce and K. Tsotsos. Saliency based on information maximization. In *NIPS*, 2005.

[5] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[6] W. Einhäuser, and P. König. Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 2003.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, 2010.

[8] C. Guo, Q. Ma, and L. Zhang. Spatial-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008.

[9] M. Heikkilä, and M. Pietikäinen, C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 2009.

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 1998.

[12] C. Koch, and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 1985.

[13] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM TOG*, 2007.

[14] Z. Liang, M. Wang, X. Zhou, L. Lin, and W. Li. Salient object detection based on regions. *Multimedia Tools and Applications*, 2012.

[15] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[16] J. Lu, K. Shi, D. Min, L. Lin, and M. Do. Cross-based local multipoint filtering. In *CVPR*, 2012.

[17] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE TPAMI*, 1996.

[18] V. Movahedi and J. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *POCV*, 2010.

[19] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: contrast based filtering for salient region detection. In *CVPR*, 2012.

[20] G. Sharma, F. Jurie, and C. Schmid. Discriminative Spatial Saliency for Image Classification. In *CVPR*, 2012.

[21] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.

[22] L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue. In *ICCV*, 2011.

[23] Y. Wang, C. Tai, O. Sorkine, and T. Lee. Optimized scale-and-stretch for image resizing. *ACM TOG*, 2008.

[24] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012.

[25] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *CVPR*, 2006.