Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Visual Saliency Estimation through Manifold Learning

Richard M. Jiang
Department of Computer Science,
The University of Bath,
Bath, United Kingdom
Email: m.jiang@bath.ac.uk

Danny Crookes
ECIT Institute,
Queen's University Belfast,
Belfast, United Kingdom
Email: d.crookes@qub.ac.uk

Abstract

Saliency detection has been a desirable way for robotic vision to find the most noticeable objects in a scene. In this paper, a robust manifold based saliency estimation method has been developed to help capture the most salient objects in front of robotic eyes, namely cameras. In the proposed approach, an image is considered as a manifold of visual signals (stimuli) spreading over a connected grid, and local visual stimuli are compared against the global image variation to model the visual saliency. With this model, manifold learning is then applied to minimize the local variation while keeping the global contrast, and turns the RGB image into a multi channel image. After the projection through manifold learning, histogram based contrast is then computed for saliency modeling of all channels of the projected images, and mutual information is introduced to evaluate each single channel saliency map against prior knowledge to provide cues for the fusion of multiple channels. In the last step, the fusion procedure combines all single channel saliency maps according to their mutual information score, and generates the final saliency map. In our experiment, the proposed method is evaluated using one of the largest publicly available image datasets. The experimental results validated that our algorithm consistently outperforms the state of the art unsupervised saliency detection methods, yielding higher precision and better recall rates. Furthermore, the proposed method is tested on a video where a moving camera is trying to catch up with the walking person a salient object in the video sequence. Our experimental results demonstrated that the proposed approach can successful accomplish this task, revealing its potential use for similar robotic applications.

Introduction

Visual saliency is an efficient way of capturing the most noticeable part in a scene, and can give the most usable cues for robotic vision (Butko *et al* 2008, Sarma 2006). However, visual saliency is usually a multidisciplinary topic involving cognitive psychology (Teuber 1955, Wolfe & Horowitz 2004), neurobiology (Desimone & Duncan 1995, Mannan *et al* 2009), and computer vision (Itti *et al* 1998, Achanta *et al* 2009). In human perception systems (Teuber 1955), higher cognitive processes in human brains can regulate signal intensity through top-down sensitivity control to influence the selection of new information and thus mediate endogenous attention. On the other hand, bottom-up saliency filters automatically enhance the response to infrequent stimuli as exogenous attention. Hence, visual saliency can be considered as a balanced response to both local stimuli (such as pixel and edges) and global contrast (such as regions or image structures).

Most early work makes more effort to build saliency models on low-level image features based on local contrast. These methods investigate the rarity of image regions with respect to (small) local neighborhoods. Koch and Ullman (1985) presented the highly influential biologically inspired early representation model, and Itti *et al* (1998) defined image saliency using central surrounded differences across multi-scale image features. Harel *et al* (2006) combine the feature maps of Itti et al. with other importance maps and highlight conspicuous parts. Ma and Zhang (2003) used an alternative local contrast analysis for saliency estimation. Liu *et al* (2011) found multi-scale contrast in a Difference-of-Gaussian image pyramid. Such methods using local contrast tend to produce higher saliency values near edges instead of uniformly highlighting salient objects, as shown in Figure 1, making it not applicable for practical image analysis.

Recent efforts have been made toward global contrast based saliency estimation, where saliency of an image region is evaluated at the global scale with respect to the entire image. Zhai and Shah (2006) defined pixel-level saliency based on a pixel's contrast to all other pixels. It can work well when color saliency is dominant, but suffers from problems when the background has similar colors. Achanta et al. (2009) proposed a frequency tuned method that directly defines pixel saliency using difference of Gaussian (DoG) features, and used mean-shift to average the pixel saliency stimuli over the whole regions. More recently, Goferman *et al* (2010) consider block-based global contrast while global image context is concerned. Instead of using fixed-size block, Cheng *et al* (2011) proposed to use the regions obtained from image segmentation methods and compute the saliency map from

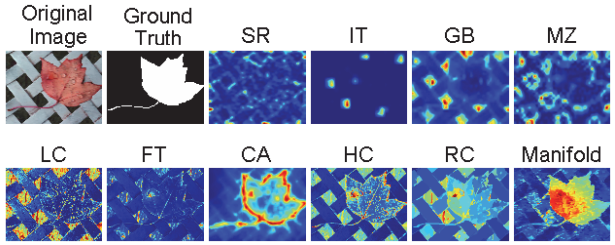


Fig.1 Saliency maps of a typical challenging case computed by various state-of-art methods, and with our proposed manifold learning approach. In the first row, from the left to the right: Original image, ground truth, SR (Hou & Zhang 2007), IT (Itti *et al* in 1998), GB (Harel *et al* 2006) and MZ (Ma & Zhang 2003). In the second row: LC (Zhai & Shah 2006), FT (Achanta *et al* 2009), CA (Goferman et al. in 2010), HC & RC (Cheng *et al* 2011), and our manifold approach.

the region-based global contrast. However, it has been revealed (Cheng *et al* 2010) that producing a correct salient map is sensitive to the size of regions, and a manual fine tuning of the segmentation is a prerequisite for some challenging images. In summary, it has been observed that most global approaches depend on either regions from image segmentation, or blocks with specified sizes.

While image structures at global scale are usually an important factor for producing salient stimuli, in this paper we present a novel approach for saliency modeling, namely manifold based saliency estimation. In our approach, we propose the balancing of local pixel-level stimuli with global contrast, and learn long range salient stimuli through unsupervised manifold learning, which provides a local-to-global abstraction for further saliency detection.

In experiments, we extensively evaluated our methods on publicly available benchmark data sets, and compared our methods with several state-of-the-art saliency methods as well as with manually produced ground truth annotations. Our experiments show significant improvements over previous methods in both precision and recall rates. Encouragingly, our approach also provided a convenient way for unsupervised saliency detection in video sequence.

Fig.1 demonstrates a typical challenging case (raised by Goferman et al. in 2010) for all state-of-the-art approaches. Our approach not only robustly detected the red leaf, but also provided a drastic contrast between the foreground and the background. With this advantage, the proposed approach can easily be extended to robotic vision tasks, such as salient object tracking.

Saliency-Aware Manifold Learning

Nearly three decades ago, Koch and Ullman (1985) proposed a theory to describe the underlying neural mechanisms of vision and bottom-up saliency. They posited that the human eye selects several features that pertain to a stimulus in the visual field and combines these features into a single topographical 'saliency map'. In the retina, photoreceptors, horizontal, and bipolar cells are the processing elements for edge extraction. After visual input is passed through a series of these cells, edge information is delivered to the visual cortex. In addition, a neural circuit in the retina creates opponent cells which receive inhibitory and excitory responses from various cones in the eye. These systems combine with further processing in the lateral geniculate nucleus (LGN), which plays a role in detecting shape and pattern information such as symmetry, as a preprocessor for the visual cortex to find a saliency region. Hence, saliency can be considered as a biological response to various stimuli.

In this paper, we consider the saliency originating from the global image structure. The human visual perception system usually ignores noise-like ephemeral or evanescent stimuli. Instead, more attention is paid to considerably longer-lasting stimuli that may have more energy. Bearing this in mind, obviously, a structure-aware saliency could be locally salient (e.g. a sharp contrast), and on the other hand, globally consistent in comparison with other evanescent stimuli. To attain this purpose, we introduce a manifold-based learning scheme to emulate this biological process.

Considering an image as a set of pixel signals $\{x_i\}$ distributed on the manifold over a 2D grid, as shown in Fig.2, the saliency computation will naturally tend to minimize the local stimuli and maximize the long-range stimuli. This means we need to pull neighboring pixels together while keeping their long-range contrast that stands for salient image structures.

Given that we have input signals $\{x_i\}$ their connectivity graph matrix S can be computed as,

$$S_{ij} = \begin{cases} 1, & \left\|\vec{r}_i - \vec{r}_j\right\|^2 < \varepsilon \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Here, $\|.\|$ is the Frobenius norm, S is a similarity matrix, r_i and r_j denote the spatial location of two pixels, and ε defines the radius of the local neighborhood that is sufficiently small, and greater than zero. When ε is set to be 1.5, each pixel will have eight connected nearest neighbor pixels, namely K_{NN}=4.

To optimize the local connected stimuli against the global image structure, we define an objective function to project $\{x_i\}$ into $\{y_i\}$, as follows:

$$\arg\min_{W} \sum_{ij} \left(y_i - y_j\right)^2 S_{ij} \quad (2)$$

where

$$x \Rightarrow y = W^T x \quad (3)$$

Here W is the projection matrix. The above target function is similar to the Laplaican Eigenmap one (Belkin & Niyogi 2003), a typical manifold learning approach. The only difference is that instead of using the distance between x_i and x_j, we use their spatial location to define the connectivity matrix S from Eq.(1).

The objective function with the choice of symmetric

weights S_{ij} incurs a heavy penalty if two pixels x_i and x_j within a small distance are mapped far apart with a large distance $||y_i-y_j||$ in their subspace projection. Therefore, minimizing the expression in (2) is an attempt to ensure that, if two pixels in the image, x_i and x_j, are "close" in term of their location on the spatially connected manifold, their projection y_i and y_j should then be close as well. Thus, this strategy using S_{ij} in Eq.(1) sets up a spatial confinement to suppress local stimuli while leaving long-distance stimuli as they are.

Following some simple algebraic steps, we can have,

$$\begin{aligned}&\frac{1}{2}\sum_{ij}\left(y_i - y_j\right)^2 S_{ij}\\ &= \frac{1}{2}\sum_{ij}\left(W^T x_i - W^T x_j\right)^2 S_{ij}\\ &= \sum_{ij} W^T x_i S_{ij} x_i^T W - \sum_{ij} W^T x_i S_{ij} x_j^T W\\ &= W^T X L X^T W\end{aligned}$$

where W is the data projection matrix, and

$$L = D - S, \text{ with } Dij = \begin{cases}\sum_j S_{ij}, i = j\\ 0, i \neq j\end{cases}$$

L is the Laplacian matrix. Then the problem becomes:

$$\arg\min_W W^T X L X^T W \qquad (4)$$

Here, the Laplacian graph model is embedded to convert the nonlinear problem into a linear problem.

Fig.2 illustrates the projection results for saliency abstraction. The left one shows the image, in which the pebbles look quite like a noisy- or texture-style scene. The third one shows the manifold base saliency-targeted image projection results. It can be seen that the salient yellow candy is kept as global stimuli, and other local differences among pebbles are drastically smoothed. Fig.2 also uses a 3D plot to show the comparison between the original image in one (red) color channel and the abstraction result in the primary projection channel. Obviously, manifold learning provides a context-aware saliency preservation. This idea is somewhat similar to the purpose of the method by Goferman et al. in 2010, which compared the local block against its global contrast, though our way is mathematically different.

It is noted that in the proposed manifold learning, the abstraction can be sensitive to the parameter ε or K_{NN}, which define how many nearest neighbors a pixel can have. Fig.3 shows the comparison using different K_{NN} from 4 to 32. We can see that the more neighbors are allowed, the higher the local-global contrast could be. However, it may blur the edges to smooth using more neighbors for each pixel. In this paper, we typically set K_{NN} as 8.

Histogram-based Saliency Detection

Through the above manifold learning, an RGB image is actually turned into a multi-channel image with up to more than ten projected dimensions. The multiplication of

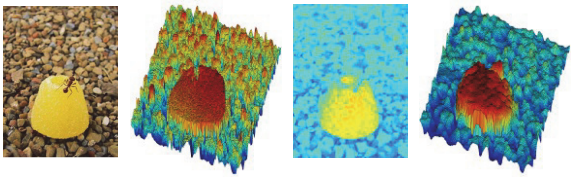


Fig.2 Manifold-based image abstraction. From left to right: Original image, terrain view of its red channel, results after abstraction and its terrain.

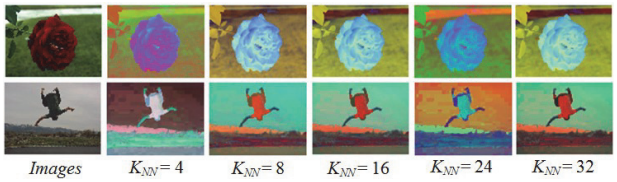
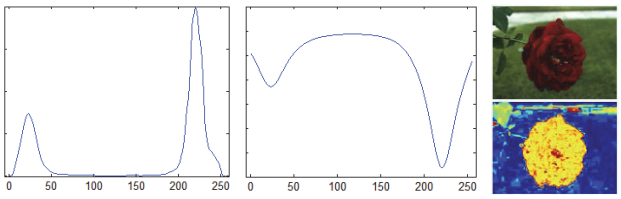
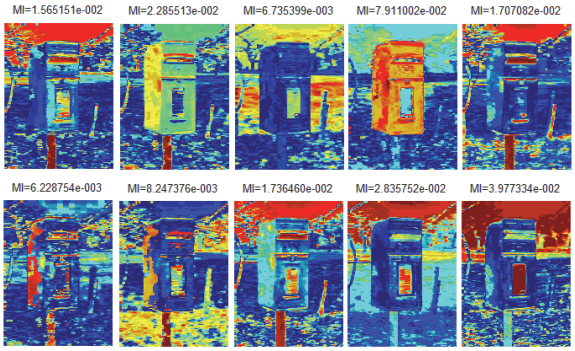


Fig.3 Manifold-based image abstraction with different K_{NN}. From left column to right: Original images, K_{NN} =4, K_{NN} =8, K_{NN} =16, K_{NN} =24, and K_{NN} =32.



(a) The computation of histogram saliency φ_k. From left to right: Single-channel 1D histogram; Computed bin's histogram saliency; Original image and one of its single-channel saliency maps.



(b) Single-channel saliency map estimation in each channel (10 channels listed).

(c) Priori map and the fusion result

Fig.4 Saliency estimation per channel and the fusion of all results using mutual information.

channels can give the saliency detection scheme more information and thus better accuracy. On the other hand, it also results in more data to process.

As it has been discussed, global contrast [Achanta *et al* 2009] has been proved to give better accuracy than most local methods. Cheng *et al* (2011) proposed to model the saliency using a 3D histogram from Lab color channels. With its simplicity and reliability, we choose to use histogram contrast estimation for our saliency detection. However, given the number of dimensions after manifold projection, it is unlikely to put all channels together, since the number of histogram bins will be overwhelming. Instead, in our proposed scheme, we first estimate saliency per channel, and then combine them together to attain robust saliency estimation.

It is noticed that region-based saliency estimation (Cheng *et al* 2011) has attained great success in its recall-precision performance. However, as has been shown, this technique needs a manual-tuning of image segmentation, making it not applicable for robotic applications, where automatic detection is the primary concern.

In mathematics, the pixel-level saliency can be formulated by the contrast between histogram bins,

$$\varphi_k = \sum_j \|k - j\| N_j \qquad (5)$$

where, N_j is the number of pixels in the j-th bin, and φ_k is the initially computed saliency for the k-th bin. Fig.4-a) shows the procedure for computing the histogram saliency φ_k. First, the histogram N_j is computed, as shown on the left. Then with the above equation, φ_k is obtained accordingly, as shown on the right.

With the above simple scheme, we can easily obtain the initial saliency map for each channel of a projected image. Fig.4-b) shows an example, where the saliency maps are computed for the first ten channels, respectively. However, it is obvious these initial estimations are far from accurate. We then introduce a mutual information scheme to refine these initial results by a weighted fusion procedure.

Fusion of Multiple Channels

To attain an accurate and coherent fusion of multiple channels, we introduce mutual information to weight the initially estimated saliency maps. Basically, we can assume a priori knowledge that human perception always pays attention to the objects around the center of a scene. We can model this using a centered anisotropic Gaussian distribution, as shown in the left image in Fig.4-c). With this expectation, we can then evaluate the initial estimation against this priori map.

Mutual information (MI) can be considered a statistic for assessing independence between a pair of variables, and has a well-specified asymptotic distribution. To calculate the MI score between a saliency map in Fig.4-a) and the priori map in Fig.4-b), we first convert these two 2D matrices into

Fig.5 Examples of single-channel saliency estimation and their fusion. The last two columns are the original images and the fusion results, and the foregoing columns are the estimated single channel saliency maps. (In the figure, 'jet' colormap was applied to make the saliency contrast visually easy to evaluate).

vectors, namely S_k and P_r, and we compute their MI score as:

$$H(S_k, P_r) = -\sum_b p(h_{S_k}, h_{P_r}) \log\{h_{S_k}, h_{P_r}\} \qquad (6)$$

Here, h_X stands for the histogram of the variable X. Details on MI can be found in the survey by Verdu *et al* (1999). In Fig.4-b), the computed MI score is tagged on every single-channel saliency map.

Once we have the estimated mutual score, it becomes simple to fuse the multi-channel saliency map, which can be expressed as a weighted totaling,

$$S_{Total}(x, y) = \sum_k H(S_k, P_r) S_k(x, y) \qquad (7)$$

where, (x, y) stands for the coordinates of a pixel in its saliency map.

The right image in Fig.4-c) shows the final added-up saliency map for the original image in the middle. In Fig.5, several more examples are demonstrated. We can see that in comparison to single-channel saliency maps, the fusion results computed by the proposed fusion scheme have been greatly enhanced, and a robust performance is attained.

Experimental Comparison

In our experiment, we evaluated our approach on the publicly available database provided by Achanta *et al* (2009). To the best of our knowledge, this database is the largest of its kind, and has ground truth in the form of

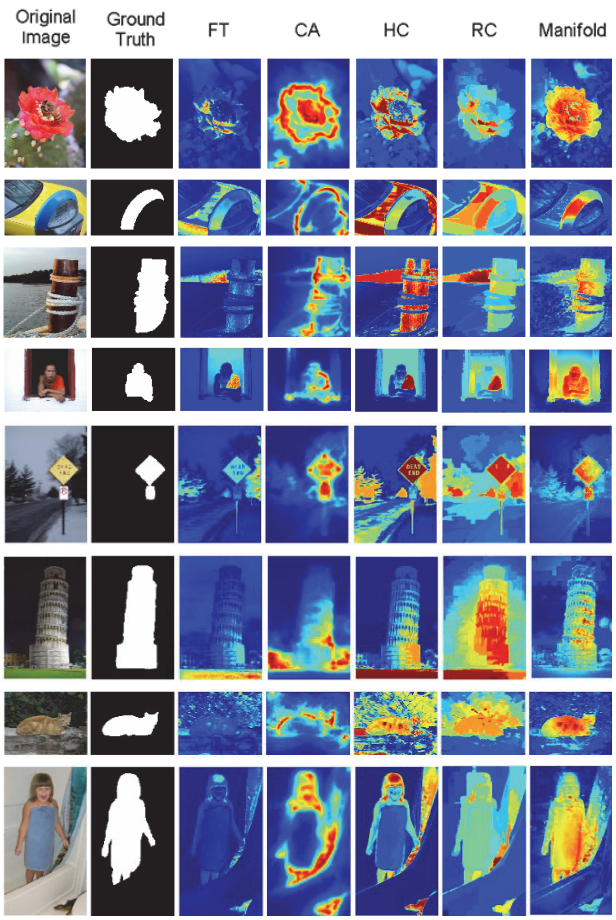


Fig.6 Visual comparison of saliency maps. From left to right columns: 1) original image, 2) ground truth, 3) FT (Achanta et al 2009), 4) CA (Goferman et al 2010), 5) HC(Cheng *et al* 2011), 6) RC(Cheng *et al* 2011), and our manifold-based method. It can be observed that our algorithm can robustly tackle challenging cases which most algorithms failed to tackle.

accurate human-marked labels for salient regions.

The average size of test images in the dataset is around 400×300 pixels each. Our system was implemented in MATLAB, on a PC with 2GB RAM and a 3GHz dual-core CPU. The test images were input by the standard interface provided by the MATLAB library.

We compared the proposed method with eight state-of-the-art saliency detection methods, namely: 1) SR (Hou & Zhang 2007); 2) IT (Itti *et al* in 1998); 3) GB (Harel *et al* 2006); 4) MZ (Ma & Zhang 2003); 5) LC (Zhai & Shah 2006); 6) FT (Achanta *et al* 2009); 7) CA (Goferman *et al* 2010); 8) HC(Cheng *et al* 2011). While our algorithm is implemented in MATLAB, the average computation time for each image is around 1.056 seconds.

For the other methods, we took the authors' published results provided from Cheng *et al* 2011 and Achanta *et al* 2009 for our evaluation and comparison. We did not compare our approach with the RC method by Cheng *et al* in 2011. This RC method was shown (Cheng *et al* 2011) to attain high precision, but sometime it depends on the user's interactive tuning of its image segmentation procedure, which is sensitive in producing correct regions for saliency estimation. It may fit well for image editing applications but not so useful for robotic applications. In the later case, unsupervised saliency estimation is required.

Fig.6 demonstrates the qualitative comparison results on several challenging cases that most existing methods failed to extract the correct saliency map from. We can clearly see that the proposed manifold-based method can robustly tackle these cases with high saliency contrast ratio between the salient regions and the background.

Fig.7 shows the statistical results of precision-recall curves. The curves were obtained in the same way as Achanta (2009) proposed, where naïve thresholding was applied from 0 to 255 to obtain successively a list of both precision and recall rates when subtracting the binarized saliency maps with their corresponded ground truth. The experimental results clearly validate that our proposed approach (the red curve in Fig.7) has consistently outperformed all state-of-the-art approaches. With the proposed manifold learning and fusion scheme, we can see that a robust unsupervised saliency estimation scheme has been successfully developed and validated.

Additional Experimental Results

In most robotic vision applications, the input signals are consecutive frames from the video camera. This means a robust saliency detector needs to tackle video-style visual

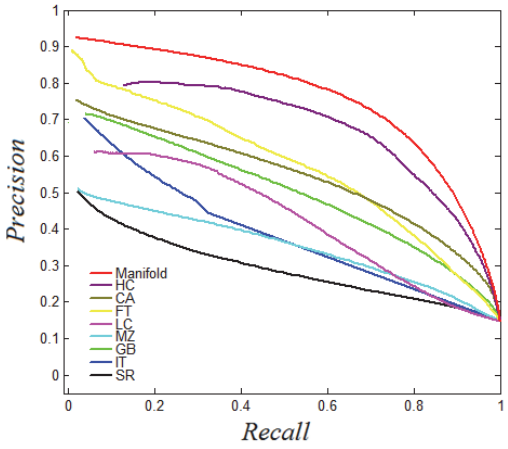


Fig.7 Precision-recall curves for naive thresholding of saliency maps using 1000 publicly available benchmark images. Our method is compared with 1) SR (Hou & Zhang 2007); 2) IT (Itti et al in 1998); 3) GB (Harel *et al* 2006); 4) MZ (Ma & Zhang 2003); 5) LC (Zhai & Shah 2006); 6) FT (Achanta et al 2009); 7) CA (Goferman et al 2010) ; 8) HC(Cheng *et al* 2011). It is shown that the proposed approach can consistently outperform the state-of-art approaches.

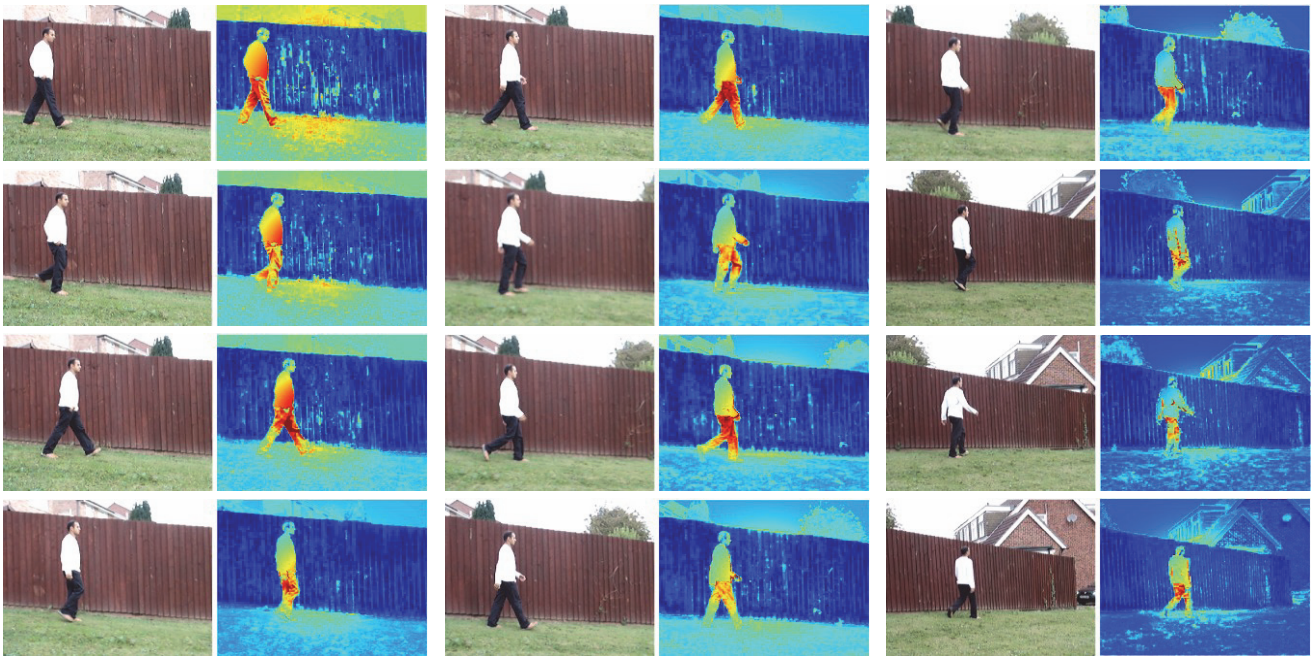


Fig.8 Video saliency detection in the challenging camera-moving case. In the test video, the camera tracks the person walking in front of the textured fence from the left to the right. From the test, the detected saliency is stably allocated on the walking person, though the background is moving in the reverse direction of camera motion.

data. A challenging problem for many robotic tracking applications is the demand to analyze the scene instantly with a moving background, while the camera equipped on a robot usually moves arbitrarily due to the random motion of the robot.

For most state-of-the-art computer vision approaches, it is still very tricky to detect a moving object consistently in front of an arbitrary moving background. Conventional approaches such as GMM-based motion segmentation and optical flow can easily fail in these challenging cases. They are also compute-intensive, making it hard to be implemented on embedding systems that can be accommodated on robots. Instead, saliency detection can instantly capture the salient object with no need for pixel-level motion field analysis or motion segmentation, making it a promising solution provided for robotic vision to overcome this sort of challenge.

While our initial evaluation on static images has demonstrated the advantages of our approach, we have further tested our approach on test videos. Fig.8 shows our estimated saliency maps of consecutive frames in a test video, where the camera tracks the person walking in front of the textured wooden fence from the left to the right.

As shown in Fig.6, our algorithm can easily detect the walking person across the whole video shot. In all 180 frames, the algorithm detected the person in all frames with no object-level false positive detection. Such unsupervised salient object detection can greatly facilitate the robotic vision to cope with various tasks in practical applications.

Conclusion

In conclusion, a robust manifold-based saliency estimation method has been proposed for robotic vision to capture the most salient objects in the observed scenes. In the proposed approach, an image is considered as a manifold of visual signals (stimuli) spreading over a connected grid, and projected into a multi-channel format through manifold learning. Histogram-based saliency estimation is then applied to extract the saliency map for each single channel, respectively, and a fusion scheme based on mutual information is introduced to combine all single-channel saliency maps together according to their mutual information score. In our experiment, the proposed method is evaluated using a well-known large image dataset. The experimental results validated that our algorithm attained the best prevision and recall rates among several state-of-art saliency detection methods. Furthermore, the proposed method has been demonstrated on a test video to show its potential use for robotic applications, such as tracking a moving target in an arbitrary scene while the camera is moving with a robot. The experimental results show that the proposed approach can successfully accomplish this sort of tasks, revealing its potential use for similar robotic applications.

References

R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency tuned salient region detection. In CVPR, pages 1597 1604, 2009.

M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Computation, vol. 15, no. 6, pp. 1373 1396, 2003.

Butko, N.J., Zhang, L. Cottrell, G.W. and Movellan, J.R. (2008) Visual saliency model for robot cameras. In International Conference on Robotics and Automation (ICRA 2008).

Ming Ming Cheng, Guo Xin Zhang, Niloy J. Mitra, Xiaolei Huang, Shi Min Hu. Global Contrast based Salient Region Detection. IEEE CVPR, p. 409 416, Colorado Springs, USA, June 21 23, 2011. http://cg.cs.tsinghua.edu.cn/people/~cmm/saliency/

R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1):193 222, 1995.

S. Goferman, L. Zelnik Manor, and A. Tal. Context aware saliency detection. In CVPR, 2010.

J. Harel, C. Koch, and P. Perona. Graph based visual saliency. In NIPS, pages 545 552, 2006.

X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In CVPR, pages 1 8, 2007.

L. Itti, C. Koch, and E. Niebur. A model of saliency based visual attention for rapid scene analysis. IEEE TPAMI, 20(11):1254 1259, 1998.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurbiology, 4:219 227, 1985.

T. Liu, Z. Yuan, J. Sun, J.Wang, N. Zheng, T. X., and S. H.Y. Learning to detect a salient object. IEEE TPAMI, 33(2):353 367, 2011.

Y. F. Ma and H. J. Zhang. Contrast based image attention analysis by using fuzzy growing. In ACM Multimedia, pages 374 381, 2003.

S. K. Mannan, C. Kennard, and M. Husain. The role of visual salience in directing eye movements in visual object agnosia. Current biology, 19(6):247 248, 2009.

J. Reynolds and R. Desimone. Interacting roles of attention and visual salience in v4. Neuron, 37(5):853 863, 2003.

C. Rother, V. Kolmogorov, and A. Blake. "Grabcut" Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph., 23(3):309 314, 2004.

U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom up attention useful for object recognition? In CVPR, pages 37 44, 2004.

Subramonia Sarma, Yoonsuck Choe. 2006. Salience in orientation filter response measured as suspicious coincidence in natural images. in Proceedings of AAAI 2006 Volume 1.

H. Teuber. Physiological psychology. Annual Review of Psychology, 6(1):267 296, 1955.

A. M. Triesman and G. Gelade. A feature integration theory of attention. Cognitive Psychology, 12(1):97 136, 1980.

J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience, pages 5:1 7, 2004.

Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In ACM Multimedia, pages 815 824, 2006.

S. Verdu, S. W. McLaughlin, editors. Information Theory: 50 Years of Discovery. IEEE Press, 1999.