Visual Attention Modeling in Compressed Domain: From Image Saliency Detection to Video Saliency Detection

FANG Yuming and ZHANG Xiaoqiang

(Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China)

Abstract: Saliency detection models, which are used to extract salient regions in visual scenes, are widely used in various multimedia processing applications. It has attracted much attention in the area of computer vision over the past decades. Since most images or videos over the Internet are stored in compressed domains such as images in JPEG format and videos in MPEG2 format, H.264 format, and MPEG4 Visual format, many saliency detection models have been proposed in the compressed domain recently. We provide a review of our works on saliency detection models in the compressed domain in this paper. Besides, we introduce some commonly used fusion strategies to combine spatial saliency map and temporal saliency map to compute the final video saliency map.

Keywords: saliency detection; computer vision; compressed domain; visual attention; fusion strategy

DOI: 10.12142/ZTECOM.201901006

http://kns.cnki.net/kcms/detail/34.1294.TN.20190314.1717.004.html, published online March 14, 2019

Manuscript received: 2018-07-19

1 Introduction

The human visual system (HVS) has limited capacity and cannot process everything that falls onto the retina [1]. Visual attention would selectively bring important information into focus while filtering other parts to reduce the complexity of scene analysis. Saliency detection model, which simulates visual attention mechanism, could identify regions of interest in images or videos. There are two visual attention mechanisms: bottom-up and top-down approaches. The bottom-up attention [2] is determined by characteristics of a visual scene (stimulus-driven), while the top-down attention [3] is determined by cognitive phenomena like expectations, current goals, and knowledge (goal-driven). Saliency estimation from one computational model is shown in **Fig. 1**, where the brighter the region is, the more salient it is.

Currently, saliency detection has been applied to the important preprocessing step in various multimedia processing applications, such as object tracking [4], [5], image retargeting [6], object detection [7], object recognition [8], person re-identification [9], image compression [10], quality assessment [11]–[13], abstraction [14], segmentation [15], and so on.

Saliency detection models can be divided into pixel-domain models and compressed-domain models. Early research on saliency detection mostly focuses on feature extraction in the pixel domain [16]–[34]. However, most images or videos over the Internet are basically stored in the compressed domain. For example, images over the Internet are stored in Joint Photographic Experts Group (JPEG) format, while videos are stored in H.264 and Moving Picture Experts Group (MPEG2) format.

▲**Figure 1. Saliency estimation results [16] on the public database Densely Annotated Video Segmentation (DAVIS) [17]. From the first column to the last column: original images, saliency maps, and ground truth maps.**

Compressed images or videos are widely used in various multimedia applications over the Internet, because they can reduce storage space and improve transmission efficiency. The current saliency detection models have to decompress the compressed images or videos into the pixel domain for feature extraction, which is time consuming. To avoid the process, some saliency detection models are proposed in the compressed domain [6], [35]–[43].

As a pioneer, Itti et al. [18] proposed a conceptually computational model for saliency detection based on multiscale image features including intensity, color, and orientation. Harel et al. [19] introduced a bottom-up visual saliency model (GBVS) with the new definition of dissimilarity to extract saliency information. After that, Yang et al. [20] computed visual saliency by ranking the similarity of the image elements (pixels or regions) with foreground cues or background cues via graph-based manifold ranking. However, many graph-based models [19], [20] heavily depend to the performance of the superpixel segmentation preprocessing. Therefore, Li et al. [21] introduced a saliency detection approach that considers the advantages of both region-based features and image details by the regularized random walk ranking. Tu et al. [24] proposed a method for measuring the image boundary efficiently based on the minimum spanning tree. The method established by Qin et al. [22] calculates saliency by Cellular Automata-a dynamic evolution model, which can obtain the relevance of similar regions. Tong et al. [23] exploited both weak and strong models for saliency detection by developing a bootstrap learning algorithm. Recently, deep learning based methods become more and more popular in saliency detection. Wang et al. [25] estimated saliency by integrating local features and global features extracted by two deep neural networks, respectively. Based on auto-encoder neural network, Zhang [26] presented a saliency detection model by learning uncertain convolutional features.

Recently, some video saliency detection models were also explored [28]–[31] in the pixel domain. Kim et al. [28] introduced the approach of random walk with restart to detect spatially and temporally salient regions. They calculated spatiotemporal saliency by finding the steady-state distribution of the walker. In [29], temporal background priors are combined with spatial background priors to generate spatiotemporal background priors. Then, saliency estimation is conducted by a dual-graph based structure using spatiotemporal background priors. A spectral foreground extraction algorithm, Quantum Cuts (QCUT), is applied to estimate the saliency probability of regions [30]. Chen et al. [31] designed a video saliency detection model based on the spatiotemporal saliency fusion and low-rank coherency guided saliency diffusion. In [44], Li et al. proposed an unsupervised approach for video saliency object detection by using stacked auto-encoder neural network. In that approach, three saliency cues including pixel, superpixel, and object levels are extracted based on the algorithms of [45], [46], and [47]. Then the three saliency cues are fed into stacked auto-encoders to infer a saliency score for each pixel.

The models mentioned above are all saliency detection models in the pixel domain. Recently, there have been some works explored on saliency detection in the compressed domain. Muthuswamy et al. [35] used discrete cosine transform (DCT) coefficients [6] and motion vectors [48] as features for MPEG2 video saliency detection. Khatoonabadi et al. [36] proposed a new feature, operational block description length (OBDL), as a measure of saliency. The OBDL represents the minimum number of bits required to encode a given video block under a certain distortion criterion [36]. In [37], Khatoonabadi et al. introduced two video features called Motion Vector Entropy and Smoothed Residual Norm extracted from the compressed video bitstream. Using the statistics of these two features in videos, they proposed a visual saliency detection model for compressed video. Two compressed domain features called residual DCT coefficient norms and operational block description are extracted from video bitstream [38], [39]. Then Li et al. [38], [39] used a fusion algorithm whose fusion coefficients vary with quantization parameters to fuse the two feature maps for saliency estimation. Xu et al. [40] first extracted High Efficiency Video Coding (HEVC) features in the HEVC domain and then those features are integrated by the learned support vector machine for video saliency detection. Jian et al. [41] introduced a saliency detection model by extracting three features including Quaternionic Distance Based Weber Descriptor (QDWD), pattern distinctness, and local contrast. By exploiting MPEG4 AVC compression principles, Ammar et al. [42] calculated the intensity, color, orientation, and motion feature maps by extracting the energy of luma coefficients, energy of chroma coefficients, gradient of the prediction modes, and amplitude of motion vectors. Finally, spatiotemporal saliency map is obtained by an average fusion algorithm. We proposed a saliency detection model in the compressed domain for images [6] and video [43].

The remaining of this paper is organized as follows. Section 2 describes our works in the compressed domain. Section 3 compares our fusion strategies of spatial and temporal saliency with those from other existing fusion strategies. The final Section 4 concludes the paper.

2 Compressed-Domain Visual Saliency Models

There are two works in computational modeling of visual attention in the compressed domain [6], [43]. In [6], the saliency detection model in compressed domain is built for 2D images, while the model is established for visual saliency modeling of video sequences in [43].

2.1 Saliency Detection Model for Compressed-Domain Image

We proposed a saliency detection model for images in com-

pressed domain [6]. The general framework of the proposed model is shown in **Fig. 2**. Three kinds of features (including intensity, color, and texture features) are firstly extracted from DCT coefficients. Then four visual saliency maps (one intensity saliency map, two color saliency maps, and one texture saliency map) are estimated by calculating feature contrast based on small DCT blocks weighted by a Gaussian model. Finally, by using coherent normalization-based fusion method to fuse these saliency maps, the final saliency map is obtained. Some saliency detection results of the proposed model [6] are shown in **Fig. 3**.

This work [6] in saliency detection mainly includes two contributions: the first one is how to extract features (intensity, color, and texture features) directly from the JPEG bitstream; the second is to design a new computational model of visual attention based on DCT blocks in the compressed domain. The details of the model are described as follows.

(1) Feature Extraction from the JPEG Bitstream

The color space of the input JPEG images is converted from RGB color space to YCbCr color space. YCbCr color space could be used to extract the three kinds of features mentioned above. Specifically, L channel contains the intensity and texture information while Cb and Cr channels contain color information. Each channel is divided into 8×8 blocks, and the DCT is carried out for each small block. DCT coefficients in each block include the DC coefficient and AC coefficients. Please note that DC coefficient is a measure of the average energy of this block, while the remaining 63 AC coefficients represent high frequency information of this block. Therefore, we could use DC coefficient in L channel to extract intensity feature L. DC coefficients in Cb and Cr channels are used to extract two color features (C_1 and C_2). AC coefficients in L channel are used to extract texture feature T.

(2) Saliency Estimation in the Compressed Domain

Four saliency maps (one intensity saliency map, two color saliency map, and one texture saliency map) are computed by feature contrast based on DCT blocks. The saliency value of each DCT block in each feature map is determined by two factors, including the block differences and weights between this block and all other blocks of the input image. Intensity and color feature differences of each DCT block are calculated by L1-norm distance, while texture difference of each DCT block is estimated by Hausdorff distance. The saliency value for each block is proportional to the block difference. The human eye is more sensitive about the differences between the current block and nearer blocks compared with the relatively distant area. A Gaussian model is thus used to weight the block differences for saliency detection. The saliency map for the nth feature can be calculated as follows:

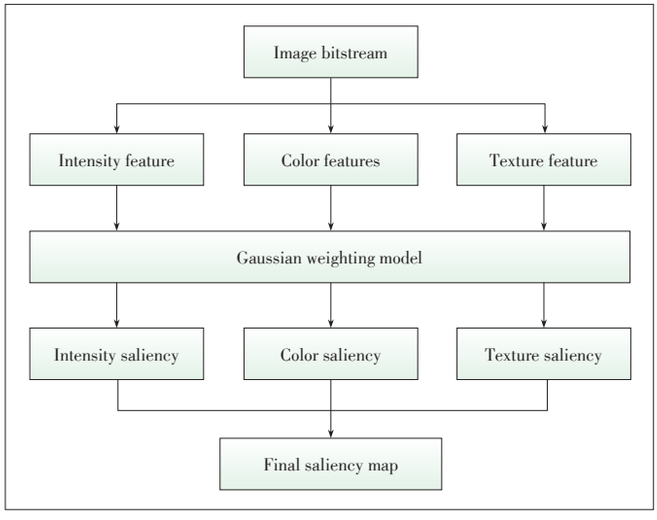
$$S_i^n = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} D_{ij}^n, \quad (1)$$

where d_{ij} represents the Euclidean distance between DCT blocks i and j; $n \in \{L, C_1, C_2, T\}$ and D_{ij} is DCT block difference; σ is a parameter of the Gaussian model. In the study [6], σ is set to 5.

According to Eq. (1), different saliency maps are calculated based on different features. These saliency maps include one intensity saliency map, two color saliency maps, and one texture saliency map. The final saliency map S for a given JPEG image can be calculated by fusing these four saliency maps. In [6], the coherent normalization-based fusion method is used to combine these four saliency maps as follows:

$$S = \beta \sum N(k) + \gamma \prod N(k), \quad (2)$$

where β and γ are parameters determining the weights for each components. In [6], the two parameters are both set to 1/5. N is the normalization operation; $k \in \{S^n\}$.



▲**Figure 2. The framework of the model proposed in [6].**

▲**Figure 3. Saliency estimation results [6] on the public database in [49]. The first row is original images, while the second row represents saliency maps.**

2.2 Saliency Detection Model for Compressed-Domain Video

Similar with images, videos over the Internet are almost stored in the compressed domain such as H.264 and MPEG2.

In the study [43], a video saliency detection model is proposed based on feature contrast in the compressed domain.

The framework of the model [43] is shown in **Fig. 4**. That video saliency detection model could be roughly divided into two stages, including spatial saliency and temporal saliency estimation and fusion of these two kinds of visual saliency. Specifically, spatial saliency is calculated based on the three features (including luminance, color, and texture) extracted from the video bitstream. Temporal saliency is calculated based on the motion features extracted from the motion vectors in video bitstream. In the second stage, based on a new fusion method of parameterized normalization, sum and product (PNSP), the final saliency map for the video frame is calculated. Some saliency detection results of the proposed model [43] are shown in **Fig. 5**. The details of the model will be described as follows.

(1) Feature Extraction in the Video Bitstream

In MPEG4 advanced simple profile (ASP) video, there are two kinds of predicted frames: P frames use motion compensated prediction from a past reference frame, while B frames are bidirectionally predictive-coded by using motion compensated prediction from a past and/or a future reference frame. As there are two kinds of frames, there are two kinds of ways to calculate the motion feature. The motion vector MV is used to represent the motion feature for P frames. The motion feature for B frames can be calculated by both motion prediction from the past and future reference frames. Assume the motion compensated prediction from the past reference and the future reference frames are MV_1 and MV_2. The motion feature V of B frames is computed as follows:

$$V = MV_1 - MV_2. \tag{3}$$

Please note that no matter what kinds of frames are used, motion features are computed based on DCT blocks. And the motion feature of P/B frames can be obtained as V. For spatial features include intensity, color, and texture, they can be extracted as [6].

(2) Saliency Estimation in the Compressed Domain

Based on the motion feature V, the feature map of each video frame is computed as follows:

$$S_i^v = \sum_{j \neq i} w_{ij} D_{ij}^v, \tag{4}$$

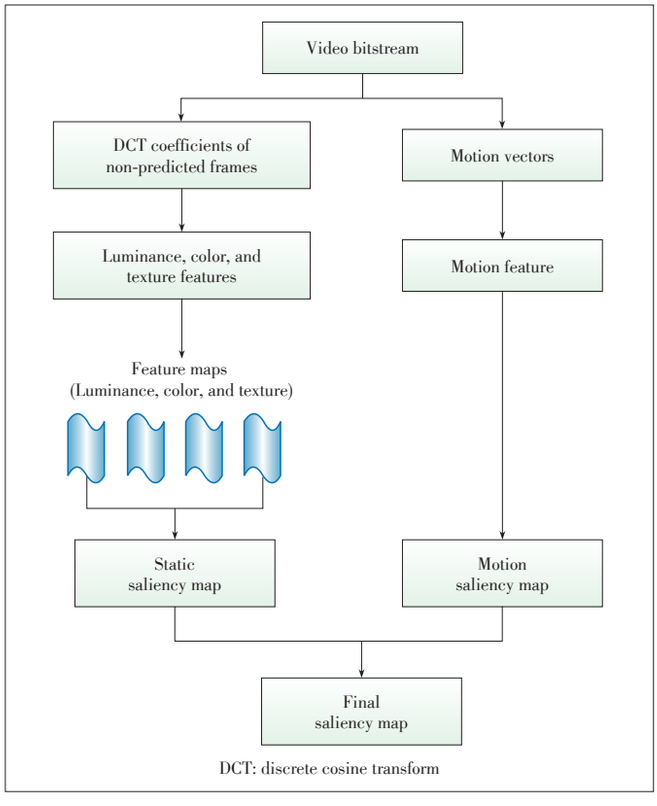
$$w_{ij} = \frac{1}{\sigma_v \sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma_v^2}}, \tag{5}$$

where S_i^v represents temporal saliency value of the ith DCT block in the motion feature map; D_{ij}^v is the motion feature difference between DCT blocks i and j; σ_v is a parameter of the Gaussian model. The spatial saliency map S^s is calculated by linearly combining the four spatial feature maps from intensity, color, and texture features (L, C_1, C_2, T).

In [43], based on the characteristics of the spatial saliency map and temporal saliency map, a new fusion method called PNSP is proposed. The final saliency map for video frame is calculated as follows:

$$S^f = \beta_1 S^s + \beta_2 S^v + \beta_3 S^s S^v, \tag{6}$$

where S^f denotes the final saliency map for the video frame; β_1, β_2, and β_3 are the parameters determining the weights of each component; S^s is the spatial saliency map and S^v is the temporal saliency map.



▲**Figure 4. The framework of the model proposed by [43].**

3 Spatiotemporal Weighting Strategy

Here, we introduce our works on spatiotemporal weighting strategy [16], [43]. In [16], based on Gestalt theory, the spatial

▲**Figure 5. Saliency estimation results [43] on the public database Densely Annotated Video Segmentation (DAVIS) [17]. The first row is original images, while the second row represents saliency maps.**

PNSP is designed to combine the spatial and temporal saliency to obtain the final saliency map. We also introduce several common fusion algorithms in [50] for integrating spatial saliency and temporal saliency map.

3.1 Common Fusion Approaches

(1) Normalization and Sum (NS)

The most simple and direct method for fusing spatial saliency and temporal saliency is to normalize these two salient maps to the same dynamic range (between 0 and 1) and then sum the two maps to get the final saliency map as follows:

$$S = \sum_n N(S_n), \tag{7}$$

where S is the final saliency map; $n \in \{1,2\}$ and S_n is the spatial saliency map or temporal saliency map.

(2) Normalization and Maximum (NM)

The fusion algorithm tries to normalize spatial saliency map and temporal saliency map to the same dynamic range and then uses the maximum value as the final saliency value at each location.

$$S = \max_n N(S_n), \tag{8}$$

where max is the maximum operator.

(3) Normalization and Product (NP)

Compared to NS and NM methods, the summation and maximum are replaced by the product operator in NP.

$$S = \prod_n N(S_n). \tag{9}$$

3.2 The Fusion Approach Based on Uncertainty Weighting

Compared with image saliency detection, video saliency detection is a more challenging problem due to its complex background and utilization of motion information. So far, only a few video saliency detection models have been proposed [51]–[53]. In [16], a novel method is proposed to estimate video saliency by using Gestalt theory and uncertainty weighting.

The algorithm [16] could be divided into two main stages including the spatial and temporal saliency estimation stage and the fusion stage of the two saliency maps. Spatial saliency is calculated by extracted spatial features including luminance, color, and texture features from a given video frame using DCT coefficients [6]. Temporal saliency can be measured based on a psychological study of human visual speed perception [54]. Based on uncertainty weighting strategy, we can fuse the spatial saliency map and temporal saliency map for obtaining the final saliency map. Spatial uncertainty estimation is conceptually rooted in the Gestalt theory including the law of proximity and the law of continuity [55], [56]. Temporal uncertainty estimation is calculated based on the psychovisual studies in [54]. Some saliency detection results of the proposed model [16] are shown in Fig. 1.

The proximity law of Gestalt theory states that elements which are close to each other tend to be perceived as a group, while the continuity law of Gestalt theory indicates that elements which are connected to each other tend to be perceived as a group. These two laws can be applied to saliency detection as follows: first, the spatial location which is closer to the saliency center in an image is more likely to be a salient location; second, a spatial location which is more connected to other saliency regions is more likely to be a salient location. Then the spatial uncertainty for each pixel in the spatial saliency map is calculated as follows:

$$U^s = U^d + U^c, \tag{10}$$

where U^s is the spatial uncertainty map; U^d is the probability of a pixel being salient given its distance from saliency center; U^c represents the probability of a pixel being salient given its connectedness to other salient pixels.

When the background motion is very large in the video, or the local contrast increases, the system cannot detect motion of the object accurately. This temporal uncertainty evaluation U^t is conducted based on the psychovisual studies in [54]. Therefore, the spatial and temporal saliency map can be integrated into spatiotemporal saliency map of the given video sequence by using these two uncertainty weighting map as follows:

$$S_{sp} = \frac{U^t S_s + U^s S_t}{U^t + U^s}, \tag{11}$$

where S_{sp} is the spatiotemporal saliency map; U^t is the temporal uncertainty map; U^s represents the spatial uncertainty map; S_s is the spatial saliency map calculated by DCT coefficients; S_t is the temporal saliency map calculated by optical flow algorithm.

3.3 The Fusion approach Based on PNSP Weighting

Another fusion method [43] has already been described in Section 2.2. And we can find that this fusion method is the combination of the NS method and NP method mentioned above. If the salient regions have low spatial saliency value with high temporal saliency value, the NS method can highlight the saliency by summation operation. If the non-salient regions have low spatial saliency value with high temporal saliency value, the NP method can suppress the saliency by product operation. Therefore, this algorithm can combine the advantages of these two algorithms.

4 Conclusions

In this paper, we review some works in the pixel-domain and compressed-domain. We first attempt to provide a comprehensive description of two compressed-domain saliency detection models. These two models are designed to handle with different tasks including compressed-domain 2D images and compressed-domain 2D video saliency detection. The difficulty of

the tasks continues to increase since increasing complexity of scene. Then we exhaustively review our two fusion strategies of spatial saliency map and temporal saliency map. The PNSP fusion algorithm considers the advantages of NS algorithm and NP algorithm. Another fusion algorithm is designed based on proximity law and proximity law of Gestalt theory.

Specifically, According to image saliency detection or video saliency detection, feature contrast for each block is calculated by the differences between the features of this block and other blocks in the whole image. In the future, we hope that we could propose more effective methods to handle the computational modeling for visual attention.

References

[1] JAMES W. The Principles of Psychology [M]. England: Read Books Ltd, 2013
[2] NOTHDURFT H C. Salience of Feature Contrast [M]//NOTHDURFT H C. eds. Neurobiology of Attention. Amsterdam, Netherlands: Elsevier, 2005: 233–239. DOI: 10.1016/b978-012375731-9/50042-2
[3] ITTI L, KOCH C. Computational Modelling of Visual Attention [J]. Nature Reviews Neuroscience, 2001, 2(3): 194–203. DOI: 10.1038/35058500
[4] MAHADEVAN V, VASCONCELOS N. Saliency-Based Discriminant Tracking [C]//IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 1007–1013. DOI: 10.1109/CVPR.2009.5206573
[5] MA C, MIAO Z J, ZHANG X P, et al. A Saliency Prior Context Model for Real-Time Object Tracking [J]. IEEE Transactions on Multimedia, 2017, 19(11): 2415–2424. DOI: 10.1109/tmm.2017.2694219
[6] FANG Y M, CHEN Z Z, LIN W S, et al. Saliency Detection in the Compressed Domain for Adaptive Image Retargeting [J]. IEEE Transactions on Image Processing, 2012, 21(9): 3888–3901. DOI: 10.1109/tip.2012.2199126
[7] GUO M W, ZHAO Y Z, ZHANG C B, et al. Fast Object Detection Based on Selective Visual Attention [J]. Neurocomputing, 2014, 144: 184–197. DOI: 10.1016/j.neucom.2014.04.054
[8] REN Z X, GAO S H, CHIA L T, et al. Region-Based Saliency Detection and Its Application in Object Recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(5): 769–779. DOI: 10.1109/tcsvt.2013.2280096
[9] ZHAO R, WANLI O Y, WANG X G. Unsupervised Salience Learning for Person Re-Identification [C]//IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3586–3593. DOI: 10.1109/CVPR.2013.460
[10] HOU Y H, WANG P C, XIANG W, et al. A Novel Rate Control Algorithm for Video Coding Based on fuzzy-PID Controller [J]. Signal, Image and Video Processing, 2015, 9(4): 875–884. DOI: 10.1007/s11760-013-0518-2
[11] CULIBRK D, MIRKOVIC M, ZLOKOLICA V, et al. Salient Motion Features for Video Quality Assessment [J]. IEEE Transactions on Image Processing, 2011, 20(4): 948–958. DOI: 10.1109/tip.2010.2080279
[12] LIU H T, HEYNDERICKX I. Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(7): 971–982. DOI: 10.1109/tcsvt.2011.2133770
[13] FENG X, LIU T, YANG D, et al. Saliency Inspired Full-Reference Quality Metrics for Packet-Loss-Impaired Video [J]. IEEE Transactions on Broadcasting, 2011, 57(1): 81–88. DOI: 10.1109/tbc.2010.2092150
[14] JI Q G, FANG Z D, XIE Z H, et al. Video Abstraction Based on the Visual Attention Model and Online Clustering [J]. Signal Processing: Image Communication, 2013, 28(3): 241–253. DOI: 10.1016/j.image.2012.11.008
[15] MISHRA A K, ALOIMONOS Y, CHEONG L F, et al. Active Visual Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 639–653. DOI: 10.1109/tpami.2011.171
[16] FANG Y M, WANG Z, LIN W S, et al. Video Saliency Incorporating Spatiotemporal Cues and Uncertainty Weighting [J]. IEEE Transactions on Image Processing, 2014, 23(9): 3910–3921. DOI: 10.1109/tip.2014.2336549
[17] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 724–732. DOI: 10.1109/CVPR.2016.85
[18] ITTI L, KOCH C, NIEBUR E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254–1259. DOI: 10.1109/34.730558
[19] HAREL J, KOCH C, PERONA P. Graph-Based Visual Saliency [M]//Advances in Neural Information Processing Systems 19. Cambridge, USA: The MIT Press, 2007: 545–552. DOI: 10.7551/mitpress/7503.003.0073
[20] YANG C, ZHANG L H, LU H C, et al. Saliency Detection Via Graph-Based Manifold Ranking [C]//IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3166–3173. DOI: 10.1109/CVPR.2013.407
[21] LI C Y, YUAN Y C, CAI W D, et al. Robust Saliency Detection Via Regularized Random Walks Ranking [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 2710–2717. DOI: 10.1109/CVPR.2015.7298887
[22] QIN Y, LU H C, XU Y Q, et al. Saliency Detection Via Cellular Automata [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 110–119. DOI: 10.1109/CVPR.2015.7298606
[23] TONG N, LU H C, RUAN X, et al. Salient Object Detection Via Bootstrap Learning [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 1884–1892. DOI:10.1109/CVPR.2015.7298798
[24] TU W C, HE S F, YANG Q X, et al. Real-Time Salient Object Detection with a Minimum Spanning Tree [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 2334–2342. DOI: 10.1109/CVPR.2016.256
[25] WANG L J, LU H C, RUAN X, et al. Deep Networks for Saliency Detection Via Local Estimation and Global Search [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 3183–3192. DOI: 10.1109/CVPR.2015.7298938
[26] ZHANG P P, WANG D, LU H C, et al. Learning Uncertain Convolutional Features for Accurate Saliency Detection [C]//IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 212–221. DOI:10.1109/ICCV.2017.32
[27] YUAN Y C, LI C Y, KIM J, et al. Reversion Correction and Regularized Random Walk Ranking for Saliency Detection [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1311–1322. DOI: 10.1109/tip.2017.2762422
[28] KIM H, KIM Y, SIM J Y, et al. Spatiotemporal Saliency Detection for Video Sequences Based on Random Walk with Restart [J]. IEEE Transactions on Image Processing, 2015, 24(8): 2552–2564. DOI: 10.1109/tip.2015.2425544
[29] XI T, ZHAO W, WANG H, et al. Salient Object Detection with Spatiotemporal Background Priors for Video [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3425–3436. DOI: 10.1109/tip.2016.2631900
[30] AYTEKIN C, POSSEGGER H, MAUTHNER T, et al. Spatiotemporal Saliency Estimation by Spectral Foreground Detection [J]. IEEE Transactions on Multimedia, 2018, 20(1): 82–95. DOI:10.1109/tmm.2017.2713982
[31] CHEN C, LI S, WANG Y G, et al. Video Saliency Detection Via Spatial-Temporal Fusion and Low-Rank Coherency Diffusion [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3156–3170. DOI: 10.1109/tip.2017.2670143
[32] FANG Y M, LIN W S, LEE B S, et al. Bottom-Up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum [J]. IEEE Transactions on Multimedia, 2012, 14(1): 187–198. DOI:10.1109/tmm.2011.2169775
[33] FANG Y M, WANG J L, NARWARIA M, et al. Saliency Detection for Stereoscopic Images [J]. IEEE Transactions on Image Processing, 2014, 23(6): 2625–2636. DOI: 10.1109/tip.2014.2305100
[34] FANG Y M, ZHANG C, LI J, et al. Visual Attention Modeling for Stereoscopic Video: A Benchmark and Computational Model [J]. IEEE Transactions on Image Processing, 2017, 26(10): 4684–4696. DOI: 10.1109/tip.2017.2721112
[35] MUTHUSWAMY K, RAJAN D. Salient Motion Detection in Compressed Domain [J]. IEEE Signal Processing Letters, 2013, 20(10): 996–999. DOI: 10.1109/lsp.2013.2277884
[36] KHATOONABADI S H, VASCONCELOS N, BAJICI V, et al. How Many Bits does it Take for a Stimulus to be Salient? [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 5501–5510. DOI: 10.1109/CVPR.2015.7299189
[37] KHATOONABADI S H, BAJICI V, SHAN Y F. Compressed-Domain Correlates of Human Fixations in Dynamic Scenes [J]. Multimedia Tools and Applications, 2015, 74(22): 10057–10075. DOI: 10.1007/s11042-015-2802-3
[38] LI Y J, LI Y S. A Fast and Efficient Saliency Detection Model in Video Compressed-Domain for Human Fixations Prediction [J]. Multimedia Tools and Applications, 2017, 76(24): 26273–26295. DOI: 10.1007/s11042-016-4118-3
[39] LI Y J, LI Y S, LIU W J, et al. Human Fixation Detection Model in Video Compressed Domain Based on Markov Random Field [J]. Journal of Electronic Imaging, 2017, 26(1): 013008. DOI: 10.1117/1.jei.26.1.013008

[40] XU M, JIANG L, SUN X Y, et al. Learning to Detect Video Saliency with HEVC Features [J]. IEEE Transactions on Image Processing, 2017, 26(1): 369-385. DOI: 10.1109/tip.2016.2628583
[41] JIAN M W, QI Q, DONG J Y, et al. Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection [J]. Journal of Visual Communication and Image Representation, 2018, 53: 31-41. DOI: 10.1016/j.jvcir.2018.03.008
[42] AMMAR M, MITREA M, HASNAOUI M, et al. MPEG-4 AVC Stream-Based Saliency Detection. Application to Robust Watermarking [J]. Signal Processing: Image Communication, 2018, 60: 116-130. DOI: 10.1016/j.image.2017.09.007
[43] FANG Y M, LIN W S, CHEN Z Z, et al. A Video Saliency Detection Model in Compressed Domain [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(1): 27-38. DOI: 10.1109/tcsvt.2013.2273613
[44] LI J, XIA C Q, CHEN X W. A Benchmark Dataset and Saliency-Guided Stacked Autoencoders for Video-Based Salient Object Detection [J]. IEEE Transactions on Image Processing, 2018, 27(1): 349-364. DOI: 10.1109/tip.2017.2762594
[45] ZHANG J M, SCLAROFF S, LIN Z, et al. Minimum Barrier Salient Object Detection at 80 FPS [C]//IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 1404-1412. DOI: 10.1109/ICCV.2015.165
[46] PENG H W, LI B, LING H B, et al. Salient Object Detection via Structured Matrix Decomposition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 818-832. DOI: 10.1109/tpami.2016.2562626
[47] LI J, LEVINE M D, AN X J, et al. Visual Saliency Based on Scale-Space Analysis in the Frequency Domain [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(4): 996-1010. DOI: 10.1109/tpami.2012.147
[48] AGARWAL G, ANBU A, SINHA A. A Fast Algorithm to Find the Region-of-Interest in the Compressed MPEG Domain [C]//International Conference on Multimedia and Expo. (ICME'03), Baltimore, USA, 2003: 133-136. DOI: 10.1109/ICME.2003.1221571
[49] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-Tuned Salient Region Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 1597-1604. DOI: 10.1109/CVPR.2009.5206596
[50] CHAMARET C, CHEVET J C, LE MEUR O. Spatio-Temporal Combination of Saliency Maps and Eye-Tracking Assessment of Different Strategies [C]//IEEE International Conference on Image Processing, Hong Kong, China, 2010: 1077-1080. DOI: 10.1109/ICIP.2010.5651381
[51] GUO C L, ZHANG L M. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression [J]. IEEE Transactions on Image Processing, 2010, 19(1): 185-198. DOI: 10.1109/tip.2009.2030969
[52] LE MEUR O, LE CALLET P, BARBA D. Predicting Visual Fixations on Video Based on Low-Level Visual Features [J]. Vision Research, 2007, 47(19): 2483-2498. DOI: 10.1016/j.visres.2007.06.015
[53] MAHADEVAN V, VASCONCELOS N. Spatiotemporal Saliency in Dynamic Scenes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(1): 171-177. DOI: 10.1109/tpami.2009.112
[54] STOCKER A A, SIMONCELLI E P. Noise Characteristics and Prior Expectations in Human Visual Speed Perception [J]. Nature Neuroscience, 2006, 9(4): 578-585. DOI: 10.1038/nn1669
[55] BANERJEE J C. Gestalt Theory of Perception (M)//Encyclopaedic Dictionary of Psychological Terms. New Delhi, India: MD Publications Pvt. Ltd., 1994: 107-109
[56] STEVENSON H. Emergence: The Gestalt Approach to Change [EB/OL]. (2012). http://www.clevelandconsultinggroup.com/articles/emergence-gestalt-approach-to-change.php

Biographies

FANG Yuming (fa0001ng@e.ntu.edu.sg) received his Ph.D. degree from Nanyang Technological University, Singapore, M.S. degree from Beijing University of Technology, China, and B.E. degree from Sichuan University, China. Currently, he is a professor in the School of Information Management, Jiangxi University of Finance and Economics, China. He serves as an associate editor of *IEEE Access* and is on the editorial board of *Signal Processing: Image Communication*. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3D image/video processing, etc.

ZHANG Xiaoqiang is currently pursuing the master's degree with the School of Information Technology, Jiangxi University of Finance and Economics, China. His research interests include saliency detection, computer vision, machine learning, and deep learning.