Visual Conspicuity Index: Spatial Dissimilarity, Distance, and Central Bias

Lijuan Duan, *Member, IEEE*, Chunpeng Wu, Jun Miao, *Member, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—We propose an image conspicuity index that combines three factors: spatial dissimilarity, spatial distance and central bias. The dissimilarity between image patches is evaluated in a reduced dimensional principal component space and is inversely weighted by the spatial separations between patches. An additional weighting mechanism is deployed that reflects the bias of human fixations towards the image center. The method is tested on three public image datasets and a video clip to evaluate its performance. The experimental results indicate highly competitive performance despite the simple definition of the proposed index. The conspicuity maps generated are more consistent with human fixations than prior state-of-the-art models when tested on color image datasets. This is demonstrated using both receiver operator characteristics (ROC) analysis and the Kullback–Leibler distance metric. The method should prove useful for such diverse image processing tasks as quality assessment, segmentation, search, or compression. The high performance and relative simplicity of the conspicuity index relative to other much more complex models suggests that it may find wide usage.**

***Index Terms*—central bias, conspicuity, dissimilarity, spatial distance, visual saliency.**

I. Introduction

THE selective attention mechanism makes it possible to rapidly understand visual scenes by dynamically changing the point of fixation. Via the ballistic saccades of the eyes, the limited resources of the visual apparatus are directed to points of attentional awareness. In principle, computational models of attention and of visual fixations have great potential to enhance algorithms that accomplish visual tasks such as image retargeting, object detection, object recognition and nonphotorealistic rendering. Central to the development of such goals are models of visual attention, which have gained heightened interest in recent years.

Daly [1] presented a visible differences predictor that relates contrast sensitivity to the attention mechanism. Several later models have been proposed of visual attentional strategies [2]–[4]. While visual tasks dominate where the gaze is cast [5], various lower-level visual features also appear to play a role. Models for "fixation prediction" have utilized quite a large variety of features and indeed, overall philosophies. The majority of such methods could be categorized among three general types: i) perceptual feature matching, whereby the model involves finding occurrences of features that match cortical neuronal response models, the idea being similar to the matched filter principle [3]; typically the idea is to identify center-surround patches in the image [6]; ii) statistical fixation analysis, whereby the statistics of images at the point of gaze are measured from real eyetracking data, or human annotations, and used to model likely fixation locations [7], [8]; and iii) local information maximization, where the idea is to collect as much unique information as possible at each fixation [4], [9].

In some regard all these three general approaches are directed towards accomplishing different goals, but are then measured against similar criteria (coincidence with human fixation selection). In all of these the method seeks to measure some aspect(s) of visual "saliency" that tends to draw fixations.

The notion of saliency as conspicuity takes the idea that salient image locations that markedly visually differ from their surroundings by some measureable property, hence will tend to visually "stand out," thereby drawing attention [10]. The models i)–iii) above do not necessarily imply that the salient locations sought after are necessarily highly conspicuous, as information gathering and feature matching are different things.

We utilize the term "conspicuity" to imply that a location in an image exhibits different properties from its surroundings, where different is taken to mean measurably dissimilar. The concept is similar to that of "surprise" [11] models in videos, where occurrences of statistically unusual events are made using a measure of sudden change of information over time using a statistical model. We develop a model called *VIsual Conspicuity Index* (VICI) that extends our initial concept in [12] in a number of ways, e.g., by Gaussian fitting the spatial distance in (2), rather than using the L_2-norm, using a generalized Gaussian to model central bias, consistent with an average human fixation map [13], and by application to video-based saliency detection.

VICI seeks locations in an image that are measurably dissimilar from their surroundings, along with some other relevant factors. It does not utilize image features that ostensibly match cortical response profiles, nor does it rely on statistics measured at visual fixations, nor does it seek to capture as much new information as possible at each selected location.

Manuscript received June 09, 2011; revised August 24, 2011; accepted September 01, 2011. Date of publication September 12, 2011; date of current version October 13, 2011. This work was supported in part by the National Basic Research Program of China (2009CB320902), the Beijing Natural Science Foundation (4102013 and 4072023), and by the Natural Science Foundation of China (61175115, 60702031, 60970087, 61070116, 61070149, and 60802067). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrizio Campisi.

L. J. Duan and C. P. Wu are with the College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: ljduan@bjut.edu.cn; wuchunpeng@emails.bjut.edu.cn).

J. Miao is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jmiao@ict.ac.cn).

A. C. Bovik is with the Laboratory of Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-0240 USA (e-mail: bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/LSP.2011.2167752

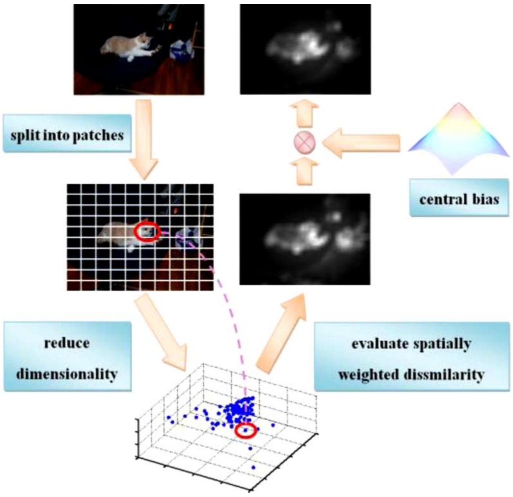

Fig. 1. Framework of VIsual Conspicuity Index (VICI).

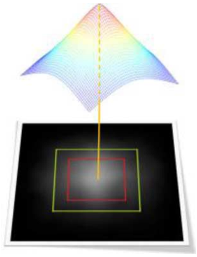

Fig. 2. Generalized Gaussian fit to average human fixation map [13].

Instead, VICI seeks image locations that are different from their surround. This philosophy is also similar to the idea of center-surround filtering of luminance or color, but is not necessarily identified with discontinuities or edges, which need not be "visually attractive" [14].

Secondly, we assign greater importance to plausible next fixation locations by assigning a spatial distance weight.

Thirdly, we account for the tendency of the gaze to return to center [15].

Thus, VICI integrates three elements: spatial dissimilarity, spatial distance and central bias. Spatial dissimilarities are evaluated in a reduced dimensional space. Measured increases in spatial distance between patches cause the influence of dissimilarity between them to decrease; dissimilarity is inversely weighted by distance. The inclusion of a model for the tendency to return the gaze to central is accomplished by a weighting mechanism that biases VICI towards conspicuous nearer the center of the image.

II. Proposed Visual Conspicuity Index

The framework of VICI is shown in Fig. 1 and involves four main stages. First nonoverlapping patches are drawn from an image, and are mapped into a reduced dimensional space. A spatially weighted dissimilarity measure is computed for each patch relative to the other patches. A weighting mechanism that imposes a bias towards the image center is used in the next step. Finally, the saliency map is normalized, resized to the scale of the original image, and smoothed with a unit-energy Gaussian function ($\sigma = 3$).

A. Splitting an Image Into Patches

The input $M \times N$ image is split into nonoverlapping patches. The size of each patch is $l \times l$, so the total number of patches is $\lfloor M/l \rfloor \cdot \lfloor N/l \rfloor$. A patch is denoted as $p_{i,j}$ where $i = 1, 2, \ldots, \lfloor M/l \rfloor$ and $j = 1, 2, \ldots, \lfloor N/l \rfloor$. All of the color channels are stacked to represent each image patch as a column vector of pixel values.

B. Reducing Dimensionality

PCA is used to represent the patches in a reduced dimensional space. Unlike [16], the PCs are sampled from patches in the current image, not from a large number of images. The patch PCA coefficients, which are maximally decorrelated, thereby tend to emphasize spatial dissimilarity within the image. As shown in Fig. 1, a patch is mapped to a point in the reduced dimensional space. By using PCA, the length of the vector corresponding to patch $p_{i,j}$ is reduced to d. More specifically, the patch $p_{i,j}$ is represented as a vector $f_{i,j}$.

C. Spatially Weighted Dissimilarity (SWD)

As the spatial distance between two patches increases, the degree of influence of dissimilarity between them is taken to be decreased. In this way, the dissimilarity of a patch from its neighbors is increased if more similar patches to it are placed more distantly. Thus dissimilarities are inversely weighted by their spatial distance. The Spatially Weighted Dissimilarity (SWD) of the patch $p_{i,j}$ is

$$\mathrm{SWD}(p_{i,j}) = \sum_{s,t} \alpha(p_{i,j}, p_{s,t}) \cdot \varphi(p_{i,j}, p_{s,t}). \tag{1}$$

where the spatial distance weighting $\alpha(p_{i,j}, p_{s,t})$ between $p_{i,j}$ and $p_{s,t}$ is

$$\alpha(p_{i,j}, p_{s,t}) = \exp\left(-\|p_{i,j} - p_{s,t}\| / \sigma_2^2\right) \tag{2}$$

and where $\|p_{i,j} - p_{s,t}\|$ is the Euclidean distance between the patches $p_{i,j}, p_{s,t}$ and $\sigma_2 = 0.4$. We fixed the parameter σ_2 using the learning procedure and dataset in [8]. The "salient" regions of images in this dataset were manually labeled by a large number of subjects [8]. We fixed the parameter $\sigma_2 = 0.4$ simply to be able to cover as many pixels as possible of all the labeled regions in all images in the database. The dissimilarity $\varphi(p_{i,j}, p_{s,t})$ between patches $p_{i,j}$ and $p_{s,t}$ in the reduced dimensional space is defined

$$\varphi(p_{i,j}, p_{s,t}) = \|f_{i,j} - f_{s,t}\|. \tag{3}$$

D. Weighting the SWD Based on Central Bias

The SWD of each patch is also decreased by a factor that weights it as a function of the distance between each patch and the image center, thus accounting for an increase in conspicuity owing to central bias [15]. Therefore, the conspicuity of image patch $p_{i,j}$ is defined

$$C(p_{i,j}) = \beta(p_{i,j}) \cdot \mathrm{SWD}(p_{i,j}) \tag{4}$$

where

$$\beta(p_{i,j}) = \exp[-(\|p_{i,j}\| / \sigma_1)^{r_1}] \tag{5}$$

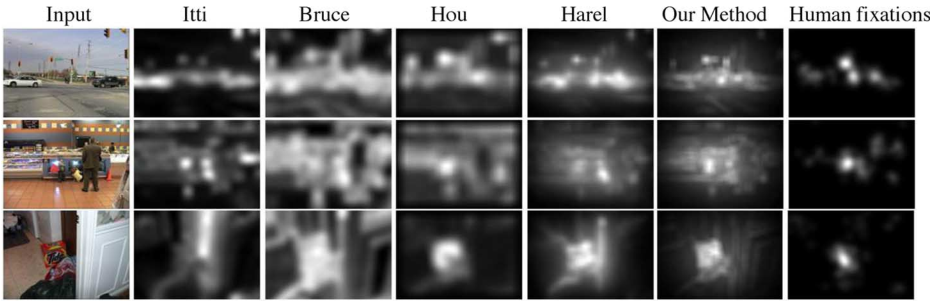

Fig. 3. Images for qualitative comparison between VICI and the four other approaches on the color image dataset 1. The columns from the left to the right are: the input images, the saliency maps from [3], [4], [18] and [17], from VICI, and lastly the human fixation maps.

TABLE I
PERFORMANCE ON THE COLOR IMAGE DATASETS

Attention Model	AUC on Dataset1	AUC on Dataset2
Bruce et al. [4]	0.6949	0.7181
Itti et al. [3]	0.7049	0.7640
Hou et al. [18]	0.7923	0.7625
Harel et al. [17]	0.8021	0.8172
VICI index	**0.8328**	**0.8351**

where $\|p_{i,j}\|$ is the Euclidean distance between the center of the patch $p_{i,j}$ and the image center, and where $r_1 = 1.3$ and $\sigma_1 = cZ$, where $c = 1.7$ and where $Z = \sqrt{(M/2)^2 + (N/2)^2}$ is the maximum possible distance from a pixel to the center of the image. The factors r_1 and c as used in VICI were determined as follows. In [13], the authors calculate an average human fixation map from all images of their database (shown in Fig. 4 of their paper). They find that this map indicates the central bias. To model the central bias, they fit a Gaussian function to the "average human fixation map"; likewise, we fit a generalized Gaussian (5) to the "average human fixation map" from [13] by selecting the parameters r_1 and c (Fig. 2).

III. EXPERIMENTAL VALIDATION

We applied our method on three public image datasets and one video clip to evaluate its performance. VICI was compared with different state-of-the-art saliency detection and fixation selection models based on a commonly-used validation approach. We fixed 14 as the size of each patch and 11 as the number of reduced dimensions. This patch size is consistent with those of the "salient" regions derived from human fixations on Bruce's dataset [4]. We used the same parameter settings on all datasets. We assumed a common YCbCr color space for color images and the video.

A. Results on Color Image Datasets

We tested VICI on two color image datasets. The first dataset is introduced in [4]. It contains 120 images including indoor and outdoor scenes, along with 20 subjects' fixations that were recorded for each image. The second dataset [13] contains 1003 natural images of different scenes with recorded fixations.

To compare the conspicuity maps obtained by the VICI index with human fixations, we use the validation approach from [4]. Specifically, the area under the ROC, i.e., AUC, was used to quantitatively evaluate model performance. We generate ROC curves using code from [17].

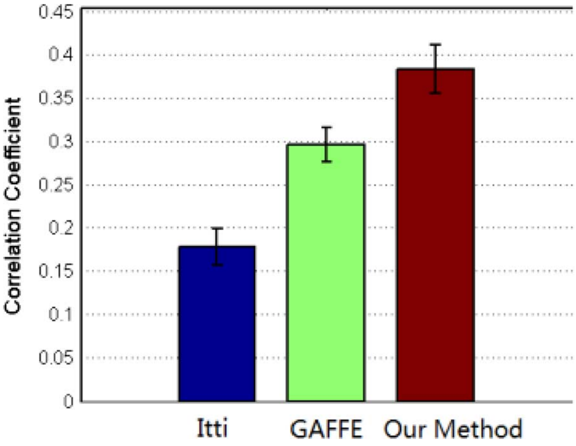

Fig. 4. Quantitative comparison between correlations of VICI index and two other approaches [3], [7] against human fixations on the gray image dataset.

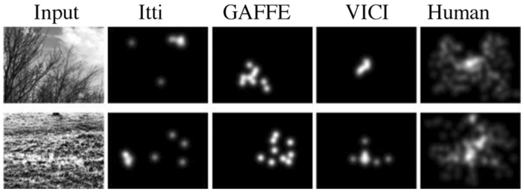

Fig. 5. Qualitative comparison of VICI and two other approaches [3], [7] on the DOVES dataset [19]. Each model generated ten fixations.

As demonstrated in Table I, VICI significantly outperforms the four other methods relative to measured human fixations. To illustrate the efficiency of VICI, Fig. 3 compares it with four highly competitive state-of-the-art approaches [3], [4], [17], [18]. Clearly, VICI yields a highly competitive degree of consistency with human fixation maps.

B. Results on Gray Scale Image Dataset

The gray scale image dataset used is DOVES which is described in [19]. The DOVES dataset includes visual eye movement data from 29 human observers viewing 101 naturalistic calibrated images. We remove the first fixation of each eye movement trace since this fixation is forced [19]. To evaluate performance, the comparison method introduced in [7] is used: the spatial correlations between algorithmic and human fixations is calculated. In order to match the fixations in this dataset, we generate ten fixations from each VICI conspicuity map by selecting the most ten conspicuous regions. We also calculate the correlation for the approach in [3] and for GAFFE [7], with the quantitative results shown in Fig. 4. The first

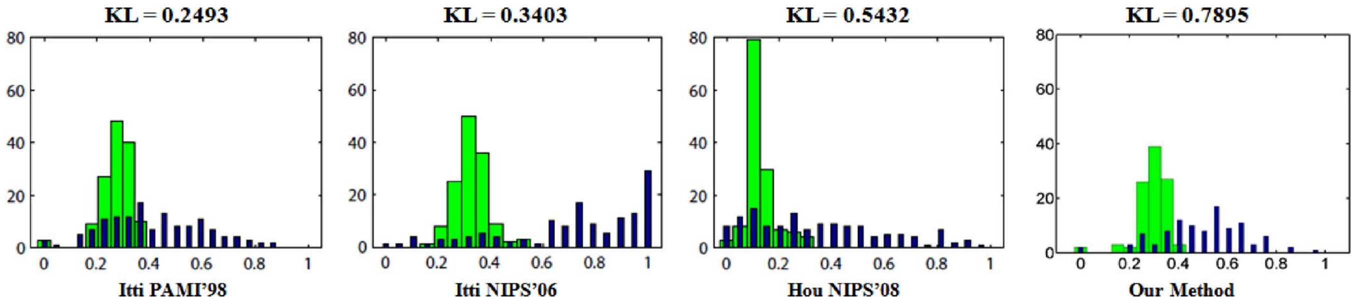

Fig. 6. Performance comparison between VICI and three other methods [3], [11], [18] on a video clip from [11]. The computed conspicuity or saliency distributions are shown at human saccade locations (narrow dark blue bars) and random locations (wide light green bars). The KL divergence between these distributions indicates the performance of each model.

ten fixations generated by these algorithms are also used. As shown in Fig. 4, the points selected by the VICI index are more correlated with the human fixations than are the salient points of [3] or the statistically predicted fixations of [7]. The qualitative results are shown in Fig. 5. The fixations "predicted" by VICI are more compact.

C. *Results on Video Dataset*

To compare with the results reported in Fig. 4 of Hou *et al.* [18], we use the same video clip "beverly03" as Hou did. This video clip is from the dataset introduced in [11]. We use the evaluation method proposed in [11] to compare the performance of the VICI index with the other models. Fig. 6 compares VICI with three classical and competitive methods [3], [11], [18]. The conspicuity or saliency distribution at human saccade locations are represented by narrow blue bars. The random locations are represented by wide green bars. The horizontal axes represents rescaled "saliency" values (dark blue bars) and random locations (light green bars), while the vertical axes represents the rescaled number of pixels corresponding to the saliency value. The KL divergence of these two distributions indicates the performance of each model. The ranking of KL distances indicates that our method achieves better performance than the other three methods.

IV. Discussions and Conclusions

We proposed a visual conspicity index method by integrating three elements: dissimilarity, spatial distance and central bias. Such a conspicuity index can be used to guide image compression [20] or image quality assessment [21].

References

[1] S. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[2] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Foveated analysis of image features at fixations," *Vision Res.*, vol. 47, no. 25, pp. 3160–3172, Nov. 2007.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, Nov. 1998.

[4] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Conf. on Neural Information Processing Systems*, Vancouver, BC, Canada, Nov. 2005.

[5] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe, "Task and context determine where you look," *J. Vis.*, vol. 7, pp. 1–20, 2007.

[6] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vison model," in *IEEE Int. Conf. Comput. Vision Pattern Recogn*, Colorado Springs, CO, Jun. 2011.

[7] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, Apr. 2008.

[8] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[9] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, pp. 387–391, 2005.

[10] L. Itti and C. Koch, "Computational modeling of visual attention," *Nat. Rev. Neurosci.*, vol. 2, pp. 194–203, Mar. 2001.

[11] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Conf. Neural Info. Process Syst.*, Vancouver, Nov. 2005.

[12] L. J. Duan, C. P. Wu, J. Miao, L. Y. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *IEEE Int. Conf. Comput. Vision Pattern Recogn*, Colorado Springs, CO, Jun. 2011.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Int. Conf. Comput. Vision*, Kyoto, Sep. 2009.

[14] Y. Liu, L. K. Cormack, and A. C. Bovik, "Dichotomy between luminance and disparity features at binocular fixations," *J. Vis.*, vol. 10, pp. 1–17, Dec. 2010.

[15] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.

[16] U. Rajashekar, L. K. Cormack, and A. C. Bovik, "Image features that draw fixations," in *IEEE Int. Conf. Image Process.*, Barcelona, Sep. 2003.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Conf. Neural Info. Process Syst.*, Vancouver, Dec. 2006.

[18] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Conf. on Neural Information Processing Systems*, Vancouver, BC, Dec. 2008.

[19] I. van der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "DOVES: A database of visual eye movements," *Spatial Vis.*, vol. 22, no. 2, pp. 161–177, Feb. 2009.

[20] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[21] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–200, Apr. 2009.