Multi-Camera Saliency

Yan Luo, Ming Jiang, *Student Member, IEEE*, Yongkang Wong, *Member, IEEE*, and Qi Zhao, *Member, IEEE*

Abstract—A significant body of literature on saliency modeling predicts where humans look in a single image or video. Besides the scientific goal of understanding how information is fused from multiple visual sources to identify regions of interest in a holistic manner, there are tremendous engineering applications of multi-camera saliency due to the widespread of cameras. This paper proposes a principled framework to smoothly integrate visual information from multiple views to a global scene map, and to employ a saliency algorithm incorporating high-level features to identify the most important regions by fusing visual information. The proposed method has the following key distinguishing features compared with its counterparts: (1) the proposed saliency detection is global (salient regions from one local view may not be important in a global context), (2) it does not require special ways for camera deployment or overlapping field of view, and (3) the key saliency algorithm is effective in highlighting interesting object regions though not a single detector is used. Experiments on several data sets confirm the effectiveness of the proposed principled framework.

Index Terms—Multi-camera saliency, global saliency, region competition, high-level feature saliency, label consistent K-SVD, multi-camera eye tracking data set

1 Introduction

One bottleneck of many visual systems is the information overload problem. In the biological domain, humans and other primates shift their gaze to allocate resources to the most relevant part of the visual world. This ability allows them to process the input data and react in real-time. In the computational domain, the same problem exists, especially with the ever increasing resolution of visual sensors and the volume of visual data. Inspired by the human attentional mechanism, computational saliency models [1], [2], [3], [4], [5], [6] that predict where people look in a visual input identify the most important information from a visual input and have straightforward applications to a variety of real-world tasks such as target detection, video compression, and so on.

While the saliency literature focuses on predicting important regions in a single visual image or video, many real-world problems involve multiple cameras, and it is of great interest to identify regions of interest with information from all camera sources in an integrated way. In a surveillance site, for example, multiple Close-Circuit Television (CCTV) cameras are mounted to have a large field of view. While the conventional saliency prediction methods detect regions of interest in each single view, highlighted regions from one visual source (e.g., simply background areas with bright colors, etc.) can be much less important than those from another source (e.g., humans, etc.). The fact that there is no communication or integration between multiple cameras makes the conventional attentional system local thus limited in terms of information processing or resource allocation at the global level. In reality, human operators often sit in front of tens of screens to monitor the environment and make decisions, which is prone to human boredom and fatigue. In addition, there is a limit in attentional capacity that would worsen human performance when watching multiple views at the same time.

Despite the great practical needs of multi-camera saliency predictions, there are several challenges that make the generalization of single view saliency detection to a multiple views setting non-trivial. First, the placement of cameras can be random thus the perspective views or lighting conditions may differ wildly. Second, a large body of saliency models focuses on low-level information while ignoring higher-level semantics of a scene. Although a couple of object detectors (e.g., face detector [7]) have been added into a saliency model to address this problem [3], [4], [5], [6], performance degenerates in multiple views cases as the commonly used detectors are not entirely view-invariant. Furthermore, adding object detectors does not scale well to the many object categories in practice.

To address the challenges, this paper proposes a principled framework to integrate image features from multiple visual sources for global saliency computation and to globally identify important regions with high-level saliency features. A conceptual example of the proposed principled framework is shown in Fig. 1. Briefly, we first extract the features from the visual sources, followed by transforming the feature channels to a common plane and align them with calibration information. Each group of spatial-temporal synchronized local feature channels are then combined losslessly to obtain global feature channels. Saliency prediction is then performed on the global feature channels with region competition, and finally the predicted results are transformed back to the original views for illustration.

- *Y. Luo, M. Jiang and Q. Zhao are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583. E-mail: {luoyan, mjiang, eleqiz}@nus.edu.sg.*
- *Y. Wong is with the Interactive & Digital Media Institute, National University of Singapore, Singapore 119613. E-mail: yongkang.wong@nus.edu.sg.*

Manuscript received 3 Apr. 2014; revised 12 Dec. 2104; accepted 5 Jan. 2015. Date of publication 14 Jan. 2015; date of current version 4 Sept. 2015.
Recommended for acceptance by Y. Matsushita.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2392783

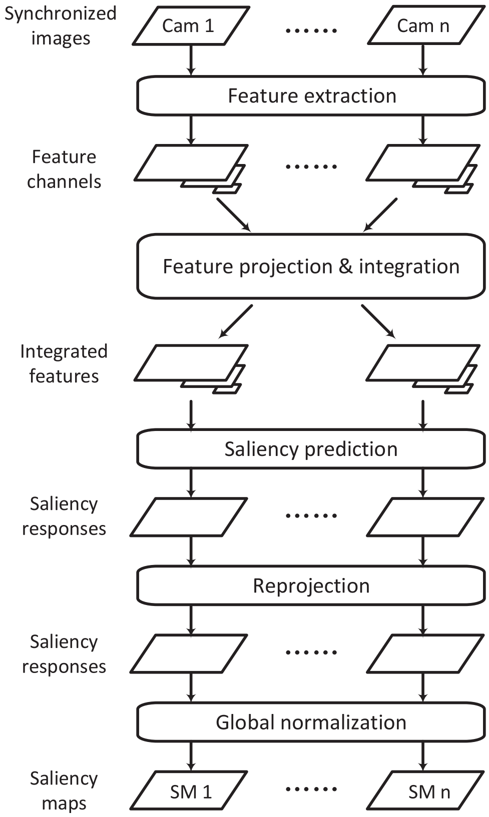


Fig. 1. Flowchart of the multi-camera saliency framework.

Major computational modules here include (1) A feature integration mechanism to losslessly combine the feature channels from multiple visual sources, and (2) A sparse coding based saliency algorithm [8] using supervised information from eye-tracking experiments to enhance the discriminative power of the sparse codes.

The contributions of the paper are summarized as follows:

- We propose a principled framework for conventional computational saliency models to integrate image features from multiple visual sources for enabling region competition mechanism in global saliency computation. Experiments over multiple saliency models show consistent improvement when compared to standalone local saliency computation.
- In our knowledge, this is the first time multi-camera saliency has been applied with unrestricted camera placement and overlapping in field of view for unlimited number of visual sources.
- We introduce a new multi-camera image data set, termed Multi-Camera Image and Eye tracking data set (MCIE). This data set is designed for multi-camera saliency computation under real-world conditions using existing technologies.
- The key saliency prediction method leverages human fixations and learns from where humans look. High-level features have been used, so the model is able to highlight interesting objects rather than regions with distinct low-level features.

The remaining of the paper is organized as follows. Section 2 describes related work. Section 3 describes our saliency prediction algorithm and Section 4 elaborates the details of the proposed principled framework for multi-camera visual saliency. Section 5 demonstrates promising qualitative and quantitative results, and Section 6 concludes the paper.

2 Related Works

2.1 Saliency Model

Modeling visual attention has received increasing interest in both psychology and computer vision fields [1], [2], [3], [4], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. The conventional computational visual saliency models can be generalized into three categories: (1) bottom-up approach [12], [24], [25], [26], [27], [28], [29], (2) top-down approach [30], [31], [32], [33], and (3) hybrid approach [34], [35], [36]. The bottom-up approach only considers the use of early visual features in the saliency prediction, whereas the top-down approach applies task-specific features to model the goal-directed attention. The hybrid approach models both the bottom-up and the top-down factors.

Computational saliency models predict important locations of a visual scene and focus limited resources to the identified regions. Based on the feature integration theory by Treisman and Gelade [9], the first saliency model was proposed by Koch and Ullman [1], and later implemented by Itti et al. [2]. Along the same line, there are a number of biologically-inspired algorithms to predict where humans look at in images [3], [4], [11]. In these models, low-level features (i.e., color, intensity, and orientation) were extracted and feature channels were computed through center-surround filtering at numerous spatial scales. The features were later combined by a linear mechanism for saliency computation. Color, intensity and orientation have proved to be effective attributes to guide visual search [37] and attention-based computational model [9], [38].

Based mostly on low-level features, various computational algorithms were developed to infer saliency of different feature channels. Most commonly adopted feature integration algorithms include Bayesian framework [27], Markov chains [39], information maximization [40], [41], and spectral analysis in the frequency domain [22], [42]. With the improved integration algorithms, these models perform better than the classic one [2]. However, a well recognized problem of the low-level-feature-based models is that they fail to encode the higher-level statistics in a visual scene. As recent psychophysical [43], [44] and computational studies [3], [6], [45] suggest, visual attention is attracted by semantically interesting regions or objects, especially in complex visual scenes like crowds [46]. Therefore, in this work, we propose to extract high-level image features for a better saliency detection.

To fill the *semantic gap* between computational saliency models and human performance, specifically-trained object detectors have been incorporated into saliency models. For example, faces have been shown to attract attention independent of tasks, and several recent models [3], [4], [5] combined face detection as a separate visual cue with traditional low-level features to improve saliency detection. Furthermore, Judd et al. [6] proposed a Support Vector

Machine (SVM) based learning approach to linearly combine face, pedestrian and car detectors with low- and mid-level features. To some extent, the integration of multiple object detectors increases the prediction performance, yet it is barely possible to scale such algorithms to the large number of object categories in real life. To approach this challenge, this paper leverages human data with supervised sparse coding and a set of features to effectively represent low-level and high-level information. The saliency maps learned directly from the human data are therefore capable of encoding interesting objects that are not limited to any specific categories.

2.2 Saliency Model with Multiple Visual Sources

Modeling visual attention with the stereo camera has been explored to take advantage of the depth and disparity information [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57]. The existing literature mainly focus on stereoscopic configuration or combining 2D and 3D visual sources. Frintrop et al. [48] presented a bimodal attention system for robotic applications which are capable of processing data from different sensor modes simultaneously. The intensity maps are weighted and added to the orientation maps to compute the conspicuity maps. Maki et al. [47] fused stereo disparity, flow maps and motion to predict visual selection. Bruce and Tsotsos [49] presented a stereo model of visual attention based on the Selective Tuning model, which extends naturally to the binocular domain. Several binocular feature channels can be combined to compute the stereo saliency. Jeong et al. [50] introduced a binocular stereo attention model which integrates the static and dynamic features together. Similarly, Zhang et al. [51] proposed a stereoscopic visual attention model to simulate human visual system, which simultaneously combines the image saliency, motion saliency and depth map. Niu et al. [52] explored stereopsis for saliency analysis by considering color contrast-based and disparity contrast saliency together. Lang et al. [53] discussed whether and how depth information influences visual saliency, and extended some existing methods to include the learned depth priors. The influence on human visual attention based on the visual field contiguity and depth contiguity are discussed in [56]. Yuan et al. [57] introduced a model which can discover the thematic object in a given collection of images or a video sequence.

Most of the aforementioned works focus on stereoscopic or binocular configuration to obtain additional cues (e.g., depth) for saliency prediction. In contrast, the proposed principled framework aims to integrate visual information from multiple visual sources to a holistic common plane. The integrated visual information allows the visual saliency to be predicted in a global and cognitively natural space. In this sense, the motivation of the proposed work is quite different from the aforementioned ones. As a result, the configuration, methodology and focuses to achieve the objectives are also quite different. For example, the visual sources discussed in the aforementioned literature are always assumed to be closely positioned (i.e., stereoscopic, binocular, etc.), and have a large overlap in the field of view; while the proposed method does not require special ways for camera deployment or overlapping field of view. In addition, the aforementioned works focus on the setting of two visual sensors as it is the conventional configuration to estimate depth information. The proposed work, on the other hand, is designed to be applied to an arbitrary number of sensors.

In this work, we aim to simultaneously analyze the visual information from multiple visual sources. The most relevant work that fall under our proposed scenario are in [58]. The authors proposed to stitch images from two image sources, followed by predicting the salient region using existing saliency model [59]. The proposed method requires strict camera placement such that each visual source share some degree of overlap region with each other, and having small degree of differences in the corresponding view perspective. In contrast to [58], the proposed principled framework combines the visual input of multiple visual sources for saliency prediction in a global, and cognitively natural space to enable region competition mechanism. In our knowledge, this is the first time multi-camera saliency has been applied with unrestricted camera placement and overlapping in field of view for unlimited number of visual sources.

3 Learning a Discriminative Dictionary for Saliency Prediction

This section provides a general framework for saliency prediction. In particular, we aim to learn high-level information to fill the *semantic gap* between computational saliency models and human behavior. Two distinctions from conventional object detection methods are that: (1) Interesting objects highlighted by this method are not restricted to specific categories, and (2) instead of using pre-defined image sets with object labels, the training data are sampled from images viewed by human subjects. A conceptual example of the general framework is shown in Fig. 2. We extract low-level image features from salient and non-salient patches at various scales. Our saliency model is learned with a Label Consistent K-SVD (LC-KSVD) approach proposed by Jiang et al. [60]. The discriminative sparse codes learned with LC-KSVD can be seen as higher-level features to best differentiate salient objects or image structures from the non-salient ones.

3.1 Feature Extraction and Sampling

3.1.1 Center-Surround Features

Following the conventional saliency model by Itti et al. [2], an input image is first sub-sampled into a Gaussian pyramid of S scales from $1/1$ (scale 0) to $1/256$ (scale 8). At each scale, the image is decomposed into seven feature channels, including red-green and blue-yellow color contrast channels (C_{RG} and C_{BY}), intensity channel (I), and four local orientation channels ($O_\theta,\ \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$) computed using Gabor filters. Center-surround differences are computed and normalized following [2].

3.1.2 Histograms of Oriented Gradients (HOG)

HOG features have been widely used in object detection [61], for its ability of capturing object texture and contour information against noises or environmental changes. To encode image statistics as a complementary cue to the

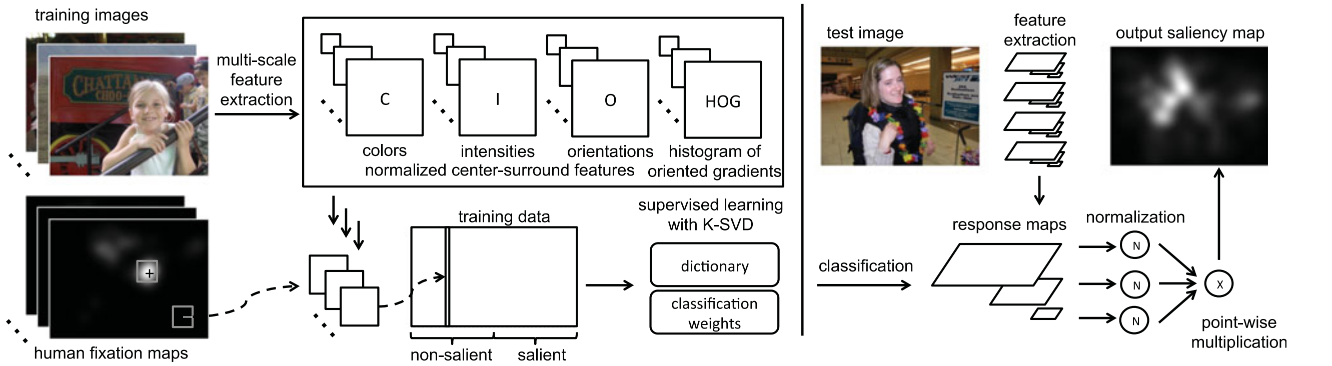


Fig. 2. An overview of the LC-KSVD saliency model. In the training phase, center-surround and HOG features are first extracted from a Gaussian pyramid of each training image. Then, using the ground-truth saliency map generated with human fixations, salient and non-salient image patches are sampled, whose features are later fed into a dictionary learning algorithm to jointly learn a discriminative dictionary and a linear classifier. In the testing phase, the dictionary and weights are used to generate multi-scale saliency maps of a test image. These maps are finally normalized and combined with a point-wise multiplication into the final saliency map.

pixel-level center-surround features, locally normalized HOG representation with both contrast-sensitive and contrast-insensitive orientation bins is incorporated. We follow the construction in [62] to define a dense representation of an image at each particular scale.

3.1.3 Feature Sampling

In this work, a dictionary of saliency features is learned from randomly sampled image patches labeled with ground-truth saliency. Particularly, given an image, we randomly select p pixels from the top 30 percent salient regions and q pixels from the bottom 30 percent salient regions. The thresholding is based on the ground-truth fixation map derived from human eye-tracking data. Particularly, each fixated location is represented as a white pixel (and non-fixated as black ones). The fixation map is then blurred with a Gaussian kernel to generate the ground-truth saliency map. The intensities of the blurred saliency map indicate the fixation density at each particular image pixel. For each selected pixel, we extract its $r \times r$ neighborhood from each scale (we use $r = 7$ in all experiments) and concatenate all the center-surround and HOG features to form a feature vector.

3.2 Dictionary Learning with Class and Scale Consistency

Sparse coding has found support in the biological domain where sparsity is not only the response property of neurons in area V1, but also that of areas deeper in the cortical hierarchy [63]. In this work, sparse coding approach is employed to learn an efficient representation of image features in relation to saliency.

In the context of sparse representation, the objective is to approximate a given sample as a linear combination of a small number of vectors, where these vectors form the subspaces of a feature space. This feature space is thought to be overcomplete such that any given sample can be represented with a relatively small set of vectors. Under the formal mathematical formulation, let us suppose that $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_K] \in \mathbb{R}^{N\times K}$ is a real matrix where each column, $\boldsymbol{d}_i$, is a N-dimensional vector with unit Euclidean norm. The matrix $\boldsymbol{D}$ is generally referred to as a *dictionary* and each column $\boldsymbol{d}_i$ is known as a *basis*. Given a set of training feature samples, $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M] \in \mathbb{R}^{N\times M}$, extracted from labeled salient or non-salient image patches. We aim to obtain discriminative sparse codes $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M] \in \mathbb{R}^{K\times M}$ and the dictionary $\boldsymbol{D}$. The objective of this dictionary learning problem can be formulated as:

$$<\boldsymbol{D}, \boldsymbol{X}> = \arg\min_{\boldsymbol{D},\boldsymbol{X}} \|\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \; s.t. \; \forall\, i, \|\boldsymbol{x}_i\|_0 \le T, \tag{1}$$

where the term $\|\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{X}\|_F^2$ represents the reconstruction error. The notation $\|\boldsymbol{M}\|_F$ stands for the Frobenius norm, where $\|\boldsymbol{M}\|_F^2$ is defined as $\sum_i \sum_j |m_{i,j}|^2$. T is a sparsity constraint factor that stands for the maximum number of nonzero entries in each sparse code $\boldsymbol{x}_i$.

Saliency prediction is casted as a binary classification problem, where each class corresponds to a class label (i.e., salient or non-salient). In this work, we follow the LC-KSVD to simultaneously learn a set of discriminative sparse codes and a linear classifier. Specifically, this is done by adding two regularization terms to (1). One term enforces the discrimination capabilities for salient versus non-salient image patches at different scales, which encourages the input data sampled from the same class (salient or non-salient) and the same scale to have very similar sparse representations. The other is a classification error term, which allows the learned sparse codes to be predictive with a linear classifier. Intuitively, sparse codes learned at different scales capture various aspects of the visual input. In addition, similar features learned at different scales can also differ in their ability to attract attention. For example, larger faces tend to attract attention more strongly than smaller ones [46], possibly as they are closer to the viewer.

The objective function can now be re-formulated as:

$$\begin{aligned}<\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{w}> = \arg\min_{\boldsymbol{D},\boldsymbol{A},\boldsymbol{X},\boldsymbol{w}} & \|\boldsymbol{Z}–\boldsymbol{D}\boldsymbol{X}\|_F^2 + \alpha\|\boldsymbol{U} - \boldsymbol{A}\boldsymbol{X}\|_F^2 \\ & + \beta\|\boldsymbol{v}^T - \boldsymbol{w}^T\boldsymbol{X}\|_2^2 \;\; s.t. \; \forall\, i, \|\boldsymbol{x}_i\|_0 \le T,\end{aligned} \tag{2}$$

where the terms $\|\boldsymbol{U} - \boldsymbol{A}\boldsymbol{X}\|_F^2$ and $\|\boldsymbol{v}^T - \boldsymbol{w}^T\boldsymbol{X}\|_2^2$ represents the discriminative sparse code error, and the linear classification error, respectively. The coefficients α and β control the relative contribution of the corresponding terms and are both 0.5 in this work. $\boldsymbol{v}$ is the saliency labels. The discriminative sparse code error term forces feature z that belong to the same class to have similar sparse representation. The linear classification error term learns an optimal classifier which infers the saliency label from the sparse representation. The two terms enables LC-KSVD algorithm to take discrimination capability of the dictionary and connect sparse representation to the saliency label.

Here the matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_M] \in \{0,1\}^{K\times M}$ are the discriminative sparse codes of input $\boldsymbol{Z}$ for classification. Each column $\boldsymbol{u}_i$ is a "discriminative" sparse code corresponding to an input sample z_i. $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ is a linear transformation matrix which transforms the original sparse codes in $\boldsymbol{X}$ to be most discriminative. To explain this in the saliency context, assuming the input data $\boldsymbol{Z} = (\boldsymbol{Z}_0^1, \ldots, \boldsymbol{Z}_0^s, \boldsymbol{Z}_1^1, \ldots, \boldsymbol{Z}_1^s)$ is a set of image features sampled at s scales where $s = 1, \ldots, S$. $\boldsymbol{Z}_0^s$ and $\boldsymbol{Z}_1^s$ respectively represent non-salient and salient sub-matrices of $\boldsymbol{Z}$ at scale s. Now, the matrix $\boldsymbol{U}$ can be defined as:

$$\boldsymbol{U} \equiv \begin{pmatrix} \boldsymbol{U}_0^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{U}_0^S & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{U}_1^1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{U}_1^S \end{pmatrix}, \quad (3)$$

where $\boldsymbol{U}_l^s, l \in \{0,1\}$ are all matrices of ones. For example, given $S = 1$, the diagonal entries of $\boldsymbol{U}$ are $\boldsymbol{U}_0^1$ and $\boldsymbol{U}_1^1$ where $\boldsymbol{U}_0^1, \boldsymbol{U}_1^1 \in \mathbb{R}^{\frac{K}{2}\times\frac{M}{2}}$ are all-ones matrices. Thus, the discriminative sparse code error term enforces that the sparse codes $\boldsymbol{X}$ can approximate the discriminative sparse codes $\boldsymbol{U}$ with a linear transformation $\boldsymbol{A}$.

In the linear classification error term $\|\boldsymbol{v}^T - \boldsymbol{w}^T\boldsymbol{X}\|_2^2$, $\boldsymbol{w} = [w_1, w_2, \ldots, w_K]^T \in \mathbb{R}^K$ represents the classification weights to reconstruct the ground-truth saliency labels $\boldsymbol{v} = [v_1, v_2, \ldots, v_M]^T \in \{0,1\}^M$ with the sparse representations $\boldsymbol{X}$. Note that instead of using binary labels for classification, v_i represents the ground-truth saliency value of the ith input sample, which is the pixel intensity at the coordinate of the patch center in the ground-truth saliency map.

To find the optimal solution for all parameters simultaneously, (2) can be rewritten as:

$$<\boldsymbol{D}', \boldsymbol{X}> = \arg\min_{\boldsymbol{D}', \boldsymbol{X}} \|\boldsymbol{Z}' - \boldsymbol{D}'\boldsymbol{X}\|_F^2 \; s.t. \; \forall\, i, \|\boldsymbol{x}_i\|_0 \leq T, \quad (4)$$

where $\boldsymbol{Z}'$ and $\boldsymbol{D}'$ are denoted as:

$$\boldsymbol{Z}' = (\boldsymbol{Z}^T, \sqrt{\alpha}\boldsymbol{U}^T, \sqrt{\beta}\boldsymbol{v})^T$$
$$\boldsymbol{D}' = (\boldsymbol{D}^T, \sqrt{\alpha}\boldsymbol{A}^T, \sqrt{\beta}\boldsymbol{w})^T.$$

As a generalization of data clustering, the above dictionary learning problem can be efficiently solved by the K-SVD algorithm [64].

3.3 Saliency Prediction

The obtained dictionary $\boldsymbol{D}$, linear transformation parameters $\boldsymbol{A}$ and classification weights $\boldsymbol{w}$ in the supervised training phase can be used to predict the saliency map of a new image. Note that $\boldsymbol{D}$, $\boldsymbol{A}$ and $\boldsymbol{w}$ cannot be directly used for testing since they are jointly normalized in $\boldsymbol{D}'$ in the LC-KSVD algorithm, *i.e.*,$\forall\, k$, $\|(\boldsymbol{d}_k^T, \sqrt{\alpha}\boldsymbol{a}_k^T, \sqrt{\beta}w_k)^T\|_2 = 1$. Instead, given a test feature vector $\boldsymbol{z}$, sparse code $\boldsymbol{x}$ and saliency value v can be computed as follows:

$$\boldsymbol{x} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{z} - \hat{\boldsymbol{D}}\boldsymbol{x}\|_2^2 \quad s.t. \quad \|\boldsymbol{x}\|_0 \leq T, \quad (5)$$

$$v = \exp(\hat{\boldsymbol{w}}^T\boldsymbol{x}) - 1, \quad (6)$$

where $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{w}}$ are denoted as:

$$\hat{\boldsymbol{D}} = \left\{\frac{\boldsymbol{d}_1}{\|\boldsymbol{d}_1\|_2}, \cdots \frac{\boldsymbol{d}_k}{\|\boldsymbol{d}_k\|_2}, \cdots \frac{\boldsymbol{d}_K}{\|\boldsymbol{d}_K\|_2}\right\} \quad (7)$$

$$\hat{\boldsymbol{w}} = \left\{\frac{w_1}{\|\boldsymbol{d}_1\|_2}, \cdots \frac{w_k}{\|\boldsymbol{d}_k\|_2}, \cdots \frac{w_K}{\|\boldsymbol{d}_K\|_2}\right\}. \quad (8)$$

For each scale of features, a sliding window approach is employed to compute the saliency value of every pixel to generate a saliency map. The saliency maps of all scales are then normalized and combined to generate the master saliency map. Empirically, we find that using a pixel-wise multiplication instead of summing up across all scales leads to better prediction performance and visualization results.

4 Multi-Camera Saliency Framework

This section elaborates the proposed principled framework for multi-camera visual saliency. Intuitively, the feature channels obtained from each camera are transformed and integrated into a common plane with pre-calibrated parameters, followed by visual saliency prediction. Finally, the predicted saliency are reprojected back to the original views followed by global normalization. In the following sections, we first present the generalization to conventional visual saliency models and the details of each component. Then, we describe the procedure to apply the proposed principled framework with our LC-KSVD saliency model [8].

4.1 Overview of Multi-Camera Saliency

As discussed in Section 2, one common property of the conventional saliency models is that these models are designed for single visual source. The visual saliency predicted on a single visual source is considered as *local* and often isolated from other sensors' field of view, which disregards the influences of global event or object-of-interest on another visual source. In addition, using only the local saliency may suppress the responses of semantically informative regions. To overcome such shortcoming, a number of global rarity based models [27], [28], [65], [66], [67] are proposed, where saliency computation is a result of spatial competition of the reference images. However, the resulting saliency prediction is still confined to a single visual source. In order to perform genuine multi-camera saliency prediction, we generate global feature channels by integrating visual features from all available visual sources. Therefore, the inter-feature's weights and spatial competition are more comprehensive when compared to single visual source scenario.

Based on the above discussions, we formulate the multi-camera saliency framework for n cameras as:

$$\boldsymbol{m}_i = f_H^{-1}(\mathcal{S}(f_H(\boldsymbol{F}_i) \oplus f_H(\boldsymbol{F}_{j\neq i}))), \tag{9}$$

where $\boldsymbol{m}_i$ is the saliency response of ith camera, $\mathcal{S}$ is an arbitrary saliency model, f_H and f_H^{-1} are homogeneous mapping function and the inverse of homogeneous mapping function, respectively. $\oplus$ is the integration operation. $\boldsymbol{F}_i$ is the feature channels from ith camera.

The flowchart of the multi-camera saliency framework is shown in Fig. 1. First, the feature channels are extracted from the synchronized images of n cameras. Then, they are projected onto the common image plane and integrated together to construct the global feature channels. Next, based on the global feature channels, the global saliency responses are generated by the saliency model for each camera. Finally, the global saliency responses are reprojected back to the respective original views, followed by global normalization.

4.2 Geometric Transformation

In the real-world environment, cameras are positioned in various locations with different visual perspective with respect to the common image plane. Specifically, the common image plane is the floor plane in this work. Therefore, the view of each visual source has to be perspectively transformed to obtain an unified image perspective using a learned homogeneous mapping function. We assume that the relationship between the source views' image plane, denoted as *local image plane*, and common image plane are none, and manually calibrate this relationship with labeled reference points. Specifically, we marked a set of reference points based on the intersection points of a grid reference map, and only the visually correspondence points are used for calibration. Given a set of correspondence points captured in the local image plane, $\{(x_1^l, y_1^l), (x_2^l, y_2^l), \ldots, (x_n^l, y_n^l)\}$, and the respective shared points from common image plane, $\{(x_1^g, y_1^g), (x_2^g, y_2^g), \ldots, (x_n^g, y_n^g)\}$, the transformation can be modeled as a projectivity transformation with eight degrees of freedom [68]. The plane projection can be modeled with a 3×3 non-singular homogeneous matrix $\boldsymbol{H}$, which can be estimated via

$$\begin{bmatrix} x^g \\ y^g \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x^l \\ y^l \\ 1 \end{bmatrix}. \tag{10}$$

In order to project all local image planes to the common image plane, we need to estimate the homography matrix for each camera. In the following sections, we denote the homography matrix for the ith camera as $\boldsymbol{H}_i$.

Once $\boldsymbol{H}_i$ is learned, we can project the ith camera's feature channels, $\boldsymbol{F}_i$, to the common image plane to generate the transformed feature channels, $\bar{\boldsymbol{F}}_i$, by:

$$\bar{\boldsymbol{F}}_i = f_{H_i}(\boldsymbol{F}_i) \tag{11}$$

$$(\bar{x}, \bar{y}, 1)^T = \boldsymbol{H}_i(x, y, 1)^T, \tag{12}$$

where $\forall\ (x, y) \in \boldsymbol{F}_i$ and $(\bar{x}, \bar{y}) \in \bar{\boldsymbol{F}}_i$.

4.3 View and Feature Integration

The geometric image transformation in Section 4.2 raises two potential problems to the feature extraction process for convention saliency models. First, the unobserved region (black region in Fig. 4) in the common image plane will introduced artifact for saliency model with holistic feature. This also applies to the patch-based feature extraction approach for patches surrounding image's boundary. Despite the possibility to apply a mask to avoid these edges, some useful information near these region will not contribute to visual saliency prediction. Second, the aforementioned geometric image transformation does not consider 3D model estimation for foreground objects (e.g., human, bag, chair, etc.). Therefore, pixels correspond to any foreground objects will have visible image distortion, which poses a threat for the quality of the extracted features (see Fig. 4). To overcome these problems, the proposed principled framework perform the geometric transformation on the feature channels. This approach will also guarantee the generalization of this framework to most conventional saliency models.

Given the feature channels extracted from n cameras, i.e., $\boldsymbol{F}_i$ where $i = 1, \ldots, n$, and the corresponding feature maps on the common image plane $\bar{\boldsymbol{F}}_i$. The integrated feature channels with respect to the ith camera, $\hat{\boldsymbol{F}}_i$, can be obtained via:

$$\hat{\boldsymbol{F}}_i = \bar{\boldsymbol{F}}_i \oplus \bar{\boldsymbol{F}}_{j\neq i} = \bar{\boldsymbol{F}}_i \cup \left(\bigcup_{j\neq i} (\bar{\boldsymbol{F}}_j \backslash (\bar{\boldsymbol{F}}_j \cap \bar{\boldsymbol{F}}_i)) \right), \tag{13}$$

where $\bigcup$ is the intersection operator for a sequence of feature channels. In other words, to compute the integrated feature channels $\hat{\boldsymbol{F}}_i$, we conserve all related features on $\bar{\boldsymbol{F}}_i$ to $\hat{\boldsymbol{F}}_i$ and integrate the non-overlapping information of $\bar{\boldsymbol{F}}_{j\neq i}$ into $\hat{\boldsymbol{F}}_i$. An conceptual example is shown in Fig. 5. In this example, $\hat{\boldsymbol{F}}_1$ is composed of the union of $\bar{\boldsymbol{F}}_1$ and $\bar{\boldsymbol{F}}_2$ where the overlapping region is selected from $\bar{\boldsymbol{F}}_1$. For the overlapped region on $\bar{\boldsymbol{F}}_{j\neq i}$, we will assign higher priority to the camera with smaller visual distortion, which can be derived with the visual perspective with respect to the common image plane.

4.4 Saliency Prediction and Global Normalization

After we obtain the integrated feature channels, $\hat{\boldsymbol{F}}_i$, we can generate the respective saliency response, $\hat{\boldsymbol{m}}_i$, with a given saliency model, $\mathcal{S}(\cdot)$, via

$$\hat{\boldsymbol{m}}_i = \mathcal{S}(\hat{\boldsymbol{F}}_i). \tag{14}$$

Now, the saliency response $\boldsymbol{m}_i$ on the ith camera can be obtained by reprojecting $\hat{\boldsymbol{m}}_i$ to the original view by the inverse of homogeneous mapping $f_{H_i}^{-1}$, followed by prune out the region which is not within the original view. To represent the conspicuity at every location in the visual field by a scalar quantity and simulate the field of view of human attention, saliency response $\boldsymbol{m}$ of each image are convoluted with a Gaussian kernel $\boldsymbol{g}$. The global normalized saliency map of ith camera, $\tilde{\boldsymbol{m}}_i$, is formulated as follows:

$$\tilde{\boldsymbol{m}}_i = \frac{\boldsymbol{m}_i * \boldsymbol{g} - \min(\boldsymbol{m}_j * \boldsymbol{g})}{\max(\boldsymbol{m}_j * \boldsymbol{g}) - \min(\boldsymbol{m}_j * \boldsymbol{g})}, \tag{15}$$

where $j = 1, \ldots, n$ and $*$ represents the convolution operator. By perform the global normalization, small number of

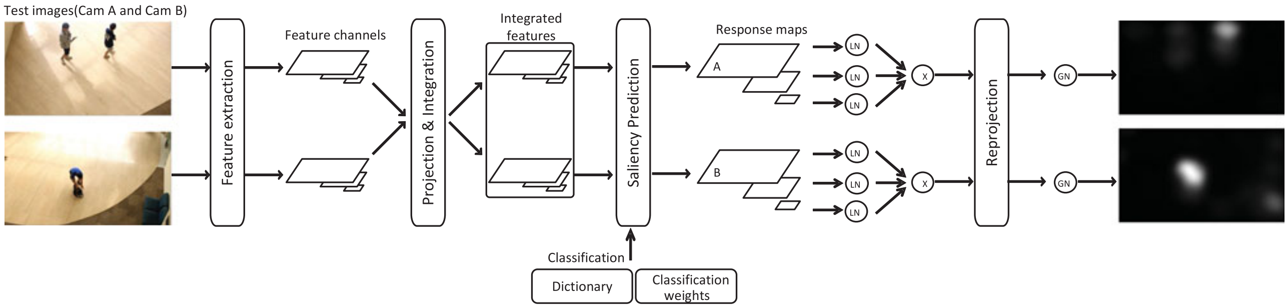


Fig. 3. The conceptual example of the proposed principled framework with LC-KSVD saliency model. First, the feature channels obtained from local views are transformed and integrated into a common plane with pre-calibrated parameters. Then, we compute the global saliency map the learned overcomplete dictionary and classification weights. The predicted saliency are then re-projected to the local views followed by global normalization step. Node **LN**, **GN** and **X** local normalization, global normalization, and point-wise multiplication, respectively.

strong peaks in these response maps are promoted with the same global normalized parameters.

4.5 Multi-Camera Saliency with LC-KSVD Model

In this section, we delineate the procedure to apply the proposed principled framework with our LC-KSVD saliency model (see Section 3). Specifically, we detail the dictionary learning stage and visual saliency prediction stage. The details of the remaining components is the same as previous sections. A conceptual example is shown in Fig. 3.

4.5.1 Dictionary Learning

The discriminative dictionary and the classification weights for the multi-camera saliency framework are learned with training patches extracted from the projected feature channels, $\bar{F}$. Due to the projectivity transformation, the shape of the transformed feature channel is in trapeziform and the patches extracted from the edge of the image will contains pixels from the unobserved regions. Under this scenario, the feature sampling mechanism in Section 3.1 might fail. To address this, each training patch must satisfies a selection condition, where every pixels in the extracted patch has a corresponding pixel in the original image. Now, given the training data, $\bar{Z} \in \mathbb{R}^N$, extracted with the center-surround and HOG feature patches from the projected feature channels, (2) can now be re-formulated as:

$$\begin{aligned} < \bar{\boldsymbol{D}}, \bar{\boldsymbol{A}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{w}} > &= \arg\min_{\bar{\boldsymbol{D}},\bar{\boldsymbol{A}},\bar{\boldsymbol{X}},\bar{\boldsymbol{w}}} \|\bar{\boldsymbol{Z}} - \bar{\boldsymbol{D}}\bar{\boldsymbol{X}}\|_F^2 \\ &\quad + \alpha\|\bar{\boldsymbol{U}} - \bar{\boldsymbol{A}}\bar{\boldsymbol{X}}\|_F^2 + \beta\|\bar{\boldsymbol{v}}^T - \bar{\boldsymbol{w}}^T\bar{\boldsymbol{X}}\|_2^2 \qquad (16) \\ s.t. \quad & \forall\ j,\ \|\bar{\boldsymbol{x}}_j\|_0 \le T. \end{aligned}$$

Note that the dictionary learning mechanism of LC-KSVD model is a patch-based learning method. The patches extracted from various cameras within the overlapped region are treated equally, whereas the spatial competition discussed in Section 4.1 is not applicable. Therefore, the dictionary can be trained with the transformed feature channel instead of integrated feature channel.

4.5.2 Saliency Prediction

Given the learned $\bar{D}$ and a test feature vector $\hat{\boldsymbol{z}}$, the corresponding saliency value $\hat{v}$ can be computed with (6). To compute the saliency response for the ith camera, $\hat{\boldsymbol{m}}_i$, we first perform feature integration as delineated in Section 4.3, followed by computing the saliency value for each valid feature patch. Now, the local saliency response $\boldsymbol{m}_i$ can be obtained via:

$$\boldsymbol{m}_i = f_{H_i}^{-1}(\hat{\boldsymbol{m}}_i). \qquad (17)$$

Finally, the globally normalized saliency map $\tilde{\boldsymbol{m}}_i$ can be calculated using (15).

5 Experiments

In this section, we first delineate the evaluation metrics and the new multi-camera data set used in this work, followed by human behavioral analysis on two-view configuration to study the impact on human eye-fixations. The proposed principled framework is evaluated with the learning based LC-KSVD model and four state-of-the-art saliency models. We refer reader to [8] for the details

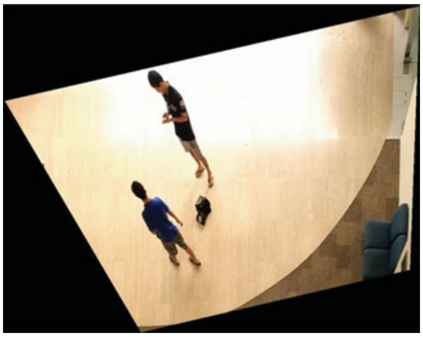


Fig. 4. Example of a transformed image obtained with the pre-calibrated parameters.

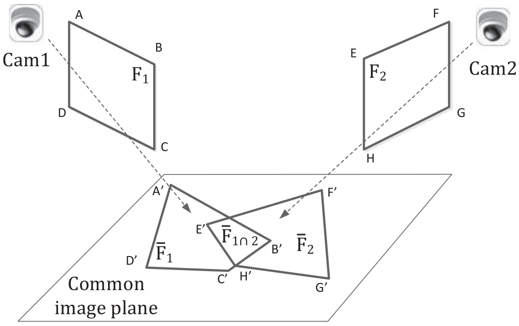


Fig. 5. A conceptual example of projectivity transformation from source view to common image plane.

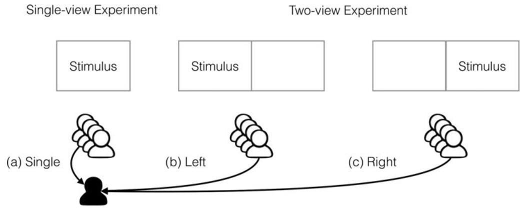


Fig. 6. Experimental configurations and comparisons between single-view and two-view fixations.

performance of the LC-KSVD model under single-view configuration. For all computational experiments, we provide qualitative and quantitative results on two-view and three-view configurations.

5.1 Evaluation Metrics and Saliency Models

In the saliency literature, there are several widely used criteria to quantitatively evaluate the performance of saliency models by comparing the saliency prediction with eye movement data. One of the most common evaluation metrics is the area under the Receiver Operator Characteristic (ROC) curve (i.e., AUC) [69]. ROC refers to a curve obtained by varying the threshold values on the predicted saliency map, and for each value, plotting the true positive rate on the y-axis against the false positive rate on the x-axis, and AUC is the area under the ROC curve with respect to the increase of the false positive rate. A problem with this metric is that it is significantly affected by the center bias effect [70], so the shuffled AUC was then introduced [27] to address this problem. Particularly, to calculate the shuffled AUC, negative samples are selected from human fixation locations from all training images, instead of uniformly sampling from all image locations.

In addition, the Correlation Coefficient (CC) [71] and the Normalized Scanpath Saliency (NSS) [72] are also used to measure the statistical relationship between the saliency prediction and the ground truth. NSS is defined as the average saliency value at the fixation locations in the normalized predicted saliency map which has zero mean and unit standard deviation, while the CC measures the linear correlation between the saliency map and the ground-truth map. The three metrics are complementary and provide a more objective evaluation of the various models. All the reported performance is the mean accuracy with 10-fold cross validations.

In this work, we evaluate the performance of the learning based LC-KSVD model, as well as four state-of-the-art saliency models that are publicly available (i.e., Itti's model [2] (denoted as Itti) implemented by Harel [39], the GBVS model [39], the SUN model [27] and the Image Signature model [22]).

5.2 Data Sets

In order to evaluate the multi-camera saliency framework, a data set which contains multiple synchronized visual sources with unrestricted camera placements, perspective views and lighting conditions is required. Existing data sets with multiple visual sources have the following limitations: (i) large overlapping and confined views [73], (ii) highly controlled conditions [74], and (iii) insufficient synchronized images [75]. Due to the above limitations, we collected a new multi-camera data set, termed Multi-Camera Image and Eye tracking data set,[1] designed for multi-camera saliency experiment under real-world conditions using existing technologies. The new MCIE data set incorporates two image subsets: two-view subset and three-view subset.

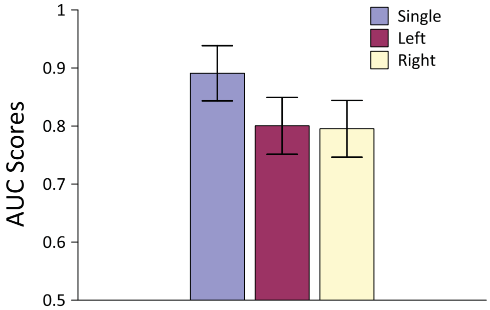


Fig. 7. Inter-subject AUC scores (mean and standard deviation) in the single-view and two-view experiments. *Left* and *Right* indicate the image position under two-view experiment.

The two-view subset was recorded with two digital SLR cameras with image resolution of 1280×720 pixels at 25 frames per second (FPS). Each camera is positioned to provide maximal coverage of the scene with small overlap view between cameras. It consists of two scenes: *indoor scene*: captured from the building lobby, and *outdoor scene*: captured from the entrance of a building. Both scenes contains six video sequences with different scenarios and content (e.g., a pedestrian walks pass the area, two individuals enter the scene and have a conversation, etc.). We manually selected 450 pairs of synchronized images, The three-view subset was recorded with three AXIS P5512 network cameras with image resolution of 704×576 pixels at 25 FPS. This subset is recorded from the lobby of an auditorium and the pedestrians' behavior is uncontrolled. It is composed of 3×450 synchronized images, which contains 200 pairs from indoor scenes and 250 pairs from outdoor scenes, to ensure diversity in the semantic contains in each image pair. For all scenes, we manually label reference points on the overlapped region as reference grid map.

For both subsets, we collected eye tracking data with 16 human subjects[2] free-viewing the synchronized images for 2 seconds. As subjects free-viewed the images, we used Eyelink 1,000 (SR Research, Osgoode, Canada) eye tracking device to record eye movements at a sample rate of 2,000 Hz. The screen resolution was set to $1{,}920 \times 1{,}080$ pixels, and the synchronized images were displayed in a random order, and uniformly scaled to full-screen when presented on a 22 inches display. The synchronized images from the two-view subset were horizontally aligned side-by-side manner, where images from the three-view subset were aligned as a quadtree with one quad of the quadtree leaves blank. The display was placed at 66.04 cm from the subjects, and the screen size was 47.39×29.62 cm, therefore the visual angle of the stimuli was about $40.5^\circ \times 25.3^\circ$. A chin-rest and a forehead-rest were used to stabilize the subjects head. In the experiments, each pair of images was presented for 2 seconds followed by a drift correction, which required subjects to fixate at the

[1] Available via https://github.com/NUS-VIP/MCIE
[2] The subjects for each subset were recruited independently.

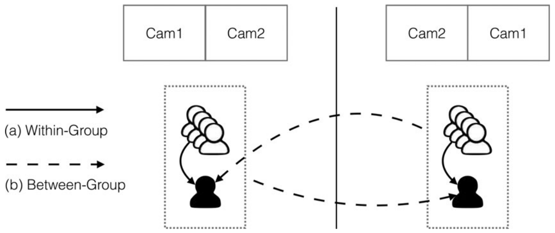


Fig. 8. Experimental configuration and comparison between the Cam1-Cam2 and Cam2-Cam1 settings. Within-group and between-group AUC scores are computed for each setting.

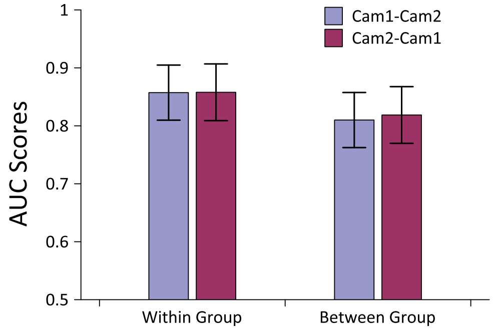


Fig. 9. Within-group and between-group AUC scores (means and standard deviations) in the two-view experiment.

screen center and press a key to continue. For both subset, the 450 synchronized images were randomly permuted into three sections of human fixation collection. Before each section, a nine-point target display was used for calibration and a second one was used for validation. Subjects took a short break after each section.

5.3 Human Behavioral Analysis

To compare the human eye-fixations between different experimental settings, we computed the inter-subject AUC scores by evaluating each subject's fixations against the compared subject groups. Averaging across all images led to an overall AUC for each subject. Specifically, to compute the AUC score for an image, positive samples are the fixated pixels, and negative samples are the other pixels in the compared fixation map. For the following discussions, the mean AUC scores and the corresponding standard deviations (SD) are reported, and the statistical significance of their differences is tested with paired t-test.

First, compared with the eye-fixations in single images, subjects behaved differently when two or more images were viewed together. To investigate this difference, as illustrated in Fig. 6, we conducted a single-view eye-tracking experiment using the same set of images as the two-view one, in which all pairs of images were presented in separate trials. In addition, for each image stimulus in the single-view experiment, we categorized the two-view eye-tracking data into two groups (left and right) according to the position where the image was viewed. We computed the inter-subject AUC scores by comparing each subject's fixations with (a) the fixation map generated from all other subjects' fixations in the single-view experiment, (b) left and (c) right fixations in the two-view experiment.

As shown in Fig. 7, across all subjects, the two-view fixations (left: AUC $= 0.80 \pm 0.05$; right: AUC $= 0.80 \pm 0.05$) do not outperform the single-view ones (AUC $= 0.89 \pm 0.05$), suggesting that the viewing patterns in the two experiments are significantly different (left versus single: $t(16) = -41.97$, $p = 8.51 \times 10^{-18}$; right versus single: $t(16) = -41.36$, $p = 1.07 \times 10^{-17}$).

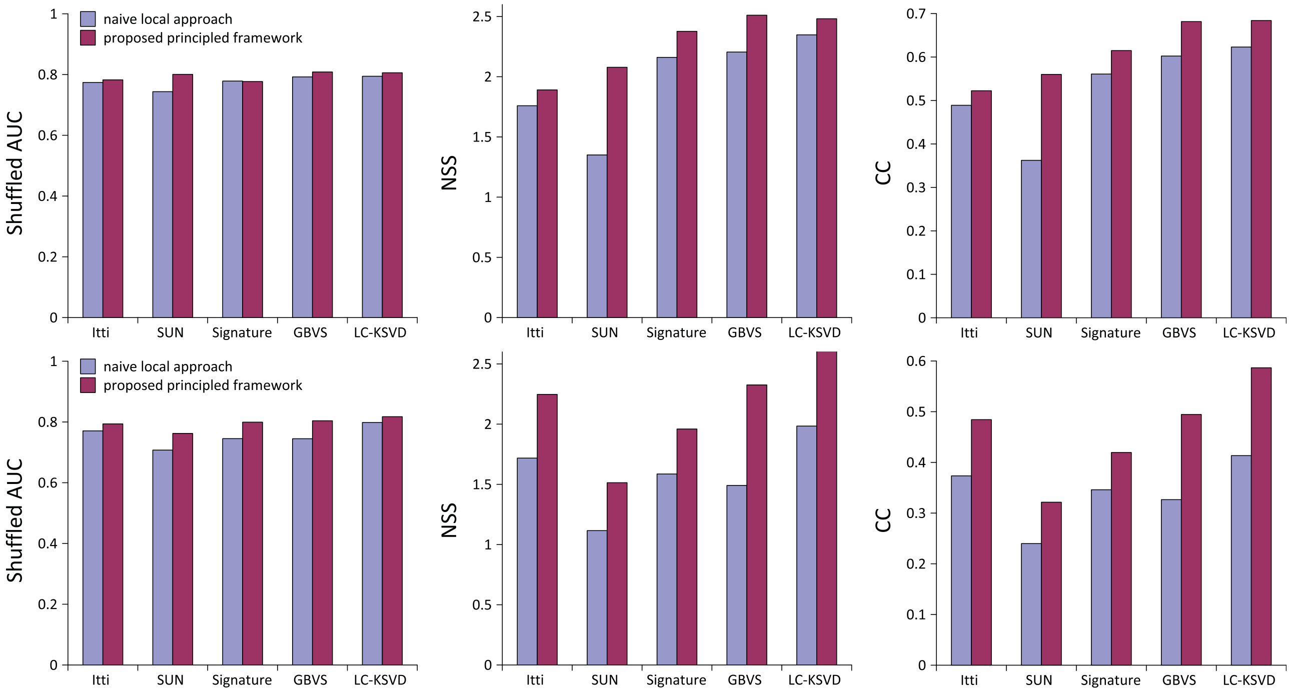


Fig. 10. Quantitative comparison of various saliency models on MCIE data set. **TOP:** two-view subset, and **BOT:** three-view subset. The prediction accuracy is measured with the shuffled AUC, NSS and CC scores. The reported performance is the mean accuracy with 10-fold validations.

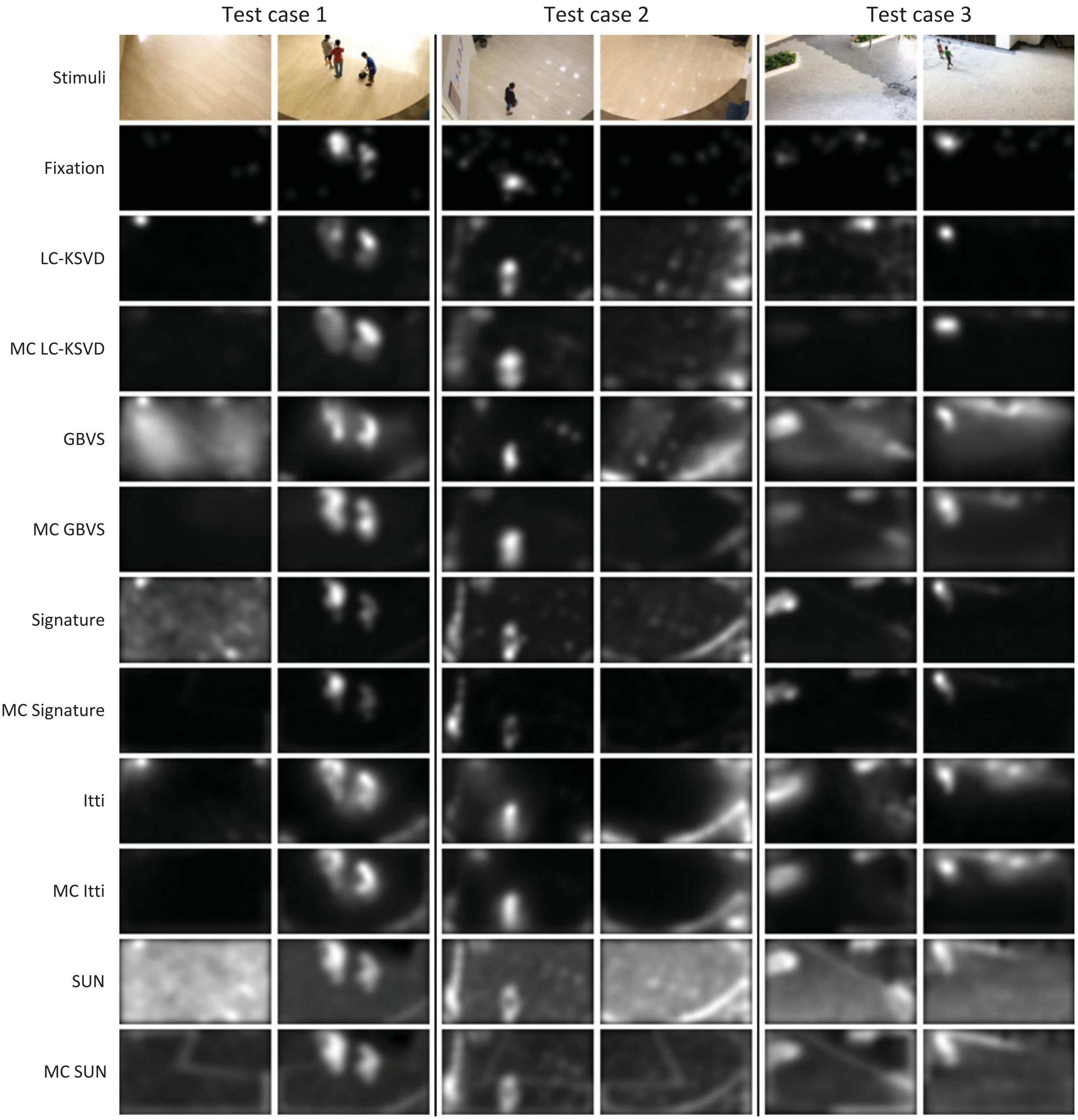


Fig. 11. Qualitative results of the proposed principled multi-camera saliency framework with the state-of-the-art models over samples from MCIE two-view subset. MC indicates the saliency prediction with the proposed principled framework.

Further, to investigate the effect of the image placement on the viewing patterns, we categorized the two-view fixations into two groups, according to the order of the two views, i.e., Cam1-Cam2 and Cam2-Cam1. As illustrated in Fig. 8, each subject's fixations were evaluated against both (a) other fixations in the same group, and (b) all fixations in the other group. AUC scores were computed for each subject in both of the aforementioned conditions, and averaged across all images.

As shown in Fig. 9, the within-group AUC scores (Cam1-Cam2: AUC $= 0.86 \pm 0.05$; Cam2-Cam1: AUC $= 0.86 \pm 0.05$) were not significantly different between the two settings ($t(15) = -0.17,\ p = 0.87$). However, for both settings, the between-group evaluations (Cam1-Cam2: AUC $= 0.81\pm 0.05$; Cam2-Cam1: AUC $= 0.82 \pm 0.05$) scored lower than the within-group ones (Cam1-Cam2: AUC $= 0.86 \pm 0.05$; Cam2-Cam1: AUC $= 0.86 \pm 0.05$). Both differences were statistically significant (Cam1-Cam2: $t(15) = 8.21,\ p = 6.29\times 10^{-7}$; Cam2-Cam1: $t(15) = 8.11,\ p = 7.26 \times 10^{-7}$). This is mostly due to the central fixation bias caused by the experimental setup that requires the subjects to fixate at the screen center before the onset of the stimuli, as well as strategic advantages in looking at the image center. Therefore, by randomizing the display order of the two-view images, the center bias is reduced in our data set for a fair comparison of the saliency models.

5.4 Two-View Evaluations

In this section, we evaluate the performance of the proposed principled framework on the two-view subset. The computed saliency map of the synchronized image set are

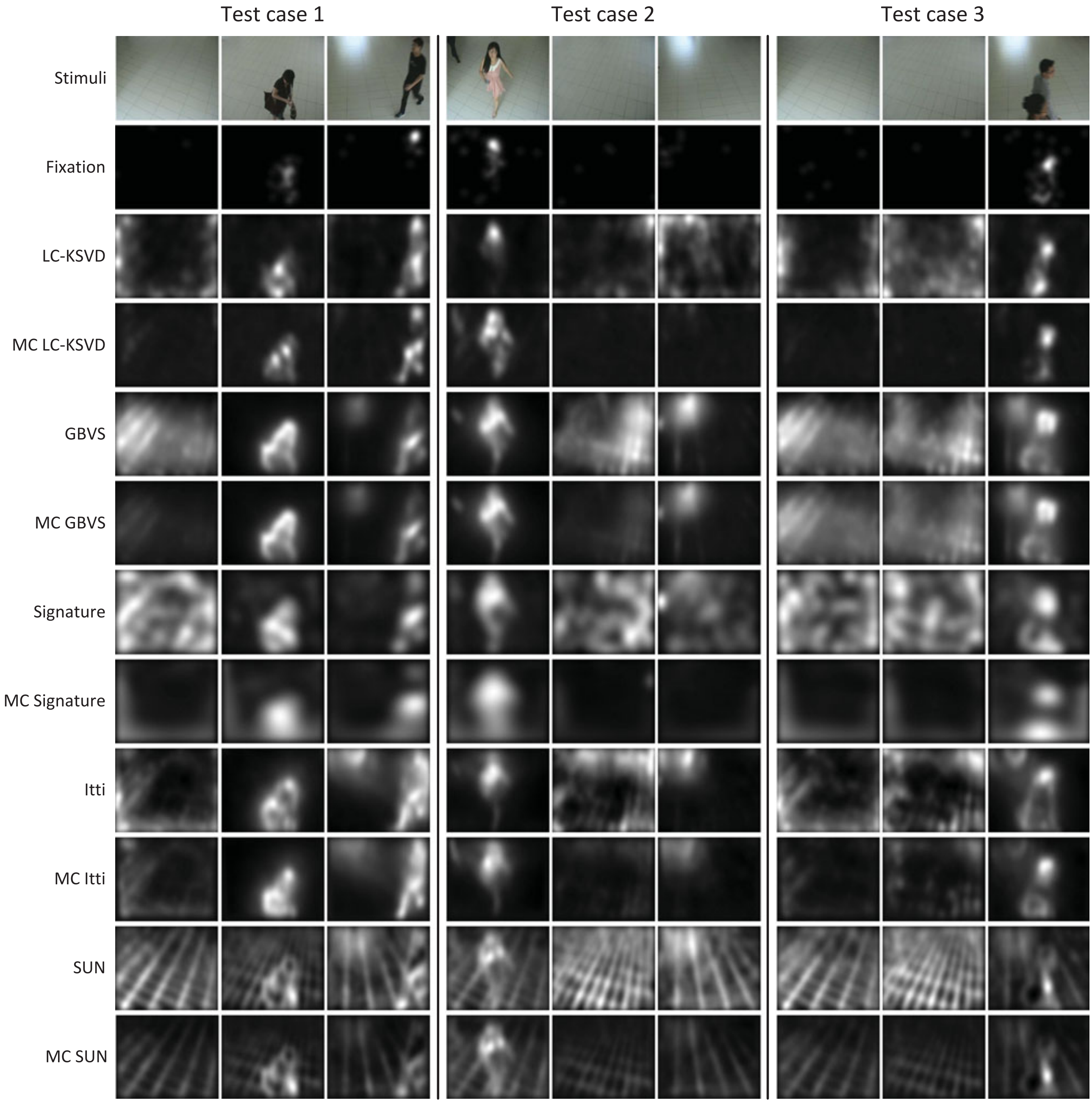


Fig. 12. Qualitative results of the proposed principled multi-camera saliency framework with the state-of-the-art models over samples from MCIE three-view subset. MC indicates the saliency prediction with the proposed principled framework.

stitched horizontally to compute the CC, NSS and shuffled AUC with corresponding human fixation maps.

Fig. 10 illustrates the quantitative results. We conducted two sets of experiments for each comparison method. First, we conducted native local saliency prediction on each image individually, denoted as naive local approach. Second, we applied the proposed principled framework on each comparison method to simultaneously predict the saliency map for both images. Overall, all comparison methods show that the proposed principled framework is better, if not the same, than the naive local approach across all three evaluation metrics. The most significant improvement is on SUN model with 53.9 and 54.6 percent on NSS and CC, respectively.

The qualitative results on three test cases are shown in Fig. 11. The first and second test cases are extracted from the indoor scene and the third test case is obtained from the outdoor scene. For each comparison method, we show the saliency prediction results with both the naive local approach and the proposed principled framework. As shown in Fig. 11, the saliency predicted with the proposed principled framework are closer to the human fixation maps. The efficacy of the proposed principled framework are dramatically shown in GBVS model and Image Signature model, where the relatively less salient image (with no subjects) does not contains highlighted salient regions in both models. Also, the LC-KSVD model gives lower saliency response over the flowerbed (test case 3) with the proposed principled framework, which indicates that the person is more saliency in the global perspective. Based on the results, we conclude that the advantages of the

TABLE 1
Computational Cost of the Proposed Principled Framework with the LC-KSVD Saliency Model

Operation	Cost (seconds)
Feature extraction	0.330
Feature projection (per channel)	0.244
Feature integration (per channel)	0.018
Saliency prediction	3.939
Saliency reprojection	1.521
Global normalization	0.006

proposed principled framework are two folds. First, we can penalize the saliency objects or regions which is salient in local view but not in the global context. Second, learning to detect salient objects directly from eye tracking data makes the proposed principled framework more scalable than explicitly adding object detectors [3], [4], [5].

5.5 Three-View Evaluations

In this experiment, we observed significant improvement on all comparison methods across shuffled AUC, NSS, and CC (see Fig. 10). The improvement on the proposed LC-KSVD is around 42.7 and 41.9 percent when compared with 5.7 and 9.78 percent in the two-view experiment on NSS and CC, respectively. Three qualitative results are shown in Fig. 12 test case 2. Overall, the results agree with the observation on the two-view subset, where the locally salient regions are now suppressed in the global view perspective. In test case 2, the strong reflection of light source on tiles in the third camera results in high saliency response for naive local approaches, while the saliency response of the same region is strongly penalized with the proposed principled framework.

The proposed principled framework and all comparison models were implemented in Matlab 2013b, running on a 64-bits Window 7 machine with 3.4 GHz Intel i5-3570 CPU. The computational time of dictionary learning is about 340 seconds. The average computational time for each component on one image can be found in Table1.

6 Conclusions and Future Works

In this work, we have presented a sparse coding based algorithm to learn a discriminative dictionary for high-level saliency prediction with the LC-KSVD algorithm, and proposed a principled framework to effectively integrate local visual sources and predict global visual saliency. The proposed principled framework has the following key distinguishing features compared with its counterparts: (1) the proposed algorithm detects important regions in the global context, (2) it does not require certain layout of camera deployment or overlapping fields of view, and (3) the key saliency algorithm is aware of high-level feature though not a single detector is used. Comprehensive evaluation over a number of data sets confirm the efficacy of the key saliency algorithm and the multi-camera saliency framework. In addition, the multi-camera saliency framework can be directly adapted by the conventional saliency models and shows good performance on MCIE data set.

For future work, we plan to employ 3D model fitting and depth estimation to improve the quality of image projection, and employ automated camera calibration method to eliminate the manual labeling task. Another direction is to apply the multi-camera saliency framework to existing video saliency models, we are now working on the extension of LC-KSVD saliency model to video and will present it in the future Last but not least, we would like to conduct study on the human eye-fixations using multi-camera data set with a wide variety of object categories and scenes.

Acknowledgments

The research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office, and the Defense Innovative Research Programme (No. 9014100596). Q. Zhao is the corresponding author.

References

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Matters Intell.*, vol. 188, pp. 115–141, 1987.

[2] 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, no. 11, pp. 1254–1259, Nov. 1998.

[3] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, pp. 1–15, 2009.

[4] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[5] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *J. Vis.*, vol. 12, no. 6, pp. 1–15, 2012.

[6] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[7] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[8] M. Jiang, M. Song, and Q. Zhao, "Leveraging human fixations in sparse coding: Learning a discriminative dictionary for saliency prediction," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2013, pp. 2126–2133.

[9] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[10] A. Torralba, "Contextual modulation of target saliency," in *Proc. Adv. Neural Inform. Process. Syst.*, 2001, pp. 1303–1310.

[11] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.*, vol. 42, no. 1, pp. 107–123, 2002.

[12] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[13] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, Jun. 2009.

[14] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[15] J. Fan and Y. Wu, "Contextual saliency," *IEEE Vis. Commun. Image Process.*, pp. 1–4, Nov. 2011.

[16] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[17] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[18] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.

[19] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *J. Vis.*, vol. 11, no. 4, pp. 1–20, 2011.

[20] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[21] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 853–860.

[22] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[23] S. Ramenahalli and E. Niebur, "Computing 3D saliency from a 2D image," in *Proc. Conf. Inform. Sci. Syst.*, 2013, pp. 1–5.

[24] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE 48th Annu. Int. Symp. Optical Sci. Technol.*, 2004, vol. 5200, pp. 64–78.

[25] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Am. A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[26] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[27] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[28] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inform. Process. Syst.*, 2005, pp. 155–162.

[29] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inform. Process. Syst.*, 2008, pp. 681–688.

[30] A. Borji, D. N. Sihite, and L. Itti, "Computational modeling of top-down visual attention in interactive environments." in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.

[31] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 61–76, 2011.

[32] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," *Vis. Res.*, vol. 50, no. 14, pp. 1338–1352, 2010.

[33] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2751–2758.

[34] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[35] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, 2010.

[36] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A stochastic model of selective visual attention with a dynamic Bayesian network," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 1073–1076.

[37] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, 2004.

[38] J. M. Wolfe, "Guided search 4.0: Current progress with a model of visual search," *Int. Models Cognit. Syst.*, pp. 99–119, 2007.

[39] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inform. Process. Syst.*, 2006, pp. 545–552.

[40] L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," in *Proc. Adv. Neural Inform. Process. Syst.*, 2004, pp. 1121–1128.

[41] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 1–24, 2009.

[42] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[43] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, pp. 1–26, 2008.

[44] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *J. Vis.*, vol. 10, no. 8, pp. 1–19, 2010.

[45] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.

[46] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. 13th Eur. Conf. Comput. Vis.*, vol. 8695, 2014, pp. 17–32.

[47] A. Maki, P. Nordlund, and J.-O. Eklundh, "Attentional scene segmentation: Integrating depth and motion," *Comput. Vis. Image Understanding*, vol. 78, no. 3, pp. 351–373, 2000.

[48] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann, "A bimodal laser-based attention system," *Comput. Vis. Image Understanding*, vol. 100, no. 1-2, pp. 124–151, 2005.

[49] N. D. B. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," in *Proc. 2nd Canadian Conf. Comput. Robot Vis.*, 2005, pp. 88–95.

[50] S. Jeong, S.-W. Ban, and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment," *Neural Netw.*, vol. 21, no. 10, pp. 1420–1430, 2008.

[51] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3D video," in *Proc. 16th Int. Conf. Adv. Multimedia Model.*, 2010, pp. 314–324.

[52] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 454–461.

[53] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. S. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 7573, 2012, pp. 101–115.

[54] E. Ekmekcioglu, H. K. Arachchi, A. Kondoz, C. G. Gurler, and S. Sedef Savas, "Content aware delivery of visual attention based scalable multi-view video over P2P," in *Proc. Int. Packet Video Workshop*, 2012, pp. 71–76.

[55] R. Horaud, D. Knossow, and M. Michaelis, "Camera cooperation for achieving visual attention," *Mach. Vis. Appl.*, vol. 16, no. 6, pp. 331–342, 2006.

[56] U. Rashid, M. A. Nacenta, and A. J. Quigley, "Factors influencing visual attention switch in multi-display user interfaces: A survey," in *Proc. Int. Symp. Pervasive Displays*, 2012.

[57] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.

[58] C. W. H. Ngau, L.-M. Ang, and K. P. Seng, "Multi camera visual saliency using image stitching," in *Proc. Int. Conf. Telecommun. Technol. Appl.*, 2011, pp. 93–98.

[59] F. Urban, B. Follet, C. Chamaret, O. L. Meur, and T. Baccino, "Medium spatial frequencies, a strong predictor of salience," *Cogn. Comput.*, vol. 3, no. 1, pp. 37–47, 2011.

[60] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1697–1704.

[61] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[62] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[63] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inform. Process. Syst.*, 2007, pp. 873–880.

[64] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[65] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

[66] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[67] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inform. Process. Syst.*, 2009, pp. 681–688.

[68] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K: Cambridge Univ. Press, 2004.

[69] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, 2005.

[70] 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.

[71] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri, "Empirical validation of the saliency-based model of visual attention," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 3, no. 1, pp. 13–24, 2004.

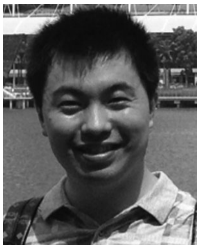
[72] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.

[73] D. Thirde, L. Li, and F. Ferryman, "Overview of the pets2006 challenge," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, 2006, pp. 47–50.

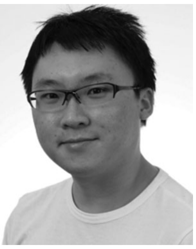
[74] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 441–444.

[75] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey, "A database for person re-identification in multi-camera surveillance networks," in *Proc. Int. Conf. Dig. Image Comput. Tech. Appl.*, 2012, pp. 1–8.

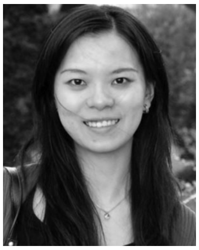
Yan Luo received the BSc degree in computer science from Xi‘an, China, in 2008. After that, he worked in the industry for several years on distributed system. In 2013, he joined the Sensor-enhanced Social Media (SeSaMe) Centre at the Interactive and Digital Media Institute, National University of Singapore, as a research assistant. In 2014, he joined the Visual Information Processing Laboratory at the National University of Singapore.



Ming Jiang received the bachelor of engineering degree in 2004 and the master of engineering degree in 2008 from Zhejiang University, China, and he is working towards the PhD degree at the National University of Singapore. His research interests include the areas of computer vision, computational visual cognition, and computational neuroscience. He is a student member of the IEEE.



Yongkang Wong received the bachelor of engineering degree in 2006 from the University of Adelaide, Australia, and the PhD degree in 2012 from the University of Queensland, Australia. He is a research fellow at the Sensor-enhanced Social Media (SeSaMe) Centre in the Interactive and Digital Media Institute, National University of Singapore. He has worked as a graduate researcher at NICTA's Queensland laboratory from 2008 to 2012. His current research interests are in the areas of computer vision, machine learning, multi-camera analysis, and video surveillance. He is a member of the IEEE.



Qi Zhao received the MSc and PhD degrees in computer engineering from the University of California, Santa Cruz, in 2007 and 2009, respectively. He is an assistant professor in the Electrical and Computer Engineering Department at the National University of Singapore (NUS) and the principal investigator at the Visual Information Processing Lab (http://www.ece.nus.edu.sg/stfpage/eleqiz), working on computational vision and cognitive neuroscience. She also holds an appointment in the Interactive and Digital Media Institute at NUS. Prior to joining NUS, she was a postdoctoral researcher in the Computation & Neural Systems, and Division of Biology at the California Institute of Technology from 2009 to 2011. Her main research interests include computational vision, machine learning, computational cognition, and neuroscience. She has published more than 30 journal and conference papers in top computer vision, cognitive neuroscience, and machine learning venues, and is editing a book with Springer, titled *Computational and Cognitive Neuroscience of Vision*, that provides a systematic and comprehensive overview of vision from various perspectives, ranging from neuroscience to cognition, and from computational principles to engineering developments. She is a member of the IEEE.