Biologically Inspired Media Quality Modeling

Luming Zhang[†‡], Meng Wang[†], Liqiang Nie[‡], Richang Hong[†],
Yingjie Xia[*] and Roger Zimmermann[†]
[†]Department of CSIE, Hefei University of Technology, China
[‡]School of Computing, National University of Singapore, Singapore
[*]College of Computer Sciences, Zhejiang University, China
{zglumg,eric.mengwang,nieliqiang,hongrc.hfut}@gmail.com,
xiayingjie@zju.edu.cn rogerz@comp.nus.edu.sg

ABSTRACT

In this paper, we propose a biologically inspired quality model, focusing on interpreting how humans perceive visually and semantically important regions in an image (or a video clip). Particularly, we first extract local descriptors (graphlets in this work) from an image/frame. They are projected onto the perceptual space, which is built upon a set of low-level and high-level visual features. Then, an active learning algorithm is utilized to select graphlets that are both visually and semantically salient. The algorithm is based on the observation that each graphlet can be linearly reconstructed by its surrounding ones, and spatially nearer ones make a greater contribution. In this way, both the local and global geometric properties of an image/frame can be encoded in the selection process. These selected graphlets are linked into a so-called biological viewing path (BVP) to simulate human visual perception. Finally, the quality of an image or a video clip is predicted by a probabilistic model. Experiments shown that 1) the predicted BVPs are over 90% consistent with real human gaze shifting paths on average; and 2) our quality model outperforms many of its competitors remarkably.

Keywords

Biological; Gaze shifting; Path; Active learning; Geometry, Preservation; Human perception

1. INTRODUCTION

The volume of media data[1] we handle on a daily basis is growing exponentially due to the availability of ubiquitous and cheap sensors, sharing platforms, and new social trends. Artificial intelligence has proven useful for interpreting this preponderance of data. In the last decades, many models have been proposed to evaluate the quality of an image or a video clip. A successful quality model can facilitate many multimedia applications. For example, by quantifying the saliency of different image regions, a higher accuracy can be achieved if we employ only the salient regions from each image for retrieval. Moreover, video summarization algorithms typically extract key frames from one or multiple video clips. A well-designed quality model can generate a set of semantically representative and low redundant key frames.

In the literature, various media quality models have been proposed by utilizing both low-level (*e.g.*, texture and structure) and high-level visual features, (*e.g.*, portrait and architecture). As far as we know, however, none of them can capture the movement of the human eye and the mechanism of of visual cortex in aesthetic perception. Recent reports from both neuroscience and computer vision have demonstrated that biologically plausible features perform impressively in visual recognition. Thus, we propose a biological inspired quality model containing two modules: 1) engineering biological viewing paths (BVPs) that reflect how human perceive regions salient in an image/frame; and 2) a mathematical model that combines the BVPs of multiple users to assess media quality. It is worth emphasizing that developing such two modules encounters the following challenges:

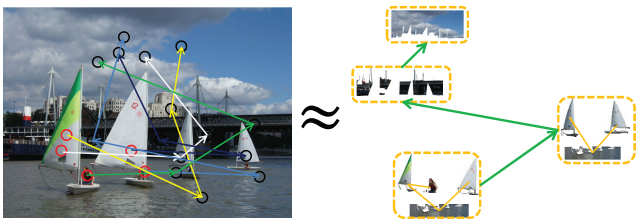


Figure 1: The gaze shifting paths recorded from five observers (left) and the BVP predicted by our algorithm (right)

- To mimick human gaze shifting, the predicted BVP should maximally reflect human perception of image/frame geometric attributes. However, current models cannot ensure that the detected salient regions can best recover an image/frame both locally and globally. This is because of the difficulty to establish a mathematical model that optimizes the local and global descriptiveness of a BVP simultaneously.

[1]The media data in this work specifically denotes images or video clips. The training media means the training images or video clips; and the test media denotes a test image or video clip.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
MM '15, October 26–30, 2015, Brisbane, Australia.
Copyright 2014 ACM 978-1-4503-3459-4/15/10 ...$15.00.
http://dx.doi.org/10.1145/2733373.2806255.

- Psychophysics studies have demonstrated that object-level visual cues dominate over the deployment of attention [1]. That is to say, observers are more likely to attend to the "interesting" objects, rather than the highly contrasted non-object regions. However, current computational saliency models have limited capabilities to exploit object-level cues. One common scheme is to integrate object detectors, but this does not scale well and in practice, only a few pre-specified categories such as human faces can be well detected.

- We target a quality model that can simulate human visual perception. That means the model should convey some subjective factors. As shown in Fig. 1, observers with different backgrounds, education, and nationalities might generate slightly different gaze shifting paths on a same image (or a video clip). Toward a fair gaze shifting prediction, we expect that a BVP is maximally similar to the gaze shifting paths recorded from multiple observers. The challenge is how to implement such a "maximally similar" mechanism.

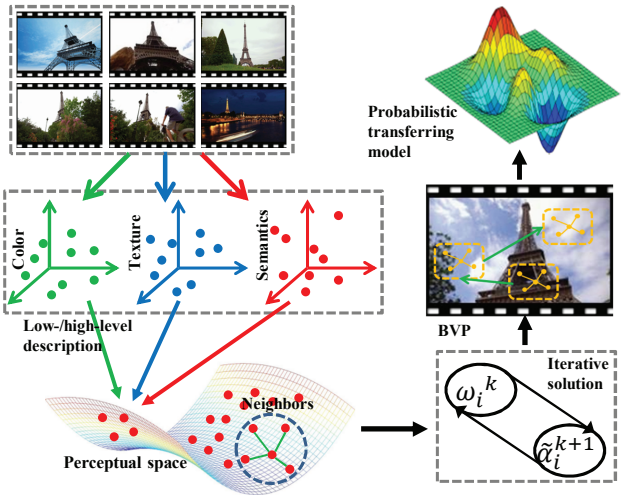


Figure 2: The pipeline of the proposed biological inspired media quality model

To solve the above problems, we propose a novel media quality model. The key is a geometry-preserving active learning (GPAL) that constructs BVPs in the perceptual space, which reflects both the low-level and high-level features. An overview of the proposed framework is presented in Fig. 2. By transferring semantics of media tags (*i.e.*, image/video-level labels) into different graphlets in an image, we represent each graphlet by a number of low-level and high-level visual features. Thereby, each graphlet can be deemed as a point in the perceptual space. To select those salient graphlets, a geometry-preserving active learning (GPAL) algorithm is utilized based on [9]. GPAL assumes that each graphlet can be linearly reconstructed by its spatially neighboring ones. To encode image/frame geometric properties both locally and globally, the nearer neighbors contribute a greater effect. To solve GPAL, an efficient two-stage iterative scheme is proposed. These detected salient graphlets are then linked into a biological viewing path (BVP) to mimick human gaze shifting. Based on the BVP, the media quality is predicted by probabilistically combining the BVPs learned from multiple observers. More specifically, we use GPAL to predict BVPs from both the training media recorded from multiple observers and the test media. Afterward, the media quality is quantified by the amount of BVPs that can be transferred from the training media data into a test one based on a probabilistic model.

2. RELATED WORK

Our approach assesses video quality by combining the quality scores of its constituent frames. Therefore, it is closely related to two topics in multimedia: image quality measurements based on human perception and human gaze estimation from an image.

2.1 Perceptual Image Quality Measure

Recently many image quality evaluation methods have been proposed, aiming at simulating how humans perceive an image. Ke *et al.* [6] developed a group of high-level visual features, such as the image simplicity based on the spatial distribution of edges, to imitate human perception of photo quality. Datta *et al.* [4] proposed 58 low-level visual features, *e.g.*, shape convexity, to capture photo quality. Dhar *et al.* [5] proposed a set of high-level attribute predictors to evaluate photo aesthetic quality. In [8], Luo *et al.* evaluated photo quality by utilizing a GMM-based hue distribution and a prominent line-based texture distribution. In [58], Cheng *et al.* proposed the omni-range context, *i.e.*, the spatial distribution of arbitrary pairwise image patches, to model photo aesthetic quality. Nishiyama *et al.* [10] assesses image quality by combining the SVM classifiers corresponding to a photo's internal subject regions. In [15], Nishiyama *et al.* proposed a color harmony-based photo quality model. The patch-level color distribution is converted into a bag-of-patches histogram, which is then classified by an SVM to identify photo quality. Luo *et al.* [29] attempted to use both photo-based aesthetic features and motion features for evaluating video quality. Moorthy *et al.* [30] introduced an approach to identify video quality based on 1) image-quality features, 2) motion features, and 3) single-photo aesthetic features. Yeh *et al.* [32] proposed to model video aesthetics by two modules: aesthetic features construction and temporal integration. Zhang *et al.* [7] learn human gaze shifting paths from image sub-regions to evaluate the quality of a candidate cropped photo. This model, however, discovers salient regions only in the semantic space. Comparatively, our approach detects salient regions in the perceptual space, reflecting both the low-level and high-level visual cues.

2.2 Human Gaze Prediction Techniques

The current gaze estimation can be categorized into two groups: model-based and appearance-based methods. Model-based methods use 3D eyeball models and estimate gaze direction using the geometric eye features [17, 44, 47]. They typically use infrared light sources and a high-resolution camera to locate the 3D eyeball position. Although this approach can estimate gaze directions accurately, its heavy reliance on the specialized hardware is a limitation. There exist methods relaxing this requirement by adopting only eye images to calculate the line of sight from the iris contour [42]. However, this is only effective within a short distance where high-resolution observations are available.

Appearance-based methods compute the non-geometric image features from the input eye image and then estimate gaze directions. The eye position is pre-computed for estimating the gaze target in the world coordinate system. With the popularity of monocular head pose tracking [49] and RGB-D head-tracking cameras [50], head poses can be captured accurately. Some appearance-based gaze estimation techniques use head poses as an auxiliary input for gaze estimation [51, 52]. Appearance variation of the eye images caused by head pose changes is another challenge. Usually they are tackled by a compensation function [51] or warping training images to new head poses [52]. While most of the existing appearance-based models adopt a person-dependent data set, Funes *et al.* proposed a cross subject training method for gaze estimation [53]. An RGB-D camera warps the training and test images to the frontal view. Then, an adaptive linear regression function is applied to compute gaze directions.

3. THE PROPOSED APPROACH

3.1 Perceptual Space Construction

3.1.1 The Concept of Graphlets

There are usually a number of of components within a photo. Among them, a few spatially neighboring ones and their spatial interactions capture the local features of a photo. Since graph is a powerful tool to describe the relationships among objects, we use it to model the spatial interactions of components in a photo. Particularly, we segment a photo into a set of atomic regions[2], and then construct graphlets to capture the local features of this photo. Formally, a graphlet is a small-sized graph defined as:

$$G = (V, E), \tag{1}$$

where V is a set of vertices representing those spatially neighboring atomic regions; and E is a set of edges, each of which connects pairwise spatially adjacent atomic regions. We call a graphlet with t vertices a t-sized graphlet. It is worth emphasizing that the number of graphlets within a photo is exponentially increasing with the graphlet size. Therefore, only small graphlets (*i.e.*, vertex number less than 10) are employed.

In this work, we characterize each graphlet in both color and texture channels. Given a t-sized graphlet, each row of matrix $\mathbf{M}_r^c$ represents the 9-dimensional color moment [2] and each row of matrix $\mathbf{M}_r^t$ denotes the 128-dimensional HOG [3] of an atomic region. To describe the spatial interactions of atomic regions, we employ a $t \times t$ adjacency matrix as:

$$\mathbf{M}_s(i,j) = \begin{cases} \theta(R_i, R_j) & \text{if } R_i \text{ and } R_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $\theta(R_i, R_j)$ is the horizontal angle of the vector from the centroid of atomic region R_i to that of atomic region R_j. Based on the three matrices $\mathbf{M}_r^c$, $\mathbf{M}_r^t$, and $\mathbf{M}_s$, we can describe a graphlet by $\mathbf{M} = [\mathbf{M}_r^c, \mathbf{M}_r^t, \mathbf{M}_s]$.

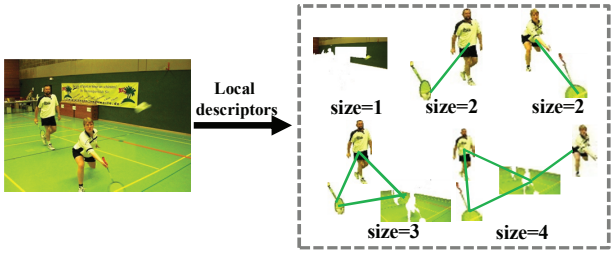


Figure 3: An example of differently sized graphlets

[2]The atomic regions are superpixels segmented using SLIC [46].

3.1.2 Semantically Encoding Graphlets

In addition to the color and texture channels description, high-level semantic cues should also be exploited for predicting human gaze shifting paths. In this paper, the semantic cues are integrated based on a weakly supervised algorithm. We transfer the semantics of image/video-level labels into different graphlets in an image/frame[3]. Particularly, the weakly supervised algorithm is implemented based on a manifold embedding described as:

$$\begin{aligned} &\arg\min_{\mathbf{Y}} [\sum_{i,j} ||y_i - y_j||^2 l_s(i,j) - \sum_{i,j} ||y_i - y_j||^2 l_d(i,j)] \\ &= \arg\min_{\mathbf{Y}} \mathrm{tr}(\mathbf{Y}\mathbf{R}\mathbf{Y}^T), \end{aligned} \tag{3}$$

where $\mathbf{Y} = [y_1, y_2, \cdots, y_n]$ contains a collection of post-embedding graphlets; $\mathbf{R} = [\vec{\mathbf{e}}_{n-1}^T, -\mathbf{I}_{n-1}]\mathbf{W}_1[\vec{\mathbf{e}}_{n-1}^T, -\mathbf{I}_{n-1}] + \cdots + [-\mathbf{I}_{n-1}, \vec{\mathbf{e}}_{n-1}^T]\mathbf{W}_n[-\mathbf{I}_{n-1}, \vec{\mathbf{e}}_{n-1}^T]$; $\mathbf{W}_i$ is an $n \times n$ diagonal matrix whose h-th diagonal element is $[l_s(h,i) - l_d(h,i)]$.

$l_s(\cdot,\cdot)$ and $l_d(\cdot,\cdot)$ are functions measuring the semantic similarity and difference between graphlets respectively. Denoting $\mathbf{b}_i$ as a C-dimensional row vector containing the multiple labels of the media from which graphlet G_i is extracted; and $\vec{\mathbf{n}} = [n^1, n^2 \cdots, n^C]^T$ where n^c is the number of images (or video clips) with label c, then l_s and l_d are defined as:

$$l_s(i,j) = \frac{[\mathbf{b}_i \cap \mathbf{b}_j]\vec{\mathbf{n}}}{\sum_c n^c}, \tag{4}$$

$$l_d(i,j) = \frac{[\mathbf{b}_i \oplus \mathbf{b}_j]\vec{\mathbf{n}}}{\sum_c n^c}, \tag{5}$$

Based on the semantically aware embedding, we project graphlets onto the perceptual space, where each graphlet is described by $[\phi(\mathbf{M}_r^c), \phi(\mathbf{M}_r^t), y]$. $\phi(\cdot)$ is a row-wise matrix stacking operator. $\phi(\mathbf{M}_r^c)$, $\phi(\mathbf{M}_r^t)$, and y are three vectors describing each graphlet in color, texture, and semantic channels, respectively. For ease of expression, we again use $\mathbf{Y} = \{y_1, y_2, \cdots, y_n\}$ to represent the n graphlets in the perceptual space.

3.2 Biological Viewing Path Learning

Our proposed salient graphlets discovery is motivated by the transductive experimental design (TED) proposed by Yu *et al.* [18]. The key idea is to minimize the prediction variance of salient graphlets by a regularized linear regression function. In a geometrical view, it is equivalent to find

[3]Note that image/video-level labels are cheaply available nowadays. For example, many Flickr images and Youtube videos are associated with semantic tags. Also, image and video labels can be accurately predicted by an existing classification model, such as the spatial pyramid matching (SPM) [59].

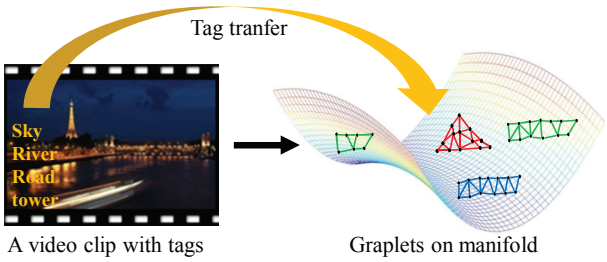


Figure 4: Projecting graphlets onto manifold by semantically encoding image/video-level labels

m salient (representative) graphlets $\mathbf{Z} = \{z_1, z_2, \cdots, z_m\} \in \mathbf{Y}$ in the perceptual space to maximally retain the information of $\mathbf{Y}$. This objective can be formulated as:

$$\min_{\mathbf{Z},\mathbf{A}} \sum_{i=1}^{n} (||x_i - \mathbf{Z}a_i||_2^2 + \alpha||a_i||_2^2)$$
$$s.t. \quad \mathbf{Z} = [z_1, z_2, \cdots, z_m] \in \mathbf{Y},$$
$$\mathbf{A} = [a_1, a_2, \cdots, a_n] \in \mathbb{R}^{m \times n}, \tag{6}$$

where α is a regularization parameter controlling the speed of shrinkage. To solve this problem, a sequential greedy strategy is proposed by Yu *et al.* [18].

3.2.1 *Mathematical Formulation of GPAL*

Obviously, TED reconstructs each graphlet via a linear combination of all the selected salient ones. Thus, it is reasonable to approximate a graphlet y_i by the linear combination of those salient graphlets, where the local and global geometric properties are preserved simultaneously as shown in Fig. 5. For any selected salient graphlet $z_j \in \mathbf{Z}$, we denote function $d(z_j, y_i)$ as the distance between z_j and y_i, where $d(\cdot,\cdot)$ can be any distance such as the geodesic distance. Intuitively, the smaller $d(z_j, y_i)$ is, the greater effect z_j will have for the reconstruction of y_i.

Motivated by this, we propose a new method called

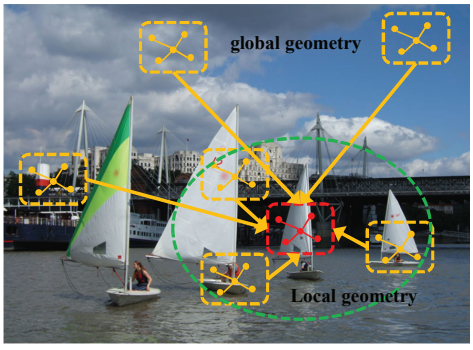


Figure 5: An illustration of the geometry-preserving mechanism in GPAL. The distance between a salient graphlet (red) and the rest non-salient ones (yellow) are well preserved in the salient graphlets discovery.

Geometry-Preserving Active Learning (GPAL) to select a few salient graphlets from an image/frame. For each graphlet y_i in an image/frame, we assume that the reconstruction is built upon the remaining graphlets. By penalizing the coefficients of the reconstruction, we formulate an objective function as:

$$\min_{\mathbf{Z},\mathbf{A}} \sum_{i=1}^{n} (||y_i - \mathbf{Z}a_i||_2^2 + \mu||a_{ji}||d(z_j, y_i))$$
$$s.t. \quad \mathbf{Z} = [z_1, \cdots, z_m] \in \mathbf{Y},$$
$$\mathbf{A} = [a_1, \cdots, a_n] \in \mathbb{R}^{m \times n}, \tag{7}$$

where a_{ji} is the j-th element of vector a_i and μ is a regularization parameter. In this objective function, the first term $||y_i - \mathbf{Z}a_i||_2^2$ means that y_i should be close to its approximation $\mathbf{Z}a_i$. The second term $\sum_{j=1}^{n} |a_{ji}|d(z_j, y_i)$ reflects that salient graphlets closer to y_i contribute more than those distant ones. This term encodes both the local and global geometric properties implicitly. The solution of the above algorithm is based on [9]'s theory.

The convergence can be guaranteed by the block-wise coordinate descent scheme. After obtaining the m salient graphlets, we link them into a biological viewing path (BVP). The first BVP vertex denotes the most salient graphlet, the second vertex represents the second most salient one, and so on.

3.3 Probabilistic Media Quality Measure

The learned BVPs capture the local and global geometric properties of an image/frame, at both the low-level and high-level. To effectively integrate them for measuring the quality of an image or a video clip, a probabilistic model is proposed.

As shown in Fig. 6, given a set of training images[4] $\{I^1, I^2, \cdots, I^L\}$ and a test media, they are highly correlated through their respective BVPs $\mathcal{P}$ and $\mathcal{P}^*$. The probabilistic model contains two types of nodes: observable nodes (green rectangles) and hidden nodes (blue rectangles). These two types of nodes form four layers. The first layer corresponds to all the training images with high quality; the second layer denotes all the BVPs learned from training media; the third layer represents all the BVPs learned from test media; and the last layer denotes the test media (I_*: an image, or $\{I_*^1, I_*^2, \cdots, I_*^F\}$: a video clip with F frames). The correlation between the first and the second layers is $p(\mathcal{P}|I^1, I^2, \cdots, I^L)$. The correlation between the second and the third layers is $p(\mathcal{P}_*|\mathcal{P})$. The correlation between the third and the fourth layers is $p(I_*|\mathcal{P}_*)$ (an image) or $p(I_*^1, I_*^2, \cdots, I_*^F|\mathcal{P}_*)$ (a video clip).

Intuitively, media quality can be quantified as the number of BVPs can be probabilistically transferred from the training media into the test one. Thus, the quality of a test video clip[5] can be formulated as a posterior probability as:

$$\begin{aligned}\gamma &= p(I_*^1, I_*^1, \cdots, I_*^F|I_*^1, I_*^1, \cdots, I_*^L) \\ &= p(I_*^1, I_*^1, \cdots, I_*^F|\mathcal{P}_*) \cdot p(\mathcal{P}_*|\mathcal{P}) \cdot p(\mathcal{P}|I^1, I^2, \cdots, I^L) \\ &= p(\mathcal{P}_*|\mathcal{P}),\end{aligned} \tag{8}$$

where the probability $p(\mathcal{P}_*|\mathcal{P})$ is calculated as:

$$\begin{aligned}p(\mathcal{P}_*|\mathcal{P}) &= \prod_{i=1}^{F} \prod_{j=1}^{L} p(P_*^i|P^j) \\ &= \prod_{i=1}^{F} \prod_{j=1}^{L} \prod_{k=1}^{K} p(G_*^i(k)|G^j(k)),\end{aligned} \tag{9}$$

where P_*^i and P^j denote the i-th and j-th BVP from the training and test media respectively; $G_*^i(k)$ and $G^i(k)$ are the k-th graphlet in BVP P_*^i and P^j respectively.

[4] A video clip can be treated as a sequence of frames.
[5] We simply set $F = 1$ for predicting the quality of an image.

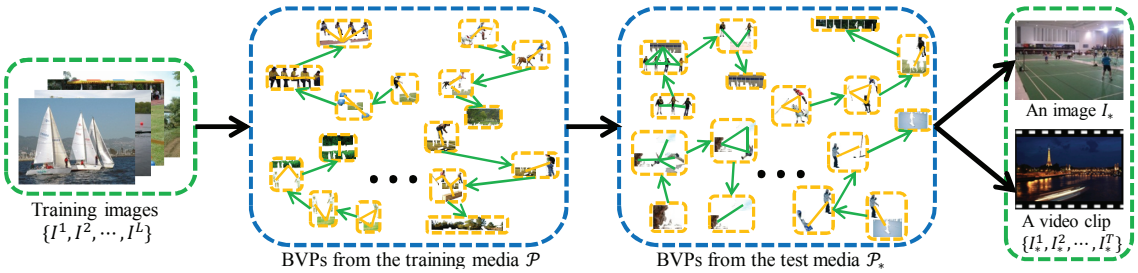


Figure 6: An illustration of the quality model by probabilistically transferring BVPs

Following many algorithms such as [20], we define the similarity between graphlets as a Gaussian kernel:

$$p(G|G') \propto \exp\left(-\frac{||y(G) - y(G')||^2}{2\sigma^2}\right), \qquad (10)$$

where $y(G)$ is the vector corresponding to graphlet G in the perceptual space, as elaborated in Sec 3.1.

4. EXPERIMENTS

This section evaluates the effectiveness of our biological inspired quality model based on four experiments. The first experiment compares our predicted BVP with the previous computational saliency models, as well as the real human gaze shifting paths. The second experiment compares our method with well-known quality models. The third part conducts user studies to evaluate our method. Finally, we analyze the influence of different parameters.

All the experiments were conducted on a PC equipped with an Intel X5482 CPU and 8GB RAM. The algorithms were implemented on the Matlab 2012 platform.

4.1 Effectiveness of the BVP

As the key component of our quality model, it is important to assert whether the learned BVPs can capture human visual saliency accurately. To the best of our knowledge, there are only a few public saliency data sets containing images with ground-truth human fixations. We experiment on the NUSEF [54] and the recent OSIE [1] data sets since they contain appropriate numbers of different semantic objects. The NUSEF consists of 758 images with 75 different objects, and the ground-truth human fixations are also provided. The OSIE is comprised of 700 images with eye-tracking data recorded from 15 viewers. The accompanied annotation data consists of 5,551 segmented objects with fine contours and 12 types of semantics. As the NUSEF and OSIE contain only hundreds of images, it is difficult to affirm our approach on them comprehensively. We thus collect 4,819 images from the PASCAL VOC series [55] and the Lotus Hill (LHI) data set [56]. For each image, we manually assign multiple semantic tags and collect the human fixation from five volunteers.

In our first experiment, we compare our learned BVP with ten well-known visual saliency models, including five low-level feature-based saliency models: Itti *et al.*'s saliency model [28], graph-based visual saliency by Harel *et al.* [26], image signature by Hou *et al.* [27], information-theory-based saliency model by Burce *et al.* [37], and hypercomplex fourier transform (HFT) saliency model by Li *et al.* [35]; as well as five high-level feature-based saliency models proposed by Judd *et al.* [33], Coferman *et al.* [25], Yang *et al.* [22], Yan *et al.* [23], and Zhang *et al.* [24] respectively. Toward a fair comparison, we have to convert our BVP into a saliency map. Specifically, we extract $m = 5$ salient graphlet. Then, we learn a Gaussian mixture model to describe the BVP distribution. Lastly, we calculate the probability of each graphlet $p(G_i)$, and then the saliency of the i-th pixel ρ is:

$$s(\rho_i) = \max_{G \supset \rho_i} p(G), \qquad (11)$$

where $G \supset \rho_i$ is the set of graphlets containing pixel ρ_i, and the "max" term mimicks the "winner-taken-all" mechanism in biological vision [16].

The AUC (area under the curve) scores of the NUSEF, the OSIE, and our own compiled data sets are presented in Table 2. Accordingly, the three ROC curves are shown in Fig. 7. To decide the threshold of saliency maps, for the training images, we use the Dice similarity coefficient to evaluate the overlap between the thresholded saliency map and the ground truth, wherein the peak value corresponds to the optimal threshold. Pixels whose saliency values falls below this threshold are deemed as non-salient, and vice versa. As shown in Table 2 and Fig. 7, the best performance is always achieved by our proposed BVP.

4.2 Comparison with the State-of-the-Art

4.2.1 *Image Quality Evaluation*

To the best of our knowledge, there exist three data sets for evaluating photo quality: the CUHK [6], the Photo.net [4], and the AVA [38]. A high-level description of the three data sets is as follows: 1) the CUHK [6] contains 12,000 photos collected from DPChallenge.com. We use a standard split of training/test sets on this data set; 2) the Photo.net consists of 3,581 images. Only URLs of the original photos are provided. Nearly half of the images have been removed from the websites, leaving only about 1,700 images available. They are randomly split into equal partitions, one for training and the rest for testing; and 3) the AVA [38] data set contains 25,000 highly- and low-aesthetic photos, each of which is associated with two semantic tags. The training and test photos of the AVA data set are pre-specified.

The first experiment compares our approach with five photo perceptual quality methods, including three global feature-based approaches by Dhar *et al.* [5], Luo *et al.* [8], and Marchesotti *et al.* [11], respectively; and two local patch integration-based methods by Cheng *et al.* [58] and Nishiyama *et al.* [15], respectively.

The source codes of the above five compared methods are not provided and some implementation details are ob-

Table 1: Comparative AUC Scores on the NUSEF, the OSIE, and our own compiled data set

Data set	Gofer. *et al*	Hou *et al*	GBVS	Judd *et al*	Yang *et al*	Itti *et al.*	Harrel *et al.*	Zhang *et al.*	Yan *et al.*	Li *et al.*	PM
NUSEF	57.64%	53.54%	50.37%	50.43%	53.19%	52.76%	52.11%	53.12%	54.22%	56.32%	**64.48%**
OSIE	67.65%	61.43%	58.89%	56.77%	53.37%	56.69%	54.42%	57.69%	55.64%	54.49%	**65.93%**
Ours	61.43%%	55.47%	56.32%	55.98%	54.12%	59.45%	56.64%	52.27%	57.76%	59.45%	**67.11%**

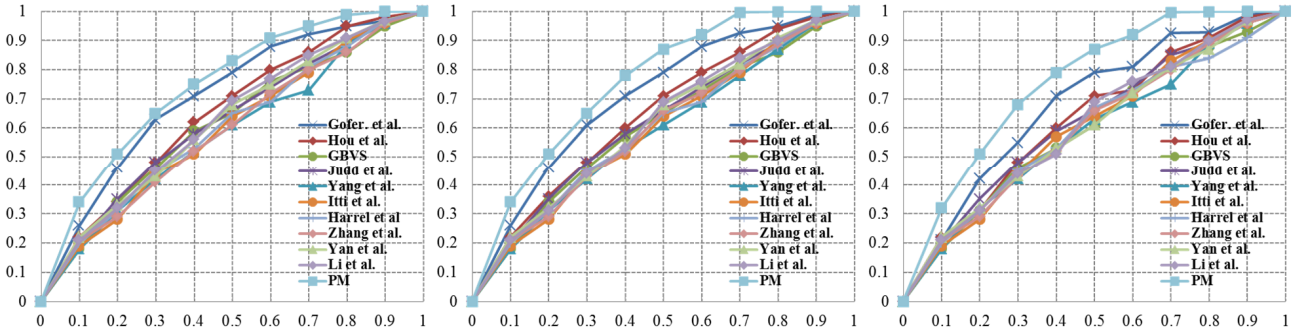


Figure 7: The ROC curves on the NUSEF (left), the OSIE (middle), and our own compiled data sets

Table 2: Comparison of aesthetics prediction accuracies

	CUHK	PNE	AVA
Dhar *et al.*	0.7386	0.6754	0.6435
Luo *et al.*	0.8004	0.7213	0.6879
Marchesotti *et al.* (FV-Color-SP)	0.8767	0.8114	0.7891
Cheng *et al.*	0.8432	0.7754	0.8121
Nishiyama *et al.*	0.7745	0.7341	0.7659
The proposed method	**0.9103**	**0.8595**	**0.8531**

scure. Therefore, it is difficult to strictly implement them. Our experiment adopts the following setups. For Dhar's approach, we use the public codes of Li *et al.* [12] to extract the attributes from each photo. These attributes are combined with the low-level features proposed by Yeh *et al.* [13] to train the quality classifier. For Luo *et al.*'s approach, not only the low-level and high-level features in their publication are implemented, but also the six global features from Getlter *et al.* [14]'s work are used to strengthen the quality prediction. For Marchesotti *et al.*'s approach, similar to the implementation of Luo *et al.*'s method, the six additional features are also employed. Cheng *et al.*'s approach is implemented by adopting only 2-sized graphlets for aesthetic quality measure. Noticeably, for the probabilistic model-based quality evaluation methods (*i.e.*, Cheng *et al.*'s method, Nishiyama *et al.*'s method, and our model), if the score is larger than 0.5, then an image is considered as a high quality one, and vice versa.

We present the aesthetics prediction accuracies on the CUHK, the PNE, and the AVA in Table 2. On the three data sets, our approach outperforms Marchesotti *et al.*'s method by nearly 4%, and exceeds the rest of the compared methods by over 6%, which demonstrates the advantage of our approach.

4.2.2 *Video Quality Evaluation*

The only publicly available benchmark for perceptual video quality evaluation is the Telefonica data set [30]. It contains 160 rated videos crawled from YouTube and is grouped into 16 categories. The categories include "baby laughing", "sky diving", and so on. Each video is cropped into a 15-seconds clip to reduce the potential biases of video length. The rating values of the Telefonica range from -2 to 2. If the quality score is above 0, then the video is deemed as high quality, and vice versa.

In the experiment, we compare our approach with three

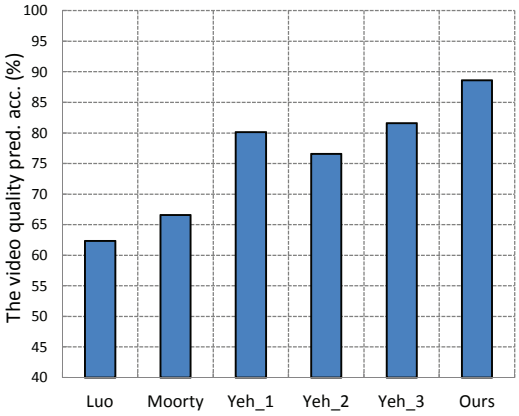


Figure 8: Performance comparison between our method and the other video quality models. Yeh_1, Yeh_2, and Yeh_3 denote the three best features selected in [32].

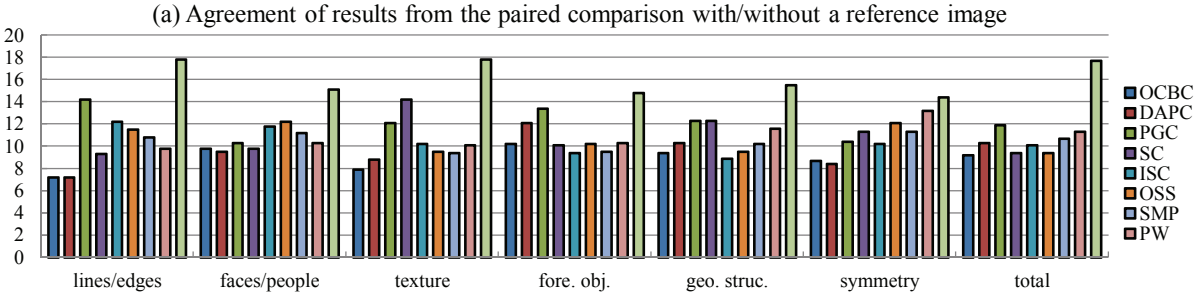
video quality models proposed by Luo *et al.* [29], Moorthy *et al.* [30], and Yeh *et al.* [32] respectively. We use half of the videos from Telefonica for training while leaving the rest for testing. The experiments were repeated five times and the average accuracies are presented in Fig. 8. Our approach achieves the best performance, which confirms the necessity of exploiting human gaze shifting paths in video quality prediction.

4.3 User Study

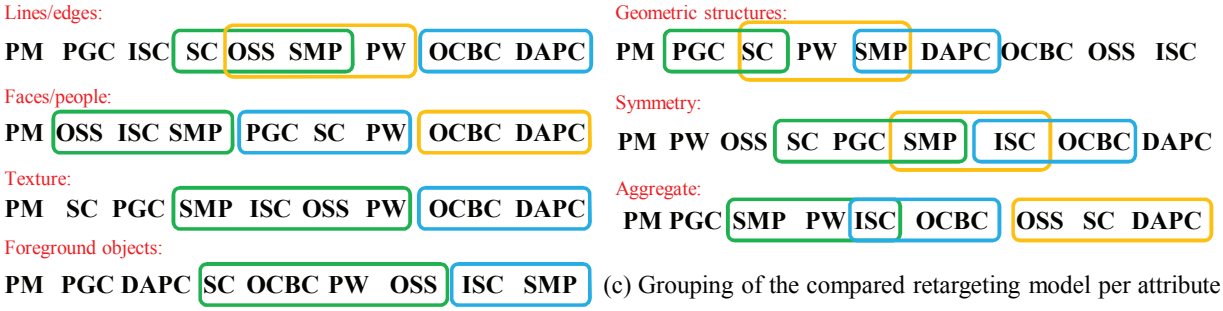
This experiment introduces subjective analysis. We apply different quality models for photo retargeting on the AVA data set [38], and then employ 30 volunteers to appraise the retargeted photos from different aspects. Specifically, as shown in Fig. 9, we learn the distribution of BVPs from the

	Lines/ edges	Faces/ people	texture	Fore. objects	Geometric structure	Symmetry	Aggregate
μ (with ref.)	0.103	0.191	0.193	0.154	0.123	0.212	0.111
μ (without ref.)	0.924	0.181	0.191	0.144	0.111	0.201	0.994

(a) Agreement of results from the paired comparison with/without a reference image



(b) The percentage of votes and total ranking of the nine compared method per attribute



(c) Grouping of the compared retargeting model per attribute

Figure 10: A detailed subjective analysis of the comparative retargeted photos on the AVA data set

well-aesthetic training photos on the AVA [38], afterward the learned distribution guides the shrinking of 50 randomly selected photos. The baseline retargeting algorithms are from the survey article [36], including three cropping methods: omni-range context-based cropping (OCBC), probabilistic graphlet-based cropping (PGC), describable attribute for photo cropping (DAPC), and five content-aware retargeting algorithms: seam carving (SC) and its improved version (ISC), optimized scale-and sketch (OSS), saliency-based mesh parametrization (SMP), and the patch-based wrapping (PW).

The results in Fig. 10 are based on the subjective evaluation of retargeting algorithms in [57]. First, we evaluate the degree of agreement when the volunteers vote for their favorite retargeted photos, where a high disagreement reflects the difficulty in decision making. In our experiment, we use the coefficient of agreement defined by Kendall and Babington-Smith [34]. The coefficients over all the four retargeted photos are shown in the last column of Fig. 10(a). Besides, we also collect the volunteers' votes on each attribute of the 50 sets of retargeted photos. As shown from the second column to the seventh column in Fig. 10(a), the volunteers are highly agreeable on the face/people, the texture, and the symmetry because these attributes are well preserved. Then, we present the votes on each attribute based on the four retargeted photos. As shown in Fig. 10(b), the proposed method consistently receives the most votes on all the attributes. Also, the probabilistic graphlet cropping proposed by Zhang *et al.* [31] performs competitively on several attributes. Based on the above votes, we rank all the compared retargeting algorithms as shown in Fig. 10(c), where the algorithms within a rectangle are statistically indistinguishable with respect to volunteer preferences. This again demonstrates the competitiveness of our model.

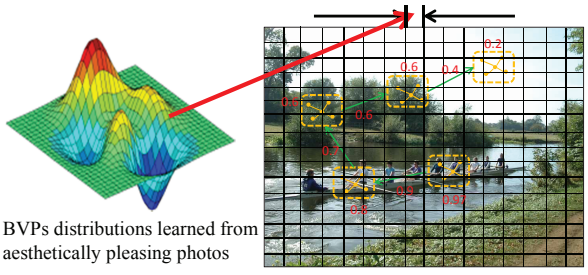


Figure 9: A illustration of photo retargeting based on the proposed BVP

Next, we compare the proposed BVPs with real human gaze shifting paths quantitatively and qualitatively. We record the eye fixations from five observers by using EyeLink II[6], and then connect the fixations into a path sequentially. As can be seen from Fig. 11, in most images the proposed BVPs are consistent with human gaze shifting paths. Furthermore, the average proportion of overlap between the human gaze shifting path and the BVP is 90.89%. This shows that the proposed BVP can accurately predict the human gaze shifting process, leading to an excellent media quality prediction.

[6]http://www.sr-research.com/EL II.html

Figure 11: Comparison between the learned BVPs (yellow paths) and real human gaze shifting paths recorded from five observers. The vertices in each BVP indicate the centroid of the BVP's constituent graphlets.

4.4 Influence of Parameter Settings

The final experiment evaluates the influence of the three important parameters in our quality model: the graphlet number in a learned BVP m, and μ, λ in the GPAL optimization.

We first evaluate the quality prediction by varying m. As shown in Fig. 12, on all the four data sets, the quality prediction accuracy improves significantly when m increases from 1 to 4. Afterward, the accuracy fluctuates when m is tuned from 4 to 15. This is because humans typically fix on only $3 \sim 5$ objects when they view an image (free viewing without specific search tasks). Therefore, we set m to 4 in our model.

Then, we evaluate the performance under different values of μ, the regularization parameter. We tune the value of μ from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$. As shown on the left of Fig. 13, on all the four data sets, the best accuracies were achieved when $\mu = 0.01$. Neither a too heavy nor a very slight penalty on the geometry preservation term yields a good quality prediction. Next, we evaluate the performance of our model by tuning the value of λ, a regularization parameter controlling the sparsity of $\hat{a}$. Similarly, we tune the value of λ from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$. As shown on the right of Fig. 13, the best accuracies were achieved when $\lambda = 0.1$ on all the four data sets.

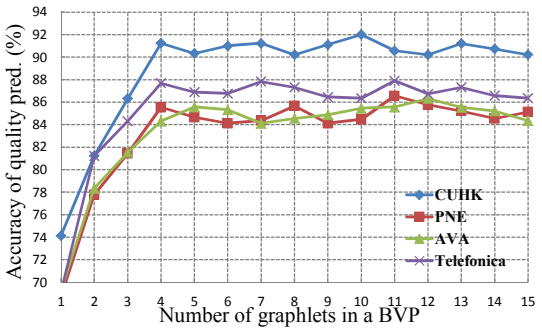


Figure 12: The quality prediction accuracies by tuning m on the four data sets

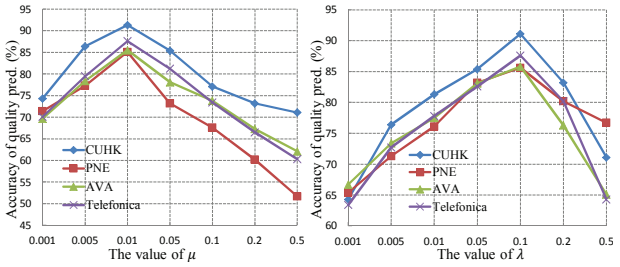


Figure 13: The quality prediction accuracies by tuning μ and λ respectively

5. CONCLUSIONS

This paper presents a perceptually aware quality model by mimicking how humans perceive low-level and high-level visual features from media. We first projected graphlets onto a pre-defined perceptual space. Then, an active learning algorithm GPAL is utilized to select graphlets both visually and semantically salient, and image/frame geometric attributes can be preserved optimally. Finally, these discovered graphlets are linked into a BVP, which is further integrated into a probabilistic model for predicting media quality. We also demonstrate that applications such as photo retargeting can be enhanced significantly by the learned BVPs.

6. ACKNOWLEDGMENTS

Thanks for the constructive comments of the three reviewers. Dr. Yingjie Xia is the correspondence author of this article.

7. REFERENCES

[1] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, Qi Zhao, Predicting Human Gaze beyond Pixels, *Journal of Vision*, 14(1): 28, pages: 1–20 2014.

[2] Markus Stricker, Markus Orengo, Similarity of Color Images, *Storage and Retrieval of Image and Video Databases*, 1995.

[3] Navneet Dalal, Bill Triggs, Histograms of Oriented Gradients for Human Detection, *in Proc. of CVPR*, 2005.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang, Studying Aesthetics in Photographic Images using a Computational Approach, *in Proceedings of ECCV*, pages: 288–301, 2006.

[5] Sagnik Dhar, Vicente Ordonez, Tamara L. Berg, High Level Describable Attributes for Predicting Aesthetics and Interestingness, *in Proceedings of CVPR*, pages: 1657–1664. 2011.

[6] Yan Ke, Xiaoou Tang, Feng Jing, The Design of High-level Features for Photo Quality Assessment, *in Proceedings of CVPR*, pages: 419–426, 2006.

[7] Luming Zhang, Yue Gao, Rongrong Ji, Qionghai Dai, Xuelong Li, Actively Learning Human Gaze Shifting Paths for Photo Cropping, *IEEE T-IP*, 21(5), pages: 2235–2245, 2014.

[8] Wei Luo, Xiaogang Wang, Xiaoou Tang, Content-based Photo Quality Assessment, *in Proceedings of ICCV*, pages: 2206–2213. 2011.

[9] Yao Hu, Debing Zhang, Zhongming Jin, Deng Cai, Xiaofei He, Active Learning Based on Local Representation, *IJCAI*, 2013.

[10] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, Imari Sato, Sensation-based Photo Cropping, *in Proceedings of ACM Multimedia*, pages: 669–672, 2009.

[11] Luca Marchesotti, Florent Perronnin, Diane Larlus, Gabriela Csurka, Assessing the Aesthetic Quality of Photographs using Generic Image Descriptors, *in Proceedings of ICCV*, pages: 1784–1791, 2011.

[12] Fei-Fei Li, Pietro Perona, A Bayesian Hierarchical Model for Learning Natural Scene Categories, *in Proceedings of CVPR*, pages: 524–531, 2005.

[13] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, Ming Ouhyoung, Personalized Photograph Ranking and Selection System, *in Proceedings of ACM Multimedia*, pages: 211–220, 2010.

[14] Peter Gehler, Sebastian Nowozin, On Feature Combination for Multiclass Object Classification, *in Proceedings of ICCV*, pages: 221–228, 2009.

[15] Masashi Nishiyama, Takahiro Okabe1, Imari Sato, Yoichi Sato, Aesthetic Quality Classification of Photographs based on Color Harmony, *in Proceedings of CVPR*, pages: 33–40, 2011.

[16] C. Koch, S. Ullman, Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry, *Human Neurobiology*, 4, pages: 219–227, 1985.

[17] Elias Daniel Guestrin, Moshe Eizenman, General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections, *IEEE T-BE*, 53(6), pages: 1124–1133, 2006.

[18] Kai Yu, Jinbo Bi, Volker Tresp, Active Learning via Transductive Experimental Design, *in Proc. of ICML*, 2006.

[19] Yurii Nesterov, Gradient methods for minimizing composite objective function, *Technical Report*, 2007.

[20] Mingli Song, Dacheng Tao, Chun Chen, Jiajun Bu, Jiebo Luo, Chengqi Zhang, Probabilistic Exposure Fusion, *IEEE T-IP*, 21(1): 341–357, 2012.

[21] Honglak Lee, Alexis Battle, Rajat Raina, Andrew Y. Ng, Efficient Sparse Coding Algorithms, *in Proc. of NIPS*, 1996.

[22] Jimei Yang, Ming-Hsuan Yang, Top-down Visual Saliency via Joint CRF and Dictionary Learning, *in Proc. of CVPR*, pages: 2296–2303, 2012.

[23] Qiong Yan, Li Xu,Jianping Shi, Jiaya Jia, Hierarchical Saliency Detection, *in Proc. of CVPR*, pages: 1155–1162, 2013.

[24] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, Garrison W. Cottrell, SUN: A Bayesian Framework for Saliency using Natural Statistics, *Journal of Vision*, 8(7), article 32, 2008.

[25] Stas Goferman, Lihi Zelnik-Manor, Ayellet Tal, Context-Aware Saliency Detection, *in Proc. of CVPR*, pages: 2376–2383, 2010.

[26] Jonathan Harel, Christof Koch, Pietro Perona, Graph-Based Visual Saliency, *in Proc. of NIPS*, pages: 545–552, 2007.

[27] Xiaodi Hou, Jonathan Harel, Christof Koch, Image Signature: Highlighting Sparse Salient Regions, *IEEE T-PAMI*, 34(1), pages: 194–201, 2011.

[28] Laurent Itti, Christof Koch, Ernst Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE T-PAMI*, 20(11), pages: 1254–1259, 1998.

[29] Yiwen Luo, Xiaoou Tang, Photo and Video Quality Evaluation: Focusing on the Subject, *in Proc. of ECCV*, 2008.

[30] Anush K. Moorthy, Pere Obrador, Nuria Oliver, Towards Computational Models of the Visual Aesthetic Appeal of Consumer Videos, *in Proc. of ECCV*, 2010.

[31] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, Chun Chen, Probabilistic Graphlet Transfer for Photo Cropping, *IEEE T-IP*, 22(2), pages: 802–815, 2013.

[32] Hsin-Ho Yeh, Chun-Yu Yang, Ming-Sui Lee, Chu-Song Chen, Video Aesthetic Quality Assessment by Temporal Integration of Photo- and Motion-Based Features, *IEEE T-MM*, 15(8): 1944–1957, 2013.

[33] Tilke Judd, Krista A. Ehinger, Frédo Durand, Antonio Torralba, Learning to Predict Where Humans Look, *in Proc. of ICCV*, pages: 2106–2113, 2009.

[34] Michael Rubinstein, Ariel Shamir, Shai Avidan, Multi-operator Media Retargeting, *ACM TOG*, 28(3), 23, 2009.

[35] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, Hangen He, Visual Saliency Based on Scale-Space Analysis in the Frequency Domain, *IEEE T-PAMI*, 35(4), pages: 996–1010, 2013.

[36] Ariel Shamir, Olga Sorkine, Alexander Hornung, Modern Approaches for Media Retargeting, *SIGGRAPH Asia Courses*, 2012.

[37] Neil D. B. Bruce, John K. Tsotsos, Saliency, Attention, and Visual Search: An Information Theoretic Approach, *Journal of Vision*, 9(3), article 5, 2009.

[38] Naila Murray, Luca Marchesotti, Florent Perronnin, AVA: A Large-Scale Database for Aesthetic Visual Analysis, *in Proceedings of CVPR*, pages: 2408–2415, 2012.

[39] Jixu Chen, Qiang Ji, Probabilistic Gaze Estimation Without Active Personal Calibration, *in Proc. of CVPR*, 2011.

[40] Atsushi Nakazawa, Christian Nitschke, Point of Gaze Estimation through Corneal Surface Reflection in an Active Illumination Environment, *in Proc. of ECCV*, 2012.

[41] Han Liu, Mark Palatucci, Jian Zhang, Blockwise Coordinate Descent Procedures for the Multitask Lasso, with Applications to Neural Semantic Basis Discovery, *in Proc. of ICML*, 2009.

[42] Takahiro Ishikawa, Simon Baker, Iain Matthews, Takeo Kanade, Passive Driver Gaze Tracking with Active Appearance Models, *in Proc. of the 11th World Congress on Intelligent Transportation Systems*, 2004.

[43] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *in Proceedings of CVPR*, 2006.

[44] Jixu Chen, Qiang Ji, Probabilistic Gaze Estimation Without Active Personal Calibration, *in Proc. of CVPR*, 2011.

[45] Peng Zhao, Guilherme Rocha, Bin Yu, The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection, *Annals of Statistics*, 37(6A): 3468–3497, 2009.

[46] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine SÍźsstrunk, SLIC Superpixels Compared to State-of-the-art Superpixel Methods, *IEEE T-PAMI*, 34(11): 2274–2282, 2012.

[47] Atsushi Nakazawa, Christian Nitschke, Point of Gaze Estimation through Corneal Surface Reflection in an Active Illumination Environment, *in Proc. of ECCV*, 2012.

[48] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, Francis Bach, Proximal Methods for Hierarchical Sparse Coding, *in Proc. of ICML*, 2010.

[49] Erik Murphy-Chutorian, Mohan Manubhai Trivedi, Head Pose Estimation in Computer Vision: A Survey, *IEEE T-PAMI*, 31(4), pages: 607–626, 2009.

[50] Qin Cai, David Gallup, Cha Zhang, Zhengyou Zhang, 3D Deformable Face Tracking with a Commodity Depth Camera, *in Proc. of ECCV*, 2010.

[51] Feng Lu, Takahiro Okabe, Yusuke Sugano, Yoichi Sato, A Head Pose-free Approach for Appearance-based Gaze Estimation, *in Proc. of BMVC*, 2011.

[52] Kenneth Alberto Funes Mora, Jean-Marc Odobez, Gaze Estimation from Multimodal Kinect Data, *CVPR Workshop*, 2012.

[53] Kenneth Alberto Funes Mora, Jean-Marc Odobez, Person Independent 3d Gaze Estimation from Remote RGB-D Camera, *in Proc. of ICIP*, 2013.

[54] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, Tat-Seng Chua, An Eye Fixation Database for Saliency Detection in Images, *in Proc. of ECCV*, 2012.

[55] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman, The PASCAL Visual Object Classes (VOC) Challenge, *IJCV*, 88(2), pages: 303–338, 2010.

[56] Benjamin Yao, Xiong Yang, Song-Chun Zhu, Introduction to a Large Scale General Purpose Ground Truth Dataset: Methodology, Annotation Tool, and Benchmarks, *EMMCVPR*, 2007.

[57] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, Ariel Shamir, A Comparative Study of Image Retargeting, *ACM TOG*, 29(5),160, 2010.

[58] Bin Cheng, Bingbing Ni, Shuicheng Yan, Qi Tian, Learning to Photograph, *ACM Multimedia*, 2010.

[59] Jianchao Yang, Kai Yu, Yihong Gong, Thomas S. Huang, Linear Spatial Pyramid Matching using Sparse Coding for Image Classification, *in Proc. of CVPR*, 2009.