Perceptually Guided Photo Retargeting

Yingjie Xia, *Member, IEEE*, Luming Zhang, *Member, IEEE*, Richang Hong, *Member, IEEE*,
Liqiang Nie, Yan Yan, and Ling Shao, *Senior Member, IEEE*

***Abstract*—We propose perceptually guided photo retargeting, which shrinks a photo by simulating a human's process of sequentially perceiving visually/semantically important regions in a photo. In particular, we first project the local features (graphlets in this paper) onto a semantic space, wherein visual cues such as global spatial layout and rough geometric context are exploited. Thereafter, a sparsity-constrained learning algorithm is derived to select semantically representative graphlets of a photo, and the selecting process can be interpreted by a path which simulates how a human actively perceives semantics in a photo. Furthermore, we learn the prior distribution of such active graphlet paths (AGPs) from training photos that are marked as esthetically pleasing by multiple users. The learned priors enforce the corresponding AGP of a retargeted photo to be maximally similar to those from the training photos. On top of the retargeting model, we further design an online learning scheme to incrementally update the model with new photos that are esthetically pleasing. The online update module makes the algorithm less dependent on the number and contents of the initial training data. Experimental results show that: 1) the proposed AGP is over 90% consistent with human gaze shifting path, as verified by the eye-tracking data, and 2) the retargeting algorithm outperforms its competitors significantly, as AGP is more indicative of photo esthetics than conventional saliency maps.**

***Index Terms*—Active graphlet path (AGP), retargeting, semantics, shrink.**

I. Introduction

With the popularity of mobile devices in recent years, photo retargeting has become an indispensable technique to adapt a high resolution photo to a low resolution screen. For example, to design an iPhone wallpaper, people usually adapt an approximately 4000 × 3000-sized photo taken by a digital single lens camera camera to a 640×960-sized iPhone screen. Nonuniform scaling often leads to unsatisfactory results if the aspect ratio of the targeting photo is significantly different from that of the original one. Simple photo cropping does not work when esthetically pleasing regions are dispersely distributed in a photo. Toward an optimal retargeting result, most of the existing works focus on content-aware photo retargeting [1], [13], [16], [18], [22], aiming at maximally preserving visually salient regions while keeping the nonsalient ones to a minimum scale. However, the existing content-aware photo retargeting algorithms suffer from the following two drawbacks.

1) Visually salient regions may not be esthetically pleasing. One competitive saliency model is Li *et al.*'s [12] method, based on which many retargeting algorithms can be built [1], [23]. However, the major limitation of this family of algorithms is that current saliency models focus on low-level features and have limited predictability power of region semantics. The problem is known as the "semantic gap" between the ground truth human data and the existing low-level features-based attention models. One common approach to fill the gap is to add object detectors, but it does not scale well. Moreover, object detectors are not reliable enough. In practice, only a few prespecified categories such as human faces can be accurately detected.
2) Eye tracking experiments show that humans gaze at important regions in a sequential manner [6]. Existing fully automatic retargeting methods fail to encode such a gaze shifting sequence, i.e., the path linking different graphlets.

To address the above problems, we propose perceptually guided photo retargeting, which shrinks each region in a photo based on its semantic significance and in a sequential manner. Our approach contains four main components. To describe the local compositions of a photo, we first extract a number of graphlets (small-sized connected subgraphs) and accordingly project them onto a low-dimensional semantic space, by exploiting the manifold structure of the graphlets. The semantic space is built upon the image-level semantics which is obtained using existing image retrieval models. Second, inspired by biological vision where

Manuscript received September 6, 2015; revised November 26, 2015; accepted January 14, 2016. Date of publication April 22, 2016; date of current version February 14, 2017. This work was supported in part by the National Basic Research Program under Grant 2013CB336500, in part by the National High-Tech Development Program under Grant 2014AA015104, in part by the Program for New Century Excellent Talents under Grant NCET-13-0764, in part by the Project from the National Nature Science Foundation of China under Grant 61472116, Grant 61572169, and Grant 61472113, in part by the Anhui Fund for Distinguished Young Scholars under Grant 1508085J04, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ13F020004 and Grant LR14F020003. This paper was recommended by Associate Editor D. Wang.

Y. Xia is with the College of Computer Sciences, Zhejiang University, Hangzhou 310027, China.

L. Zhang and R. Hong are with the Department of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: zglumg@zju.edu.cn).

L. Nie is with the School of Computer Science and Technology, Shandong University, Shandong 250100, China.

Y. Yan is with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy.

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a PDF file that contains additional graphics not included within the paper itself. The total size of the file is 2.32 MB.

humans actively deploy their attention to semantically important regions, a sparsity-constrained algorithm is designed to select several semantically important graphlets from a photo. Intuitively, this process can be formulated as a path, termed active graphlet path (AGP), wherein each directed-edge links pairwise graphlets. Then, to leverage the knowledge from a pool of experienced photographers, a probabilistic model is developed to learn the distribution of AGP from a set of esthetically pleasing training photos, which are used to further shrink a test photo. Particularly, we divide a test photo into a collection of grids; the learned priors enforce the grids covered with more semantically significant graphlets/directed-edges to shrink less and vice versa. Finally, as the proposed model transfers the AGPs from the esthetically pleasing photos into the retargeted one, its performance depends on the number and diversity of training photos. In addition, the proposed retargeting is an AGP-transferring framework, and thus its performance depends on the training data. Toward a diverse collection of training photos, an online learning scheme is developed to incrementally update our model by encoding an increasing number of esthetically pleasing photos.

The rest of this paper is organized as follows. Section II briefly reviews previous retargeting models. Sections III–VI introduce the proposed retargeting model, including manifold graphlet embedding, AGP construction, probabilistically retargeting, and online retargeting model updating. Experimental results in Section VII thoroughly demonstrate the effectiveness and the efficiency of the new model. Section VIII concludes and suggests some future work.

II. Related Work

Content-aware photo retargeting can be roughly categorized into discrete and continuous retargeting. There are many retargeting methods and we discuss the most relevant ones here. The reader can refer to [42] and [44]–[46] for more comprehensive discussions on the topic. For the former, a seam (eight-connected path of pixels from top to bottom or from left to right) is iteratively removed to preserve important content in a photo. Many seam detection algorithms have been proposed recently. Avidan and Shamir [1] formulated seam detection as dynamic programming, where a gradient energy is used as the importance map. Rubinstein *et al.* [18] introduced a forward energy criterion to improve Avidan and Shamir's [1] work. The criterion selects the seam that reintroduces the minimal amount of energy, which is solved using graph cut optimization. As a variant of seaming, Pritch *et al.* [16] proposed to discretely remove repeated patterns in homogenous image regions. For continuous retargeting, Wolf *et al.* [22] merged less important pixels to reduce distortion. However, the distortion can only be propagated along the resizing direction. To improve the distortion propagation, Wang *et al.* [23] proposed an optimized scale-and-stretch (OSS) approach, which iteratively wraps local regions to match the optimal scaling factors as close as possible. Guo *et al.* [10] presented an effective image retargeting method using saliency-based mesh parametrization (SMP), which optimally preserves image structures. Since many approaches cannot effectively preserve structural lines, Lin *et al.* [13] presented a patch-based photo retargeting model which preserves the shapes of both visually salient objects and structural lines. Note that the above content-aware photo retargeting methods depend merely on low-level feature-based saliency maps, which reflect no photo semantics. In addition, the proposed model belongs to the category of continuous retargeting. By defining a resizing space that combines multiple resizing operators, Rubinstein *et al.* [34] presented a retargeting algorithm focusing on searching the optimal path in the resizing space. The searching process can be viewed as optimizing an objective function that measures the similarity between the source and the target images. Wang *et al.* [23] introduced a scale-and-stretch warping algorithm that allows resizing images into different aspect ratios while preserving visually prominent features. The algorithm iteratively computes optimal local scaling factors for each local region and updates the warped image that can maximally match these scaling factors. Furthermore, Wang *et al.* [38] proposed to combine cropping and wrapping operators for video retargeting. The cropping process removes temporally recurring contents and the warping utilizes available homogeneous regions to mask deformations while preserving motion. Zhang *et al.* [35] proposed a content-aware dynamic video retargeting algorithm. A pixel-level shrinkable map is constructed that indicates both the importance of each pixel and its continuity, based on which a scaling function calculates the new pixel location of the retargeted video. Panozzo *et al.* [36] proposed to retarget an image in the space of axis-aligned deformations. Such a deformation space excludes local rotations, avoids harmful visual distortions, and can be parameterized in 1-D. Krähenbühl *et al.* [37] developed a content-aware interactive video retargeting system. The system combines key frame-based constraint editing with numerous automatic algorithms for video analysis, which gives content producers higher-level intervention in the retargeting process. Furthermore, Vaquero *et al.*'s [44] survey article reviewed and grouped the content-aware image retargeting algorithms into several categories: content-aware photo cropping, seam carving (SC), etc. Interested readers can refer to this survey for a comprehensive overview of the previous content-based retargeting techniques.

To integrate high-order spatial interactions of image patches for photo esthetic evaluation, Zhang *et al.* [30] introduced graphlets and further designed a probabilistic model to transfer them from the training photos into the cropped photo. However, graphlets do not reflect photo semantics or photo global spatial configurations, which are essential cues to be exploited in a cropping model. Besides, graphlets extracted from each photo are unselectively transferred from the training photos into the cropped one, while graphlets along the human gaze shifting path are much more important in capturing photo esthetics thus should be selected for esthetic feature transfers. Furthermore, Zhang *et al.* [31] proposed a weakly supervised segmentation algorithm by formulating high-order structural potentials among superpixels into graphlets. The key is a manifold embedding method [41] to transfer semantics of image-level labels into different regions in an image, which motivates a semantics transferring module in photo retargeting. The objective function of the embedding in [31] and that in our retargeting model is similar, whereas their solutions are completely different. The former is solved directly based on the coordination propagation algorithm proposed by Xiang *et al.* [32], while the latter is solved by the online updating scheme proposed by us.

Recently, Castillo *et al.* [33] evaluated the impact of photo retargeting on human fixations, by experimenting on the RetargetMe data set [19]. The authors observed that: 1) even strong artifacts in the retargeted photo cannot influence human gaze shifting if they are distributed outside the regions of interest; 2) removing contents in photo retargeting might change its semantics, which influences human perception of photo esthetics accordingly; and 3) employing eye-tracking data can more accurately reflect the regions of interest, which might be helpful for photo retargeting. Consistent with [33], our experiments also show that human gaze shifting paths can substantially increase retargeting performance. Since it is impractical to use human data for real-world retargeting tasks, our method aims to generate AGPs to mimick human gaze shifting, where experimental results show a high degree of consistency (over 90%) of our AGP and real human gaze shifting path. Yang *et al.* [50] proposed a semi-supervised batch mode multiclass learning algorithm for visual concept recognition, which exploits the whole active pool to evaluate the data uncertainty. Yang *et al.* [51] built upon the assumption that different-related tasks share common structures. Multiple feature selection functions of different tasks are simultaneously learned in a joint framework, which enables the algorithm to utilize the common knowledge of multiple tasks as supplementary information to facilitate decision-making.

III. Graphlets on the Semantic Space

A. Graphlets Onto Manifold

There are usually tens to hundreds of local components in a photo. Among these components, a few spatially adjacent ones along with their correlations determine the local composition of a photo and reflect regional esthetics in a photo [47] and are thus important in the retargeting process. To capture an arbitrary spatial structure among image components, as shown in Fig. 1, we define a graphlet as

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

where V is a set of vertices, each representing an image component, and E is a set of edges, each connecting pairwise spatially adjacent image components. The components are generated using the unsupervised fuzzy clustering-based image segmentation [25], because the tolerance bound is convenient to tune. We call a graphlet with t vertices a t-sized graphlet. Practically, only small-sized graphlets ($t \leq 5$) are adopted since the number of graphlets in a photo is exponentially increasing with their size.

Given a t-sized graphlet G, we can represent it by a $t \times (t + 128 + 9)$-sized matrix $\mathbf{M}$, that is

$$\mathbf{M} = [\mathbf{M}_c, \mathbf{M}_t, \mathbf{M}_s] \tag{2}$$

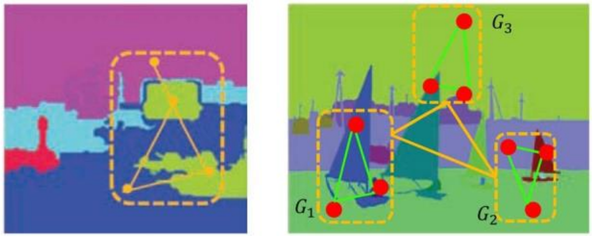


Fig. 1. Left: example graphlet. Right: preserving the three pairwise graphlets' distances implicitly keeps the global spatial layout.

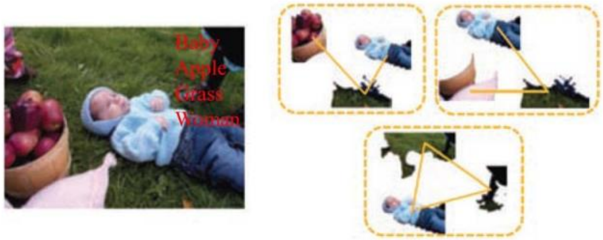


Fig. 2. Example illustrating the high correlation between a graphlet and its spatially neighboring ones. The three graphlets share the "baby" region.

where $\mathbf{M}_c$ is a $t \times 9$-sized matrix and each row is the 9-D color moment [21] from a segmented region; $\mathbf{M}_t$ is a $t \times 128$-sized matrix and each row is a 128-D HOG [8] from a segmented region; and $\mathbf{M}_s$ is a $t \times t$-sized adjacency matrix representing the structure of graphlet G.

It can be observed that spatially neighboring graphlets in a photo are partially overlapping. As shown in Fig. 2, the three spatially neighboring graphlets share the region of the baby. This brings the property of local preservation, which indicates that a graphlet and its spatially neighboring ones are highly correlated. Therefore, it is beneficial to exploit the local structure among graphlets when projecting them onto the semantic space. That is, each matrix can be deemed as a point on manifold [26] and the Golub–Werman distance [24] between identical-sized matrices is

$$d_{\mathrm{GW}}\left(\mathbf{M}, \mathbf{M}'\right) = \left\|\mathbf{M}_o - \mathbf{M}'_o\right\|_2 \tag{3}$$

where $\mathbf{M}_o$ and $\mathbf{M}'_o$ denote the orthonormal basis of $\mathbf{M}$ and $\mathbf{M}'$, respectively.

B. Semantics-Encoded Graphlet Embedding

Based on the matrix representation, a manifold embedding is derived to project graphlets onto the low-dimensional semantic space. The semantic space is constructed based on the semantics of training images, e.g., the four labels in Fig. 2. The objective function of the embedding is given as

$$\arg\min_{\mathbf{Y}} \underbrace{\sum_{ij}\left\|y_i - y_j\right\|^2 l_s(i,j) - \sum_{ij}\left\|y_i - y_j\right\|^2 l_d(i,j)}_{\text{Encode image\_level semantics into graphlets}} + \underbrace{\sum_h \sum_{ij}\left(d_{\mathrm{GW}}\left(\mathbf{M}_i^h, \mathbf{M}_j^h\right) - d_E\left(y_i^h, y_j^h\right)\right)^2}_{\text{Preserve global spatial layout into graphlets}}$$

$$\text{s.t.} \quad \mathbf{Y}\mathbf{Y}^T = \mathbf{I}_d \tag{4}$$

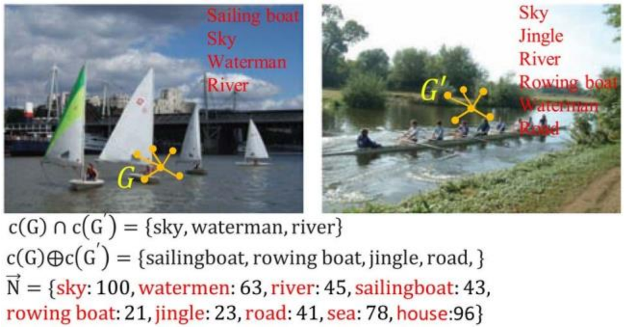

Sailing boat
Sky
Waterman
River
G
Sky
Jingle
River
Rowing boat
Waterman
Road
G'
c(G) ∩ c(G') = {sky, waterman, river}
c(G)⊕c(G') = {sailingboat, rowing boat, jingle, road, }
N⃗ = {sky: 100, watermen: 63, river: 45, sailingboat: 43,
rowing boat: 21, jingle: 23, road: 41, sea: 78, house:96}

Fig. 3. Illustration of calculating l_s and l_d. $\dot{N}$ denotes the number of training images containing each semantic object.

where $\mathbf{Y} = [y_1, y_2, \ldots, y_N]$, and y_i^h and y_j^h are column vectors standing for the d-dimensional representations of the ith and the jth graphlets from the hth photo. The first part adds image-level semantics into graphlets. The second part preserves all pairwise graphlets' Golub–Werman distances of a photo, which implicitly keeps its global spatial layout, as illustrated on the right of Fig. 1. $\mathbf{M}_i^h$ and $\mathbf{M}_j^h$, respectively, denote matrices corresponding to the ith and the jth identical-sized graphlets from the hth photo, and $d_E(\cdot, \cdot)$ represents the Euclidean distance. $\mathbf{YY}^T = \mathbf{I}_d$ is a term to uniquely determine $\mathbf{Y}$.

Particularly, for the first part of objective function (4), $l_s(\cdot, \cdot)$ and $l_d(\cdot, \cdot)$ are functions measuring the semantic similarity and difference between graphlets, respectively. Denoting $\dot{\mathbf{N}} = [N^1, \ldots, N^C]^T$, where $N^i (i \in [1, C])$ is the number of photos from the ith category, and $\mathbf{c}(\cdot)$ denotes the semantic labels of the photo from which the graphlet is extracted, then we obtain

$$l_s(i, j) = \frac{\left(\mathbf{c}(G_i) \cap \mathbf{c}\left(G_j\right)\right) \dot{\mathbf{N}}}{\sum_c N^c} \quad (5)$$

$$l_d(i, j) = \frac{\left(\mathbf{c}(G_i) \oplus \mathbf{c}\left(G_j\right)\right) \dot{\mathbf{N}}}{\sum_c N^c} \quad (6)$$

where the numerator of l_s denotes the number of photos in common categories with the photos where the two graphlets are extracted, the numerator of l_d is the number of photos in different categories with the photos where the two graphlets are extracted, and the denominator represents the total number of photos in all categories with the photos where the two graphlets are extracted. An example illustrating l_s and l_d is given in Fig. 3, in this case $l_s(G, G') = (100 + 63 + 45/100 + 63 + \cdots + 96) = 0.4078$ and $l_d(G, G') = (43 + 21 + 23 + 41/100 + 63 + \cdots + 96) = 0.251$.

Denote $\mathbf{D}^h_{GW}$ as an $N \times N$ matrix whose ijth entry is $d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h)$, i.e., the Golub–Werman distance between the ith and the jth identical-sized graphlets extracted from the hth photo. Then, the inner product matrix is obtained by $\tau(\mathbf{D}^h_{GW}) = -\mathbf{R}_{N_h} \mathbf{S}^h_{GW} \mathbf{R}_{N_h}/2$, where $(\mathbf{S}^h_{GW})_{ij} = (\mathbf{D}^h_{GW})^2_{ij}$; $\mathbf{R}_{N_h} = \mathbf{I}_{N_h} - \dot{\mathbf{e}}_{N_h} \dot{\mathbf{e}}^T_{N_h}/N$ is the centralization matrix; $\mathbf{I}_{N_h}$ is an $N_h \times N_h$ identity matrix and $\dot{\mathbf{e}}_{N_h} = [1, \ldots, 1]^T \in R^{N_h}$; and N_h is the number of graphlets from the hth photo. The first term in (4) can be rewritten as

$$\begin{aligned} &\arg\min_{\mathbf{Y}} \sum_h \sum_{ij} \left(d_{GW}\left(\mathbf{M}_i^h, \mathbf{M}_j^h\right) - d_E\left(y_i^h, y_j^h\right)\right)^2 \\ &= \arg\min_{\mathbf{Y}} \sum_h \left\| \tau\left(\mathbf{D}^h_{GW}\right) - \tau\left(\mathbf{D}^h_{Y}\right) \right\|^2 \\ &= \arg\max_{\mathbf{Y}} \operatorname{tr} \left(\mathbf{Y} \tau\left(\mathbf{D}^h_{GW}\right) \mathbf{Y}^T\right) \\ &= \arg\max_{\mathbf{Y}} \operatorname{tr} \left(\mathbf{Y} \tau(\mathbf{D}_{GW}) \mathbf{Y}^T\right) \end{aligned} \quad (7)$$

where $\tau(\mathbf{D}_{GW})$ is a block diagonal matrix with $H \times H$ blocks, and the hth diagonal block is $\tau(\mathbf{D}^h_{GW})$.

The second term in (4) can be rewritten as

$$\begin{aligned} &\arg\min_{\mathbf{Y}} \sum_{ij} \left\| y_i - y_j \right\|^2 [l_w(i, j) - l_b(i, j)] \\ &= \arg\max_{\mathbf{Y}} \operatorname{tr}\left(\mathbf{YAY}^T\right) \end{aligned} \quad (8)$$

where $\mathbf{A} = [-\dot{\mathbf{e}}^T_{N-1}, \mathbf{I}_{N-1}]^T \mathbf{W}_1 [-\dot{\mathbf{e}}^T_{N-1}, \mathbf{I}_{N-1}] + \cdots + [\mathbf{I}_{N-1}, -\dot{\mathbf{e}}^T_{N-1}]^T \mathbf{W}_N [\mathbf{I}_{N-1}, -\dot{\mathbf{e}}^T_{N-1}]$, and $\mathbf{W}_i$ is an $N \times N$ diagonal matrix whose hth diagonal element is $l_s(h, i) - l_d(h, i)$.

Based on (7) and (8), the objective function (4) can be reorganized into

$$\begin{aligned} &\arg\max_{\mathbf{Y}} \operatorname{tr}\left(\mathbf{Y}\left(\mathbf{A} + \tau(\mathbf{D}_{GW})\right)\mathbf{Y}^T\right) = \arg\max_{\mathbf{Y}} \operatorname{tr}\left(\mathbf{YBY}^T\right) \\ &\text{s.t. } \mathbf{YY}^T = \mathbf{I}_d \end{aligned} \quad (9)$$

where $\mathbf{B} = \mathbf{A} + \tau(\mathbf{D}_{GW})$ and (9) is solved using the online learning algorithm described in Section VI-A.

IV. Sparsity-Constrained Active Graphlet Path

In a human vision system, only the distinctive sensory information is selected for further processing [4]. From this perspective, only a few visually/semantically salient graphlets within a photo can be perceived by humans [48]. These salient graphlets are significantly different from those nonsalient ones, either in their appearances or in their semantics. Thus, a sparsity-constrained ranking scheme is developed to discover salient graphlets, by exploring color, texture, and semantic channels collaboratively.

Let $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ be the three feature matrices in color, texture, and semantic channels, respectively, where the columns in different matrices with the same index correspond to the same graphlet. The size of each $\mathbf{X}_i$ is $d_i \times N$, where d_i is the feature dimension and N is the number of graphlets. Then, the task is to find a weighting function to each graphlet: $S(G_i) \in [0, 1]$ by integrating the three feature matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$.

Based on the theory of visual perception [11], there are usually strong correlation among the nonsalient regions in a photo. That is to say, the nonsalient graphlets can be self-represented. This analysis suggests that feature matrix $\mathbf{X}$ ($\mathbf{X}$ can be any one of matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$) can be decomposed into a salient part and a nonsalient part, that is

$$\mathbf{X} = \mathbf{XZ}_0 + \mathbf{E}_0 \quad (10)$$

where $\mathbf{XZ}_0$ denotes the nonsalient part that can be reconstructed by itself, $\mathbf{Z}_0$ denotes the reconstruction coefficients, and $\mathbf{E}_0$ is the remaining part corresponding to the salient targets.

Without a constraint, there are an infinite number of possible decompositions with respect to $\mathbf{Z}_0$ and $\mathbf{E}_0$. Toward a unique solution that indicates those salient graphlets, we need some criteria for characterizing matrices $\mathbf{Z}_0$ and $\mathbf{E}_0$. Aiming at this, two observations are made. On the one hand, motivated by many approaches in computer vision, we assume that only a small fraction of graphlets are salient, i.e., matrix $\mathbf{E}_0$ is sparse. The connection between sparsity and saliency is also consistent with the fact that only a small subset of sensory information is selected for further processing in a human vision system. On the other hand, the strong correlation among the background graphlets suggests that matrix $\mathbf{Z}_0$ is with low rankness. Based on the above analysis, we can infer the salient graphlets by adding a sparsity and low-rankness constraint to (10), thereby the graphlet saliency detection can be formulated as a low-rank representation [20] problem

$$\min_{\mathbf{Z}_0,\mathbf{E}_0} ||\mathbf{Z}_0||_* + \lambda||\mathbf{E}_0||_{2,1}, \quad \text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}_0 + \mathbf{E}_0 \tag{11}$$

where $||\cdot||_*$ denotes the matrix nuclear norm that is a convex relaxation of the rank function, parameter $\lambda > 0$ is used to balance the effects of the two parts, and $||\cdot||_{2,1}$ is the $l_{2,1}$ norm defined as the sum of the $l_{2,1}$ norms of the columns of a matrix

$$||\mathbf{E}_0||_{2,1} = \sum_i \sqrt{\sum_j (\mathbf{E}_0(j,i))^2}. \tag{12}$$

It is noticeable that the minimization of the $l_{2,1}$ norm encourages the columns of $\mathbf{E}_0$ to be zero. It fits well to our saliency detection problem. For a column corresponding to the ith graphlet G_i, a larger magnitude implies that the corresponding graphlet is more salient in drawing the attention of human eye. Let $\mathbf{E}_0^*$ be the optimal solution (with respect to $\mathbf{E}_0$) to (10). To obtain the saliency value of graphlet G_i, we quantify the response of the sparse matrix

$$S(G_i) = \left\|\mathbf{E}_0^*(:,i)\right\|_2 = \sqrt{\sum_i \left(\mathbf{E}_0^*(j,i)\right)^2} \tag{13}$$

where $||\mathbf{E}_0^*(:,i)||_2$ denotes the l_2 norm of the ith column of $\mathbf{E}_0^*(:,i)$. A larger score of $S(G_i)$ means that graphlet G_i has a higher probability to be salient.

The objective function (10) calculates the saliency of a graphlet based on one type of visual feature, which is suboptimal since multiple visual features determine graphlet saliency collaboratively. To combine together visual features in color, texture, and semantic channels, we extend the objective function (8) into a multimodal version

$$\min_{\substack{\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3\\ \mathbf{E}_1,\mathbf{E}_2,\mathbf{E}_3}} \sum_{i=1}^{3} ||\mathbf{Z}_i||_* + \lambda||\mathbf{E}||_{2,1}, \quad \text{s.t. } \mathbf{X}_i = \mathbf{X}_i\mathbf{Z}_i + \mathbf{E}_i \tag{14}$$

where $\mathbf{E} = [\mathbf{E}_1; \mathbf{E}_2; \mathbf{E}_3]$ is formed by vertically concatenating $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ together along the column. The integration of multiple features is seamlessly performed by minimizing the $l_{2,1}$ norm of $\mathbf{E}$. That is, we enforce the columns of $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ to have jointly consistent magnitude values.

Algorithm 1 Inextact ALM-Based Solution of (14)

```
input: Data matrices $\{\mathbf{X}_i\}$, parameter $\lambda$;
output: The optimal solution $\mathbf{E}^*$;
while not converged do
    1) Fix the others and update $\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3$ by:
    $\mathbf{J}_i = \arg\min_{\mathbf{J}} \frac{1}{\mu}||\mathbf{J}||_* + \frac{1}{2}||\mathbf{J}_i - (\mathbf{Z}_i + \frac{\mathbf{W}_i}{\mu})||_F^2$.
    2) Fix the others and update $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ by:
    $\mathbf{Z}_i = \mathbf{M}(\mathbf{X}_i^T(\mathbf{X}_i - \mathbf{E}_i) + \mathbf{J}_i + \frac{\mathbf{X}_i^T\mathbf{Y}_i - \mathbf{W}_i}{\mu})$
    where $\mathbf{M} = (\mathbf{I} + \sum_{i=1}^{3}\mathbf{X}_i^T\mathbf{X}_i)^{-1}$.
    3) Fix the others and update $\mathbf{E} = [\mathbf{E}_1; \mathbf{E}_2; \mathbf{E}_3]$ by
    $\mathbf{E} = \arg\min_{\mathbf{E}} \frac{\lambda}{\mu}||\mathbf{E}||_{2,1} + \frac{1}{2}||\mathbf{E} - \mathbf{G}||_F^2$,
    where $\mathbf{G}$ is formed by vertically concatenating the matrices
    $\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i + (\mathbf{Y}_i/\mu)$, $i = 1, 2, 3$ together along column.
    4) Update the multipliers
    $\mathbf{Y}_i = \mathbf{Y}_i + \mu(\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i)$; $\mathbf{W}_i = \mathbf{W}_i + \mu(\mathbf{Z}_i - \mathbf{J}_i)$;
    5) Update the parameter $\mu$ by
    $\mu = \min(\rho\mu, 10^{10})$
    where the parameter $\rho$ controls the convergence speed. It is
    set as $\rho = 1.1$ in all experiments.
    6) Check the convergence condition: $\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i \rightarrow 0$ and
    $\mathbf{Z} - \mathbf{J}_i \rightarrow 0$, $i = 1, 2, 3$;
end while
```

Let $\{\mathbf{E}_1^*, \mathbf{E}_2^*, \mathbf{E}_3^*\}$ be the optimal solution to (14), to obtain a saliency score for graphlet G_i, we quantify the response of the sparse matrices as follows:

$$S(G_i) = \sum_{j=1}^{3} ||\mathbf{E}_j^*(:,i)||_2 \tag{15}$$

where $||\mathbf{E}_j^*(:,i)||_2$ denotes the l_2 norm of the ith column of $\mathbf{E}_j^*$. A larger score of $S(G_i)$ means that graphlet G_i has a higher probability to be salient. The details of solving (14) are described as follows.

Problem (14) is convex and can be optimized efficiently. We first convert it into the following equivalent problem:

$$\min_{\mathbf{J}_i,\mathbf{Z}_i,\mathbf{E}_i} \sum_{i=1}^{3} ||\mathbf{J}_i||_* + \lambda||\mathbf{E}||_{2,1} \quad \text{s.t. } \mathbf{X}_i = \mathbf{X}_i\mathbf{Z}_i + \mathbf{E}_i, \mathbf{Z}_i = \mathbf{J}_i. \tag{16}$$

This problem can be solved with the ALM method which minimizes an augmented Lagrange function

$$\begin{aligned} L = &\lambda||\mathbf{E}||_{2,1} + \sum_{i=1}^{3} ||\mathbf{J}_i||_* + \langle\mathbf{Y}_i, \mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i\rangle \\ &+ \langle\mathbf{W}_i, \mathbf{Z}_i - \mathbf{J}_i\rangle + \frac{\mu}{2}||\mathbf{X}_i - \mathbf{X}_i\mathbf{Z}_i - \mathbf{E}_i||_F^2 \\ &+ \frac{\mu}{2}||\mathbf{Z}_i - \mathbf{J}_i||_F^2 \end{aligned} \tag{17}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ are the Lagrange multipliers and $\mu > 0$ is a penalty parameter. The inexact ALM method, also called the alternating direction method, is illustrated in Algorithm 1. Note that the subproblems of the algorithm are convex and they have close-form solution. Step 1 is solved via the singular value thresholding operator [9], whereas step 3 is solved via [20].

By summarizing the above discussion, the proposed sparsity-constrained salient graphlet selection is presented

```
Algorithm 2 Sparsity-Constrained Graphlets Discovery
input: Graphlets from a labeled photo; the number of selected graphlets in a photo: $K$;
output: A number of graphlets ranked by their saliency values;
1) Compute the feature matrices $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ in color, texture, and semantic channels (by manifold embedding in (9)) to describe each graphlet;
2) Obtain the sparsity matrices $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ in color, texture, and semantic channels respectively, by solving the objective function (14);
3) Compute the graphlet saliency based on (15), and the $K$ top-ranked graphlets are deemed as the salient ones.
```

in Algorithm 2. To mimick the human gaze shifting process, we link the K discovered graphlets into a path, termed AGP.

V. Probabilistic Retargeting Framework

Due to the subjectivity of photo esthetic assessment, people with different backgrounds/experiences might bias for retargeted photos with certain styles. To reduce such bias in the retargeted photos, it is necessary to exploit the esthetic experiences of multiple users. To make the retargeted photo unbiased, we study the esthetic experience of professional photographers. As a widely used statistic tool, Gaussian mixture models (GMMs) have been shown to be effective for learning the distribution of a set of data. In this paper, GMMs are used to uncover the structure of AGPs from all training esthetically pleasing photos. The training photos are obtained by googling images using the keywords such as "iPhone wallpaper." For each graphlet in the path, we use an L-component GMM to learn its distribution

$$p(y|\theta) = \sum_{l=1}^{L} \alpha_l N(y|\pi_l, X_l) \tag{18}$$

where y denotes a post-embedding graphlet and $\theta = \{\alpha_l, \pi_l, X_l\}$ are the GMM parameters.

In addition to the graphlets, the directed-edges of the path determine the spatial displacement between graphlets, which can also be leveraged in the retargeting model. For all the directed edges in the training AGPs, we construct an L^{r}-component GMM to model its distribution as

$$p(e|\eta) = \sum_{l=1}^{L^{\mathrm{r}}} \beta_l N(e|v_l, \alpha_l) \tag{19}$$

where e is the concatenated 2-D vectors denoting the length and horizontal angle of all directed-edges in an AGP; and $\eta = \{\beta_l, v_l, \alpha_l\}$ are the GMM parameters. In our implementation, we set $L = L^{\mathrm{r}} = 5$, and the AGP length is fixed to 5 as well. In our model, the two GMMs in (18) and (19) are incrementally learned as detailed in Section VI-B.

After learning the two GMM priors, we shrink a test photo to make its AGP most similar to those from the training photos. That is, given an AGP of a test photo, we calculate the probability of each of its graphlets and that of the directed edges. To avoid the triangle mesh as a control mesh in shrinking, which may result in distortions in triangle orientations, we use grid-based shrinking. Particularly, we decompose a photo into equal-sized grids,[1] and the horizontal weight of grid φ is calculated as

$$w_h(\varphi) = \max_{y\cap\varphi^h \neq \emptyset,\, e\cap\varphi^h \neq \emptyset} \{p(y|\theta), p(e|\eta)\} \tag{20}$$

where $y \cap \varphi^h \neq \emptyset$ denotes graphlet y is horizontally overlap- ping with grid φ, and $e \cap \varphi^h \neq \emptyset$ means directed edge e is horizontally overlapping with grid φ.

The vertical weight of grid φ is computed as

$$w_v(\varphi) = \max_{y\cap\varphi^v \neq \emptyset,\, e\cap\varphi^v \neq \emptyset} \{p(y|\theta), p(e|\eta)\}. \tag{21}$$

After calculating the horizontal (respectively vertical) weight of each grid, a normalization operation is carried out to make them sum to one, that is

$$\bar{w}_h(\varphi_i) = \frac{w_h(\varphi_i)}{\sum_i w_h(\varphi_i)}. \tag{22}$$

Given the size of the retargeted photo $W \times H$, the horizontal dimension of the ith grid is shrunk to $[W \cdot \bar{w}_h(\varphi_i)]$, and the vertical one of the ith grid is shrunk to $[H \cdot \bar{w}_v(\varphi_i)]$, where $[\cdot]$ rounds a real number to the nearest integer. Grids covered by the rowing boat and the watermen are semantically significant, and are thus preserved in the retargeted photo without scaling. In contrast, grids covered by the tree and sky are less semantically important, so they are shrunk in both horizontal and vertical directions.

VI. Online Retargeting Model Updating

The performance of our AGP-transfer-based retargeting model depends on the training photos. Thus, increasing the number of photos is an effective way to boost the retargeting performance. Aiming at this, we propose online algorithms for incrementally training the two components in our model: 1) manifold graphlet embedding and 2) AGP distribution learning.[2]

A. Post-Embedding Graphlet Propagation

This section develops an incremental algorithm to calculate post-embedding graphlets. First, we solve an embedding using (9) under graphlets extracted from $\{I^1, \ldots, I^{(0)}\}$ initial photos. Then, we solve a new embedding under graphlets extracted from $\{I^1, \ldots, I^{(0)}, \ldots, I^{(1)}\}$ photos, wherein the objective function is given as follows:

$$\begin{aligned} &\arg\max_{\mathbf{Y}^{(1)}} \operatorname{tr}\left(\mathbf{Y}^{(1)}\mathbf{B}^{(1)}\left(\mathbf{Y}^{(1)}\right)^T\right) \\ &\text{s.t. } y_i^{(1)} = y_i^{(0)},\ i \in \{1, \ldots, P\} \end{aligned} \tag{23}$$

where $\mathbf{Y}^{(1)} = [\mathbf{Y}_L, \mathbf{Y}_U] \in \mathrm{R}^{d\times H^{(1)}}$, $\mathbf{Y}_L = \{y_1, \ldots, y_P\}$ is the embedding calculated from $\{I^1, \ldots, I^{(0)}\}$ initial photos, and $\mathbf{Y}_U = \{y_{P+1}, \ldots, y_Q\}$ is the new embedding obtained by graphlets from dynamically increased set of photos; $\mathbf{B}^{(1)}$ is the matrix constructed from the combination of initial and an

[1] Grid size is a user-tuned parameter and we set it to 20×20 (retargeted photos under different grid sizes are compared and reported in the experimental section).

[2] The proposed sparsity-constrained graphlet discovery can inherently handle the increased set of photos, as there is not a training stage.

increasing number of photos, and it can be divided into four blocks as

$$\mathbf{B}^{(1)} = \begin{matrix} \mathbf{B}_{LL} & \mathbf{B}_{LU} \\ \mathbf{B}_{UL} & \mathbf{B}_{UU} \end{matrix} . \tag{24}$$

Denote d as the dimensionality of the post-embedding graphlet, the problem in (23) can be decomposed into d subproblems. Each subproblem is a quadratic problem with linear constraints that can be iteratively solved. Let $\mathbf{Y}_L^i = [f_i(1), \ldots, f_i(d)] \in \mathrm{R}^d, i = \{1, \ldots, P\}$, we can reorganize (23) into

$$\begin{cases} \arg\max_{\mathbf{X}(1)} \mathrm{tr}\ \mathbf{X}(1)B^{(1)}\mathbf{X}(1)^T \quad \text{s.t. } x_i(1) = f_i(1) \\ \vdots \qquad\qquad\qquad\qquad \ddots \\ \arg\max_{\mathbf{X}(d)} \mathrm{tr}\ \mathbf{X}(d)\mathbf{B}^{(1)}\mathbf{X}(d)^T \quad \text{s.t. } x_i(d) = f_i(d) \end{cases} \tag{25}$$

where $i \in \{1, \ldots, P\}$ and $\mathbf{X}(i)$ is a Q-dimensional row feature vector to be solved.

Since each subproblem in (25) has the same form, we can simplify it into

$$\arg\max_{\mathbf{X}} \mathrm{tr}\ \mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T \quad \text{s.t. } x_i = f_i. \tag{26}$$

Converting the hard constraints in (23) into soft constraints and introducing a prediction term, we have the following regularization representation:

$$\arg\max_{\mathbf{X}} \quad \mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T + \mu_1 \sum_{i=1}^{P} (x_i - f_i)^2 + \mu_2 \sum_{i=P+1}^{Q} (x_i - g_i)^2. \tag{27}$$

By differentiating the objective function (27) with respect to $\mathbf{X}_U$ and setting the derivative to 0, we obtain

$$\mathbf{X}_U = \frac{(\mathbf{I} - \mathbf{B}_{UU})\mathbf{X}_U}{1 + \mu_2} - \frac{\mathbf{B}_{UL}\mathbf{X}_L}{1 + \mu_2} + \frac{\mu_2 g_U}{1 + \mu_2} \tag{28}$$

where $\mathbf{X}_U$ denotes 1-D of the post-embedding graphlets from an increasing number of photos. Thus, we can update $\mathbf{X}_U$ based on the following equation:

$$\mathbf{X}_U^{(t+1)} = \frac{(\mathbf{I} - \mathbf{B}_{UU})\mathbf{X}_U^{(t)}}{1 + \mu_2} - \frac{\mathbf{B}_{UL}\mathbf{X}_L}{1 + \mu_2} + \frac{\mu_2 g_U}{1 + \mu_2} \tag{29}$$

where the iteration is carried out repeatedly until $\mathbf{X}_U$ becomes stable. To obtain a d-dimensional embedding of $\mathbf{X}_U$, the iterative algorithm is carried out d times, and we finally obtain the low-dimensional representation of graphlets $\mathbf{Y}^{(1)} = [\mathbf{X}(1), \ldots, \mathbf{X}(d)]^T$ from this incremental embedding step. In the next incremental embedding, we solve the new embedding $\mathbf{Y}^{(2)}$ from $\{I^1, \ldots, I^{(1)}, \ldots, I^{(2)}\}$ photos, where $\{I^{(1)}, \ldots, I^{(2)}\}$ denote the increasing number of photos in this round, and the embedding process is the same as before.

Algorithm 3 Perceptually Guided Photo Retargeting

```
// Training stage:
Input: A set of training photos $\{I_1, I_2, \cdots, I_R\}$, and the number of selected graphlets $K$;
Output: The graphlet embedding model, the two online learned GMM priors;
1) Extract graphlets from each training photo; learn the embedding model and project them onto the semantic space using (4);
2) Construct an AGP for each training photo using Alg. 2;
3) Online Learn the two GMMs based on Sec 6.2.
// Test stage:
Input: The graphlet embedding model, the two learned GMMs, the size of retargeted photo $H \times W$, and a test photo $I_{test}$;
Output: The retargeted photo $I_{test}^{r}$;
1) Extract graphlets from $I_{test}$, project them onto the semantic space using the learned embedding model, and construct the active graphlet path;
2) Shrink $I_{test}$ to $W \times H$ size using (20) and (21).
```

B. Online AGP Distribution Learning

In this section, we incrementally learn the AGP distributions for the probabilistic retargeting framework.[3] First, given the post-embedding graphlets obtained from an increasing number of photos $\{I^{(0)}, \ldots, I^{(1)}\}$, we learn a GMM $\theta(I^{(1)})$ with L^{rr} components, wherein $\theta(I^{(1)})$ means that the post-embedding graphlets from $I^{(1)}$ photos are modeled by GMM. The component number L^{rr} is computed using Bayesian information criterion [15], that is

$$L^{rr} = -2 * \log \bar{p}(y|\theta) + [L(d + 1)(d + 2)/2] * \log N^{(1)} \tag{30}$$

where $\bar{p}\ (y|\theta)$ is the averaged probabilities of all new graphlets predicted using (18), $[\cdot]$ maps a real number to the smallest following integer, and d reduces the dimensionality of post-embedding graphlets. Then, we use Bregman divergence [3] to find a component-to-component match between GMMs $\theta(I^{(0)})$ and $\theta(I^{(1)})$. If the kth and the jth components from GMM $\theta(I^{(0)})$ and $\theta(I^{(1)})$ match, then a new component $\theta(I^{(0)+(1)})$ is generated by merging them, that is

$$\alpha_{\text{new}} = \frac{N^{(0)}\alpha_j + N^{(1)}}{N^{(0)} + N^{(1)}} \tag{31}$$

$$\pi_{\text{new}} = \frac{N^{(0)}\alpha_j\pi_j + N^{(1)}\pi_k}{N^{(0)}\alpha_j + N^{(1)}} \tag{32}$$

$$X_{\text{new}} = \frac{N^{(0)}\alpha_j X_j + N^{(1)} X_k}{N^{(0)}\alpha_j + N^{(1)}} + \frac{N^{(0)}\alpha_j\pi_j\pi_j^T + N^{(1)}\pi_k\pi_k^T}{N^{(0)}\alpha_j + N^{(1)}} - \pi\pi^T \tag{33}$$

where $N^{(0)}$ and $N^{(1)}$ denote the number of graphlets extracted from $I^{(0)}$ and $I^{(1)}$ photos, respectively.

Finally, for all the unmatched components in GMM $\theta(I^{(0)})$ and $\theta(I^{(1)})$, we assign them to GMM $\theta(I^{(0)+(1)})$ by normalizing the GMM component weights to make them sum to one. When newly increased photos $I^{(2)}$ arrive, their AGPs encoding process is the same as for $I^{(1)}$ photos.

By summarizing the discussions from Sections II to VI, the procedure of the method is presented in Algorithm 3.

VII. Experimental Results and Analysis

This section justifies the effectiveness of the proposed perception-based photo retargeting. First, we compare the proposed method (PM) with state-of-the-art retargeting models.

[3] Online learning the distribution of AGPs' directed edges is similar to that of their graphlets. Thus, we omit it.

A step-by-step evaluation of each component of the PM is presented subsequently. The last experiment evaluates the influence of parameter settings on the proposed approach. Due to the space limitation, in the supplementary material, we compare our AGP with prominent saliency models in describing photo esthetics and present photos retargeted using the saliency map generated by our approach. Besides, based on the eye-tracking experiment, we make a quantitative comparison between the proposed AGP with human gaze shifting path. Furthermore, we present all the 80 retargeted photos from the RetargetMe data set [19].

We select 5014 highly esthetic photos from the AVA data set [14]. AVA contains a total number of 25 000 highly and lowly ranked photos, each associating with two semantic tags. Besides, the test photos we used are from the widely used RetargetMe [19] data set, where the semantic tags of each photo are specified manually.

All experiments were carried out on a personal computer equipped with an Intel E8500 CPU and 4 GB RAM. The algorithm was implemented on the MATLAB 2011 platform.

A. Comparative Study

The proposed AGP not only captures photo esthetics, but it can also be integrated into a probabilistic model for photo retargeting. Fig. 4 compares the PM against several representative state-of-the-art approaches, including three cropping methods: 1) omni-range context-based cropping (OCBC) [5]; 2) probabilistic graphlet-based cropping (PGC) [30]; and 3) describable attribute for photo cropping (DAPC) [7], and five content-aware retargeting methods: 1) SC [1] and its improved version (ISC) [18]; 2) OSS [23]; 3) SMP [10]; and 4) the overlapping between our proposed patch-based wrapping (PW) [13]. We notice that current mobile phone screens can be either horizonal, e.g., Blackberry phones and Nokia E series, or vertical, such as iPhone and Android phones. Thus, two resolutions of the resulting photos are applied: 640 × 480 and 640 × 960.[4]

In order to make the evaluation comprehensive, we adopt a paired comparison-based user study to evaluate the effectiveness of the proposed retargeting method. This strategy was also used in [30] to evaluate the quality of a cropped photo. In the paired comparison, each subject is presented with a pair of retargeted photos from two different approaches and required to indicate a preference as of which one they would choose for a phone wallpaper. In our user study, the participants are 40 amateur/professional photographers.

As the comparative results shown in Figs. 4 and 5, we can make the following conclusions.

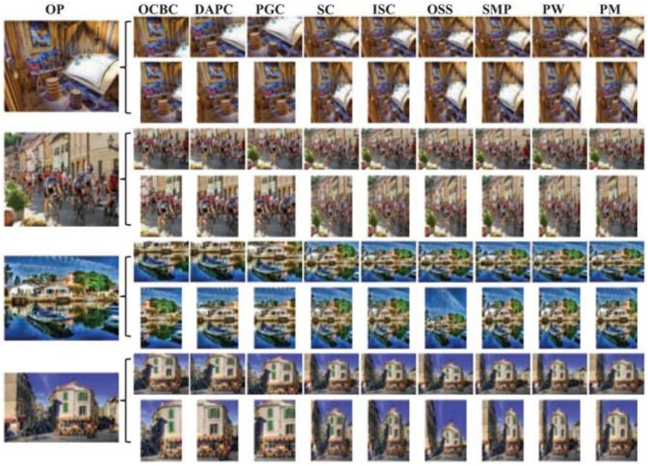


Fig. 4. Comparison of our approach with well-known photo retargeting methods. For each set of resulting photos, those from the top row are with resolution 640 × 480 and those from the bottom row are with resolution 640 × 960.

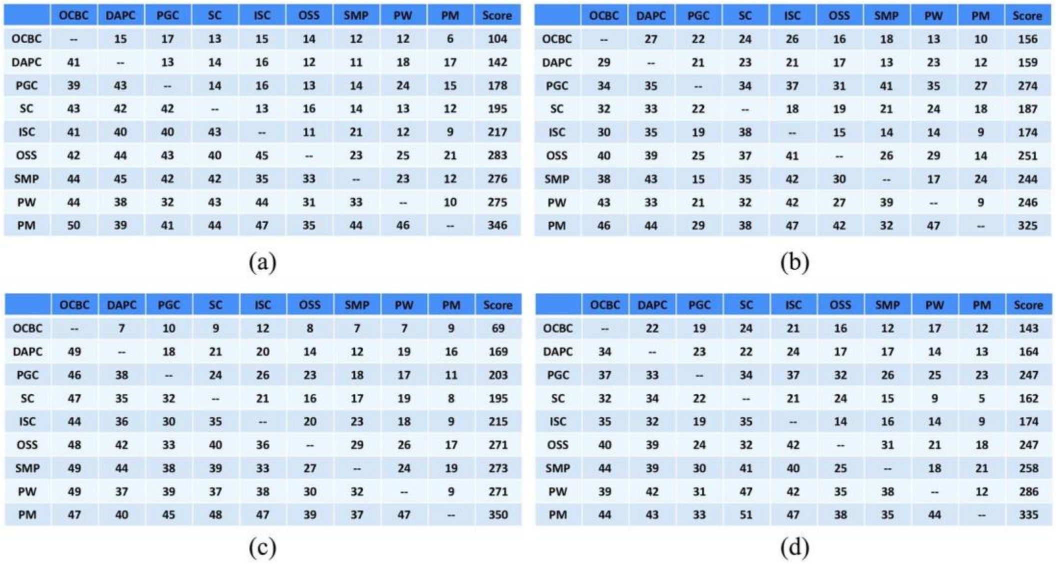
1) Compared with the three content-aware retargeting methods, our approach preserves the semantically important objects in the original photo well, such as the barrels from the first photo, the wheels from the second photo, and the boat and the central building, respectively, from the third and the fourth photos. In contrast, the compared retargeting methods may shrink the semantically important objects, such as the bicycle wheels (PW), the boat (SC, ISC, and PW), the barrels (SMP, SC, and ISC), and the bicyclists (SC, ISC, OSS, and SMP). Even worse, SC and its variant ISC, as well as OSS may result in visual distortions, i.e., the drawing paper and the barrels from the first photo, and the central building from the last photo.
2) Although cropping methods preserve important regions without visual distortions, they abandon regions that are less visually salient but still reflect the global spatial layout. For example, those lagging bikers from the second photo, the horses from the third one, and the left/right buildings from the last photo.
3) As the four preference matrices shown in Fig. 5, the user study demonstrates that our method outperforms its competitors on the resulting photos of 640 × 960.[5] It is observed that, when the resulting photos appear without distortion, the content-aware retargeting outperforms the cropping technique, and vice versa. On all the four photos, our approach produces nondistorted photos and the semantically significant objects are well preserved. Thus, the best resulting photos are consistently achieved by our approach.

To compare the time consumption of our approach with the other retargeting methods, we fix the original photo size to 1600 × 1200 and report the time consumption of retargeting it to different resolutions.[6] As shown in Table I, the efficiency of the proposed approach is competitive, especially when the resulting photo size is small. This is because our approach is based on shrinking grids, and shrinking a grid to different sizes requires similar time consumption. In this case, however, more

[4] After simple uniform scaling (supported by many mobile phones), screen resolution such as 480×320 or 320×480 can also fit a 640×480 or 640×960 wallpaper.

[5] User study is carried out on 640 × 960-sized resulting photos because this resolution is more challenging, since photos are shrunk to a greater extent.

[6] OSS is not included for the comparison of time consumption since the executive program used in our experiment is not implemented in MATLAB platform.



	OCBC	DAPC	PGC	SC	ISC	OSS	SMP	PW	PM	Score
OCBC	--	15	17	13	15	14	12	12	6	104
DAPC	41	--	13	14	16	12	11	18	17	142
PGC	39	43	--	14	16	13	14	24	15	178
SC	43	42	42	--	13	16	14	13	12	195
ISC	41	40	40	43	--	11	21	12	9	217
OSS	42	44	43	40	45	--	23	25	21	283
SMP	44	45	42	42	35	33	--	23	12	276
PW	44	38	32	43	44	31	33	--	10	275
PM	50	39	41	44	47	35	44	46	--	346

(a)

	OCBC	DAPC	PGC	SC	ISC	OSS	SMP	PW	PM	Score
OCBC	--	27	22	24	26	16	18	13	10	156
DAPC	29	--	21	23	21	17	13	23	12	159
PGC	34	35	--	34	37	31	41	35	27	274
SC	32	33	22	--	18	19	21	24	18	187
ISC	30	35	19	38	--	15	14	14	9	174
OSS	40	39	25	37	41	--	26	29	14	251
SMP	38	43	15	35	42	30	--	17	24	244
PW	43	33	21	32	42	27	39	--	9	246
PM	46	44	29	38	47	42	32	47	--	325

(b)

	OCBC	DAPC	PGC	SC	ISC	OSS	SMP	PW	PM	Score
OCBC	--	7	10	9	12	8	7	7	9	69
DAPC	49	--	18	21	20	14	12	19	16	169
PGC	46	38	--	24	26	23	18	17	11	203
SC	47	35	32	--	21	16	17	19	8	195
ISC	44	36	30	35	--	20	23	18	9	215
OSS	48	42	33	40	36	--	29	26	17	271
SMP	49	44	38	39	33	27	--	24	19	273
PW	49	37	39	37	38	30	32	--	9	271
PM	47	40	45	48	47	39	37	47	--	350

(c)

	OCBC	DAPC	PGC	SC	ISC	OSS	SMP	PW	PM	Score
OCBC	--	22	19	24	21	16	12	17	12	143
DAPC	34	--	23	22	24	17	17	14	13	164
PGC	37	33	--	34	37	32	26	25	23	247
SC	32	34	22	--	21	24	15	9	5	162
ISC	35	32	19	35	--	14	16	14	9	174
OSS	40	39	24	32	42	--	31	21	18	247
SMP	44	39	30	41	40	25	--	18	21	258
PW	39	42	31	47	42	35	38	--	12	286
PM	44	43	33	51	47	38	35	44	--	335

(d)

Fig. 5. Preference matrices from the four sets of 640 × 960-sized resulting photos in Fig. 4. Preference matrix of the (a) first sets of retargeted photos, (b) second sets of retargeted photos, (c) third sets of retargeted photos, and (d) fourth sets of retargeted photos.

TABLE I
TIME CONSUMPTION OF THE COMPARED RETARGETING METHODS

Resolution	OCBC	DAPC	PGC	SC	ISC	OSS	PW	PM
640 × 960	16.74s	24.51s	7.65s	6.54s	10.21s	24.45s	21.23s	10.12s
640 × 480	18.98s	34.11s	10.23s	10.23s	14.43s	32.25s	24.89s	12.43s
320 × 240	24.54s	40.23s	13.31s	12.37s	17.79s	46.67s	27.45s	13.11s
220 × 176	28.12s	48.14s	18.54s	17.39s	19.12s	60.22s	29.13s	14.12s

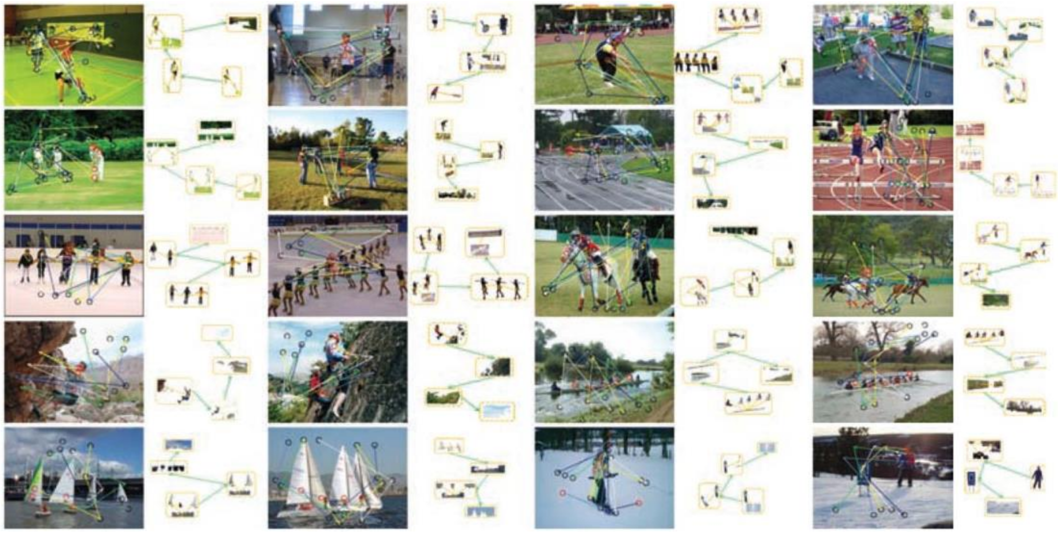


Fig. 6. Comparison of gaze shifting paths from five observers (differently colored) and the proposed AGP.

seams are detected by SC and ISC, thus the time consumption increases much more than our approach. For the cropping methods, when the size of a cropped photo is small, there are a large number of candidate cropped photos. In particular, the testing stage of our approach can be divided into three procedures: 1) graphlet extraction+embedding; 2) AGP generation; and 3) probabilistic photo shrinking. Given the size of the retargeted photo 640 × 960, the three steps take 5.34, 2.11, and 2.67 s, respectively.

Theoretically, the AGP proposed in this paper can maximally reconstruct all the graphlets extracted from a. This idea is consistent with human visual perception mechanism that sequentially allocates gazes to a few semantically important regions. To demonstrate the effectiveness of our proposed AGP, we compare it with the real human gaze shifting paths obtained from the EyeLink II eye tracker. As the examples shown in Fig. 6, our AGP can 90% overlap with the real human gaze shifting path on average.

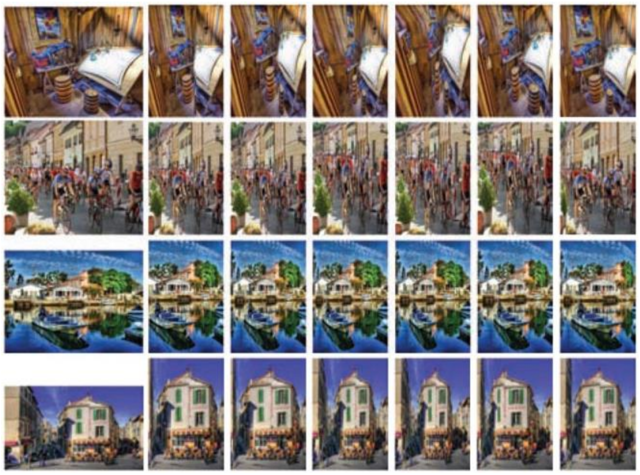


Fig. 7. Retargeted photo produced by replacing one component of the PM at a time. The first column: the original image, the second column: replacing graphlets with superpixels, the third column: removing the structure term from graphlets, the fourth column: removing image-level labels from the embedding, the fifth column: ignoring the global spatial layout from the embedding, the sixth column: replacing the probabilistic retargeting model with the probabilistic SVM-based one, and the last column: photo retargted using the PM.

B. Step-by-Step Model Illustration

The proposed retargeting framework includes four main components: 1) graphlets extraction; 2) manifold graphlet embedding; 3) probabilistic retargeting model; and 4) online retargeting model updating, which are theoretically indispensable and inseparable. To empirically demonstrate the effectiveness of each step, we replace each component with a functionally reduced counterpart and report the corresponding retargeting performance.

To demonstrate the proposed graphlet for retargeting, two experimental settings are applied to weaken the descriptive power of graphlets. First, we reduce graphlets to superpixels, i.e., one-sized graphlets which capture no spatial interaction of superpixels. Second, we remove the structure term $\mathbf{M}_s$ from the matrix-form graphlets (i.e., $\mathbf{M} = [\mathbf{M}_c, \mathbf{M}_t, \mathbf{M}_s]$). As shown in Fig. 7, compared with our approach, retargeting using superpixels or nonstructural graphlets both result in less esthetic photos. Specifically, some semantically important details are missing, such as the right barrel from the first photo, the right foremost bicyclist from the second photo, and the leftmost horses from the third and last photos. This observation shows that superpixels are not as effective for the retargeting process.

For the second component of the manifold graphlet embedding, two visual cues: 1) image-level semantics and 2) photo global spatial layout, are incorporated into the embedding. To evaluate the contributions of each cue, two experimental settings are used accordingly. First, we transform the supervised embedding into an unsupervised one, i.e., removing the image-level label integration term in (4). Then, we remove the global spatial layout term in (4). Based on the retargeted photos shown in the fourth and the fifth columns, we make the following observations.

1) As shown in the fourth column in Fig. 7, by ignoring image-level semantics, the main objects are evenly shrunk in the retargeted photo. Thus, the results are suboptimal as semantically important objects such as the barrels, the boat, and the central building should be emphasized and thus slightly shrunk.
2) As shown in the fifth column in Fig. 7, if we remove the global spatial layout term, the retargeted photos are sometimes distorted. For example, the linear woods texture from the first photo bends unnaturally; the wheels from the second photo become unharmoniously small.

To show the effectiveness of the third component, we replace the probabilistic retargeting model with a probabilistic-SVM-based one [17]. That is, we generate a large number of candidate retargeted photos by dividing an original photo into 10×10 grids and then each grid is half horizontally/vertically shrunk or nonshrunk. Thus, $2^{10+10} = 1\ 048\ 576$ candidate retargeted photo are generated and their quality is computed as follows:

$$p(I \rightarrow \text{highly esthetic}|I) = \frac{1}{1 + \exp(-f(x))} \tag{34}$$

where $f(x)$ is the linear function of SVM, and x is the concatenated global color moment and HOG (137-D) from each candidate retargeted photo. As can be seen from the sixth column in Fig. 7, retargeting based on the probabilistic SVM performs worse than our approach. The semantically important objects, such as the barrels and bicyclists, are shrunk too much. Even worse, enumeratively evaluating all candidate retargeted photos is computationally heavy and consumes too much main memory. It takes more than one hour and more than 1 GB main memory to retarget a photo,[7] while the proposed approach consumes less than 15 s and 200 MB main memory.

Third, we evaluate the effectiveness of the online model updating component. Particularly, we select for training the first 20% highly esthetic photos (5014 photos ranked by their IDs) from the AVA [14] data set, and then incrementally incorporate 20% of the photos from the remaining highly esthetics training photos into the retargeting model. As shown in Fig. 8, the proposed online learning scheme is efficient for dynamically incorporating photos. And the retargeted photos are constantly improved when more training photos are incorporated.

Besides the three components, we conduct an experiment to compare retargeting results based on AGP and a real human gaze shifting path. Particularly, we link the image segmented regions along the human gaze shifting path and compare it with our AGP. As shown in Fig. 9, the retargeted photos based on our approach are similar to those based on the real human gaze shifting paths. We further carry out a user study as described in Fig. 5, where volunteers vote nearly equally on different retargeting results. This demonstrates that our AGP performs similarly to the real human gaze shifting paths. Furthermore, as shown in Fig. 9, different persons have different gaze shifting paths, which make the corresponding

[7] As there are more than 100 000 candidate retargeted photos, it is impossible to store all of them in the main memory. In our implementation, the candidates are processed in a batch mode, wherein only 1000 candidate retargeted photos are evaluated each time.

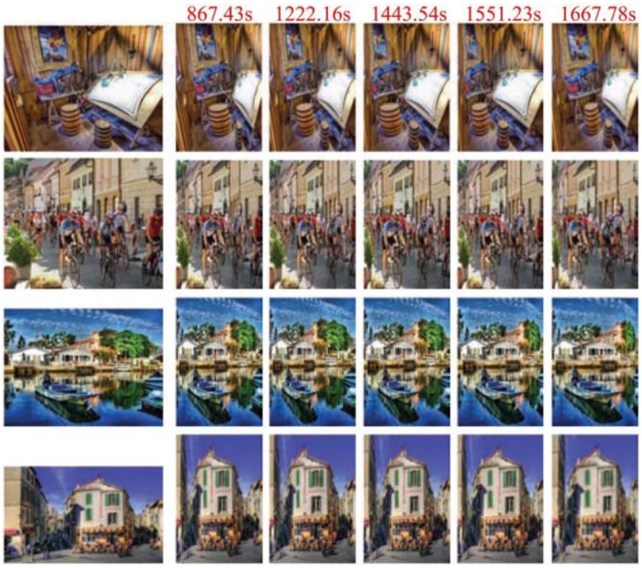


Fig. 8. Photos retargeted by incrementally incorporating 20%, 40%, 60%, 80%, and 100% training photos (from left to right). The red text denotes the training time consumption in seconds.

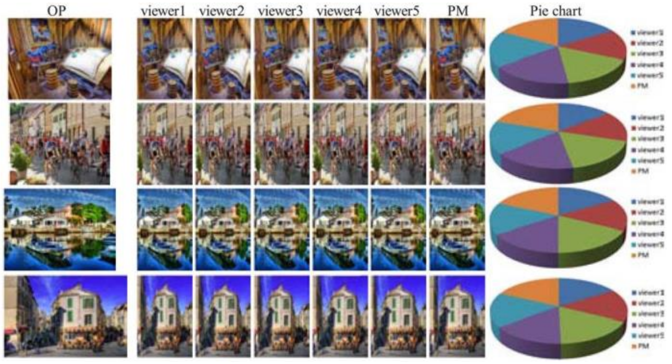


Fig. 9. Photos retargeted based on the proposed AGP (PM) and human gaze shifting paths from five different viewers.

cropped photos slightly different. Comparatively, our approach can always output a same retargeted photo and it is fully automatically, which can be conveniently applied to large-scale photo retargeting.

Last but not least, we conduct photo retargeting when replacing the L-component GMM by a Gaussian distribution, a probabilistic SVM output, and a linear regression respectively. As shown in Fig. 10, when predicting the probability of an AGP using one of the three learners, the retargeted photo is less satisfactory than those that are based on GMM. The reason is that compared with the other three learners, GMM can better model the distribution of post-embedding graphlets. Note that if we replace GMM by a more complicated model, we may expect certain improvement in performance, yet such models usually contain many parameters which are not trivial to tune.

C. Influence of Parameter Settings

This experiment evaluates the influence of important parameters on a specific photo: the number of selected graphlets K, the grid size $B \times B$, the dimensionality of the post-embedding graphlets d, and the regularization parameter μ.

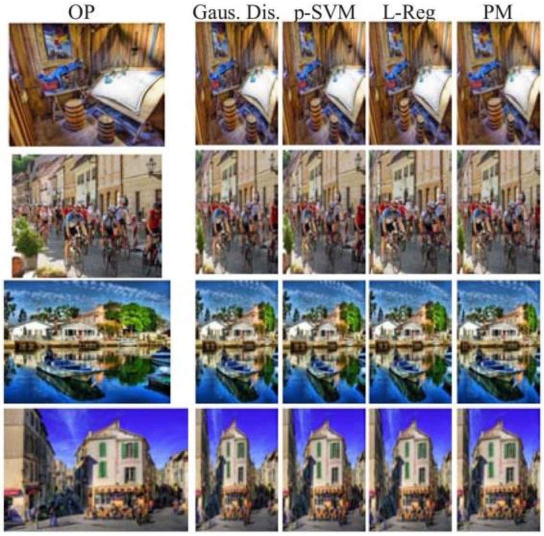


Fig. 10. Photos retargeted when replacing GMM to Gaussian distribution, probabilistic SVM output, and linear regression, respectively.

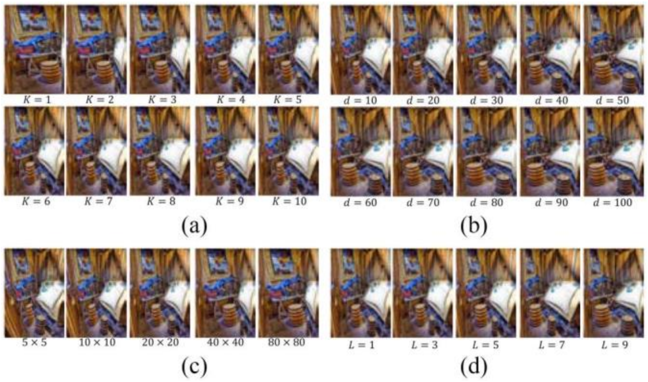


Fig. 11. Retargeted photos under different parameter settings. (a) Resulting photos by varying the selected graphlets #. (b) Resulting photos under different dimen. of post-embed. graphlets. (c) Resulting photos by varying the grid size. (d) Resulting photos by varying the Gaussian component number L.

Retargeting results under different parameter settings are shown in Fig. 11. First, we present the retargeted photos when a different number of graphlets are selected. As seen, by increasing the number of selected graphlets K from 1 to 4, the semantically significant objects, such as the barrels and the drawing board, are better preserved in the retargeted photo. For $K > 4$, however, the resulting photo remains almost unchanged. Thereby, we set $K = 4$ for this photo. Second, we change the grid size $B \times B$ and display the corresponding retargeted photo. As seen, when the gird size is set to 5×5 and 10×10, respectively, the resulting photos are both distorted. This observation is consistent with many grid/triangular-based retargeting models. When the grid size is larger than 20×20, the distortion disappears but the left barrel becomes disharmonically large. Thus, we set the gird size to 20×20. Third, we retarget a photo using different dimensional post-embedding graphlets and observe that larger dimensionality means more semantically important regions are retained in the retargeted photo. Thus, for this photo, we set the dimensionality d to 40. Last but not least, we report the retargeting results by tuning the number of Gaussian components L. As can be seen, the two barrels are visually harmonious when L is 5 or 7. Therefore, we set it to 7 in our experiment.

VIII. CONCLUSION

Photo retargeting is a useful technique in computer vision [27], [40] and multimedia [28], [29], [43]. In this paper, a new retargeting method has been proposed to shrink a photo by simulating the process of humans who sequentially perceive semantics of a photo. Particularly, an AGP is constructed to mimick the process of humans while actively looking at visually semantically important components in a photo. Furthermore, we have also proposed a probabilistic model to maximally transfer the paths from a training set of esthetically pleasing photos into the retargeted one. Extensive comparative experiments demonstrate the effectiveness of our approach.

REFERENCES

[1] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 10.

[2] S. Bhattacharya, R. Sukthankar, and M. Shah, “A framework for photo-quality assessment and enhancement based on visual aesthetics,” in *Proc. ACM Multimedia*, Florence, Italy, 2010, pp. 271–280.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[4] N. D. B. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *J. Vis.*, vol. 9, no. 3, 2009, Art. ID 5.

[5] B. Cheng, B. Ni, S. Yan, and Q. Tian, “Learning to photograph,” in *Proc. ACM Multimedia*, Florence, Italy, 2010, pp. 291–300.

[6] M. S. Castelhano, M. L. Mack, and J. M. Henderson, “Viewing task influences eye movement control during active scene perception,” *J. Vis.*, vol. 9, no. 3, 2009, Art. ID 6.

[7] S. Dhar, V. Ordonez, and T. L. Berg, “High level describable attributes for predicting aesthetics and interestingness,” in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 1657–1664.

[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.

[9] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[10] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Gleicher, “Image retargeting using mesh parameterization,” *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 856–867, Aug. 2009.

[11] 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. 2011.

[12] 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.

[13] S.-S. Lin, I.-C. Yeh, C.-H. Lin, and T.-Y. Lee, “Patch-based image warping for content-aware retargeting,” *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 359–368, Feb. 2013.

[14] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2408–2415.

[15] C. Ordonez and E. Omiecinski, “Accelerating EM clustering to find high-quality solutions,” *Knowl. Inf. Syst.*, vol. 7, no. 2, pp. 135–157, 2005.

[16] Y. Pritch, E. Kav-Venaki, and S. Peleg, “Shift-map image editing,” in *Proc. ICCV*, Kyoto, Japan, 2009, pp. 151–158.

[17] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, pp. 61–74.

[18] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. ID 16.

[19] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, “A comparative study of image retargeting,” *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. ID 160.

[20] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in *Proc. ICML*, Haifa, Israel, 2010, pp. 663–670.

[21] M. A. Stricker and M. Orengo, “Similarity of color images,” in *Proc. SPIE Stor. Retrieval Image Video Databases*, San Jose, CA, USA, 1995, pp. 381–392.

[22] L. Wolf, M. Guttmann, and D. Cohen-Or, “Non-homogeneous content-driven video-retargeting,” in *Proc. ICCV*, Rio de Janeiro, Brazil, 2007, pp. 1–6.

[23] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, “Optimized scale-and-stretch for image resizing,” *ACM Trans. Graph.*, vol. 27, no. 5, 2008, Art. ID 118.

[24] M. Werman and D. Weinshall, “Similarity and affine invariant distances between 2D point sets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 810–814, Aug. 1995.

[25] S. Xiang, C. Pan, F. Nie, and C. Zhang, “Interactive image segmentation with multiple linear reconstructions in windows,” *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 342–352, Apr. 2011.

[26] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, “Ranking with local regression and global alignment for cross media retrieval,” in *Proc. ACM Multimedia*, New York, NY, USA, 2009, pp. 175–184.

[27] L. Shao, X. Zhen, D. Tao, and X. Li, “Spatio-temporal Laplacian pyramid coding for action recognition,” *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[28] L. Shao, R. Yan, X. Li, and Y. Liu, “From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms,” *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1001–1013, Jul. 2014.

[29] L. Liu, L. Shao, X. Zhen, and X. Li, “Learning discriminative key poses for action recognition,” *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[30] L. Zhang *et al.*, “Probabilistic graphlet transfer for photo cropping,” *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.

[31] L. Zhang *et al.*, “Probabilistic graphlet cut: Exploring spatial structure cue for weakly supervised image segmentation,” in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 1908–1915.

[32] S. Xiang, F. Nie, Y. Song, C. Zhang, and C. Zhang, “Embedding new data points for manifold learning via coordinate propagation,” *Knowl. Inf. Syst.*, vol. 19, no. 2, pp. 159–184, 2008.

[33] S. Castillo, T. Judd, and D. Gutierrez, “Using eye-tracking to assess different image retargeting methods,” in *Proc. APGV*, Toulouse, France, 2011, pp. 7–14.

[34] M. Rubinstein, A. Shamir, and S. Avidan, “Multi-operator media retargeting,” *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. ID 23.

[35] Y.-F. Zhang, S.-M. Hu, and R. R. Martin, “Shrinkability maps for content-aware video resizing,” *Comput. Graph. Forum*, vol. 27, no. 7, pp. 1797–1804, 2008.

[36] D. Panozzo, O. Weber, and O. Sorkine, “Robust image retargeting via axis-aligned deformation,” *Comput. Graph. Forum*, vol. 31, no. 2, pp. 229–236, 2012.

[37] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, “A system for retargeting of streaming video,” *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. ID 126.

[38] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, “Motion-based video retargeting with optimized crop-and-warp,” *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art. ID 90.

[39] M. G. Kendall and B. B. Smith, “On the method of paired comparisons,” *Biometrica*, vol. 31, nos. 3–4, pp. 324–345, 1940.

[40] K. Zhang, Q. Liu, H. Song, and X. Li, “A variational approach to simultaneous image segmentation and bias correction,” *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1426–1437, Aug. 2015.

[41] Y. Yang *et al.*, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[42] X. Qin, J. Shen, X. Mao, X. Li, and Y. Jia, “Robust match fusion using optimization,” *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1549–1560, Aug. 2015.

[43] X. Li, L. Mou, and X. Lu, “Scene parsing from an MAP perspective,” *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1876–1886, Sep. 2015.

[44] D. Vaquero, M. Turk, K. Pulli, M. Tico, and N. Gelfand, “A survey of image retargeting techniques,” in *Proc. SPIE*, San Diego, CA, USA, 2010, pp. 953–956.

[45] A. Shamir, A. Sorkine-Hornung, and O. Sorkine-Hornung, “Modern approaches to media retargeting,” in *Proc. SIGGRAPH Asia Courses*, Singapore, 2012, pp. 534–547.

[46] F. Banterle *et al*., “Spatial image retargeting. In multidimensional image retargeting,” in *Proc. SIGGRAPH Asia Courses*, Seoul, Korea, 2011, pp. 853–856.

[47] Y. Han, X. Wei, X. Cao, Y. Yang, and X. Zhou, "Augmenting image descriptions using structured prediction output," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1665–1676, Oct. 2014.

[48] Y. Han, J. Zhang, Z. Xu, and S.-I. Yu, "Discriminative multi-task feature selection," in *Proc. AAAI*, Bellevue, WA, USA, 2013, pp. 41–43.

[49] L. Zhang *et al.*, "Perception-guided multimodal aesthetics discovery for photo quality assessment," in *Proc. ACM Multimedia*, 2014, pp. 237–246.

[50] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.

[51] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

Yingjie Xia (M'12) received the Ph.D. degree from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2009.

He was a Research Scientist with the National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, Champaign, IL, USA. He is currently an Associate Professor of Zhejiang University. His currrent research interests include transportation operation algorithms, parallel/distributed computing, and its applications in intelligent transportation systems.

Luming Zhang (M'14) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China.

He is currently a Faculty Member with the Hefei University of Technology, Hefei, China. His current research interests include visual perception analysis, image enhancement, and pattern recognition.

Richang Hong (M'12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008.

He is a Professor with the Hefei University of Technology, Hefei. He is a Research Fellow with the School of Computing, National University of Singapore, Singapore, from 2008 to 2010. His current research interests include multimedia content analysis and social media. He has co-authored over 70 publications in the above areas.

Dr. Hong was a recipient of the Best Paper Award in the ACM Multimedia 2010, the Best Paper Award in the ACM ICMR 2015, and the Honorable Mention of the IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award. He served as an Associate Editor for *Information Sciences* and *Signal Processing* (Elsevier), and the Technical Program Chair of the MMM 2016. He is a member of ACM and the Executive Committee Member of the ACM SIGMM China Chapter.

Liqiang Nie received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2009, and the Ph.D. degree from the National University of Singapore, Singapore, in 2013.

He was a Research Fellow with the School of Computing, National University of Singapore. Various parts of his work have been published in top forums, including ACM SIGIR, ACM SIGMM, TOIS, IIST and the IEEE TRANSACTIONS ON MULTIMEDIA. His current research interests include information retrieval and healthcare analytics.

Dr. Nie has served as a Reviewer for various journals and conferences.

Yan Yan received the Ph.D. degree from the University of Trento, Trento, Italy, in 2014.

He is currently a Post-Doctoral Researcher with the MHUG Group, University of Trento. His current research interests include machine learning and its application to computer vision and multimedia analysis.

Ling Shao (M'09–SM'10) received the B.Eng. degree in electronic and information engineering from the University of Science and Technology of China, Hefei, China, and the M.Sc. degree in medical image analysis and the Ph.D. (D.Phil.) degree in computer vision from the Robotics Research Group, University of Oxford, Oxford, U.K.

He was a Senior Lecturer with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., from 2009 to 2014, and a Senior Scientist with Philips Research, The Netherlands, from 2005 to 2009. He is the Chair of Computer Vision and the Head of the Computer Vision and Artificial Intelligence Group with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. He is an Advanced Visiting Fellow with the Department of Electronic and Electrical Engineering, University of Sheffield. He has edited three books and several special issues for journals, such as TNNLS and PR. He has authored/co-authored over 200 papers in refereed journals/conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, TNNLS, IJCV, ICCV, CVPR, IJCAI, and ACM MM, and holds over 10 EU/U.S. patents. His current research interests include computer vision, image/video processing, pattern recognition, and machine learning.

Dr. Shao is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, and several other journals. He has organized a number of international workshops with top conferences, including ICCV, ECCV, and ACM Multimedia. He is/was an Area Chair for WACV 2014, BMVC 2014/2015, and ICME 2015, has been serving as a Program Committee Member for many international conferences, including ICCV, CVPR, ECCV, BMVC, and ACM MM, and a Reviewer for many leading journals. He is a fellow of the British Computer Society and IET, and a Life Member of ACM.