Learning Sparse Representation for Objective Image Retargeting Quality Assessment

Qiuping Jiang, Feng Shao, *Member, IEEE,* Weisi Lin, *Fellow, IEEE,* and Gangyi Jiang, *Member, IEEE*

***Abstract*—The goal of image retargeting is to adapt source images to target displays with different sizes and aspect ratios. Different retargeting operators create different retargeted images, and a key problem is to evaluate the performance of each retargeting operator. Subjective evaluation is most reliable, but it is cumbersome and labor-consuming, and more importantly, it is hard to be embedded into online optimization systems. This paper focuses on exploring the effectiveness of sparse representation for objective image retargeting quality assessment. The principle idea is to extract distortion sensitive features from one image (e.g., retargeted image) and further investigate how many of these features are preserved or changed in another one (e.g., source image) to measure the perceptual similarity between them. To create a compact and robust feature representation, we learn two overcomplete dictionaries to represent the distortion sensitive features of an image. Features including local geometric structure and global context information are both addressed in the proposed framework. The intrinsic discriminative power of sparse representation is then exploited to measure the similarity between the source and retargeted images. Finally, individual quality scores are fused into an overall quality by a typical regression method. Experimental results on several databases have demonstrated the superiority of the proposed method.**

***Index Terms*—Global context information (GCI), image retargeting quality assessment (IRQA), local geometric structure (LGS), sparse representation.**

I. Introduction

OVER the past decades, the emerging progress in digital imaging technology along with the advance in display devices have imposed demands of media adaption to displays with different sizes or aspect ratios. Toward this end, many media retargeting algorithms have been proposed. Traditional media retargeting algorithms, such as linear scaling and manual cropping, usually have poor quality of experience due to serious geometric structure distortion and semantic information loss. Although many content-aware image retargeting operators have been developed [1]–[10], there is no one universal retargeting operator that can handle all kinds of images with diverse contents and structures. Therefore, designing objective quality metrics to faithfully evaluate the quality of retargeted images is particularly desirable.

Traditional full reference image quality assessment (FR-IQA) metrics [11]–[14] may encounter several significant issues when directly applying them into image retargeting quality assessment (IRQA). First, in traditional FR-IQA, the source and distorted images are normally assumed to be well-aligned. However, it is always opposed to the fact that the sizes of the source and retargeted images in IRQA are inconsistent. Second, traditional FR-IQA metrics mainly focus on evaluating the perceptual similarities between a source image and its corresponding nongeometrically distorted version. Nevertheless, geometric distortion is a significant issue in image retargeting, which makes the problem of IRQA to be quite different with general image quality assessment (IQA). Third, in traditional IQA, the involved image distortions, such as noise, blurring, and compression artifacts, mostly lead to intensity change and the substantial difference can be measured by subtraction between two images. While in IRQA, it is impossible to measure the undergone artificial retargeting modification on images simply by subtraction. In fact, the design of IRQA metrics that can evaluate image contents under varying sizes or aspect ratios is extremely challenging. One critical ingredient to the success of IRQA is to build benchmark databases for performance validation. Recently, several IRQA databases, such as RetargetMe [15], CUHK [16], and NRID [17], have been created. Due to the availability of these databases, recent years have witnessed some progresses in IRQA.

Currently, the vast majority of IRQA metrics share a general evaluation architecture: 1) establishing the interimage dense correspondence and 2) estimating the distance between the matched pixels/regions as the quality measure. For example, Simakov *et al.* [18] proposed a bidirectional similarity (BDS) metric that seeks one image to identify the counterpart of an image patch in the other image by using the minimal intensity-related error as the search criteria. Pele and Werman [19]

Manuscript received October 2, 2016; revised January 3, 2017 and March 5, 2017; accepted March 30, 2017. This work was supported in part by the Natural Science Foundation of China under Grant 61622109 and Grant 61271021, in part by the Scientific Research Foundation of Graduate School of Ningbo University, and in part by K. C. W. Magna from Ningbo University. This paper was recommended by Associate Editor Y. Yuan. *(Corresponding author: Feng Shao.)*

Q. Jiang is with the School of Information Science and Engineering, Ningbo University, Ningbo 315211, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jqp910707@126.com).

F. Shao and G. Jiang are with the School of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: shaofeng@nbu.edu.cn; jianggangyi@nbu.edu.cn).

W. Lin is with the Centre for Multimedia and Network Technology, School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2690452

proposed an earth mover's distance (EMD) metric based on the minimal cost that must be paid to transform one distribution into the other. Liu *et al.* [20] developed a robust scale-invariant feature transform flow (SIFT-flow) matching algorithm to generate dense correspondence map between two images. The similarity is simply estimated on the matched SIFT key points. Manjunath *et al.* [21] found that edge distortion is particularly critical to IRQA and presented a low-level edge histogram (EH) metric for similarity evaluation by comparing the spatial distribution between local edges in two images. Liu *et al.* [22] proposed an objective IRQA metric based on global geometric structures and local pixel correspondence. Recently, Hsu *et al.* [17] found that geometric distortion and information loss are two major distortion types affecting the retargeting quality. The perceptual geometric distortion and information loss in a retargeted image are measured based on the SIFT-flow algorithm. Fang *et al.* [23] devised a simple yet effective IRQA approach called IR-SSIM by creating an SSIM quality map to indicate how the structure information of each pixel in a source image is preserved in the retargeted image. Similarly, the pixel correspondence is also created by the SIFT-flow algorithm. Most recently, Zhang *et al.* [24] investigated the retargeted images correspondence estimation and developed an aspect ratio similarity metric by evaluating the local block quality changes.

The performances of the aforesaid IRQA metrics are highly dependent on the accuracy of the used dense correspondence algorithms. Since dense correspondence between two images (especially for images with different sizes or aspect ratios) with high accuracy remains as open issues in computer vision [25], existing dense correspondence-based IRQA metrics may lead to unsatisfactory performance. For example, for images with lots of repeated texture patterns or very smooth areas, the SIFT-flow algorithm cannot work well for these areas since it may suffer from incorrect correspondences. Generally, inaccurate SIFT-flow correspondences for extremely smooth areas do not have much impact on the perceptual quality, while for those texture patterns, such inaccuracy really does matter. Based on such consideration, in this paper we attempt to address the problem of IRQA that does not need to create interimage dense correspondence, so that the issue of incorrect correspondence can be avoided. For this purpose, we opt to measure the similarity between the source and retargeted images from a global perspective. The basic assumption is that quality degradation in image retargeting process can be quantified by the fidelity in terms of some specific distortion sensitive features. Based on this assumption, the problem of IRQA is decomposed into two modules: 1) distortion sensitive feature extraction and 2) feature similarity measurement. How to extract effective features and how to design powerful feature similarity metric are the key problems to be addressed in this paradigm.

In this paper, we propose a novel objective IRQA metric that first extracts distortion sensitive features from an image and then investigates how many of these features are preserved or changed in another image. To specify, we first learn two overcomplete dictionaries from an image to represent its corresponding distortion sensitive features. These two learned dictionaries are used to represent the local geometric structure (LGS) and global context information (GCI) of an image, respectively. Then, for each feature modality (taking LGS and GCI as modality information), by concatenating the dictionaries from the retargeted and source images, a joint overcomplete dictionary is derived for sparse representation. For feature similarity measurement, we decompose an image (the source or retargeted image) via sparse representation with respect to the previously derived joint dictionary and measure the similarity by taking advantage of the discriminative power of sparse representation. Overall, the main contributions of this paper are summarized as follows.

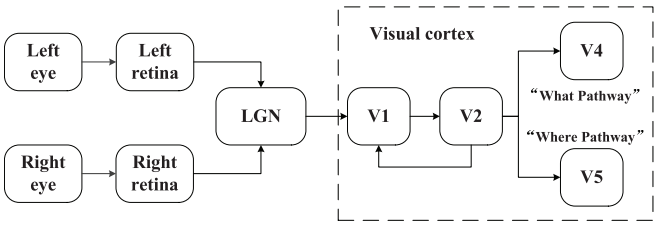


Fig. 1. Hierarchical structure of visual cortex in human brain.

1) Instead of creating dense correspondence between the source and retargeted images, we decompose the problem of IRQA into two separate modules including distortion sensitive feature extraction and feature similarity measurement.
2) To create a compact and robust feature representation, we learn two overcomplete dictionaries to represent the distortion sensitive features of an image. The intrinsic discriminative power of sparse representation is further exploited for similarity measurement.
3) Extensive experiments are conducted on three benchmark datasets and the experimental results demonstrate the proposed metric can well evaluate the retargeting quality. To our best knowledge, it is the first attempt that has conducted such comprehensive validations on all these datasets.

The rest of this paper is organized as follows. In Section II, we illustrate some related works and techniques. In Section III, we detail the proposed method. In Section IV, we show the experimental results and also discuss the limitations of our method. Finally, the conclusions are drawn in Section V.

II. Related Work

A. Discrepancies Between IRQA and Traditional IQA

From the visual physiology viewpoint, a powerful solution for quality assessment is to simulate the behavior of visual perception. Thus, it is necessary to understand the structure and processing mechanism of human visual system (HVS) as possible. Previous discoveries have revealed that visual cortex is highly hierarchical [26]. As shown in Fig. 1, the visual information of an image that projected onto the retina is first transmitted to the lateral geniculate nucleus (LGN). After processing by the LGN and primary visual cortex ($V1$), the processed information will be sent to two separate pathways diverged from $V2$, known as ventral and dorsal streams, respectively. The ventral stream generally follows the path

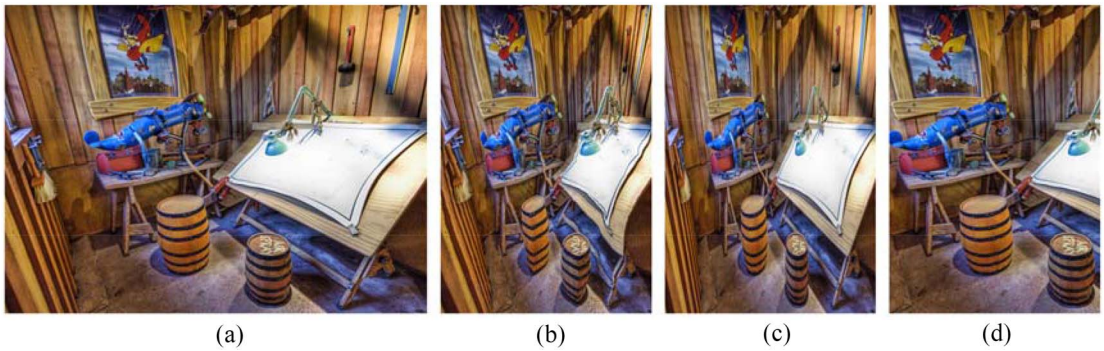


Fig. 2. Examples of typical distortions during image retargeting: LGS distortion and GCI loss. (a) Source image "ArtRoom." (b)–(d) Created retargeted images of (a) by applying different retargeting operators with a scaling ratio of 0.5. Retargeted image by (b) seam carving, (c) linear scaling, and (d) manual cropping.

$V1\rightarrow$ $V2\rightarrow$ $V4\rightarrow$ temporal lobe and is also known as "what pathway," as the processing is greatly implicated with shape recognition and object representation. On the contrary, the dorsal stream follows the path $V1\rightarrow$ $V2\rightarrow$ $V5\rightarrow$ parietal lobe and is known as "where pathway," which is relevant with motion computation, object location, and trajectory.

It is known that $V1$ is mainly responsible for the perception of general image distortions, such as noise, blurring, and compression artifacts [27]. That is, simulating receptive fields (RFs) of simple and complex cells in $V1$ will contribute greatly to the success of traditional IQA task. Many IQA metrics have been proposed from the perspective of simulating RFs in area $V1$ [27]–[31]. However, as described before, the undergone artificial retargeting modifications are quite different with the distortions involved in traditional IQA task. Fig. 2 shows some typical distortions evoked by image retargeting: LGS distortion and GCI loss. Fig. 2(b)–(d) shows the retargeted images with a scaling ratio of 0.5[1] generated by seam carving [1], linear scaling, and manual cropping, respectively. Obviously, the dominant factor of distortions in Fig. 2(b) and (c) is the LGS distortion, while in Fig. 2(d) it is the GCI loss because the salient regions (e.g., the black board) are partially discarded. Once the salient regions are cropped, visual perceptions will be inevitably affected and changed. Due to the different types of distortions, different visual areas in human brain will play different roles in quality assessment tasks. In this process, only using the $V1$-inspired visual features may not work well to reflect the visual perception of the retargeting-related distortions. Therefore, how to understand the role of visual cortex in perceiving and responding the retargeting-related distortions deserves to be further explored.

Recent studies in visual cognition have revealed that area $V4$ plays an important role in shape recognition and the neurons in $V4$ exhibit high sensitivity for the orientation and curvature of boundary fragments [32]. Based on these findings, we believe that the perception behaviors responding to the geometric distortions [as shown in Fig. 2(b) and (c)] are mainly occurred in area $V4$. Meanwhile, area $V4$ also shows strongly selective visual attention over spatial location in a scene [32]. As known, visual attention is an important mechanism of HVS to efficiently understand the context information, which is also another important aspect in IRQA [as shown in Fig. 2(d), the salient regions are partially discarded]. All these physiological evidences indicate that area $V4$ is highly implicated with the visual perception of LGS distortion and GCI loss in retargeting.

[1]In this paper, a scaling ratio of 0.5 indicates that a source image is scaled to a resolution on either horizontal or vertical direction with a ratio of 0.5.

B. Sparse Representation-Based Classification

The goal of sparse representation is to seek compact representations for signals and it is primarily suitable for denoising [33], super-resolution [34], and visual tracking [35]. Recently, the representative work on face recognition has showed that sparse representation has intrinsic discriminative power [36]. Theoretically, it only selects those basis elements that can most compactly represent a signal and therefore it is powerful for classification. Given a dictionary corresponding to the cth class $\mathbf{D}_c = [\mathbf{d}_{c,1}, \mathbf{d}_{c,2}, \ldots, \mathbf{d}_{c,n}]$, a test sample $\mathbf{x}$ from class c can be well represented by $\mathbf{D}_c$ along with sparse coefficients $\mathbf{z}_c$. Particularly, $\mathbf{D}_c$ can be generated by directly combining the raw training samples or be learned from a set of training samples. Previous studies have found that the learned overcomplete dictionaries are more effective for sparse representation-based classification and more efficient in l_1-norm minimization [37]. Since the label for $\mathbf{x}$ is unknown, $\mathbf{z}_c$ can be estimated based on the dictionary integrated from training samples of all k classes by

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \|\mathbf{z}\|_1, \text{ subject to } \|\mathbf{D}\mathbf{z} - \mathbf{x}\|_2^2 \leq \xi \tag{1}$$

where $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k]$ is the combination of k sub-dictionaries from all classes, and $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k]$ is the sparse coefficient matrix. Then, the class label for $\mathbf{x}$ is determined by projecting the test sample on each class with the minimum reconstruction error, that is

$$\hat{c} = \arg\min_{c} \|\mathbf{D}\delta_c(\mathbf{z}) - \mathbf{x}\|_2^2 = \arg\min_{c} \|\mathbf{D}_c\mathbf{z}_c - \mathbf{x}\|_2^2 \tag{2}$$

where $\delta_c(\cdot)$ is a vector indicator that only keeps the elements associated with the cth class.

C. Motivation for IRQA

Based on the above discussions, we are motivated to address the IRQA by simultaneously accounting for the physiological evidences in area $V4$ and taking advantages

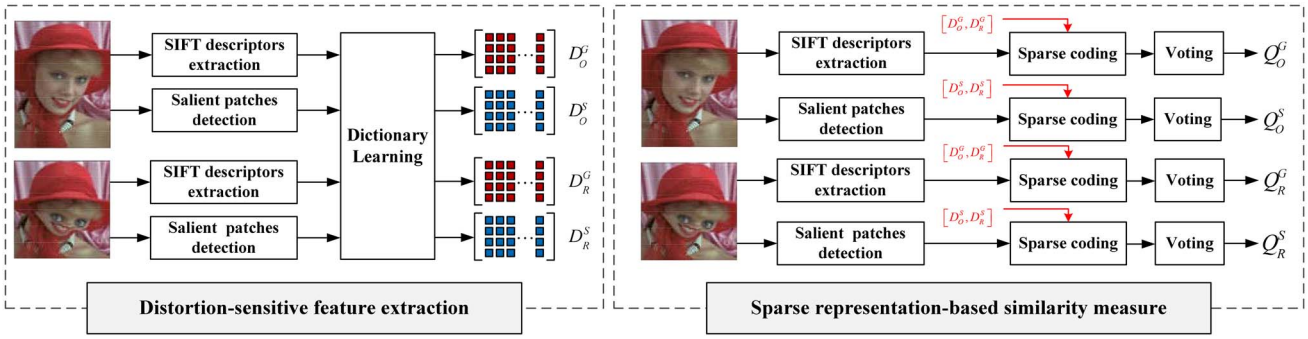


Fig. 3. Framework for the proposed sparse representation-based IRQA method.

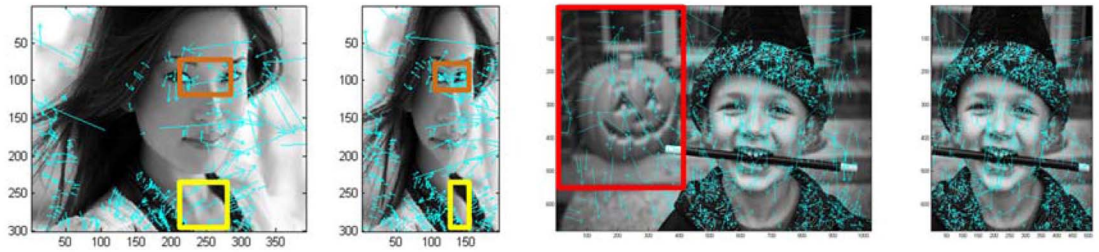


Fig. 4. Example of the SIFT key points detection results. (a) and (c) Source images. (b) and (d) Retargeted images of (a) and (c), respectively. Best viewed in color.

of the sparse representation-based classification technique. Specifically, overcomplete dictionaries are learned to interpret the distortion-sensitive features, so that the LGS distortion and GCI loss can be well characterized. Furthermore, inspired by the intrinsic discriminative power of sparse representation, the similarity between the source and retargeted images is measured based on sparse representation-based classification paradigm. To be more specific, if most of contents in a retargeted image can be better represented by the dictionary learned from the retargeted image itself rather than the one learned from the source image, the two images are visually different (having poor visual quality). Otherwise, the retargeted image has good visual quality.

III. Sparse Representation-Based IRQA

Our proposed sparse representation-based IRQA framework is depicted in Fig. 3. In the method, how to build effective distortion-sensitive feature representation and how to make full use of sparse representation for similarity measurement are the key challenges to the final success. Considered the roles of LGS distortion and GCI loss in IRQA and also inspired by the physiological evidences in area $V4$, in this paper we extract SIFT feature descriptors and salient image patches to represent the distortion-sensitive information contained in an image. In order to obtain a compact and robust feature representation, we further learn overcomplete dictionaries for each feature modality (taking SIFT descriptors and salient patches as modality information) as the final distortion-sensitive features. Then, we utilize the intrinsic discriminative power of sparse representation to represent an image over the learned dictionaries, and then apply a simple voting scheme to derive a similarity score between the source and retargeted images. Finally, different qualities are fused into an overall quality score using support vector regression (SVR). In what follows, we elaborate on each step of our proposed method.

A. Distortion-Sensitive Feature Extraction

In this section, we extract SIFT feature descriptors and context-aware (CA) salient patches for distortion-sensitive feature representation. The extracted SIFT feature descriptors are shown to be beneficial to capture the LGS of an image, while the extracted CA salient patches are useful to characterize the GCI of an image. Note that both LGS and GCI are two important and complementary aspects in measuring the visual quality of the retargeted images [17].

1) SIFT Key Point Detection: In [38], each SIFT key point is represented by a 128-dimensional feature vector. Generally, thousands of such key points can be detected from an image. In Fig. 4, we show an example of SIFT key point detection results for both source and retargeted images. From the figure, it is quite evident that the SIFT key point detection results are different for the source and retargeted images. More importantly, the differences mainly appear on the regions with locally geometric deformation [marked by yellow and red rectangles in Fig. 4(a) and (b)]. While in Fig. 4(c) and (d), the differences mainly appear on the regions only contained in the source image (these regions are discarded in the retargeted image, hereby changing the intrinsic structure of the source image). The example demonstrates that the extracted SIFT key points are effective to capture the LGS information of an image. Although the extracted SIFT key points are not accurate enough in practice [e.g., the chin is seriously

deformed in Fig. 4(b), while the extracted SIFT key points fail to capture this distortion], similarity measurement from the SIFT key points are still expected to provide a reasonable estimation of LGS distortion between the source and retargeted images.

Given an image, let $\mathbf{G} = \{\mathbf{f}_i^G\}_{i=1}^N$ denote the SIFT feature descriptor set with each column $\mathbf{f}_i^G \in \mathbb{R}^{128\times 1}$ representing the SIFT feature vector of the ith key point, and N is the number of SIFT key points. All the extracted SIFT feature descriptors will be used for subsequent dictionary learning.

2) Salient Patch Detection: As mentioned above, the differences of the extracted SIFT key points in Fig. 4(c) and (d) mainly appear on the regions that are only contained in the source image. From the viewpoint of scene context, these discarded regions will inevitably change the initial semantic information that the photographers want to convey, leading to context information loss. This observation naturally indicates that the GCI loss can also be an important factor deteriorating the quality of a retargeted image. However, how to characterize the GCI of an image remains an open issue in computer vision. In this paper, we resort to select salient patches from an image using state-of-the-art saliency detection models [39]–[43]. The definition of salient patches in this paper is similar with the terminology of region of interest in [44]. These salient patches are to be perceived and processed by HVS more finely, so as to provide richer high level context information.

Instead of generating a bi-level map in [44], the salient patch selection in this paper is guided by estimated saliency map. In the literature, a variety of saliency detection models have been proposed for different applications [39]–[43]. Compared with those saliency models that focus on detecting a single salient object, saliency maps created by the CA method [39] are more suitable for context-related applications. Examples are shown in Fig. 5. Fig. 5(b) provides context descriptions for corresponding scenes in Fig. 5(a). In general, humans tend to describe the scene context rather than only the most salient object. Compared with the results that only extract the most salient single object shown in Fig. 5(c), the CA saliency maps shown in Fig. 5(d) better comply with the context descriptions and can highlight most meaningful regions. However, these salient regions may be discarded when a large adjustment is applied to the source image, making it difficult to understand the initial context information.

The procedures of salient patch selection are summarized as follows. First, the CA method is applied to estimate the saliency map (Fig. 6 shows an example of CA-based salient detection results). Then, guided by the estimated CA saliency map of an image, a certain 8×8 image patch will be selected if the corresponding patch-level saliency value (computed by the mean saliency value of all the pixels within the patch) is among the top 70% over all the patches in the image (the impact of different patch sizes and patch selection percentages will be analyzed in Section IV-C). Given an image, all the selected patches finally result in a salient patch set which will be used for dictionary learning to obtain a GCI-aware dictionary.

Let $\mathbf{S} = \{\mathbf{f}_i^S\}_{i=1}^K$ be the salient patch set with each column $\mathbf{f}_i^S \in \mathbb{R}^{192\times 1}$ representing the column vector by scanning the intensity values in the ith patch row-by-row over three image channels (RGB color space is used in this paper), and K is the number of selected salient patches. Instead of using other color spaces and texture descriptors such as local binary pattern, we directly use the RGB color space due to its good capacity in appearance representation.

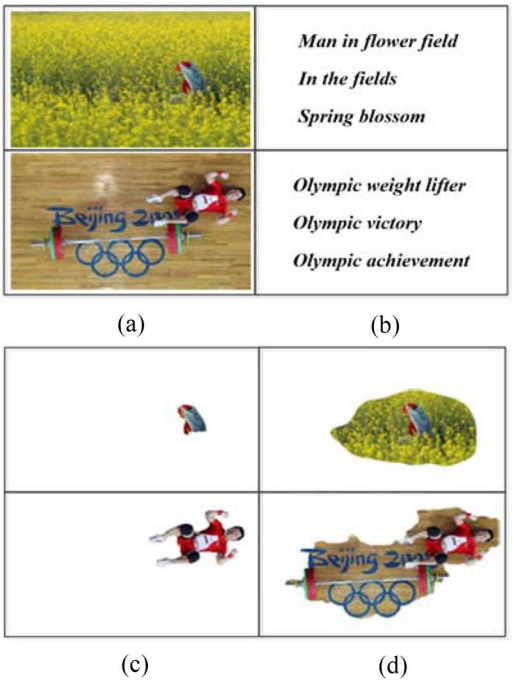


Fig. 5. CA saliency versus single salient object detection. (a) Input scenes. (b) Human subjective description. (c) Single salient object detection. (d) CA saliency [38].

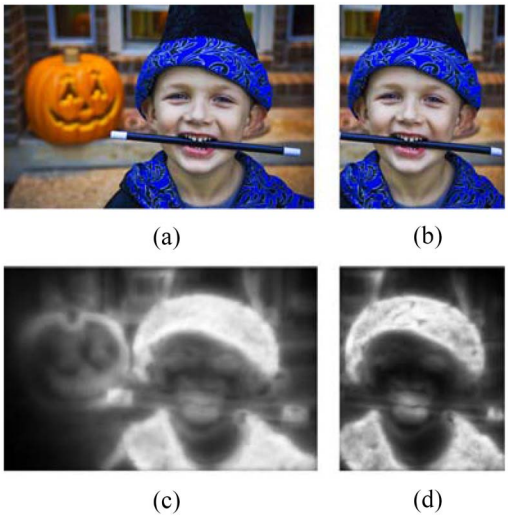


Fig. 6. Example of CA-based saliency maps. (a) Input image. (b) Retargeted image of (a). (c) Saliency map of (a). (d) Saliency map of (b).

3) Overcomplete Dictionary Learning: Recall the effort we make in this paper is to explore the effectiveness of sparse representation for IRQA, while the performance of sparse representation-based classification technique is highly dependent on the construction of an overcomplete dictionary. Here, we opt to learn overcomplete dictionaries for each feature modality (LGS or GCI) so as to obtain a compact and

robust feature representation. Taking the LGS modality as an example, given $\mathbf{G} = \{\mathbf{f}_i^G\}_{i=1}^N \in \mathbb{R}^{128\times N}$ as input, we aim to simultaneously learn an overcomplete dictionary with M basic atoms $\mathbf{D}^G = [\mathbf{d}_1^G, \mathbf{d}_2^G, \ldots, \mathbf{d}_M^G] \in \mathbb{R}^{128\times M}$ and their corresponding sparse coefficient matrix $\mathbf{X}^G = \{\mathbf{x}_i^G\}_{i=1}^N \in \mathbb{R}^{M\times N}$ by seeking a sparse representation for each feature vector under specific sparsity constraint τ. Formally, this problem can be formulated as [45]

$$\left\{\mathbf{D}^G, \mathbf{X}^G\right\} = \min \sum_{i=1}^{N} \left\|\mathbf{f}_i^G - \mathbf{D}^G\mathbf{x}_i^G\right\|_2^2 \quad \text{s.t. } \forall i, \left\|\mathbf{x}_i^G\right\|_0 \le \tau \tag{3}$$

where $\mathbf{x}_i^G$ is the sparse coefficient vector of $\mathbf{f}_i^G$, $\|\|_2$ is the l_2-norm operator, $\|\|_0$is the ℓ_0-norm operator indicating the number of nonzero elements in a vector, and τ is a predefined most desired number of nonzero coefficients. Although the ℓ_0-norm gives a straightforward measurement of sparsity, the use of ℓ_0-norm sparsity constraint makes this problem NP-hard. Recent literature has demonstrated that the solution optimized by the ℓ_1-norm minimization constraint is equivalent to the solution obtained by ℓ_0-norm minimization with full probability [46]. Therefore, the above optimization problem is rewritten as

$$\{\mathbf{D}^G, \mathbf{x}^G\} = \arg\min\left[\sum_{i=1}^{N}\left\|\mathbf{f}_i^G - \mathbf{D}^G\mathbf{x}_i^G\right\|_2^2 + \lambda\left\|\mathbf{x}_i^G\right\|_1\right] \tag{4}$$

where λ is a small positive parameter to balance the relative importance between the reconstruction residual and sparsity constraint terms. By replacing the ℓ_0 term with a l_1 penalty, the NP-hard problem defined in (3) can be transformed into a convex one. We apply the online dictionary learning (ODL) algorithm implemented in the SPArse Modeling Software [47] to solve (4), obtaining an LGS-aware dictionary (denoted by $\mathbf{D}^G \in \mathbb{R}^{128\times M}$). Similarly, a GCI-aware dictionary (denoted by $\mathbf{D}^S \in \mathbb{R}^{192\times L}$) can be learned by taking $\mathbf{S} = \{\mathbf{f}_i^S\}_{i=1}^K \in \mathbb{R}^{192\times K}$ as input.

Note that the learned dictionaries ($\mathbf{D}^G$ and $\mathbf{D}^S$) are overcomplete, containing M basic atoms as the column vectors in $\mathbf{D}^G$ and $\mathbf{D}^S$. Thus, each feature vector $\mathbf{f}_i^G$ (or $\mathbf{f}_i^S$) can be sparsely represented as a linear combination of the basic atoms contained in $\mathbf{D}^G$ (or $\mathbf{D}^S$), leading to a compact and robust feature representation.

B. Feature Similarity Measurement

1) Sparse Representation Based on Joint Dictionary: Let $\mathbf{G}_R = \{\mathbf{f}_{R,i}^G\}_{i=1}^{N_1}$ and $\mathbf{G}_O = \{\mathbf{f}_{O,j}^G\}_{j=1}^{N_2}$ denote two SIFT feature sets, where $\mathbf{f}_{R,i}^G \in \mathbb{R}^{128\times 1}$ and $\mathbf{f}_{O,j}^G \in \mathbb{R}^{128\times 1}$ represent the SIFT feature vectors of the ith key point in the retargeted image I_R and the jth key point in its associated source image I_O, respectively, and N_1 and N_2 be the number of SIFT key points in I_R and I_O, respectively. By applying the ODL algorithm on $\mathbf{G}_R$ and $\mathbf{G}_O$, we obtain the LGS-aware dictionaries $\mathbf{D}_R^G \in \mathbb{R}^{128\times M_1}$ and $\mathbf{D}_O^G \in \mathbb{R}^{128\times M_2}$ for I_R and I_O, respectively, where M_1 and M_2 are the number of the atoms in $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$, respectively. Typically, we have

$$\mathbf{f}_{R,i}^G \doteq \mathbf{D}_R^G \cdot \mathbf{x}_{R,i}^G \tag{5}$$

$$\mathbf{f}_{O,j}^G \doteq \mathbf{D}_O^G \cdot \mathbf{x}_{O,j}^G \tag{6}$$

where $\mathbf{x}_{R,i}^G \in \mathbb{R}^{M_1\times 1}$ and $\mathbf{x}_{O,j}^G \in \mathbb{R}^{M_2\times 1}$ denote the sparse coefficient vectors for $\mathbf{f}_{R,i}^G$ and $\mathbf{f}_{O,j}^G$, respectively. Obviously, if $\mathbf{f}_{R,i}^G$ and $\mathbf{f}_{O,j}^G$ are matched, $\mathbf{f}_{O,j}^G$ can be well sparsely represented by $\mathbf{D}_R^G$, and inversely $\mathbf{f}_{R,i}^G$ can be well sparsely represented by $\mathbf{D}_O^G$.

We cast the feature similarity measurement between I_R and I_O into a problem of quantifying how much the distortion-sensitive feature present in one image is preserved in the another one. For this, the classical sparse representation-based classification technique is used. By using a joint dictionary $\mathbf{D}_{R,O}^G$, sparse representation performed in representing each SIFT feature vector $\mathbf{f}_{R,i}^G$ is formulated as follows:

$$\hat{\mathbf{x}}_{R,i}^G = \arg\min\left[\sum_{i=1}^{N_1}\left\|\mathbf{f}_{R,i}^G - \mathbf{D}_{R,O}^G\mathbf{x}_{R,i}^G\right\|_2^2 + \lambda\left\|\mathbf{x}_{R,i}^G\right\|_1\right] \tag{7}$$

where $\mathbf{D}_{R,O}^G = [\mathbf{D}_R^G, \mathbf{D}_O^G] \in \mathbb{R}^{128\times(M_1+M_2)}$ is the joint dictionary by concatenating $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$, and $\hat{\mathbf{x}}_{R,i}^G \in \mathbb{R}^{(M_1+M_2)}$ is the sparse coefficient vector of $\mathbf{f}_{R,i}^G$ represented by $\mathbf{D}_{R,O}^G$. In this paper, the batch orthogonal matching pursuit (batch-OMP) algorithm [48] is employed to solve the problem. Note that the batch-OMP algorithm provides an efficient solution for the many-input/single dictionary sparse representation problem.

According to the sparse representation-based classification technique in [36], it is not difficult to know that the nonzero elements in $\hat{\mathbf{x}}_{R,i}^G$ will be highly concentrated on $\mathbf{D}_R^G$ when I_R and I_O are visually different. That is, in this case, dictionary $\mathbf{D}_R^G$ is more suitable than $\mathbf{D}_O^G$ to represent $\mathbf{f}_{R,i}^G$. However, this conclusion is no longer true when I_R and I_O are visually similar since their corresponding learned dictionaries are also similar in this case, making the involved atoms spread over all the elements in $\mathbf{D}_{R,O}^G$. To meet the demand for similarity measurement, as an inverse case, we expect that more atoms in $\mathbf{D}_O^G$ will be selected and used to represent $\mathbf{f}_{R,i}^G$ when I_R and I_O are visually similar. To achieve this goal, three basic rules are designed to tune the parameters involved in the ODL algorithm for learning these two dictionaries [49].

Rule 1: The number of atoms in $\mathbf{D}_O^G$ should be more than that in $\mathbf{D}_R^G$, i.e., $M_2 > M_1$.

Rule 2: The sparsity for learning $\mathbf{D}_O^G$ should be larger than that for learning $\mathbf{D}_R^G$, i.e., $\lambda_2 > \lambda_1$.

Rule 3: The number of iterations in the ODL algorithm for learning $\mathbf{D}_O^G$ should be more than that for learning $\mathbf{D}_R^G$.

Based on these rules, when I_R and I_O are visually similar, the l_1-minimization solver for (4) will be enforced to seek more primitives from $\mathbf{D}_O^G$ than $\mathbf{D}_R^G$ to represent $\mathbf{f}_{R,i}^G$. Meanwhile, when I_R and I_O are visually different, the sparse coefficients of $\mathbf{f}_{R,i}^G$ will still be gathered on the primitives associated with $\mathbf{D}_R^G$ which is learned from I_R itself.

2) Sparse Reconstruction Residual-Based Voting: Based on the above analyses, we believe that the similarity between I_R and I_O can be measured by the percentage of the amount that $\hat{\mathbf{x}}_{R,i}^G$ is associated with $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$, respectively. Specifically,

we compute the residuals by only using the atoms from $\mathbf{D}_R^G$ or $\mathbf{D}_O^G$ to reconstruct $\mathbf{f}_{R,i}^G$, to distinguish which subdictionary is more suitable for representing $\mathbf{f}_{R,i}^G$. Toward this end, we define a vector indicator $\delta(\cdot) : \mathbb{R}^{(M_1+M_2)} \to \mathbb{R}^{(M_1+M_2)}$ that only preserves the sparse coefficients corresponding to one subdictionary unchanged while sets all other elements to zero. For example, given $\hat{\mathbf{x}}_{R,i}^G$ as input, $\delta_R(\hat{\mathbf{x}}_{R,i}^G)$ generates a new coefficient vector in which only the entries associated with $\mathbf{D}_R^G$ are remained unchanged while the other coefficients are set to zero, and $\delta_O(\hat{\mathbf{x}}_{R,i}^G)$ generates a new coefficient vector in which only the entries associated with $\mathbf{D}_O^G$ are remained unchanged. Therefore, the residuals only using the atoms from $\mathbf{D}_R^G$ or $\mathbf{D}_O^G$ for reconstructing $\mathbf{f}_{R,i}^G$ are computed as

$$E_R\left(\mathbf{f}_{R,i}^G\right) = \left\|\mathbf{f}_{R,i}^G - \mathbf{D}_{R,O}^G \cdot \delta_R\left(\hat{\mathbf{x}}_{R,i}^G\right)\right\|_2^2 \tag{8}$$

$$E_O\left(\mathbf{f}_{R,i}^G\right) = \left\|\mathbf{f}_{R,i}^G - \mathbf{D}_{R,O}^G \cdot \delta_O\left(\hat{\mathbf{x}}_{R,i}^G\right)\right\|_2^2 \tag{9}$$

where $E_R(\mathbf{f}_{R,i}^G)$ and $E_O(\mathbf{f}_{R,i}^G)$ denote the sparse reconstruction residuals associated with $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$, respectively. Based on the reconstruction residuals, a simple voting strategy is then performed [49]

$$U_R\left(\mathbf{f}_{R,i}^G\right) = 1, \text{ if } E_O\left(\mathbf{f}_{R,i}^G\right) \geq E_R\left(\mathbf{f}_{R,i}^G\right) \tag{10}$$

$$U_O\left(\mathbf{f}_{R,i}^G\right) = 1, \text{ if } E_O\left(\mathbf{f}_{R,i}^G\right) < E_R\left(\mathbf{f}_{R,i}^G\right). \tag{11}$$

Taking all SIFT feature vectors of I_R into account ($\mathbf{G}_R = \{\mathbf{f}_{R,i}^G\}_{i=1}^{N_1}$), the probabilities of the votes corresponding to $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$ are then computed as

$$V_{R,R}^G = \frac{1}{N_1}\sum_{i=1}^{N_1} U_R\left(\mathbf{f}_{R,i}^G\right) \tag{12}$$

$$V_{R,O}^G = \frac{1}{N_1}\sum_{i=1}^{N_1} U_O\left(\mathbf{f}_{R,i}^G\right) \tag{13}$$

where $V_{R,R}^G$ and $V_{R,O}^G$ denote the probabilities of the votes corresponding to $\mathbf{D}_R^G$ and $\mathbf{D}_O^G$, respectively, $0 \leq V_{R,R}^G, V_{R,O}^G \leq 1$, and $V_{R,R}^G + V_{R,O}^G = 1$.

It should be emphasized that a higher value of $V_{R,R}^G$ implies that I_R and I_O are visually much different in terms of the LGS, while a higher value of $V_{R,O}^G$ implies I_R and I_O are visually much similar (as discussed in Section III-B1). Based on such voting strategy, we define the similarity between I_R and I_O as

$$Q_R^G = \frac{V_{R,O}^G - V_{R,R}^G + 1}{2} \tag{14}$$

where $Q_R^G \in [0, 1]$ is an LGS-aware quality score indicating how much the LGS-aware descriptors presented in the retargeted image I_R can be extracted from its source image I_O.

In order to achieve a more robust similarity measurement, the above implemented sparse coding and voting processes are repeatedly performed on the source image I_O so as to obtain another LGS-aware quality score $Q_O^G \in [0, 1]$, indicating how much the LGS-aware descriptors presented in the source image I_O can be extracted from its retargeted image I_R. It should be noticed that, in this case, $\mathbf{D}_R^G$ should be enforced to be finer than $\mathbf{D}_O^G$ in the dictionary learning stage. Details about the sparse reconstruction residual computation and voting process performed on I_O are not repeated here. Finally, by considering all the SIFT feature vectors of I_O ($G_O = \{\mathbf{f}_{O,j}^G\}_{j=1}^{N_2}$), the probabilities of the votes corresponding to $\mathbf{D}_O^G$ and $\mathbf{D}_R^G$ are similarly computed as

$$V_{O,O}^G = \frac{1}{N_2}\sum_{j=1}^{N_2} U_O\left(\mathbf{f}_{O,j}^G\right) \tag{15}$$

$$V_{O,R}^G = \frac{1}{N_2}\sum_{j=1}^{N_2} U_R\left(\mathbf{f}_{O,j}^G\right) \tag{16}$$

and the corresponding quality score is similarly computed as

$$Q_O^G = \frac{V_{O,R}^G - V_{O,O}^G + 1}{2}. \tag{17}$$

Also, the GCI-aware quality scores can be estimated by the above implemented process. We denote these two quality scores by Q_R^S and Q_O^S, respectively. As a result, given a retargeted image I_R and its associated source image I_O, we can obtain four quality scores (i.e., Q_R^G, Q_O^G, Q_R^S, and Q_O^S) to quantify the visual quality of I_R.

C. Quality Score Fusion

The final quality score of a retargeted image is estimated based on a pretrained predict function $f(\cdot)$. That is, the final quality score is given by

$$Q_f = f(\mathbf{Q}) \tag{18}$$

where $f(\cdot)$ is trained in advance using ε-sensitive SVR (ε-SVR) [50], and $\mathbf{Q} = [Q_R^G, Q_O^G, Q_R^S, Q_O^S]$. Here, $f(\cdot) : \mathbb{R}^4 \to \mathbb{R}^1$ takes $\mathbf{Q}$ as input and produces the output $Q_\mathbf{f}$ as a final quality score. Of course, other machine learning techniques can also be used here. In addition, we find the best performance is obtained by using SVR with polynomial kernel and radial basis function (RBF) kernel while the polynomial kernel-based model consumes less time costs than the one with RBF kernel (the impact of different fusion schemes will be analyzed in Section IV-F).

IV. Performance Evaluation

A. Benchmark Databases

In this paper, three benchmark databases including RetargetMe [15], CUHK [16], and NRID [17], are used for performance evaluation in our experiments. All these databases contain retargeted images and the associated subjective rating scores. The basic information of these datasets is summarized in Table I and a brief introduction is also described as follows.

1) RetargetMe: The RetargetMe database is the first released benchmark database for IRQA, which contains 37 source images. Eight retargeted results are generated for each source image by using eight different retargeting operators: 1) seam carving (SEAM) [1]; 2) nonhomogeneous warping (WARP) [2]; 3) scale and stretch (SCST) [4]; 4) multi operator (MULTI) [5]; 5) homogeneous scaling (SCAL); 6) shift map (SHIF) [6]; 7) manual cropping (CROP); and 8) streaming

TABLE I
BASIC INFORMATION OF THE USED BENCHMARK DATABASES

Datasets	RetargetMe [15]	CUHK [16]	NRID [17]
Resizing Ratio	0.5; 0.75	0.5; 0.75	0.75
Source Image No.	37	57	35
Retargeted Image No.	296	171	175
Subject No.	210	30	30
Retargeting Operator No.	8	10	5
Subjective Score Type	Pair-wise	MOS	Pair-wise

video (STVI) [7]. As a result, a total number of 296 retargeted images are created. The subjective study is conducted in a paired comparison manner [51]. Subjects are shown two retargeted images obtained by two different retargeted operators performed on a same source image, and are requested to vote for the better quality one. The number of times that a retargeted image is favored over other retargeted images is recorded as the subjective rating score.

2) CUHK: The CUHK database contains 57 source images. Three retargeted results are generated for each source image by using three different retargeting operators. For each source image, the applied retargeting operators may be different which are randomly selected from ten representative methods, including the eight methods used in RetargetMe database and other two operators namely optimized seam carving and scale [8] and energy-based deformation [3]. As a result, a total number of 171 retargeted images are included in the database. As for subjective study, different from the paired comparison scheme used in RetargetMe, the subjective evaluation involved in this database employs a five-category discrete quality scale (e.g., "bad," "poor," "fair," "good," and "excellent") to derive a final mean opinion score (MOS) for each retargeted image.

3) NRID: The NRID database contains 35 source images. Each source image is associated with five retargeted images obtained by applying five different retargeting operators including SEAM [1], WARP [2], MULTI [5], SCAL, and SHIF [6]. A total number of 175 retargeted images are included in NRID. The subjective study for this database is similar with that applied in RetargetMe.

B. Experimental Protocols

As stated before, the subjective studies for the above three databases are conducted by using different methodologies (RetargetMe and NRID are based on paired comparison methodology while CUHK is based on ACR methodology), making the derived subjective scores have different properties. For CUHK database, the five-grade quality scale quantifies the absolute perceptual quality of each retargeted image in the form of MOS, while for RetargetMe and NRID databases, the paired comparison only indicates the relative quality score against a same source image. That is, a retargeted image with a higher subjective rating score in RetargetMe and NRID databases may have worse perceptual quality than the one with a lower rating score if these two retargeted images are created from different source images.

Based on such consideration, it is reasonable for us to learn a quality fusion model on the CUHK database instead of the other two. Note that the CUHK database contains two subsets: 1) session-1 subset containing 69 retargeted images from 23 source images and 2) session-2 subset containing 102 retargeted images from 34 source images. In the experiments, we select session-1 subset as the training set to train the fusion model by SVR. Once the fusion model is trained, it is used for testing on other databases. For the CUHK, we only evaluate on the session-2 subset (to ensure the samples used for training and testing are independent).

Since the provided MOSs in CUHK are similar with the ones applied in traditional IQA study, performance evaluation on this dataset can be conducted in a standardized way [52], i.e., by measuring Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank order correlation coefficient (KROCC), and root mean square error (RMSE) between the objective and subjective scores as performance criteria. For a perfect objective IRQA model, we have PLCC = SROCC = KROCC = 1 and RMSE = 0. However, subjective scores in RetargetMe and NRID databases only reflect the relative quality between the images retargeted from the same source image, making the actual perceptual quality still unclear. In this case, similar to [15] and [17], we use the Kendall τ distance to measure the correlation between the subjective and objective rankings for the images retargeted from the same source image and take the mean and variance of the τ distribution over all images to quantify the performance of an objective IRQA metric. Theoretically, we have $\tau \in [-1, 1]$, in which $\tau = 1$ corresponds to the perfect agreement case while $\tau = -1$ corresponds to the perfect disagreement case. In the case of $\tau = 0$, the subjective and objective rankings are considered to be independent.

C. Impact of Parameter Variations

Note that there are several parameters need to be tuned in the proposed method, including the size of finer and coarse dictionaries, the patch size, and the patch selection percentage. Among these parameters, the sizes of the finer and coarse dictionaries are determined according to previous sparse representation studies. Specifically, the sizes of overcomplete dictionary for the source and retargeted images are different (refer to Rule 1 in Section III-B). We set the size of the finer dictionary to be half of the number of training samples, and the size of the coarse dictionary to be quarter of the number of training samples, ensuring a sufficient overcompleteness given the dimensionality of the input. Once these parameters are predetermined, we further investigate the influence of patch size and patch selection percentage. In Fig. 7, we show the experimental results based on different combinations of patch size and patch selection percentage. From the results, we find that the combination of {70%, 8×8} results in the best performances among all combinations on the CUHK and NRID databases, while for the RetargetMe database, the combination of {70%, 8×8} can obtain a comparable performance with the optimal combination of {60%, 6×6}.

Furthermore, we have the following observations. First, the result is highly influenced by the patch size. Specifically, by

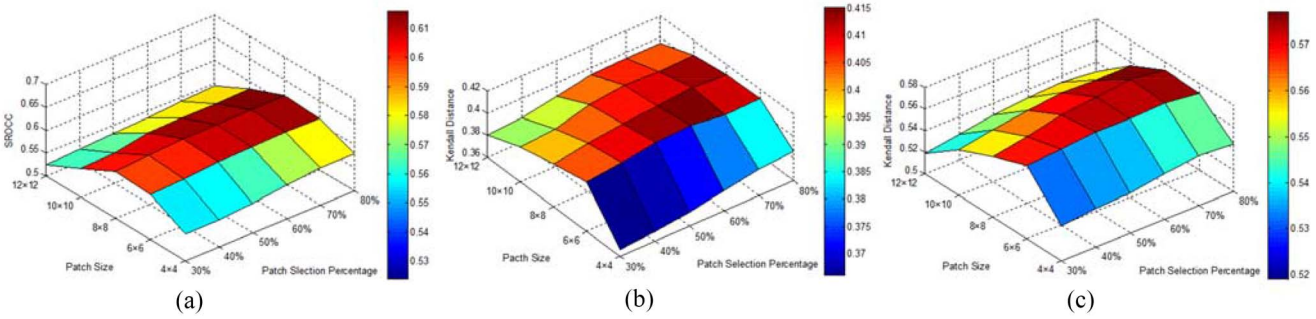


Fig. 7. Performance results based on different parameter combinations on (a) CUHK, (b) RetargetMe, and (c) NRID.

Fig. 8. (a) and (d) Source images. (b) and (c) Retargeted from (a) by using the STVI [7] and SEAM [1] algorithms, respectively. (e) and (f) Retargeted from (d) by using the SHIF [6] and WARP [2] algorithms, respectively.

TABLE II
OVERALL PERFORMANCE RESULTS OF DIFFERENT METHODS ON CUHK DATABASE (THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE)

Database	Criteria	BDS	EMD	SIFT-flow	EH	ME1	Proposed
CUHK	PLCC	0.3077	0.3062	0.3529	0.3618	0.3225	**0.644**
	SROCC	0.2662	0.2814	0.3206	0.3265	0.3013	**0.616**
	KROCC	0.1591	0.1763	0.2135	0.2187	0.1969	**0.521**
	RMSE	12.854	12.889	12.522	12.485	12.751	**10.763**

TABLE III
OVERALL PERFORMANCE RESULTS OF DIFFERENT METHODS ON RETARGETME AND NRID DATABASES
(THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE)

Database	Criteria	BDS	EMD	SIFT-flow	EH	ME1	Proposed
RetargetMe	Mean	0.083	0.251	0.145	0.004	0.182	**0.413**
	Standard deviation	0.268	0.272	0.262	0.334	**0.258**	0.282
NRID	Mean	0.131	0.362	-0.011	0.108	0.154	**0.577**
	Standard deviation	0.527	0.361	0.502	0.556	0.512	**0.334**

increasing the patch size, the performance (e.g., SRCC or Kendall τ distance) increases first and then decreases, because fewer patches are generated with a larger patch size. As a result, such few patches may be insufficient to learn an overcomplete dictionary from an image. Second, the result is also slightly influenced by the patch selection percentage. Generally, the performance measure increases when the patch selection percentage increases. As a tradeoff on all databases, we select the combination of {70%, 8×8} in the experiment.

D. Consistency With Subjective Rating Results

The main innovation of this paper is to address the problem of IRQA from a global perspective, where the process of local dense correspondence is not required. In order to evaluate the overall performance of our method, we compare it with five state-of-the-art local correspondence-based methods, including BDS [18], EMD [19], SIFT-flow [20], EH [21], and the metric in [22] (denoted by ME1). Table II shows the comparison results of different methods on CUHK database. Since only a subset (session-2 images in CUHK) is used for performance evaluation in the experiment, the evaluation results are slightly different with the ones reported in [16]. It can be observed that the proposed method performs much better than other competing methods on this database. Table III shows the average rank correlation measured by the Kendall τ distance and the standard deviation of correlation on the RetargetMe and NRID databases. These results show that correlation values between the rankings predicted by the proposed method and the subjective rankings is larger than 0.41 and 0.57 on RetargetMe and NRID databases, respectively

TABLE IV
MULTIPLE QUALITY SCORES COMPARISON FOR DIFFERENT RETARGETED IMAGES

Score \ Retargeted image	Source image "Jon"		Source image "Butterfly"	
	(b)	(c)	(e)	(f)
MOS	47.3535	24.2544	67.5727	39.0605
BDS	0.169	0.318	0.205	0.276
Q_R^G	0.354	0.211	0.318	0.312
Q_R^S	0.727	0.682	0.433	0.292
Q	43.0964	23.1165	59.7435	34.1938

TABLE V
PERFORMANCE RESULTS ASSOCIATED WITH EACH QUALITY COMPONENT ON CUHK, RETARGETME, AND NRID DATABASES

Dataset	Criteria	EH	SIFT-flow	EMD	Q_R^G	Q_O^G	Q_R^S	Q_O^S	Proposed
CUHK	PLCC	0.3618	0.3529	0.3062	0.5448	0.5429	0.2826	0.2817	**0.644**
	SROCC	0.3265	0.3206	0.2814	0.5215	0.5231	0.2335	0.2326	**0.616**
	KROCC	0.2187	0.2135	0.1763	0.4316	0.4340	0.1251	0.1244	**0.521**
	RMSE	12.485	12.522	12.889	11.299	11.309	13.002	13.007	**10.763**
RetargetMe	Kendall τ distance	0.004	0.145	0.251	0.378	0.375	0.115	0.109	**0.413**
NRID	Kendall τ distance	0.108	-0.011	0.362	0.485	0.483	-0.018	-0.016	**0.577**

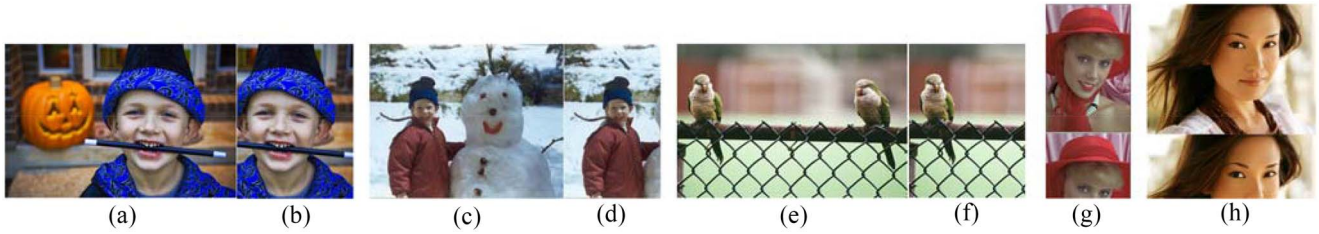


Fig. 9. Examples of retargeted images generated by manually labeled cropping windows with a scaling ratio of 0.5. (a), (c), (e), and (g) Source images. (b), (d), (f), and (h) Their corresponding retargeted images, respectively.

(also the best among all competing methods). We analyze the possible reason may be that, the proposed method evaluates the quality of retargeted images without depending on local correspondence, and thus can avoid the incorrect correspondence occurred in the local correspondence-based methods. The results further demonstrate the advantage of the proposed global-based framework in assessing the quality of retargeted images compared with existing local correspondence-based methods.

We further show a case study to analyze why the proposed method performs better. For this, five different quality scores for the images in Fig. 8(b), (c), (e), and (f) are given for comparison. These quality scores include the MOS value, the score estimated by the BDS metric, the LGS distortion-aware quality score (Q_R^G), the GCI loss-aware quality score (Q_R^S), and the score estimated by the proposed metric (Q), as shown in Table IV. For a better comparison, all the quality scores are converted into the form of similarity scores such that a higher value indicates a better quality. As observed, the retargeted image in Fig. 8(b) [or Fig. 8(e)] presents a better quality than the one in Fig. 8(c) [or Fig. 8(f)], as demonstrated by their corresponding MOS. However, the BDS metric fail to capture such relative quality as the BDS value for Fig. 8(b) [or Fig. 8(e)] is smaller than the one for Fig. 8(c) [or Fig. 8(f)]. On the contrary, the LGS-aware quality score can successfully characterize the relative quality for Fig. 8(b) and (c), while the GCI-aware quality score can make correct prediction for Fig. 8(e) and (f). This is because geometric distortion is the dominant factor affecting the quality of retargeted images in Fig. 8(b) and (c), while context information loss is the dominant factor for the images in Fig. 8(e) and (f). By jointly considering the factors of LGS distortion and GCI loss, the proposed IRQA method can make a more perceptually consistent evaluation.

E. Impact of Each Quality Component

This section aims to investigate the impact of each quality component. Table V shows the comparison results when single quality component is used for performance test and the results of EH [19], SIFT-flow [18], EMD [17], and the proposed method are also incorporated for comparison as benchmark. From the table, three important observations can be derived.

1) The proposed LGS-aware quality measures (Q_R^G and Q_O^G) both perform better than the EH and SIFT-flow metrics because their performances are highly dependent on the accuracy of dense correspondence. However, the

TABLE VI
PERFORMANCE RESULTS BY DIFFERENT QUALITY FUSION SCHEMES ON CUHK, RETARGETME, AND NRID DATABASES

Dataset	Criteria	Average	Multiply	Linear regression	Logistic regression	Linear-SVR	Poly-SVR	RBF-SVR
CUHK	PLCC	0.478	0.466	0.584	0.619	0.593	**0.644**	**0.644**
	SROCC	0.449	0.428	0.552	0.582	0.552	**0.616**	**0.617**
	KROCC	0.360	0.327	0.458	0.488	0.454	**0.521**	**0.522**
	RMSE	11.938	12.021	11.144	10.891	10.977	**10.763**	**10.762**
RetargetMe	Kendall τ distance	0.285	0.294	0.384	0.410	0.396	**0.413**	**0.416**
NRID	Kendall τ distance	0.343	0.349	0.523	0.558	0.538	**0.577**	**0.572**

TABLE VII
AVERAGE TIME COST FOR PER IMAGE IN RETARGETME DATABASE (IN SECONDS)

Criteria	EMD	SIFT-flow	EH	ME1	Proposed
Time (s)	158.2	6.4	0.9	286.5	58.3

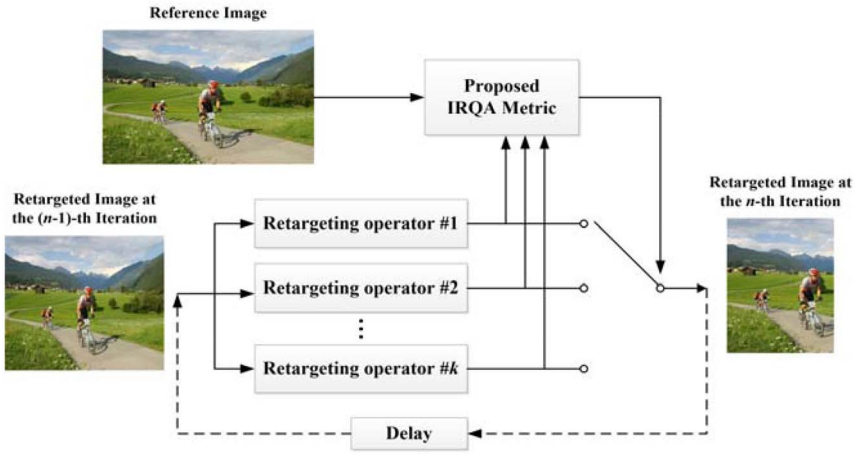


Fig. 10. Diagram of the proposed IRQA metric-guided (MO) image retargeting algorithm.

proposed method addresses this problem from a global perspective which makes the evaluation much more accurate and robust.

2) The LGS-aware quality measures are much more important than the GCI-aware measures whose performances are actually the worst among these estimators except for on the RetargetMe database. This is due to the fact that the GCI loss measure is important only when a great amount of size adjustment is performed. However, there are no sufficient retargeted images with extremely large retargeting size adjustment in both CUHK and NRID databases. As shown in Fig. 9, all the retargeted images are generated by simple image cropping operator with a scaling ratio of 0.5. In this case, the geometrical distortion becomes secondary while the GCI loss is a dominant factor affecting the retargeting visual quality.
3) The best performance is obtained when all single qualities are fused for final quality prediction.

F. Impact of Different Quality Fusion Schemes

In this section, we conduct experiments to investigate the impact of different quality fusion schemes. Toward this end, we design seven different schemes to fuse the individual quality components, including linear average, direct multiplication, linear regression, logistic regression, SVR with linear kernel (linear-SVR), SVR with polynomial kernel (poly-SVR), and SVR with RBF kernel (RBF-SVR). Note that the latter five schemes are learning-based models where the optimal fusion weights are optimized by minimizing the errors on the training set (i.e., session-1 subset in CUHK database). Table VI shows the performance results for the fusion schemes.

Since linear average and direct multiplication schemes do not require learning process, their performances are unsurprisingly the worst among these models. For the learning-based schemes, the fusion model can either be linear (linear regression and SVR with linear kernel) or nonlinear (logistic regression, SVR with polynomial, and RBF kernels). As observed from Table VI, the nonlinear models always perform better than the linear ones. It is expectable because linear schemes assume all the quality components to contribute independently on the final quality, which is not completely true if considering the interactions between the LGS distortion and GCI loss. For example, when the LGS distortion becomes severe, the GCI loss may be suppressed to some extent in some cases. Obviously, the best two results are obtained by using SVR with polynomial and RBF kernels, but the computational overhead of the model with polynomial kernel is lower than the one with RBF kernel. Therefore, we select SVR with polynomial as the fusion scheme in the experiment.

G. Related Applications

The application of IRQA metrics can be broadly categorized into two types: 1) assessing the performance of different image retargeting algorithms and 2) guiding the optimization of image retargeting operators. The first application is intuitive. Given an arbitrary retargeted image, we can use the proposed IRQA metric to assess the visual quality in a perceptually consistent manner. The second application is also valuable because it provides a new insight to design more powerful retargeting algorithms. For example, the proposed IRQA metric can be embedded into a modified multioperator (MO) image retargeting algorithm [54]. Fig. 10 shows the diagram of the proposed IRQA metric-guided MO image retargeting method which combines k types of retargeting operators. In order to resize an image with width of w into the desired one with width of w' by using k operators, the MO image retargeting method resizes the image by a specific width of C (e.g., 10 pixels) at each resizing iteration. The proposed IRQA metric is used to calculate the similarity between the source and retargeted images to select the best quality retargeted image, i.e., to determine which operator should be used at each iteration. The iteration continues until the target resizing factor is reached.

H. Further Discussions

1) Time Complexity: Time complexity is another important factor to evaluate the performance of IRQA. Table VII compares the run-time costs (the average runtime for an image in RetargetMe database) of different methods. It is seen that the proposed method is faster than EMD and ME1, but is lower than SIFT-flow and EH. Actually, the time complexity of the proposed method mainly comes from dictionary learning and sparse representation processes. For sparse representation, we opt to use the batch-OMP algorithm [48] which has been proven to be highly efficient in dealing with many-input/single dictionary sparse coding problem. The most time-consuming procedure lies in the overcomplete dictionary learning process. The elapsed time of this step will occupy about 70% of the overall time complexity. Despite of the lower time complexity of SIFT-flow and EH, their prediction accuracies are, however, remained limited as shown beforehand. The time complexity of our method can be optimized by parallel computing techniques where LGS and GCI modalities can be parallelly processed.

2) Feature Accuracy: It should be emphasized that our proposed method only utilizes SIFT descriptor to capture the LGS information, which is inaccurate in some cases. This can be seen from the example shown in Fig. 4(a) and (b), where the chin of the girl is seriously deformed but the SIFT key point detection results fail to capture this distortion. To more accurately represent the retargeting visual quality, the LGS distortions need to be characterized more precisely and comprehensively. Moreover, to what extent the saliency map can characterize the context information of a scene is also important. Therefore, it is of great importance to excavate more accurate high-level features to represent a scene. Deep learning [53], known as a powerful hierarchical feature representation tool for image classification and object recognition, may provide a new insight for the LGS and GCI representation.

V. Conclusion

In this paper, we have presented a novel IRQA method based on sparse representation. Different from traditional IRQA metrics that were designed based on specific interimage dense correspondence algorithms, the major contribution of this paper was to address the IRQA from a global perspective to eliminate the impact of incorrect correspondence. In essence, we at first extracted distortion sensitive features from an image and then investigated how many of these features were preserved or changed in another image so as to measure the similarity between them. Toward this end, we have made use of the intrinsic discriminative power of sparse representation for similarity measurement. Then, multiple quality scores from the LGS and GCI were fused to obtain a final quality score by SVR. Extensive experiments on three benchmark databases have demonstrated the superiority of the proposed IRQA method. Although the proposed method has shown better performance than other existing methods, it was still belonged to a full-reference one requiring a source image for comparison. Therefore, future work may focus on designing no-reference [55], even "completely blind" metric to provide a more practical and robust solution for IRQA. Furthermore, the metric will be used as an indicator to optimize the content production [56], [57].

References

[1] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007, Art. no. 10.
[2] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Rio de Janeiro, Brazil, 2007, pp. 1–6.
[3] Z. Karni, D. Freedman, and C. Gotsman, "Energy-based image deformation," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1257–1268, Jul. 2009.
[4] 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, Dec. 2008, Art. no. 118.
[5] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graph.*, vol. 28, no. 3, Aug. 2009, Art. no. 23.
[6] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, 2009, pp. 151–158.
[7] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, Dec. 2009, Art. no. 126.
[8] W. Dong, N. Zhou, J.-C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling," *ACM Trans. Graph.*, vol. 28, no. 5, Dec. 2009, Art. no. 125.
[9] Q. Wang and Y. Yuan, "High quality image resizing," *Neurocomputing*, vol. 131, pp. 348–356, May 2014.
[10] Q. Wang and X. Li, "Shrink image by feature matrix decomposition," *Neurocomputing*, vol. 140, pp. 162–171, Sep. 2014.
[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[12] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[13] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[14] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 347–364, Apr. 2012.

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

[16] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 626–639, Oct. 2012.

[17] C.-C. Hsu, C.-W. Lin, Y. Fang, and W. Lin, "Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 377–389, Jun. 2014.

[18] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–8.

[19] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 460–467.

[20] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[21] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[22] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu, "Image retargeting quality assessment," *Comput. Graph. Forum*, vol. 30, no. 2, pp. 583–592, Apr. 2011.

[23] Y. Fang *et al.*, "Objective quality assessment for image retargeting based on structural similarity," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 3, no. 1, pp. 95–105, Mar. 2014.

[24] Y. Zhang, Y. Fang, W. Lin, X. Zhang, and L. Li, "Backward registration-based aspect ratio similarity for image retargeting quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4286–4297, Sep. 2016.

[25] W. Lian, L. Zhang, and M.-H. Yang, "An efficient globally optimal algorithm for asymmetric point matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2016.2603988.

[26] D. C. Van Essen and J. H. R. Maunsell, "Hierarchical organization and functional streams in the visual cortex," *Trends in Neurosci.*, vol. 6, no. 9, pp. 370–375, 1983.

[27] F. Zhang, W. Jiang, F. Autrusseau, and W. Lin, "Exploring V1 by modeling the perceptual quality of images," *J. Vis.*, vol. 14, no. 1, Jan. 2014, Art. no. 26.

[28] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, Oct. 2013.

[29] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, "Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2059–2074, May 2016.

[30] F. Shao *et al.*, "Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 730–743, Mar. 2016.

[31] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2649101.

[32] H. Wei and Z. Dong, "V4 neural network model for shape-based feature extraction and object discrimination," *Cogn. Comput.*, vol. 7, no. 6, pp. 753–762, Dec. 2015.

[33] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, 2006, pp. 895–900.

[34] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[35] T. Bai, Y.-F. Li, and X. Zhou, "Learning local appearances with sparse representation for robust and fast visual tracking," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 663–675, Apr. 2015.

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[37] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Hong Kong, 2010, pp. 1601–1604.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[39] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[40] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[41] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2006, pp. 545–552.

[42] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

[43] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[44] A. Alaei, R. Raveaux, and D. Conte, "Image quality assessment based on regions of interest," *Signal Image Video Process.*, pp. 1–8, 2016, doi: 10.1007/s11760-016-1009-z.

[45] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[46] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, Mar. 2006.

[47] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Jan. 2010.

[48] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, Jan. 2008.

[49] L.-W. Kang *et al.*, "Feature-based sparse representation for image similarity assessment," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1019–1030, Oct. 2011.

[50] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

[51] R. Bradley and M. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, Dec. 1952.

[52] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding," *J. Vis. Commun. Image Representation*, vol. 33, pp. 123–133, Nov. 2015.

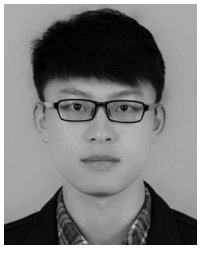
[53] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[54] Y. Fang *et al.*, "Optimized multioperator image retargeting based on perceptual similarity measure," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2016.2557225.

[55] L. Ma, L. Xu, Y. Zhang, Y. Yan, and K. N. Ngan, "No-reference retargeted image quality assessment based on pairwise rank learning," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2228–2237, Nov. 2016.

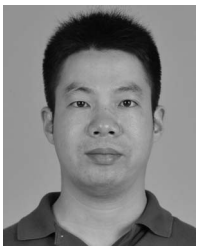
[56] F. Shao, W. Lin, Z. Li, G. Jiang, and Q. Dai, "Toward simultaneous visual comfort and depth sensation optimization for stereoscopic 3-D experience," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2615856.

[57] F. Shao, Q. Jiang, R. Fu, M. Yu, and G. Jiang, "Optimizing visual comfort for stereoscopic 3D display based on color-plus-depth signals," *Opt. Express*, vol. 24, no. 11, pp. 11640–11653, May 2016.



Qiuping Jiang received the M.Sc. degree from the School of Information Science and Engineering, Ningbo University, Ningbo, China, in 2015, where he is currently pursuing the Ph.D. degree.

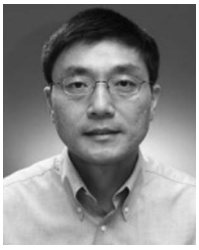
Since 2017, he has been a visiting Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include visual quality assessment, saliency detection, and visual perception.



Feng Shao (M'16) received the B.S. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively, both in electronic science and technology.

He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China. In 2012, he was a Visiting Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore, for seven months. He has published over 100 technical articles in refereed journals and proceedings in the areas of 3-D video coding, 3-D quality assessment, and image perception.

Dr. Shao was a recipient of the Excellent Young Scholar Award by the NSF of China in 2016.



Weisi Lin (M'92–SM'98–F'16) received the B.Sc. and M.Sc. degrees from Zhongshan University, Guangzhou, China, and the Ph.D. degree from King's College, London University, London, U.K.

He was the Laboratory Head of Visual Processing and the Acting Department Manager of Media Processing with the Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include image processing, perceptual modeling, video compression, multimedia communication, and computer vision. He has published over 200 refereed papers in international journals and conferences.

Prof. Lin is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS SYSTEMS FOR VIDEO TECHNOLOGY, and the *Journal of Visual Communication and Image Representation*, and was an Associate Editor for the IEEE TRANSACTION ON MULTIMEDIA and the IEEE SIGNAL PROCESSING LETTERS. He served as the Lead Guest Editor for a Special Issue on Perceptual Signal Processing of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING in 2012. He is the Chair of the IEEE MMTC Special Interest Group on Quality of Experience. He has been elected as a Distinguished Lecturer of APSIPA from 2012 to 2013. He is the Lead Technical Program Chair for the 2012 Pacific-Rim Conference on Multimedia, and the Technical Program Chair for the 2013 IEEE International Conference on Multimedia and Expo. He is a fellow of the Institution of Engineering Technology, and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He is a Chartered Engineer in U.K.

Gangyi Jiang (M'14) received the M.S. degree from Hangzhou University, Hangzhou, China, in 1992, and the Ph.D. degree from Ajou University, Suwon, South Korea, in 2000.

He is currently a Full Professor with the School of Information Science and Engineering, Ningbo University, Ningbo, China. His current research interests include 3-D/multiview video coding, multimedia communication, and visual perception.