Learning to Rank Retargeted Images

Yang Chen, Yong-Jin Liu*
Tsinghua University, China
liuyongjin@tsinghua.edu.cn

Yu-Kun Lai
Cardiff University, UK
Yukun.Lai@cs.cf.ac.uk

Abstract

Image retargeting techniques that adjust images into different sizes have attracted much attention recently. Objective quality assessment (OQA) of image retargeting results is often desired to automatically select the best results. Existing OQA methods output an absolute score for each retargeted image and use these scores to compare different results. Observing that it is challenging even for human subjects to give consistent scores for retargeting results of different source images, in this paper we propose a learning-based OQA method that predicts the ranking of a set of retargeted images with the same source image. We show that this more manageable task helps achieve more consistent prediction to human preference and is sufficient for most application scenarios. To compute the ranking, we propose a simple yet efficient machine learning framework that uses a General Regression Neural Network (GRNN) to model a combination of seven elaborate OQA metrics. We then propose a simple scheme to transform the relative scores output from GRNN into a global ranking. We train our GRNN model using human preference data collected in the elaborate RetargetMe benchmark and evaluate our method based on the subjective study in RetargetMe. Moreover, we introduce a further subjective benchmark to evaluate the generalizability of different OQA methods. Experimental results demonstrate that our method outperforms eight representative OQA methods in ranking prediction and has better generalizability to different datasets.

1. Introduction

Image retargeting refers to techniques that adjust a source image into different sizes, which has become an increasingly demanded tool with the diversification of display devices. Although a large number of retargeting methods have been developed, a single method that works well on any image still does not exist [12, 33]. Subjective quality assessment involving human judgment is usually time-consuming and laborious, and thus unpractical in many situations. As summarized in Section 2, despite recent progress, existing *objective* quality assessment (OQA) methods are still far from ideal in predicting human preference. Therefore, a good OQA method correlating well with human judgements is essential in automatically selecting the best retargeting results and helpful for developing new image retargeting methods.

(a) Source image (b) Result 1 (27) (c) Result 2 (26)

(d) Source image (e) Result 3 (30) (f) Result 4 (28)

Figure 1. Subjective scores are only comparable for retargeting results of the same source image. In each row, two retargeting results are presented and their scores are shown in parentheses. These scores provided in RetargetMe benchmark [25] are numbers of votes that people cast when comparing this image against other images. Higher scores mean better results. Although the scores of the two retargeting results in (e) and (f) are higher than the scores in (b) and (c), we cannot conclude that the results in (e) and (f) are better than the results in (b) and (c); instead, the opposite appears to be true.

Existing OQA methods compute an absolute score for every retargeting result and compare the results using these scores. A key observation that motivates the work presented in this paper is that in most cases, the scores of retargeted images are only meaningful with the same source image. Even for human subjects, it is often difficult to give consistent scores for retargeting results of different sources. An

*Corresponding author

example is shown in Figure 1, in which the two retargeting results in (b) and (c) have lower subjective scores, but appear to be more plausible than the results in (e) and (f) that have higher scores. We also notice that in majority of application scenarios, relative quality measures are sufficient, e.g., to rank a set of retargeting results. Therefore, instead of assigning every retargeted image with an absolute score, in this paper we focus on a more manageable task that predicts the ranking of a set of retargeted images with the same source image.

Given a set of retargeted images, we propose a learning-based OQA method that provides the ranking of these images as output. Our method uses the General Regression Neural Network (GRNN) [29] to model a combination of seven known OQA metrics [12] including preservation of salient regions, influence of introduced artifacts, preservation of the global structure, preservation of symmetry, and aesthetics. We train this GRNN model using the human preference data collected in the elaborate RetargetMe benchmark [25]. The GRNN model is known to work effectively with relatively few training samples, which suits our task well due to the limited availability of subjective data. In the testing stage, the GRNN model takes the features of a pair of retargeted images as input and predicts their *relative quality difference*. By computing relative quality differences of all pairs in a set of retargeted images, we propose a simple scheme to transform them into a global ranking.

The main contributions of this paper are:

- We propose a two-step OQA method to predict the ranking in a set Ω of retargeted images with the same source image: (1) at step 1, we introduce a simple yet effective machine learning framework to predict the relative quality difference in a pair of retargeted images in Ω and (2) at step 2, we propose a simple scheme to transform the relative quality differences in all pairs in Ω into a global ranking.

- We conduct a new user study using an approach similar to RetargetMe benchmark [25] with better quality control. The novel dataset obtained in this user study, as well as the source code of the proposed OQA method, will be made publicly available to provide a useful dataset for evaluating *generalizability* of different OQA methods.

Extensive experiments are presented, demonstrating that our OQA method correlates better with human judgements than eight representative OQA methods and has significantly better generalizability.

2. Related work

Image retargeting has attracted considerable attention in recent years and many content-aware methods have been developed [27]. To compare different retargeting algorithms, several quality assessment methods have been proposed. These methods can be divided into two types: subjective and objective methods.

Subjective quality assessment designs elaborate perceptual studies and systematically analyzes user preferences. RetargetMe [25] is a well-established benchmark that contains a decent number of source images and their retargeting results produced by eight representative methods. A comprehensive, comparative subjective study is also included in RetargetMe. It is the first in-depth perceptual study with a large number of users for image retargeting quality assessment. A different subjective study was proposed in [16], in which the user evaluation was carried out by simultaneous double stimulus for continuous evaluation [1] that scored only one retargeted image each time rather than pairwise comparison. Castillo et al. [3] developed an image retargeting survey using eye tracking technology. All these subjective methods can provide good evaluation, but they are laborious and very time-consuming. Nevertheless, these studies provide valuable benchmarks for developing objective quality assessment methods. Our method proposed in this paper mainly depends on the RetargetMe benchmark and we further perform an extended user study for evaluating generalizability.

Objective quality assessment (OQA) defines metrics that can be calculated from pixels of images. Edge Histogram (EH) [18] and Color Layout (CL) [10] are two image-content-based measures in the MPEG-7 standard [19]. They are low-level metrics that treat images as a whole and define image distances based on similarity of edge or color distribution. Bidirectional Similarity (BDS) [28] treats an image as a collection of patches and calculates a bidirectional mapping of these patches between two images as a measure. Bidirectional Warping (BDW) [26] is similar to BDS, but the mapping in BDW takes an asymmetric dynamic time warping, which simultaneously minimizes the warping cost and preserves the patch order. BDS and BDW are relatively easy to calculate; however, they treat every patch as equally important for the final distance and do not take salient regions or aesthetic perspectives into account. Thus their results are not always consistent with subjective ranking. Objective quality assessment methods based on SIFT flow (SFlow) [13] and Earth-Mover's Distance (EMD) [23] can capture the structural properties more robustly. Liu et al. [14] proposed a top-down simplified model of the human vision system to define a saliency-based image similarity metric in the CIE Lab color space. Recently, an aspect ratio similarity (ARS) metric [33] was proposed, which characterizes how the source image is resized into the target image by geometric changes and provides an efficient solution based on a Markov random field. Noting that human judgment of retargeting qual-

ity often involves multiple factors, several state-of-the-art methods combine several metrics that characterize different factors of image retargeting quality [16, 17, 12].

Our proposed method is inspired by the work in [12] that elaborately designs seven metrics and develops an OQA method by combining them. These seven metrics take the following factors into consideration: keeping salient image content, reducing local artifacts, preserving global structure, satisfying aesthetic requirement, and maintaining symmetry features. Liang et al. [12] make use of a linear combination of these seven metrics, with the weights learned from the RetargetMe benchmark. This method provides an all-round characterization of retargeted images. However, the linear combination is over-simplified and does not always produce a consistent prediction to human preference. To address this problem, we propose to use a machine learning approach to provide the necessary flexibility.

Artificial neural networks (ANNs) have been well studied and widely used in image processing. The universal approximation theorem [6] states that simple neural networks can represent a wide range of useful functions when given appropriate parameters. Among many types of ANNs, the RBF network is a universal approximator and is a popular alternative to the multi-layer perceptrons, due to its simpler structure and faster training process. Our work in this paper uses the general regression neural network (GRNN) [29], which is a representative RBF network and can obtain good results even with sparse data in a multidimensional measurement space, particularly suitable for our problem.

As ranking is the major needs for objective assessment of image retargeting, it is related to learning to rank techniques. Such techniques can be divided into three categories according to their loss functions, that is, pointwise (e.g., [8]), pairwise (e.g., [7]) and listwise (e.g., [2]). Pointwise methods are the earliest learning-to-rank techniques. They treat every instance separately and thus easily lose the group structure of ranking. Pairwise methods transform the ranking problem to pairwise classification or regression, and then partially protect the group structure. Listwise methods provide a more straightforward way to solve the ranking problem, which better preserve the group structure. However, since data is split into groups and sufficient instances are needed for each group, a large training dataset is required in listwise methods. In our study, given the relatively limited subjective data, the pairwise technique is more suitable and used in this paper.

3. A two-step OQA method for ranking

The quality of image retargeting depends on multiple factors and composite metrics are needed to measure them. In recent work [12], seven elaborately designed metrics $\{Q_1, Q_2, \cdots, Q_7\}$ were proposed. We briefly summarize these metrics in Section 3.1. Given a source image I_s and a retargeted image I_t, each metric $Q_i(I_s, I_t)$ computes a scale value in $[0, 1]$ to reflect the retargeting quality in one factor.

To construct an objective function $F(Q_1, Q_2, \cdots, Q_7)$ based on these seven metrics, an additive value function

$$F = \sum_{i=1}^{7} w_i Q_i \tag{1}$$

is used in [12]. The value of F is in $[0, 1]$ and a lower value of F means better quality. We argue that the linear form in Eq.(1) is over-simplified and we propose to find a better (possibly nonlinear) form for F by machine learning from human preference.

Instead of assigning an absolute value for each retargeted image, our OQA method computes a ranking to a set $\Omega = \{I_i\}_{i=1}^{n}$ of retargeted images with the same source image I_s. Our method works in two steps. In the first step, we represent each retargeted image $I_i \in \Omega$ as a six-dimensional vector

$$v(I_i) = (Q_1(I_s, I_i), Q_2(I_s, I_i), \cdots, Q_6(I_s, I_i)) \tag{2}$$

An additional dimension $Q_7(I_s, I_i)$ is only applied to symmetry images; see Section 3.3.3 for details.

To better characterize retargeted images, in Section 3.2, we transform the representation $v(I_i)$ in Eq.(2) into a more regular feature space $f(v(I_i))$ by taking a global manifold structure into consideration.

Based on the feature representation $f(v(I_i))$, we construct an objective function $\widetilde{F}(f(v(I_i)), v(I_j))$, which computes a relative score for any two retargeted images I_i and I_j, $i \neq j$, in Ω. To better predict the global ranking in the second step, we require that $\widetilde{F}$ has the following properties:

- $\widetilde{F}(f(v(I_i)), f(v(I_j))) > 0$ means I_i is of better quality than I_j and vice versa;
- The value of $\widetilde{F}$ reflects the degree of relative quality difference; e.g., $\widetilde{F}(f(v(I_i)), f(v(I_j))) = 0.01$ means I_i is slightly better than I_j and $\widetilde{F}(f(v(I_i)), f(v(I_j))) = 0.99$ means I_i is much better than I_j.

To find an appropriate nonlinear form for $\widetilde{F}$, we use a machine learning method to learn from human preference. Observing the limited availability of subjective data, we choose the GRNN model to train $\widetilde{F}$ using subjective data in the well-established RetargetMe benchmark [25]. The training details are provided in Section 3.3.

In the second step, we transform the relative scores into a global ranking by computing a ranking value r_i for each I_i, with respect to the remaining images in Ω:

$$r_i = \sum_{I_j \in \Omega \setminus \{I_i\}} \widetilde{F}(f(v(I_i)), f(v(I_j))) \tag{3}$$

Then the ranking of all the images in Ω is obtained using their ranking values. In Section 3.4, we show that this simple ranking scheme works well.

In Section 4, we demonstrate that our two-step OQA method outperforms eight representative OQA methods in RetargetMe benchmark (using leave-one-out cross validation) and a new user study with novel image dataset (to evaluate generalizability).

3.1. Seven metrics

By carefully analyzing existing retargeting methods and their outcomes, Liang et al. [12] present five key critical factors that determine image quality for a retargeting result. These factors and their related metrics are summarized below.

Preservation of salient regions. This factor is measured by two metrics Q_1 and Q_2. Q_1 considers the change of the salient areas between the source image I_s and retargeted image I_t:

$$Q_1 = |S_s - S_t| / \max(S_s, S_t), \tag{4}$$

where S_s and S_t represent the areas of the salient regions in I_s and I_t, respectively. Q_2 considers variations in content as changes in the color histogram of salient regions [21]:

$$Q_2 = \frac{1}{2}\sqrt{\sum_{i=0}^{255}(h'_s - h'_r)^2}, \tag{5}$$

where h'_s and h'_r represent the normalized color histograms in the source and retargeting salient regions, respectively.

Influence of introduced artifacts. This factor is measured by a bidirectional similarity metric Q_3 that takes into account the influence of saliency [28]:

$$\begin{aligned} Q_3 = \quad & 0.5\frac{\frac{1}{N_s}\sum_{U\subset I_s} S_U \min_{V\subset I_t} D(U,V)}{\max_{U\subset I_s}(S_U \min_{V\subset I_t} D(U,V))} + \\ & 0.5\frac{\frac{1}{N_t}\sum_{V\subset I_t} S_V \min_{U\subset I_s} D(U,V)}{\max_{V\subset I_t}(S_V \min_{U\subset I_s} D(U,V))}, \end{aligned} \tag{6}$$

where U and V are 3×3 patches from the source and retargeted images respectively, N_s and N_t are the numbers of patches in I_s and I_t, D is the distance measure between two patches as defined in [28], and S_U and S_V are saliency weights given by the average of the salience values of all pixels contained in patches U and V.

Preservation of global structure. This factor is measured by two metrics Q_4 and Q_5. Based on a structure-aware pixel mapping scheme in image scale spaces of I_s and I_t [14], both Q_4 and Q_5 evaluate the global structure similarity by a weighted summation of local similarity windows from every pair of pixel correspondence. Q_4 considers the structural similarity between two images by analyzing the degradation of structural information between corresponding windows in I_s and I_t using the SSIM metric [31]:

$$Q_4 = \sum_{i=1}^{n_t}(1 - SSIM(p_i, p'_i)), \tag{7}$$

and Q_5 applies a model VDP2 [20] of human perception to predict the overall quality of I_t, when compared to I_s:

$$Q_5 = \sum_{i=1}^{n_t}(1 - \frac{VDP2(p_i, p'_i)}{100}), \tag{8}$$

where n_t is the number of pixels in I_t, p'_i is the ith pixel of I_t and p_i is the corresponding pixel in I_s.

Aesthetics. This factor is measured by two rules in computational aesthetics [4, 15]: the rule of thirds T_{third} and visual balance V_{bal}:

$$Q_6 = 0.5T_{third}(I_s, I_t) + 0.5V_{bal}(I_s, I_t). \tag{9}$$

Preservation of symmetry. This factor is measured by accumulating all the minimum symmetry distances of symmetric regions in I_t:

$$Q_7 = \frac{1}{n_s}\sum_{r_m\in R}\min_{r_n\in R} D_{sym}(r_m, r_n), \tag{10}$$

where R is the set of symmetry regions in I_t detected by [32], n_s is the number of symmetry regions in R and D_{sym} is the symmetry distance defined in [12].

3.2. Feature space transformation

The six-dimensional representation v in Eq.(2) maps each retargeted image into a point in $\mathbb{R}^6$. Let $\mathcal{M}$ be the union of points representing all retargeted images in $\mathbb{R}^6$. We observe that the manifold structure in $\mathcal{M}$ plays an important role in quality prediction. Given a set of sample points $\mathcal{P}$ in $\mathcal{M}$ and a weighted graph $\mathcal{G}$ with edges between neighboring points in $\mathcal{P}$, the manifold structure can be characterized by the geodesic distances between all pairs of points in $\mathcal{P}$, which are approximated by the lengths of shortest paths in $\mathcal{G}$.

A transformation from the original representation space $\mathbb{R}^6$ to a more regular feature space $\mathcal{F}$ is desired, if the geodesic distances between all pair of points in $\mathcal{P}$ are better represented as Euclidean distances in $\mathcal{F}$. In our practice, this transformation can make the regression in Section 3.3 more effective, especially with limited number of training samples. We use a subspace learning technique to find such a transformation. Nonlinear methods such as ISOMAP [30] and LLE [24] are effective and insensitive to outlier; however, they do not work directly with out-of-sample data. If we use out-of-sample extension, it is only approximate and the effectiveness relies heavily on neighboring in-samples. We thus choose the neighborhood preserving embedding

(NPE) [9], which is efficient and defined everywhere (rather than only on the training samples) and thus suits our problem well. In our application, we apply NPE in the original space $\mathbb{R}^6$ and we denote by $f(v)$ the feature vector after transforming the original representation v into the feature space $\mathcal{F}$.

3.3. Training GRNN for $\widetilde{F}$

3.3.1 Training dataset

We use all the 37 groups of images (each group has one source image and eight retargeted images) in RetargetMe dataset [25] — a well-known benchmark in image retargeting — to train and evaluate our OQA model.

In RetargetMe, a comparative user study based on *linked-paired comparison design* [5] was performed to ensure balanced voting. Three complete sets were collected for each retargeted image to guarantee statistical robustness. Each time a participant was shown two retargeted images side by side, and was asked to simply choose the one he/she liked better. Each retargeted image appeared 3 times for a participant and judged by 21 participants, meaning that a retargeted image received a maximum of $21 \times 3 = 63$ votes. The number of votes for a retargeted image shows the subjective quality by human observers. As demonstrated in Figure 1, such subjective scores cannot be used to effectively compare human preference with *different* source images, but work reasonably well for retargeted images with the *same* source image.

3.3.2 $\widetilde{F}$ for non-symmetry images

We model $\widetilde{F}$ using GRNN, due to its approximate capability with relative few training samples. The input to this model is a concatenation of two feature vectors $f(v(I_i))$ and $f(v(I_j))$. We use the standard configuration for our GRNN model with the output layer being a scale that is the predicted relative score $\widetilde{F}(f(v(I_i)), f(v(I_j)))$. A spread parameter σ in GRNN controls the influence range of radial basis functions and is set to 1.4 in our experiments.

In the training stage, for each group in the training set (i.e., one source image and a set of retargeted images Ω), we take every pair of retargeted images in Ω including a retargeted image with itself. Let $n = |\Omega|$ be the number of retargeted images in Ω. Each group contributes to n^2 pairs of training input. Note that although subjective scores from user voting may not give universally comparable scores, the relative scores for the same source image are much more reliable. Based on this observation, we train the GRNN model using the input-output samples (x_k, y_k), where $x_k = (f(v(I_i)), f(v(I_j)))$ and

$$y_k = \frac{s_i - s_j}{63}, \tag{11}$$

Image	Original representation $v(I_i)$	Feature vectors $f(v(I_i))$
I_1	(0.0, 0.0, 0.12, 0.49, 0.55, 0.34)	(-0.02, -0.04, -0.10, 0.03, -0.01, 0.02)
I_2	(0.09, 0.10, 0.09, 0.63, 0.70, 0.36)	(-0.01, -0.03, -0.08, 0.03, 0.10, -0.02)
I_3	(0.13, 0.03, 0.10, 0.41, 0.54, 0.35)	(-0.02, 0.01, -0.10, 0.01, -0.02, -0.04)
I_4	(0.24, 0.07, 0.11, 0.67, 0.71, 0.36)	(-0.0, 0.02, -0.06, 0.02, 0.05, 0.04)
I_5	(0.25, 0.05, 0.10, 0.50, 0.57, 0.38)	(-0.01, 0.06, -0.09, 0.03, 0.01, 0.01)
I_6	(0.23, 0.08, 0.12, 0.50, 0.57, 0.37)	(-0.01, 0.05, -0.09, 0.05, 0.03, -0.03)
I_7	(0.66, 0.48, 0.02, 0.49, 0.51, 0.43)	(-0.01, 0.25, -0.05, 0.18, 0.43, -0.19)

Relative scores $\widetilde{F}(f(v(I_i)), f(v(I_j)))$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
I_1	0.0	0.022	0.026	0.027	0.039	0.047	0.207
I_2	−0.022	0.0	−0.004	0.004	0.017	0.025	0.185
I_3	−0.026	0.004	0.0	0.0	0.013	0.021	0.181
I_4	−0.027	−0.004	0.0	0.0	0.013	0.020	0.181
I_5	−0.039	−0.017	−0.013	−0.013	0.0	0.008	0.168
I_6	−0.047	−0.025	−0.021	−0.020	−0.008	0.0	0.161
I_7	−0.207	−0.185	−0.181	−0.181	−0.168	−0.161	0.0

Ranking values r_i

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
r_i	0.368	0.201	0.193	0.170	0.094	0.040	−1.083

Table 1. Original representation $v(I_i)$, feature vectors $f(v(I_i))$, relative scores $\widetilde{F}(f(v(I_i)), f(v(I_j)))$ and ranking values of seven retargeted images $\Omega = \{I_i\}_{i=1}^{7}$ in Figure 2.

where s_i and s_j are subjective votes that the retargeted images I_i and I_j received, and we use the maximum number of votes 63 for normalization. The relative score defined in Eq.(11) satisfies the following two desirable properties:

- $\widetilde{F}(f(v(I_i)), f(v(I_i))) = 0, \forall I_i \in \Omega$,
- Both $(f(v(I_i)), f(v(I_j)))$ and $(f(v(I_j)), f(v(I_i)))$ are in the training pairs and $\widetilde{F}(f(v(I_i)), f(v(I_j))) = -\widetilde{F}(f(v(I_j)), f(v(I_i))), \forall I_i, I_j \in \Omega$.

3.3.3 $\widetilde{F}'$ for images with symmetry

Only a small number of images are symmetric in the training set (6 images in RetargetMe). Therefore it is difficult to train a different model for images with symmetry. Inspired by the work of transfer learning [22], we build a simple composite model $\widetilde{F}'$ that combines the model $\widetilde{F}$ trained for non-symmetry images with the difference of symmetry feature values:

$$\begin{aligned}&\widetilde{F}'(f(v(I_i)), f(v(I_j))) = \\ &\widetilde{F}(f(v(I_i)), f(v(I_j))) + w(Q_7(I_j) - Q_7(I_i)),\end{aligned} \tag{12}$$

where Q_7 is the symmetry metric specified in Eq.(10) and $w > 0$ is a weight that is optimized using the symmetric images in the training set to maximize the average Kendall correlation coefficient, which is used in Section 4 to indicate the degree of agreement between objective and subjective assessments. Note that a lower value of Q_7 means better symmetry quality and then $\widetilde{F}'(f(v(I_i)), f(v(I_j))) > 0$ means I_i is of better quality than I_j.

(a) Source image I_s (b) I_1 (55) (c) I_2 (42) (d) I_3 (37) (e) I_4 (31) (f) I_5 (22) (g) I_6 (18) (h) I_7 (14)

Figure 2. A group in RetargetMe benchmark includes one source image I_s and eight retargeted images $\{I_1, I_2, \cdots, I_8\}$ (only seven are shown here, which constitute an illustrative example for effectiveness of global ranking in Section 3.4). The subjective score for each retargeting result is shown in parentheses. In our two-step OQA model, for each retargeting result I_i, $i = 1, 2, \cdots, 7$, the original representation $v(I_i)$, the feature vector after NPE transformation $f(v(I_i))$, the relative score $\widetilde{F}(f(v(I_i)), f(v(I_j)))$ and the ranking value r_i are presented in Table 1.

3.4. Effectiveness of global ranking

Note that the relative scores cannot always provide a consistent ranking of all retargeted images in a set Ω. To see this, one can easily construct an example in which I_1 is better than I_2, I_2 is better than I_3 and I_3 is better than I_1. Another real example is illustrated in Figure 2 and Table 1. In this example, $\widetilde{F}(f(v(I_1)), f(v(I_2))) = 0.022$ and $\widetilde{F}(f(v(I_1)), f(v(I_3))) = 0.026$, meaning that I_1 is better than I_2 and I_3. Since $0.022 < 0.026$, this implies I_2 is better than I_3. However, $\widetilde{F}(f(v(I_2)), f(v(I_3))) = -0.004$, meaning that I_3 is better than I_2, a contradiction.

In our two-step OQA method, we solve such potential conflict by transforming relative scores into rank values specified in Eq.(3). Our ranking strategy based on rank values is simple yet effective. Below we show that it always produces consistent ranking in the training set. By substituting Eq.(11) into Eq.(3), we have

$$r_i = \sum_{I_j \in \Omega \setminus \{I_i\}} \frac{s_i - s_j}{63} = \frac{n s_i - \sum_{i=1}^{n} s_i}{63} \qquad (13)$$

Since the value $n = |\Omega|$ and $\sum_{i=1}^{n} s_i$ are the same for each retargeted image I_i, the ranking based on r_i equals to the ranking based on the subjective votes s_i.

4. Experiments

We implemented the proposed two-step OQA method in MATLAB and the source code is publicly available[1]. We compare our method with eight representative OQA methods: BDS [28], BDW [26], EH [18], CL [10], SFlow [13], CSim [14], Liang's method [12] and ARS [33]. The comparison is performed in two experiments: one is the leave-one-out cross validation on the RetargetMe benchmark [25] (Section 4.1) and the other is a generalizability evaluation on a novel dataset constructed in a new user study (Section 4.2).

4.1. Leave-one-out cross validation on RetargetMe

RetargetMe has 37 groups of images with subjective preference scores and each group has one source image and eight retargeted images. These 37 groups are classified into six types: lines/edges (25), faces/people (15), texture (6), foreground objects (18), geometric structure (16) and symmetry (6). These classifications are not exclusive, meaning that one image can belong to more than one type.

To verify the performance of our method and compare it with eight representative methods, we apply leave-one-out cross validation (LOOCV) in RetargetMe. In each fold of LOOCV, one group is used as the test set, with the remaining groups as the training set. After 37 folds, each group has been used as a test set once.

To estimate how well the objective ranking agrees with the participants' subjective voting, we follow the method in [25] to use the Kendall correlation coefficient τ [11]. The value of τ is in $[-1, 1]$ and higher value means better agreement. The results are summarized in Table 2, classified according to six image types. We also compute a mean Kendall correlation coefficient using all the images (last column in Table 2). The results show that

- Liang's method [12], ARS [33] and our method are top three in all nine methods;
- Our method consistently produces significantly better results than [12], which uses the same seven metrics. This result demonstrates that the linear combination of

[1]Information on the data underpinning the research results presented here, including how to access them, can be found at http://47.89.51.189/liuyj and in Cardiff University's data catalogue at http://doi.org/10.17035/d.2017.0033306559.

	Lines/edges	Faces/people	Texture	Foreground objects	Geometric structure	Symmetry	All
BDS [28]	0.040	0.190	0.089	0.167	−0.004	−0.012	0.083
BDW [26]	0.031	0.048	−0.009	0.060	0.004	0.119	0.046
EH [18]	0.043	−0.076	−0.063	−0.079	0.103	0.298	0.004
CL [10]	−0.023	−0.181	−0.089	−0.183	−0.009	0.214	−0.068
SFlow [13]	0.097	0.252	0.161	0.218	0.085	0.071	0.145
CSim [14]	0.091	**0.271**	0.188	0.258	0.063	−0.024	0.151
Liang's [12]	**0.351**	**0.271**	**0.304**	**0.381**	**0.415**	**0.548**	**0.399**
ARS [33]	**0.463**	**0.519**	**0.444**	**0.330**	**0.505**	**0.464**	**0.452**
Ours	**0.437**	**0.505**	**0.429**	**0.536**	**0.438**	**0.536**	**0.473**

Table 2. The mean Kendall correlation coefficients of 37 groups of images in RetargetMe. The top three of each type are in bold and the best results are in blue.

(a) Source image (b) Checkpoint image pair 1 (c) Checkpoint image pair 2 (d) Checkpoint image pair 3

Figure 3. Three check point pairs. In (b) and (c), retargeted images on the left are obviously better than those on the right. In (d), the retargeted image on the right is obviously better than the one on the left.

these seven metrics is over-simplified and our machine learning framework can learn a better predictor from human preference;

- Our method is better than ARS [33] in the image types of foreground objects and symmetry. ARS is slightly better than our method in the image types of lines/edges, faces/people, texture and is better in geometric structure. Overall, our method is slightly better than ARS.

Our method uses two existing tools: GRNN [29] for machine learning with sparse data in a multidimensional space and NPE [9] for feature space transformation that supports out-of-samples. To demonstrate the performance of GRNN and NPE, we compare GRNN with RBFN (another artificial neural network using radial basis functions) and SVR (a typical kernel method), and compare NPE with ISOMAP [30] and LLE [24]. The experimental results in Table 4 show that the combination of GRNN and NPE achieves the best performance.

	SVR	RBFN	GRNN
None	0.438	0.421	0.438
LLE	0.110	0.077	0.064
ISOMAP	0.328	0.369	0.344
NPE	0.456	0.388	**0.473**

Table 4. The mean Kendall correlation coefficients of images in the RetargetMe benchmark for different machine learning methods and different feature space transformations (*None* means no transformation is applied). The best result is shown in bold.

4.2. Generalizability on a novel dataset

To evaluate the generalizability of OQA methods to *different* image datasets, we conducted a new user study on 26 new groups selected in RetargetMe that lack subjective scores[2]. These 26 groups are also classified into six types: lines/edges (11), faces/people (5), texture (1), foreground objects (15), geometric structure (7) and symmetry (4).

The original web-based user study in RetargetMe [25] was based on the linked-paired comparison design [5]. In the website of the survey, two retargeted images and the source image were shown simultaneously at each time. Each participant was asked to choose the retargeted image with better quality. To avoid unreliable user input such as random picking, we extend the web-based user study in RetargetMe by adding *checkpoint input* and *time check* for quality control. Any user input failed in either of these two checks is discarded.

Checkpoint input refers to three special pairs of retargeted images with obvious preference (Figure 3). In each user study session, these image pairs were randomly distributed, in which the obviously better images were located on the left in two occasions and on the right in one occasion. According to our preparatory experiments, participants with high concentration can easily choose correct images, while those who just randomly select images are likely to fail in at least one checkpoint input.

Time check is a constraint that the average selection time for an input image pair should not be shorter than 3 sec-

[2]There are 80 groups in RegargetMe. Only 37 of them have subjective preference scores. From the remaining groups, we chose all the groups without substantial similarity to those in the original 37 groups.

	Lines/edges	Faces/people	Texture	Foreground objects	Geometric structure	Symmetry	All
Liang's [12]	0.250	0.381	0.214	0.295	0.082	0.232	0.313
ARS [33]	0.351	0.345	0.571	0.371	**0.388**	**0.607**	0.313
Ours	**0.393**	**0.524**	**0.786**	**0.400**	0.347	0.339	**0.407**

Table 3. The mean Kendall correlation coefficients of 26 groups of images in our novel dataset. The best result of each type is shown in bold.

onds. In our preparatory experiments, we found that setting a fixed time limit for each image pair does not provide reliable indication as some cases are genuinely easier to decide than others. However, the average selection time is effective in differentiating reliable and unreliable user input. A participant who randomly selects images may still pass the checkpoint input test by chance, but their average selection time is likely to be much shorter than proper input.

We employed 232 participants who were postgraduate students in research labs from Australia, UK, Canada, China and USA. 168 of them passed all the checks and their subjective scores were collected for 26 groups of images.

To evaluate the generalizability of OQA methods, we use 37 groups of images from RetargetMe with provided subjective scores as the training set. The trained model is then applied to the novel dataset with 26 new groups of images. We compare the top three OQA methods (i.e., Liang's method [12], ARS [33] and our method) as indicated in Table 2. The results are summarized in Table 3, demonstrating that our method significantly outperforms Liang's method and ARS, and has better generalizability.

Since our method obtains the ranking of a group of retargeted images based on pairwise comparison, we further test the reliability of the method when only a subset of images are provided as input. In this test, we still use 37 groups of images from RetargetMe with provided subjective scores as the training set, but only use 7 of 8 retargeted images in each group of our novel dataset as the test set. The test is repeated 8 times and at each time a different retargeted image in each group is removed. The results of 8 tests are summarized in Table 5, demonstrating that our method significantly outperforms Liang's method and ARS, and has better reliability.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
Liang's [12]	0.286	0.355	0.304	0.275	0.311	0.282	0.374	0.319
ARS [33]	0.330	0.249	0.322	0.363	0.260	**0.363**	0.322	0.297
Ours	**0.410**	**0.469**	**0.399**	**0.396**	**0.432**	**0.363**	**0.447**	**0.381**

Table 5. The mean Kendall correlation coefficients of ranking 7 retargeted images in each group in our novel dataset, using Liang's method, ARS and our method. Each group has 8 retargeted images and each of them is removed in turn, resulting in eight tests. The best result in each test is shown in bold.

5. Conclusion

In this paper, we propose a simple yet effective two-step OQA method based on a machine learning framework. After representing a retargeted image in a six-dimensional representation using six metrics in [12], we transform this six-dimensional representation into a more regular feature space by applying the NPE [9]. Based on this feature representation, in the first step, we construct an objective function $\widetilde{F}$ by training a GRNN model with the subjective preference scores from RetargetMe [25]. For symmetry images, an additional feature Q_7 is further introduced into a composite model $\widetilde{F}'$ in Eq.(12). Both $\widetilde{F}$ and $\widetilde{F}'$ compute a relative score for each pair of retargeted images. In the second step, all relative scores are transformed into a global ranking. Our experiments show that our method consistently and significantly outperforms eight representative OQA methods, and correlates better with users' subjective preferences by means of a leave-one-out cross validation test in RetargetMe and a generalizability test in a new user study. In addition to the metrics defined in [12], our learning-based method is general and may benefit from including new metrics such as ARS [33] in the future work.

Acknowledgment

This work was supported by the National Key Research and Development Plan (2016YFB1001202), the Natural Science Foundation of China (61521002, 61432003, 61661130156), TNList Foundation and Royal Society-Newton Advanced Fellowship (NA150431).

References

[1] Recommendation, ITURBT and BT. 500-11. Methodology for the subjective assessment of the quality of television pictures. *ITU Telecom. Standardization Sector of ITU*, 2002. 2

[2] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *24th International Conference on Machine Learning (ICML)*, pages 129–136, 2007. 3

[3] S. Castillo, T. Judd, and D. Gutierrez. Using eye-tracking to assess different image retargeting methods. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 7–14. ACM, 2011. 2

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *9th European Conference on Computer Vision (ECCV)*, pages 288–301. Springer-Verlag, 2006. 4

[5] H. A. David. *The Method of Paired Comparisons*, volume 12. DTIC Document, 1963. 5, 7

[6] K.-L. Du and M. N. Swamy. *Neural Networks and Statistical Learning*. Springer London, 2013. 3

[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(Nov):933–969, 2003. 3

[8] N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Trans. Information Systems*, 7(3):183–204, 1989. 3

[9] X. He, D. Cai, S. Yan, and H.-J. Zhang. Neighborhood preserving embedding. In *10th IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1208–1213, 2005. 5, 7, 8

[10] E. Kasutani and A. Yamada. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 674–677, 2001. 2, 6, 7

[11] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 6

[12] Y. Liang, Y.-J. Liu, and D. Gutierrez. Objective quality prediction of image retargeting algorithms. *IEEE Trans. Vis. Comp. Graph.*, 23(2):1099–1110, 2017. 1, 2, 3, 4, 6, 7, 8

[13] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *10th European Conference on Computer Vision (ECCV)*, pages 28–42, 2008. 2, 6, 7

[14] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu. Image retargeting quality assessment. In *Comp. Graph. Forum*, volume 30, pages 583–592, 2011. 2, 4, 6, 7

[15] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *10th European Conference on Computer Vision (ECCV)*, pages 386–399. Springer-Verlag, 2008. 4

[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. Top. Signal Processing*, 6(6):626–639, 2012. 2, 3

[17] L. Ma, W. Lin, C. Deng, and K. N. Ngan. Study of subjective and objective quality assessment of retargeted images. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2677–2680, 2012. 3

[18] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, 2001. 2, 6, 7

[19] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: multimedia content description interface*, volume 1. John Wiley & Sons, 2002. 2

[20] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, 2011. 4

[21] C. L. Novak and S. A. Shafer. Anatomy of a color histogram. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 599–605, 1992. 4

[22] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, 2010. 5

[23] O. Pele and M. Werman. Fast and robust earth mover's distances. In *12th IEEE International Conference on Computer Vision (ICCV)*, pages 460–467, 2009. 2

[24] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 4, 7

[25] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. *ACM Trans. Graph.*, 29(6):160, 2010. 1, 2, 3, 5, 6, 7, 8

[26] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. *ACM Trans. Graph.*, 28(3):23, 2009. 2, 6, 7

[27] A. Shamir, O. Sorkine, and A. Hornung. Modern approaches to media retargeting. *SIGGRAPH Asia Courses*, 2012. 2

[28] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2, 4, 6, 7

[29] D. F. Specht. A general regression neural network. *IEEE Trans. Neural Networks*, 2(6):568–576, 1991. 2, 3, 7

[30] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 4, 7

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 4

[32] H. Wu, Y.-S. Wang, K.-C. Feng, T.-T. Wong, T.-Y. Lee, and P.-A. Heng. Resizing by symmetry-summarization. *ACM Trans. Graph.*, 29(6):159:1–159:10, 2010. 4

[33] 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 Processing*, 25(9):4286–4293, 2016. 1, 2, 6, 7, 8