A Comparative Study of Image Retargeting

Michael Rubinstein
MIT CSAIL

Diego Gutierrez
Universidad de Zaragoza

Olga Sorkine
New York University

Ariel Shamir
The Interdisciplinary Center

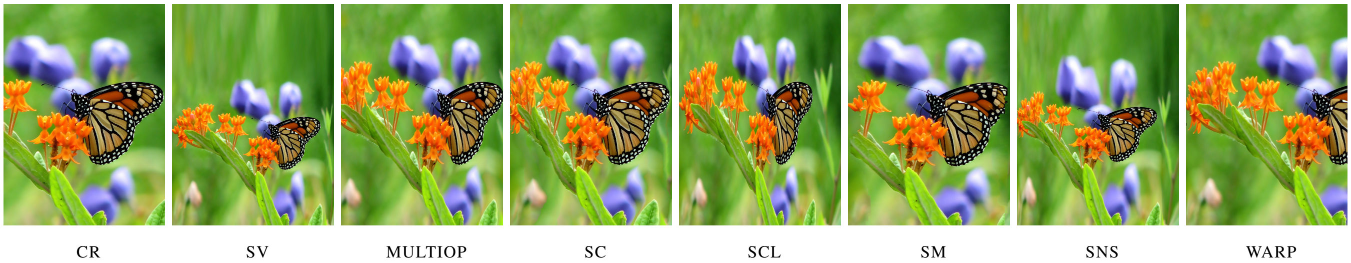


Figure 1: *Example of retargeting the* `butterfly` *image shown in Figure 2 to half its size. In this study we evaluate 8 different image retargeting methods, asking users to compare their results and examine what qualities in retargeted images mattered to them. We also correlate the users' preferences with automatic image similarity measures. Our findings provide insights on the retargeting problem, and present a clear benchmark for future research in the field.*

Abstract

The numerous works on media retargeting call for a methodological approach for evaluating retargeting results. We present the first comprehensive perceptual study and analysis of image retargeting. First, we create a benchmark of images and conduct a large scale user study to compare a representative number of state-of-the-art retargeting methods. Second, we present analysis of the users' responses, where we find that humans in general agree on the evaluation of the results and show that some retargeting methods are consistently more favorable than others. Third, we examine whether computational image distance metrics can predict human retargeting perception. We show that current measures used in this context are not necessarily consistent with human rankings, and demonstrate that better results can be achieved using image features that were not previously considered for this task. We also reveal specific qualities in retargeted media that are more important for viewers. The importance of our work lies in promoting better measures to assess and guide retargeting algorithms in the future. The full benchmark we collected, including all images, retargeted results, and the collected user data, are available to the research community for further investigation at *http://people.csail.mit.edu/mrub/retargetme*.

Keywords: media retargeting, benchmark, user study

1 Introduction

Content-aware media retargeting has drawn much attention in graphics and vision research in recent years (See [Shamir and Sorkine 2009] for a detailed background). However, little work has been done to *methodologically* evaluate the results of retargeting methods both quantitatively and qualitatively. Some works conducted small-scale user studies to support their evaluation, but most resort to simple visual comparison of results – typically involving a small set of images, and a small subset of previous methods. There is a clear need to create a benchmark and a principled evaluation framework, not only to evaluate current methods, but also to enable a more structured comparison of results in the *future*. In this paper we present such a benchmark and framework.

Collecting pure ground-truth retargeting data is challenging, and quite different from collecting user data on segmentation [Martin et al. 2001; Chen et al. 2009], line-drawing [Cole et al. 2008] or optical flow [Baker et al. 2007]. Manual retargeting requires a proficient artist. One possibility is giving an image to several artists and have them resize it to the required dimensions. Such a task is extremely laborious, and will greatly limit the size of the benchmark. It is also not clear what set of tools the artist should use in this process. For example, we can allow the artist to employ some of the suggested retargeting techniques in addition to standard editing tools, but this may clearly insert bias towards one method over another. This experiment setup is insufficiently constrained, and might end up in different results by different artists which would be difficult to compare and analyze. Moreover, we aim at collecting large scale user data, spanning a wide variety of viewers, and not just artists. We therefore choose to concentrate on a comparative study of existing retargeting methods.

Creating a benchmark image set for retargeting is not enough. Retargeting poses an inherent difficulty for evaluation. First, many times the results could depend on the media content itself: one method might work best on certain types of images, while another on different ones. Second, such evaluation is considered highly subjective. To date, there is no objective computational measure that can evaluate retargeting quality since it is mainly a perceptual criterion. Third, it is not clear if such a measure should be sought after: perhaps different people prefer different results and there is no consensus even among humans on retargeting evaluation?

Our goal in this work is to advance the understanding in all of the above questions and to provide a common ground for comparison between existing and future retargeting methods. We have conducted a comprehensive user study rating the results produced by eight different retargeting methods on a predefined set of images.

Our results clearly indicate an answer to the first and third questions: there *is* a general consensus among people regarding the evaluation of retargeting results. Our *subjective* analysis indicates that it is sensible to search for a computational measure of quality. We also provide the means to do this by offering a ground-truth ranking of various retargeting results.

When examining the smorgasbord of retargeting methods presented up to date, three main objectives are usually mentioned:

1. Preserving the important *content* of the original media.
2. Limiting visual *artifacts* in the resulting media.
3. Preserving internal *structures* of the original media.

Again, it may seem that the importance of each objective can change not only between different images, but also between different viewers. To this end, we defined a set of image attributes that could be mapped to these objectives and examined human evaluations based on these attributes. We found that viewers consistently demonstrate high sensitivity to deformation, particularly for images that include specific types of content like faces, well defined geometric structures and symmetry. Interestingly, in many cases users prefer sacrificing content over inserting deformation to the media. Our study further shows that these findings are invariant to whether or not the users are aware of the original (non-resized) content.

In an *objective* analysis, we attain somewhat surprising results concerning finding objective measures for retargeting evaluation. We compare computational measures that were suggested in this context to the labeled data, and show that they *do not* anticipate human retargeting perception well. Driven by this finding, we suggest ways to improve on those measures, and demonstrate better results using two image similarity measures which were not previously used to estimate retargeting quality. We believe our insights can help define new measures which will enable to better assess or even guide retargeting algorithms in the future.

2 The Benchmark

Content-aware retargeting methods work best on images where some content can be disposed of. These include either smooth or irregularly-textured areas such as sky, water, grass, or trees. On such images most retargeting methods would work sufficiently well. Challenge is posed in images containing either dense information or global and local structures that may be damaged during resizing. To create our benchmark set, we first gathered images from various retargeting papers. Additionally, based on insight gained from those works, we chose a set of image attributes that could be mapped to the three major retargeting objectives (preserving content, preserving structure and preventing artifacts), and gathered images containing such attributes. These attributes are: *people and faces*, *lines and/or clear edges*, evident *foreground objects*, *texture* elements or repeating patterns, specific *geometric structures*, and *symmetry*. The final benchmark is made up of 80 images having one or more of these attributes.

Retargeting Methods. Media retargeting methods can be classified as discrete or continuous [Shamir and Sorkine 2009]. Discrete approaches remove or insert pixels (or patches) judiciously to preserve content, while continuous solutions optimize a mapping (warp) from the source media size to the target size, constrained on its important regions and permissible deformations. The set of retargeting methods used in our study covers most of the recent major publications in the field, and equally samples from these two approaches. Those are: Nonhomogeneous warping (WARP) [Wolf et al. 2007], Seam-Carving (SC) [Rubinstein et al. 2008], Scale-and-Stretch (SNS) [Wang et al. 2008], Multi-operator (MULTIOP) [Rubinstein et al. 2009], Shift-maps (SM) [Pritch et al. 2009], Streaming Video (SV) [Krähenbühl et al. 2009], and Energy-based deformation (LG) [Karni et al. 2009][1].

We also used the results of a simple scaling operator (SCL), as well as manually chosen cropping windows (CR). The comparison to cropping is of particular interest in order to investigate the perceptual tradeoff between deformation and content removal. For the convenience of the reader, we supply a succinct summary of each operator in our supplemental material. Sample results produced by the methods we use are shown in Figure 1.

Retargeted Images. Given that some methods only support one-dimensional resizing, we restricted the changes to either the width or the height of the image. We concentrated on reduction in image size, and chose to use considerable resizing (25% or 50%) as most methods will work reasonably well for small changes. For accuracy of the experiment, we asked the original authors of each method to retarget the images. Note that different methods may be guided by different importance criteria on the image, as those are often not easily separable from the operator itself. For reasons of design and manageability of the experiments (see Section 3), we chose a subset of 37 images to conduct our user study.

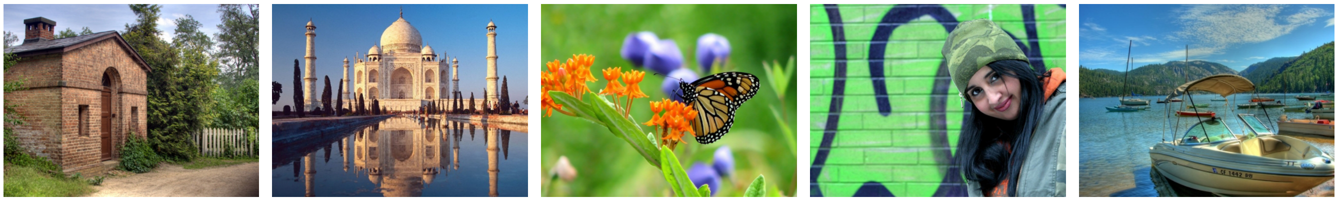
Finally, in a pilot study we classified these 37 images according to the selected attributes (the numbers in parentheses indicate how many images belong to each set): *lines/edges* (25), *faces/people* (15), *texture* (6), *foreground objects* (18), *geometric structures* (16) and *symmetry* (6). Note that one image can belong to several different sets, since it can contain several attributes. This classification sheds more light on the performance of the methods based on a high-level description of the image content. Figure 2 shows some examples of the input images used, along with the attributes assigned to each one during the pilot study. The full image set and classifications are given in the supplementary material.

3 Subjective Analysis

We aim at comparing the retargeting results from an *observer's* perspective, which requires multiple stimuli with differences between them often being quite subtle (see Figure 1). More importantly, the quality of the results that we aim to measure cannot be represented in a linear scale [Kendall and Babington-Smith 1940], which advises against ranking methods. We thus chose the *paired comparisons* technique, where the participants are shown two retargeted images at a time, side by side, and are asked to simply choose the one they like better. A web-based interface allowed them to conveniently switch between the two retargeted results in order to make the differences between them more apparent, and to view the original image as well (please refer to the supplementary material for screen shots and demonstration of the survey system).

Given our set of images and the eight methods tested, the total number of possible paired comparisons is too large: $\binom{8}{2} = 28$ per image × 37 images = 1036 comparisons. It is therefore unrealistic to ask a participant to perform a complete test while maintaining the necessary level of attention. Thus, we need to sample this space of possible comparisons in a way that ensures a solid statistical analysis. Kendall [1955] and Bose [1955] introduced the problem of what constitutes a satisfactory subset of the comparisons when the total number of comparisons is too large. Building on that, we follow the *linked-paired comparison design* [David 1963], which allows to measure not only the performance of the algorithms, but the agreement between participants as well.

[1] Due to scheduling (we received the retargeting results after the user study began) we did not use the LG method in our analysis, but the results are still included in the benchmark for the benefit of future studies.



`Brick House` (L, T, G) `Taj Mahal` (L, G, S) `butterfly` (F, G) `Fatem` (L, P, T, G) `boat` (L, F)

Figure 2: *Samples of the images used in our tests, spanning the range of attributes taken into account: lines/edges (L), faces/people (P), texture (T), foreground objects (F), geometric structures (G) and symmetry (S).*

$\mathbf{p_1}$	0-5	1-4	2-3	6-7	4-2	5-1	6-0	3-7	6-4	0-3	1-2	5-7
$\mathbf{p_2}$	1-6	2-5	3-4	0-7	5-3	6-2	0-1	4-7	0-5	1-4	2-3	6-7
$\mathbf{p_3}$	2-0	3-6	4-5	1-7	6-4	0-3	1-2	5-7	1-6	2-5	3-4	0-7
$\mathbf{p_4}$	3-1	4-0	5-6	2-7	0-5	1-4	2-3	6-7	2-0	3-6	4-5	1-7
$\mathbf{p_5}$	4-2	5-1	6-0	3-7	1-6	2-5	3-4	0-7	3-1	4-0	5-6	2-7
$\mathbf{p_6}$	5-3	6-2	0-1	4-7	2-0	3-6	4-5	1-7	4-2	5-1	6-0	3-7
$\mathbf{p_7}$	6-4	0-3	1-2	5-7	3-1	4-0	5-6	2-7	5-3	6-2	0-1	4-7

Table 1: *Linked-paired comparison design for a single image. The eight methods tested are generically numbered* $[0..7]$*, while* p_i *denotes the participant number. Each participant thus performs 12 of the total of 28 possible paired comparisons per original image, according to the parameters chosen in the design (see text for details).*

	lines/ edges	faces/ people	texture	foreground objects	geometric structures	symmetry	Aggregate
u (with ref.)	0.073	0.166	0.070	0.146	0.084	0.132	0.095
u (no ref.)	0.047	0.086	0.027	0.075	0.059	0.054	0.059
R'	107	83	53	91	85	53	129

Table 2: *Agreement of results from the paired comparison study with and without a reference image. For the reference version (the regular version of our test), there is clearly more agreement between participants for the* faces/people*,* foreground objects *and* symmetry *sets. Without a reference image, the agreement drops significantly. In both cases and for all categories, the coefficient of agreement is statistically significant at* $p < 0.01$*. The values of* R' *(Equation 2) are used for the grouping in Figure 4.*

To ensure that the experiment is balanced by comparisons and by participants, the test should be designed such that:

- Each pair is compared by the same number k of participants.
- Within the pairs compared by each participant, each stimulus appears an equal number of times β.
- Given any two participants, there are exactly λ pairs compared by both of them.

The parameters we used in our design were $k = 3$, $\beta = 3$ and $\lambda = 4$. According to these parameters, and following the derivation by David [1963] (see Table 1), each participant is assigned 12 out of the total of 28 possible paired comparisons per image. To provide a complete set of three results per pair ($\beta = 3$) seven participants are required, arriving at a total of $28 \cdot 3 = 84$ votes cast per image. Each participant would judge between three and five images (at 12 comparisons per image). To ensure more robust statistics, we collected three complete sets per image, meaning that each image was judged by 21 participants, yielding a total of 252 votes.

A total of 210 participants took part in the test, casting a total of 9324 votes. About half of the participants were volunteers and half workers from Amazon Mechanical Turk. Mechanical Turk was successfully used before [Cole et al. 2009], and in fact the comments we received from participants were very positive (they enjoyed the test and found it interesting; see supplementary material). About 40% were females and 60% males, average age was around 30, and they had varying degrees of computer graphics knowledge, being naïve as to the design and goals of the experiment. To investigate whether knowledge of the original content affects the preferred resized result, we conducted a *blind* version of the exact same test (using 210 *new* participants), where the original image was not shown. This perhaps simulates better the real-world scenario, where humans are typically exposed to edited media, and are unaware of the original (unedited) source. We refer to this version as *no reference image* test and discuss it later in this section.

Additionally, to gain more insight on the reasons to choose one result over another, the participants were occasionally asked to pick one or several items out of a proposed set of reasons for *not* choosing a result. This question appeared randomly with a probability of $1/6$, a frequency we found suitable in order to maintain the participant's attention without making the test tedious. Table 5 shows the complete list of reasons by image attribute (note that five of them are common to all six attributes).

3.1 Analysis and Discussion

Agreement. We are first interested in studying the similarity of choices between participants; all participants would be in complete agreement if they voted the same way. High disagreement, on the other hand, reflects difficulty making choices, suggesting either that the stimuli were very similar or that users tend not to agree. For this purpose, Kendall and Babington-Smith introduced the *coefficient of agreement* [1940], defined as:

$$u = \frac{2\Sigma}{\binom{m}{2}\binom{t}{2}} - 1, \quad \text{where} \quad \Sigma = \sum_{i=1}^{t}\sum_{j=1}^{t}\binom{a_{ij}}{2} \qquad (1)$$

where a_{ij} is the number of times that method i was chosen over method j, m is the number of participants (which varies depending on whether we are analyzing a single image, a set of images or the combined choices over all images), and $t = 8$ is the number of retargeting methods tested. If all participants are in complete agreement, then $u = 1$; the minimum value of u is attained by an even distribution of answers and is given by $u = -1/m$.

The coefficient over all images is $u = 0.095$, a relatively low value suggesting that the participants *in general* had difficulty judging. However, by analyzing the images according to their attributes we find (see Table 2) that the three sets defined by *faces/people*, *foreground objects* and *symmetry* clearly show greater agreement. The statistical significance of u can be determined by testing the null hypothesis that the comparisons are assigned randomly (no agreement amongst users). A χ^2 test shows that u is statistically significant at the significance level of 0.01 in all six categories.

Ranking. Figure 3 shows the eight methods, ranked by the number of votes received (number of times a method was preferred over a

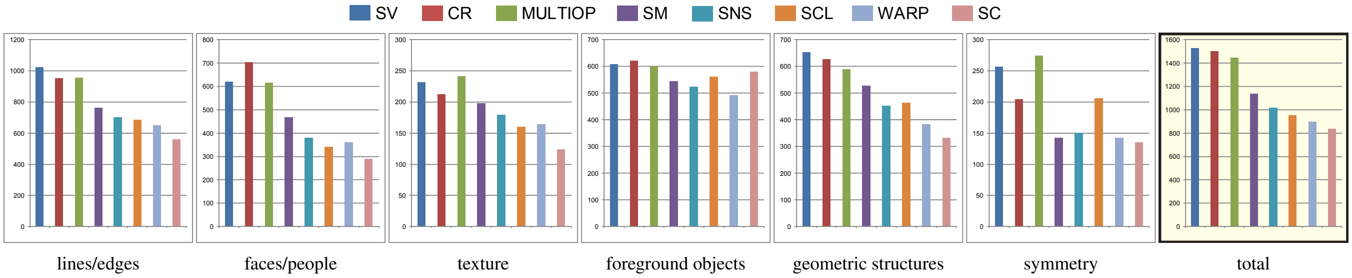


Figure 3: *The number of votes and total ranking (rightmost) of the eight methods per attribute, when the reference (original) image was shown. We notice that three operators, namely* SV, MULTIOP *and* CR *consistently rank better than the others.*

Rank	SV	MULTIOP	CR	SM	SCL	SNS	WARP	SC
Ψ (with ref.)	1.59	1.94	2.03	4.58	5.29	5.45	6.80	7.13
Rank	CR	MULTIOP	SV	SM	SNS	SCL	WARP	SC
Ψ (no ref.)	1.44	1.91	2.18	4.23	5.45	5.86	6.63	7.38

Table 3: *The eight methods sorted by their rank products, with a reference image (top row) and without (bottom row). Smaller result indicates better ranking (operator more favored by users).*

lines/ edges	faces/ people	texture	foreground objects	geometric structures	symmetry	Aggregate	Rank product
0.964	0.988	0.946	0.737	0.950	0.957	0.978	0.985

Table 4: *Correlation coefficients between the reference and no-reference tests. The high correlation between the two versions indicate that the presence of the source image in the test did not have large effect on participants' choices.*

different method). We show both the global result, as well as the results per attribute. Table 3 (top row) shows the results of the *rank product* $\Psi(\mathcal{O}) = (\prod_i r_{\mathcal{O},i})^{1/b}$, where $r_{\mathcal{O},i}$ is the specific ranking for method $\mathcal{O}$ and category i ($i = 1..b$), in all six categories.

In order to analyze the true meaning of these rankings, we perform a *significance test* of the score differences. This reveals whether any two retargeting algorithms produced results that were statistically indistinguishable (and thus can be considered to belong to the same group), or were perceived as clearly different (belonging to different groups). Following the approach of Setyawan and Lagendijk [2004], we need to find a value R' for which the variance-normalized range of scores within each group is lower or equal. The value of R' depends on the confidence level α, which means that we need to compute R' so that $P[R \geq R'] \leq \alpha$. We set again $\alpha = 0.01$. It can be shown [David 1963] that R' can be obtained from:

$$P\left(W_{t,\alpha} \geq \left(2R' - 0.5\right)/\sqrt{mt},\right) \qquad (2)$$

where the value of $W_{t,\alpha}$ has been tabulated by Pearson and Hartley [1966]. In our case, $W_{8,0.01} = 4.9884$ which yields the R' values shown in Table 2.

Figure 4 shows the resulting groups for each attribute and for the combined analysis. An interesting and important finding of this test is that three algorithms (CR, SV and MULTIOP) consistently stand out from the rest and usually yield results that can be considered perceptually similar in terms of ranking, whereas another group of algorithms (SCL, SC and WARP) was consistently ranked the lowest, also yielding statistically undistinguishable results.

Discussion. Our analysis shows a clear distinction in performance among existing methods, and additionally provides some insights that can help future design of retargeting methods. The results of the coefficient of agreement u suggest that: *(i)* detecting salient areas of the images at *object* level may be valuable in a retargeting context, and *(ii)* there is a correlation with the fact that symmetry detection is an important mechanism of human perception to identify object structure [Tyler 1996; van der Helm 2000], and it may be the most important *structural* aspect that retargeting algorithms need to maintain.

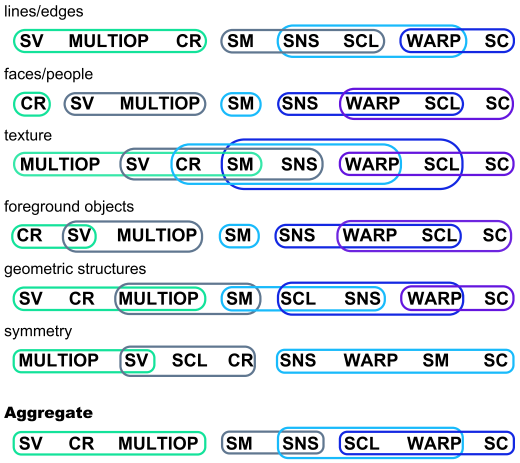

lines/edges
SV MULTIOP CR SM SNS SCL WARP SC
faces/people
CR SV MULTIOP SM SNS WARP SCL SC
texture
MULTIOP SV CR SM SNS WARP SCL SC
foreground objects
CR SV MULTIOP SM SNS WARP SCL SC
geometric structures
SV CR MULTIOP SM SCL SNS WARP SC
symmetry
MULTIOP SV SCL CR SNS WARP SM SC
Aggregate
SV CR MULTIOP SM SNS SCL WARP SC

Figure 4: *Grouping of the algorithms per attribute for the reference version of the study. Operators are ordered according to received votes from left (more votes) to right (less votes). Operators within a group are statistically indistinguishable in terms of user preference.*

The significance test of the ranking results reveals clear and consistent grouping of algorithmic performance. Interestingly, two of the content-aware methods that ranked highest use very different approaches. SV relies on complex analysis of image importance combined with various constraints. MULTIOP, on the other hand, uses simple operators and simple image features but combines them together effectively. The third one, CR, is the only operator that, by definition, does not create any artifacts. This suggests that loss of content is generally preferred over deformation artifacts. The search for optimal cropping windows (see e.g. [Liu and Gleicher 2006]), which somewhat lost its place in recent years to more sophisticated deformation-based methods, is still very much a valid and relevant research venue.

Although we allowed a certain degree of user guidance in SV, we found no statistical difference in the total number of votes between the images with user intervention (average of 41.83) and those without (average of 41.19). Similarly, it may seem that MULTIOP achieved its ranking because it was using cropping, which was

Attribute	Reason	ID
lines/edges	Lines or edges were broken	1
lines/edges	Lines or edges were distorted	2
faces/people	People or faces were squeezed	3
faces/people	People or faces were stretched	4
faces/people	People or faces were deformed	5
texture	Textures were distorted	6
foreground objects	Foreground objects were squeezed	7
foreground objects	Foreground objects were stretched	8
foreground objects	Foreground objects were deformed	9
geometric structures	Geometric structures were distorted	10
symmetry	Symmetry was violated	11
Common	Content was removed or cut-off	12
Common	Proportions in the image were changed	13
Common	Smooth image areas were destroyed or removed	14
Common	Can't put my finger on it. The other result was simply more appealing	15
Common	Other	16

Table 5: *Proposed reasons for* not *choosing a result, grouped by image attributes. The last five are common to all attributes and were always offered to the participants.*

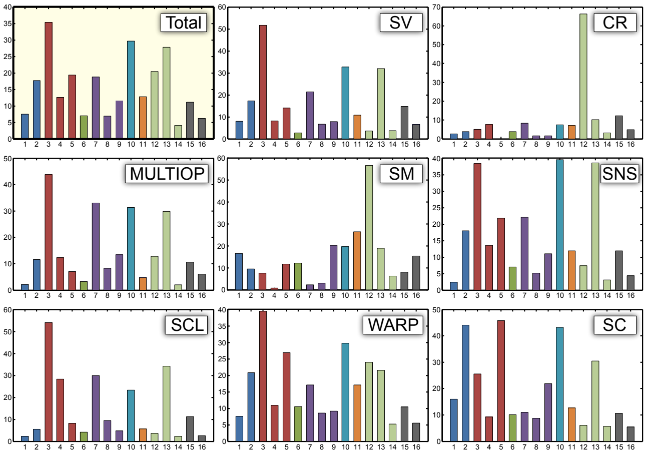


Figure 5: *Percentage of times each* reason for not picking an image *was selected, over the total number of times it was shown. We show the overall distribution and the breakdown per operator. Please refer to Table 5 for the list of reasons and their IDs.*

ranked high as well. However, the normalized means and standard deviations of the ratio of operations used in the MULTIOP results are $(0.5750, 0.1920)$ for scaling, $(0.3195, 0.1896)$ for seam carving and $(0.1055, 0.1164)$ for cropping. Clearly, it is the *combination* of the three which yields favorable results.

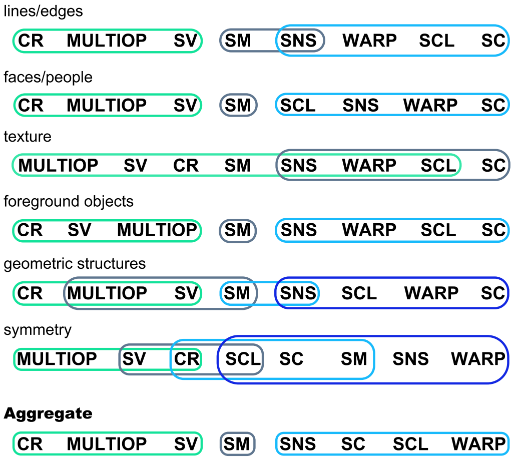
No-reference Comparison. As stated above, we repeated the experiment with a new set of participants, where the original image was not shown. As expected, the results show lower agreement between participants in general (see Table 2, bottom row), given that the comparison is less constrained without explicit knowledge of the source image. Note that although the agreement is lower, the χ^2 test still shows that u is significant at the confidence level of $\alpha = 0.01$.

Still, the second experiment provides some interesting findings. First, there is an extremely high correlation between the tests with and without a reference image (see Table 4). The rank product results in Table 3 (bottom row) show again a very similar pattern with respect to the test with reference image. Second, the results of the significance test of the score differences are also very similar: again CR, SV and MULTIOP are consistently ranked better than the rest and are perceived as similar, whereas SCL, SC and WARP are always grouped together and produce the least satisfying results (see Figure 6). The main difference, as expected, is that cropping was almost *always* the preferred choice: with no reference image to intimate the loss of content, and by not introducing any artifacts, cropping presents a clear advantage over the other methods.

In summary, except for an overall slightly better performance of CR, we found no significant differences between the two tests, which shows that the participants' preferences are independent of whether or not they are aware of the original image.

Additional Questions. Analyzing the relative frequency of the responses to our additional questions (see Table 5 and Figure 5), it can be seen how the three main reasons for rejecting an image result are: *people or faces were squeezed*, *geometric structures were distorted* and *proportions in the image were changed*. Although our choice of proposed reasons was not meant to be exhaustive, this result suggests what kind of distortions retargeting operators should avoid.

We further analyze this distribution of answers with respect to each of the operators (see Figure 5). We found that users in general did not resort to the last two reasons (that were always offered), which



lines/edges
CR MULTIOP SV SM SNS WARP SCL SC
faces/people
CR MULTIOP SV SM SCL SNS WARP SC
texture
MULTIOP SV CR SM SNS WARP SCL SC
foreground objects
CR SV MULTIOP SM SNS WARP SCL SC
geometric structures
CR MULTIOP SV SM SNS SCL WARP SC
symmetry
MULTIOP SV CR SCL SC SM SNS WARP
Aggregate
CR MULTIOP SV SM SNS SC SCL WARP

Figure 6: *No-reference version: grouping of the algorithms per attribute.*

indicates that they found a sufficiently close answer in the composed list, and that they usually knew to signify what bothered them in a result they did not select. For the SM operator, the main reason for not choosing it was removed or cut-off content. Indeed, this operator shrinks the image by gracefully removing parts of it, which in some cases were important for the users to maintain. The reasons for rejecting the SC operator results were dominated by the distortion of lines and edges, and deformation of people and objects. This operator is susceptible to such artifacts due to the discrete and local nature of its carving process. SCL results that were not chosen by the users, tend to suffer from over-squeezing, or stretching, of content. For the CR operator, almost all responses focused, not surprisingly, on removal of content. Removal of content also bothered the users in the context of the WARP operator, which might collapse regions in the image during its deformation if not enough areas of homogeneous content are found. The remaining methods tend to correlate with the global distribution of reasons. Note that SV and especially SNS were often rejected due to distortion of proportions: indeed, these two operators allow uniform scaling of image content (SNS allows the scaling factor to vary and SV fixes a global scaling factor for the entire image).

4 Objective Analysis

The main question we consider in this part of our work is whether computational distance measures between images can predict human retargeting preferences. This is important for two reasons; first, we could use such measures to compare new operators with the labeled data to see whether they improve on previous results; second, we could ideally incorporate those measures in a retargeting framework, such that optimizing them would provide satisfactory results.

Numerous image similarity measures were proposed to assess the likeness of content in two images in the context of various image processing applications, such as image compression, search, and editing. Many of the well-established image distance metrics[2] such as signal-to-noise ratio and structure similarity, have been mostly developed to compare a "ground truth" image with a modified version in terms of content, and as such they work on images of the same size or aspect ratio only. Designing a similarity metric that compares image content under varying aspect ratio is significantly more challenging, since the problem also demands semantical image analysis and content matching.

4.1 Experiment Design

We begin with the two computational measures that were suggested thus far for measuring retargeting quality: Bidirectional similarity (BDS) [Simakov et al. 2008], and Bidirectional Warping (BDW) [Rubinstein et al. 2009][3]. In contrast to these methods, which search for *mid-level* semantic correspondence between images, we additionally wanted to compare to *low-level* measures which treat the image as a whole. For this purpose we used two measures from the MPEG-7 standard [MPEG-7 2002; Manjunath et al. 2001] based on edge histogram (EH) [Manjunath et al. 2001], and color layout (CL) [Kasutani and Yamada 2001]. Both these measures are widely incorporated in content-based image retrieval systems, and can be used for retargeting analysis as they use fixed length signatures regardless of the image size.

Using these image distance measures, we compared each retargeted image to the original one. We used the same image sizes as shown on screen in the user study to adhere to the labeled stimuli. We used the original authors' implementations, as well as their suggested parameter settings, and performed basic parameter optimization to achieve the best result for each measure (i.e., best correlation with the subjective results, see Section 4.2). We refer the reader to the supplemental material for more details on the methods and the parameter settings we used.

4.2 Evaluation

We wish to estimate how well the objective metrics agree with the users' subjective preferences. In this work, we choose to formulate the rate of agreement as the correlation between the *rankings* induced by the subjective and objective measures. For every image I, we define the subjective similarity vector $s = \langle s_1, \ldots, s_n \rangle$ for $n = 8$ methods, where s_i is the number of times the retargeting result T_i using method i was favored over another result (i.e. the higher s_i the better method i is). We also define $o = \langle o_1, \ldots, o_n \rangle$ as the respective objective distance vector for the same image I calculated by one of the objective measures. For a given objective measure D, the entry $o_i = D(I, T_i)$ is the distance between I and T_i with respect to measure D (in this case, the lower o_i the better the method i is). Note that our balanced design (Section 3) guarantees that each result is used equally often over all experiments. Figure 7 (top) shows an example of the vectors s and o for one image and distance measure. Results for all images and measures can be found in the supplemental material.

Correlation. To compare between s and o we first sort them and then rank the retargeting measures according to the sorted order. The subjective vector s is sorted in descending order since it is a similarity measure, while the objective vector o is sorted in ascending order as it is a distance measure. This reduces the problem of comparing s and o to statistically determining the correlation between two rankings, $rank_{desc}(s)$ and $rank_{asc}(o)$ induced by these vectors (Figure 7, middle). We use the *Kendall τ distance* [Kendall 1938] to measure the degree of correlation between the two rankings:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{3}$$

where n is the length of the rankings, n_c is the number of concordant pairs and n_d is the number of discordant pairs over all pairs of entries in the ranking. It is easy to see that $-1 \leq \tau \leq 1$ with increasing value indicating increasing rate of agreement. Notice that $\tau = 1$ in case of perfect agreement (equal rankings), and $\tau = -1$ is case of perfect disagreement. In case $\tau = 0$, the rankings are considered independent.

Significance Test. To measure the significance of a correlation estimate, we need to consider the distribution of the τ coefficient. It turns out that the distribution of τ tends to normality for large n [Kendall 1938]. In our case, we can easily estimate the distribution of τ for $n = 8$ by considering the rank correlation of all possible permutations of 8 elements with regards to an objective order $1, 2, \ldots, 8$. We find that the distribution has normal characteristics, with zero-mean and $\sigma = 0.2887$. For a given set of observed τ coefficients, we use χ^2 test against the null hypothesis that the observed coefficients are randomly sampled from the τ distribution.

4.3 Analysis and Discussion

We gather the distribution of τ scores over all images for each measure (see Figure 7, bottom), and take the mean and variance of this distribution to represent the score of the metric in this experiment. Table 6 presents these scores, with breakdown according to image attribute, and the total score over the entire dataset. The results are shown for the full rank-vectors, and also with respect to the $k = 3$ results ranked highest by each measure. For the latter, we modify Eq. (3) such that only pairs (i, j) for which $(rank_1(i) \leq k \vee rank_1(j) \leq k) \wedge (rank_2(i) \leq k \vee rank_2(j) \leq k)$ are considered, and the denominator is modified to be the total number of such pairs. For reference, we also add in Table 6(a) the results for a random metric, RAND (will be similar for Table 6(b)). For a given pair of images, this measure simply returns a uniformly random number in $(0, 1)$.

As expected, the low-level metrics show smaller overall correspondence with the users, although EH achieves higher scores for images classified as containing apparent geometric structures or symmetries. However, both BDS and BDW show low agreement with the user data as well. The near-zero correlation for nearly all image classes suggests they cannot predict well the users' retargeting preferences. Our claim is that their unsatisfying performance has to do with both the way they construct correspondence between the

[2] We use the terms "measure" and "metric" interchangeably, as commonly done in the related literature. We do not imply, nor rely on, metric properties for any of the distance measures we discuss.

[3] Dong et al. [2009] also defined an image retargeting metric that combines BDS, dominant color and a so-called "seam carving distance"; since the latter is specifically tailored to their retargeting operator, it was difficult to use their measure in this experiment. We do experiment with BDS, as well as a color descriptor, both of which are prominent ingredients in their measure.

Metric	Attribute						Total		
	Lines/Edges	Faces/People	Texture	Foreground Objects	Geometric Structures	Symmetry	Mean	std	p-value
BDS	0.040	0.190	0.060	0.167	-0.004	-0.012	0.083	0.268	0.017
BDW	0.031	0.048	-0.048	0.060	0.004	0.119	0.046	0.181	0.869
EH	0.043	-0.076	-0.060	-0.079	0.103	0.298	0.004	0.334	0.641
CL	-0.023	-0.181	-0.071	-0.183	-0.009	0.214	-0.068	0.301	0.384
RAND	-0.046	-0.014	0.048	-0.032	-0.040	0.143	-0.031	0.284	0.693
SIFTflow	0.097	0.252	**0.119**	0.218	0.085	0.071	0.145	0.262	0.031
EMD	**0.220**	**0.262**	0.107	**0.226**	**0.237**	**0.500**	**0.251**	0.272	1e-5

(a) Complete rank correlation ($k = \infty$)

Metric	Attribute						Total		
	Lines/Edges	Faces/People	Texture	Foreground Objects	Geometric Structures	Symmetry	Mean	std	p-value
BDS	0.062	0.280	0.134	0.249	-0.025	-0.247	0.108	0.532	0.005
BDW	0.213	0.141	0.123	0.115	0.212	0.439	0.200	0.395	0.002
EH	-0.036	-0.207	-0.331	-0.177	0.111	0.294	-0.071	0.593	0.013
CL	-0.307	-0.336	-0.433	-0.519	-0.366	0.088	-0.320	0.543	1e-6
SIFTflow	0.241	**0.428**	**0.312**	**0.442**	**0.303**	0.002	0.298	0.483	1e-6
EMD	**0.301**	0.416	0.216	0.295	0.226	**0.534**	**0.326**	0.496	1e-6

(b) Rank correlation with respect to the three highest rank results ($k = 3$).

Table 6: *Correlation of objective and subjective measures for the complete rank (top) and for the three highest ranked results (bottom). In each column the mean τ correlation coefficient is shown ($-1 \leq \tau \leq 1$), calculated over all images in the dataset with the corresponding attribute. The last three columns show the mean score, standard deviation, and respective p-value over all image types. Highest score in each column appears in bold.*

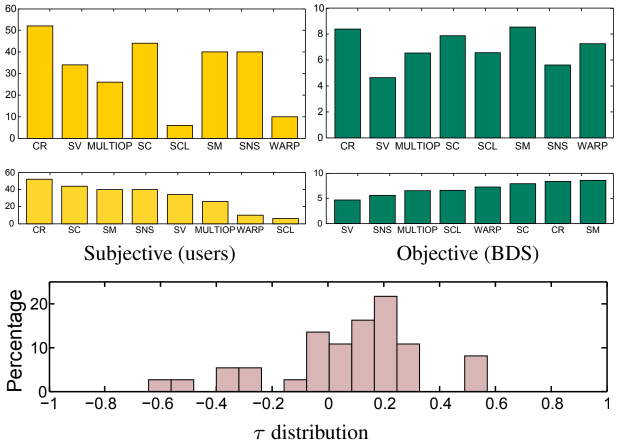


Figure 7: *Correlating the subjective and objective measures. Top left: collected user votes for the* `butterfly` *image. Top right: objective results, in this case using the BDS measure, on the same image. Middle: we measure the similarity between the two by comparing their induced rankings. Bottom: distribution of the τ rank coefficient for the BDS metric over the entire dataset. Subjective and objective measures for all images and metrics can be found in the supplemental material.*

images, and with the image features they use for measuring the distance.

Both measures use intensity-related distance between corresponding patches as indication of differences between the images. Those are strict measures which assign high penalties to patch variations caused by small local deformations (e.g small scale or rotation). Such deformations might be acceptable by human viewers, and so may be reasonable to use for retargeting purposes. As for the correspondence, since BDS uses global patch comparison, a deformed region in the result might be matched to a different part of the original image that has similar appearance. Thus, record of specific changes in content might not be reflected in the distance. BDW does constrain the correspondence such that regions in the result will be matched to approximately the same regions in the original image. However due to its one-dimensional design, it will have difficulty dealing with results produced by some of the operators.

It seems that for measuring retargeting quality, it is crucial to align the images as accurately as possible in order to capture the true modifications the image has undergone. Suggesting a novel solution for constructing better correspondence between the images (or a better descriptor) is beyond the scope of this work. Instead, we demonstrate our claim using two existing image distance measures which were not previously applied in retargeting context: SIFT-flow (SIFTflow) [Liu et al. 2008], and Earth-Mover's Distance (EMD) [Pele and Werman 2009]. Both measures use a dense SIFT descriptor [Lowe 2004], which is known for its ability to robustly capture structural properties of the image, while EMD also uses a state of the art color descriptor (ciede2000). Although the two measures take somewhat different approaches to align the images, they both encourage their solutions to small and smooth local displacements (see formulations in the supplemental material), which seem to relate well to deformations humans are tolerant to, and to operations applied by retargeting operators.

The results using these measures are more encouraging (Table 6). It is evident that EMD and SIFTflow produce rankings which generally better agree with the user labeling in comparison to the other objective measures. EMD shows somewhat better results for the full ranking, while the two are on par with respect to the top-ranked results. In general, we noticed that the measures have stronger correlation with the subjective results on images with faces or people, and evident foreground objects. Table 6 also shows the calculated p-values for this analysis. BDS, SIFTflow and EMD show significant results for $p < 0.01$, and so we can support the fact that EMD and SIFTflow have better correlation with the users with high statistics confidence. The distribution of scores for BDW and the low-level measures (available in the accompanying material) do not allow us to conclude with sufficient confidence that their results do not merely pertain to chance. For $k = 3$, the calculated correlations for all metrics are significant at $\alpha = 0.01$ confidence level.

It is therefore our belief that image descriptors such as SIFT and ciede2000 are more suitable than patch-based distances for conveying local permissible changes in content. Moreover, the constrained alignment produced by these methods also appears to model better the deformations made by retargeting operators, and so provides more reliable content matching for retargeting measures. Note that we do not claim EMD nor SIFTflow are the "correct" solution for an objective retargeting metric. They only unveil the problems in current retargeting measures and suggest a possible direction for future solution and research.

5 Using and Extending the Dataset

Our retargeting dataset, which we name *RetargetMe*, is freely available online at *http://people.csail.mit.edu/mrub/retargetme*. It contains all images and retargeted results used in this study, as well as the collected subjective and objective data. We encourage the reader to visit this web site for further technical details on accessing the data and submitting new results. It also provides a convenient synopsis of the state-of-the-art in image retargeting.

Our evaluation framework can be used in different ways. A new retargeting measure can be evaluated by comparing its rankings on the benchmark images with the ground truth user rankings we collected (Section 4). This favors the *"measure-based"* approach for retargeting, where one would first model the quality of a retargeted result by a measure that can be evaluated on the result directly; compare (and tune) it against the viewers' preferences, and then develop an operator that optimizes the result with respect to this measure. The work of Simakov et al. [2008] is one example that follows this approach [4].

Since the ground truth data we collected is on the comparisons between the results and not the results themselves, a perceptual evaluation of a new retargeting operator requires a new user study. We reiterate that this design decision is motivated mainly by the fact that it is difficult to gather large-scale user feedback on "how should this image be resized?" (as opposed to, say, "what are the segments in this image?" or "where should the lines pass?", see Section 1). On the other hand, deciding between two retargeted results is a much more feasible task to give to a human viewer. To this end we make our (web-based) survey system available, which we hope will ease conducting such studies in the future. Furthermore, depending on the reason for evaluation, a study of a new retargeting operator need not necessarily involve the full all-pair operator comparison as done in this work. For example, to show improvement of the new operator over another requires comparing between those two operators only. To show a new operator advances the current state of the art, it should suffice demonstrating that it improves upon SV, CR and MULTIOP, which significantly and consistently outperformed the rest of the methods in our experiment.

Finally, adding additional images to the dataset will require performing a study comparing all retargeting operators on the new images. We note that given different numbers of images, or different numbers of comparisons per image, other linked-paired comparison designs might serve better than the one we used here (Table 1). Designs corresponding to different parameters can be found in the literature, or derived manually [David 1963].

6 Conclusions

We have presented the first thorough study on image retargeting methods. We gathered a set of images as a benchmark and conducted a large scale user study comparing eight state-of-the-art retargeting algorithms. We further presented an analysis of the correlation between various image distance measures to user resizing preferences.

Authors of a newly suggested retargeting operator (or retargeting measure) will now be able to: *(i)* use our survey system to perform an extensive user study that compares their results and all the previous results we have gathered; *(ii)* analyze their collected data using the proposed evaluation methodology; and *(iii)* present quantitative results as to the performance of their algorithm relative to previous techniques.

Several interesting insights were discovered. In general, more recent algorithms such as SV and MULTIOP indeed outperform their predecessors. Cropping, although a relatively naïve operation, is still one of the most favored methods, most often since it does not create any artifacts. Our findings indicate that the search for an optimal cropping window, which was somewhat abandoned by researchers in the past few years, could often be favorable and should not be overlooked. These conclusions must be tuned by remembering that the images included in the study were deliberately challenging for retargeting methods and the size differences we used were rather extreme. It seems that simple operators such as uniform scaling or seam carving are better suited for small amounts of change, as suggested by the fact that when they are combined together in small amounts (MULTIOP), they produce favorable results. it is interesting to note that the two best performing methods use very distinct approaches: defining complex intelligent algorithms or combining many simple ones.

In terms of objective measures for retargeting, our results show that we are still a long way from imitating human perception. There is a relatively large discrepancy between such measures and the subjective data we collected, and in fact, the most preferred algorithm by human viewers, SV, received low ranking by almost all automatic distance measures. One possible explanation to this is that although the distance measures use multiple scales, they do not match between different resolutions – a phenomenon that may appear in retargeting when different parts of the image are scaled differently. We showed how SIFTflow and EMD, measures not used before for retargeting algorithms, generally agree better with users' preferences under our evaluation criteria. Given their better performance, it should be interesting to try and optimize retargeting with respect to those measures. Nevertheless, it is clear that further research is needed to find new measures that could better represent human perception in a retargeting context. For example, employing machine learning techniques to factor out the contribution of each of the retargeting objectives (Section 1) and training a metric according to the observed user selections could be a promising research direction.

There are many additional opportunities for further analyzing, or building upon, the current data. For instance, other classes of images or specific feature types may be defined on the image set and analyzed. This is of particular importance in retargeting, where the objectives seem to be more content-specific (e.g. text, maps, medical imaging) than in other computer vision domains. In fact, the context/application within which an image is viewed (e.g. while browsing photos on a computer, or reading a news article on a mobile device) might also affect the viewers' perception of the retargeted content. As most content-aware retargeting operators rely on different techniques for estimating the salient regions in the media, it is important to further study the effect of the saliency measure on the results. Last, more data may be added on expanding image sizes and on a more uniform sampling of the general image space (i.e. not necessarily concentrating on difficult images). A symmetric benchmark on video content retargeting is likewise essential, for which a similar methodology may apply. We believe that the benchmark and our evaluation methodology will lead to improved retargeting algorithms and measures, as well as better understanding of the retargeting problem and objectives.

[4] At the time we constructed the benchmark, running [Simakov et al. 2008] on all our images turned out to be infeasible due to its high computational cost, and so it did not take part in our experiment. In the future it would be desirable to add their results, which can now be approximated efficiently following [Barnes et al. 2009]. Note that we did consider their measure for computational analysis (Section 4).

Acknowledgements

The authors would like to thank the individuals who contributed to the collection and analysis of the data in this experiment. In particular, we wish to thank Yu-Shuen Wang, Yael Pritch, Eitam Kav-Venaki, Manuel Lang, Alexander Hornung and Zachi Karni for sharing their images and supplying the results of their retargeting operators, Ofir Pele for his support with the EMD metric, and Susana Castillo for her all-around help. We also wish to thank the anonymous SIGGRAPH Asia reviewers for their valuable comments. The images appearing in this paper were kindly shared by the flickr members extranoise (Brick House), abizeleth (buttefly), meantux (Fatem, through CC license) and John Morgan (boat). The Taj Mahal image is taken from [Rubinstein et al. 2009]. For a detailed list of contributors to our image dataset refer to the accompanying website. This work was supported in part by a Marie Curie grant from the Seventh Framework Programme (grant agreement no.: 251415), the Spanish Ministry of Science and Technology (TIN2010-21543), the Gobierno de Aragón (projects OTRI 2009/0411 and CTPP05/09), NYU URCF, and by gifts from TI, Microsoft Research, and Adobe Systems.

References

Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., and Szeliski, R. 2007. A database and evaluation methodology for optical flow. In *Proc. ICCV*, vol. 5.

Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM TOG 28*, 3.

Bose, R. C. 1955. Paired comparison designs for testing concordance between judges. *Biometrika*, 42, 113–121.

Chen, X., Golovinskiy, A., and Funkhouser, T. 2009. A benchmark for 3D mesh segmentation. *ACM TOG 28*, 3.

Cole, F., Golovinskiy, A., Limpaecher, A., Barros, H., Finkelstein, A., Funkhouser, T., and Rusinkiewicz, S. 2008. Where do people draw lines? *ACM TOG 27*, 3.

Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., and Singh, M. 2009. How well do line drawings depict shape? *ACM TOG 28*, 3.

David, H. 1963. *The Method of Paired Comparison*. Charles Griffin and Company.

Dong, W., Zhou, N., Paul, J.-C., and Zhang, X. 2009. Optimized image resizing using seam carving and scaling. *ACM TOG 28*, 5.

van der Helm, P. 2000. *Principles of Symmetry Perception*. International Congress of Psychology.

Karni, Z., Freedman, D., and Gotsman, C. 2009. Energy-based image deformation. *CGF 28*, 5, 1257–1268.

Kasutani, E., and Yamada, A. 2001. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *International Conference on Image Processing*, 674–677.

Kendall, M. G., and Babington-Smith, B. 1940. On the method of paired comparisons. *Biometrica 31*, 324–345.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 1/2 (June), 81–93.

Kendall, M. G. 1955. Reviews. *Biometrika 42*.

Krähenbühl, P., Lang, M., Hornung, A., and Gross, M. 2009. A system for retargeting of streaming video. *ACM TOG 28*, 5.

Liu, F., and Gleicher, M. 2006. Video retargeting: automating pan and scan. In *MULTIMEDIA*, ACM, 241–250.

Liu, C., Yuen, J., Torralba, A., Sivic, J., and Freeman, W. T. 2008. SIFT Flow: Dense correspondence across different scenes. In *ECCV*, 28–42.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2, 91–110.

Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., and Yamada, A. 2001. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology 11*, 703–715.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings ICCV*, vol. 2.

MPEG-7. 2002. *ISO/IEC 15938: Multimedia Content Description Interface*.

Pearson, E. S., and Hartley, H. O. 1966. *Biometrika Tables for Statisticians*, 3rd ed., vol. 1. Cambridge University Press.

Pele, O., and Werman, M. 2009. Fast and robust earth mover's distances. In *ICCV '09*.

Pritch, Y., Kav-Venaki, E., and Peleg, S. 2009. Shift-map image editing. In *ICCV*, 151–158.

Rubinstein, M., Shamir, A., and Avidan, S. 2008. Improved seam carving for video retargeting. *ACM TOG 27*, 3.

Rubinstein, M., Shamir, A., and Avidan, S. 2009. Multioperator media retargeting. *ACM Trans. Graph. 28*, 3.

Setyawan, I., and Lagendijk, R. L. 2004. Human perception of geometric distortions in images. In *Proceedings of SPIE, Security, Steganography and Watermarking of Multimedia Contents VI*, 256–267.

Shamir, A., and Sorkine, O. 2009. Visual media retargeting. In *ACM SIGGRAPH Asia Courses*.

Simakov, D., Caspi, Y., Shechtman, E., and Irani, M. 2008. Summarizing visual data using bidirectional similarity. In *CVPR*, 1–8.

Tyler, C. W., Ed. 1996. *Human Symmetry Perception and its Computational Analysis*. VSP International Science Publishers, Utrecht.

Wang, Y.-S., Tai, C.-L., Sorkine, O., and Lee, T.-Y. 2008. Optimized scale-and-stretch for image resizing. *ACM TOG 27*, 5.

Wolf, L., Guttmann, M., and Cohen-Or, D. 2007. Non-homogeneous content-driven video-retargeting. In *ICCV*.