A Comparative Study Of Image Retargeting: Supplemental

Here we give more details on the retargeting methods and computational measures used in our study. This document is designed to accompany the paper in order to make the work more self-contained, and is supplied mainly for the convenience of the reader. We strongly encourage the interested reader to additionally refer to the relevant papers for more background on the retargeting operators and measures.

1 Retargeting Methods

Scaling [SCL]: uses simple non-uniform scaling and bi-cubic interpolation to retarget the image.

Cropping [CR]: uses cropping by manually choosing an optimal window of the target size from the original image.

Seam Carving [SC]: uses seam carving “forward energy” as defined in [Rubinstein et al. 2008]. The importance map is simply the intensity gradient magnitude in L_1 metric. Seam carving removes or duplicates contiguous chains of pixels that pass through the regions of least importance in the image; the seams are computed by dynamic programming. No manual intervention was used to create the results.

Shift-maps [SM]: uses graph cut as described in [Pritch et al. 2009] to remove entire objects at a time, rather than single seams. The smoothness term is defined as the importance map that uses both differences in color and gradients between pixels. Several configurations of the algorithm were tested, varying the shift map direction, the saliency map used, and boundary conditions. The most appealing result for each image was chosen manually.

Nonhomogeneous warping [WARP]: uses warping as defined in [Wolf et al. 2007]. To reduce the image width, its columns are non-homogeneously squeezed by optimizing a one-directional warping function. The objective functional in the optimization is weighted by the importance map, such that the amount of allowed deformation is proportional to the importance. The importance map is using L_2 gradient magnitude along with face detectors. No manual intervention was used to create the results.

Scale-and-stretch [SNS]: uses warping as defined in [Wang et al. 2008]. The warping operates on both image dimensions at once, optimizing an objective functional that allows important regions to uniformly scale in order to preserve their shape. The importance map is a combination of L_2 gradient magnitude and saliency as defined by [Itti et al. 1998]. No manual intervention was used to create the results.

Energy-based deformation [LG]: uses warping as defined in [Karni et al. 2009]; similarly to SNS, salient regions are allowed to uniformly scale, and the particular local deformations introduced are granularly controlled according to the importance map (L_2 gradient magnitude). No manual intervention was used to create the results.

Multi-operator [MULTIOP]: uses a combination of seam carving, scaling and cropping as defined in [Rubinstein et al. 2009]. The importance map for seam carving is again L_1 gradient magnitude with “forward energy”. Being a multi-operator method, MULTIOP uses the Bi-Directional Warping measure (see Section 2) to choose between the three operators in a step-wise manner. The size change in each step was 5 pixels. No manual intervention was used to create the results.

Streaming Video [SV]: uses the warping method defined in [Krähenbühl et al. 2009], applying it on images. The importance map is defined by a combination of saliency based on [Guo et al. 2008], line detection and user markings of lines and other important objects. The energy optimized by the warp is a combination of SNS (but fixing the scaling factor for the entire image, whereas SNS allows the scaling to vary), WARP and additional constraints. The latter allow taking special care of lines and curves preservation, and positional constraints for marked regions are possible. Some results (10 in total, 6 in the user study) involved a small amount of manual intervention for protecting objects and straight lines.

There is great diversity in the complexity of both the importance map computation and operators of these methods. Moreover, we also allowed some degree of manual intervention as our primary goal was not to declare a winning operator, but rather to investigate human reaction to different retargeting effects. Note that for reasons of 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.

2 Computational Measures

Here we add more detailed formulations of the objective image similarity metrics [1] we used. We denote the source and target images by S and T, respectively.

Bidirectional Similarity [BDS]. Simakov et al. [2008] formulate similarity between two images as a bidirectional mapping between them. For every patch in one image a well-matching patch is sought in the other image and vice versa. The distance between the images is then defined as the mean distance in color space between corresponding patches:

$$\text{BDS}(S,T) = \frac{\alpha}{N_S}\sum_{P\subset S}\min_{Q\subset T} D(P,Q) + \frac{1-\alpha}{N_T}\sum_{Q\subset T}\min_{P\subset S} D(Q,P) \tag{1}$$

where N_S, N_T are the numbers of patches in S and T respectively, and D is the patch distance function. The parameter α controls the balance between completeness ($S \rightarrow T$) and coherency ($T \rightarrow S$) of the result. In order to capture both local and global similarity, BDS is computed on multiple image scales within a Gaussian Pyramid. For each pyramid level, the objective function in Eq. (1) is optimized using nearest neighbor search.

Barnes et al. [2009] recently proposed PatchMatch to accelerate nearest neighbor search, and demonstrated its effectiveness with BDS for retargeting. The metric is defined similarly, only that the patch correspondence is computed using a randomized algorithm based on an iterative process of random guess and refinement, and is more efficient to compute. We compared the original BDS method with one using PatchMatch-based correspondences [BDS-PM] using our own (CPU) implementation. Our experiments show that BDS and BDS-PM produce very similar results (see Section 4 and

[1] 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.

Table 7 in the paper), regardless of this approximation. We choose to focus on BDS in the paper, while further results using BDS-PM can be found in accompanying material.

Bidirectional Warping [BDW]. Rubinstein et al. [2009] define a similar objective function, with the exception that the mapping $M : (i, j) \mapsto (i', j')$ between the images is constrained to be monotonic. That is, for every two patches $P_1, P_2 \in S$,

$$\begin{aligned} i(P_1) < i(P_2) &\quad \rightarrow \quad i'(P_1) \leq i'(P_2) \\ j(P_1) < j(P_2) &\quad \rightarrow \quad j'(P_1) \leq j'(P_2) \end{aligned} \tag{2}$$

where $i(\cdot), i'(\cdot)$ and $j(\cdot), j'(\cdot)$ are used to indicate the corresponding patch position under the mapping M. The definition is similar for the direction $T \rightarrow S$. The resulting mapping will thus maintain the order of patches in the image, which is essential for properly estimating the quality of a retargeted result [Rubinstein et al. 2009]. The optimal order-preserving mapping is found using iterative evaluations of an asymmetric variation of the dynamic time warp algorithm, and the distance $BDW(S, T)$ is taken to be the mean or maximal distance between corresponding patches in color space.

Dong et al. [2009] define an image similarity measure that combines BDS, dominant color and a so-called "seam carving distance"; they then retarget images by trying to find a combination of linear scaling and seam carving that optimizes this measure. We do not use their measure in this experiment, since the seam carving distance component is specifically tailored to their retargeting operator; we do experiment with BDS, as well as a color descriptor, both of which are prominent ingredients in their measure.

SIFT Flow [SIFTflow]. Liu et al. [2008] propose an algorithm for registering a query image with its neighbors in a large image collection. Although applying this algorithm for a retargeting application has not been attempted yet, this method is attractive for two reasons. First, it uses SIFT descriptors which were shown to capture well invariant local structure information in the image [Lowe 2004]. Second, the flow field is regularized to encourage both sharp discontinuities and small displacements, which should be resilient to reasonable image operations performed by retargeting operators. The SIFTflow algorithm searches for a (dense) displacement field $w(p) = (u(p), v(p))$ between T and S, which minimizes the following energy

$$\begin{aligned} E(w) = \sum_p \|\psi_S(p) - \psi_T(p + w)\|_1 + \frac{1}{\sigma^2} \sum_p \left(u^2(p) + v^2(p)\right) + \\ \sum_{p,q \in \varepsilon} \min\left(\alpha \left|u(p) - u(q)\right|, d\right) + \min\left(\alpha \left|v(p) - v(q)\right|, d\right), \end{aligned} \tag{3}$$

where ψ_S, ψ_T are the dense SIFT fields over each image, and ε is the spatial 4-neighborhood of a pixel. Belief Propagation is used for the optimization. σ, α and d are parameters of the algorithm. We then define the distance between the images to be the resulting energy value itself: SIFTflow$(S, T) = E(w)$.

Earth Mover's Distance [EMD]. The Earth Mover's Distance is a measure of dissimilarity between two distributions. The definition involves the notion of a *ground distance*, a cost of transforming a unit of mass between the distributions. The EMD is then defined as the minimal cost that must be paid to transform one distribution into the other. In the discrete case, EMD can be cast as the well known transportation problem on a corresponding flow network, and solved by a min-flow algorithm. Pele and Werman [2009] have recently proposed EMDs with thresholded ground distances. Such saturated distances correspond to the way humans perceive distance, and are more robust to outlier noise [Pele and Werman 2009]. They demonstrate good results for image retrieval. We use their $\widehat{EMD}$ definition for non-normalized histograms. For two histograms P, Q,

$$\begin{aligned} \widehat{EMD}(P, Q) = (\min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j| \alpha \max_{i,j} d_{ij}, \ s.t. \\ f_{ij} \leq 0, \ \sum_j f_{ij} \leq P_i, \ \sum_i f_{ij} \leq Q_j, \ \sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j) \end{aligned} \tag{4}$$

where f_{ij} denotes the flow from the ith supply to the jth demand. For images S and T of size N_S and N_T pixels respectively, we construct histograms of length $(N_S + N_T)$ as

$$\begin{aligned} P_S &= [\overbrace{1, 1, \ldots, 1}^{N_S}, \overbrace{0, 0, \ldots, 0}^{N_T}] \\ Q_T &= [0, 0, \ldots, 0, 1, 1, \ldots, 1] \end{aligned}$$

and use their definition of thresholded ground distance with the CIEDE2000 color distance Δ_{00} in L*a*b colorspace. Each pixel i in the histograms has its spatial position and color attributes $(x_i, y_i, L_i, a_i, b_i)$, and the ground distance between two pixels i and j is defined as

$$d_{ij} = \min\big(\|(x_i, y_i) - (x_j, y_j)\|_2 + \tag{5}$$

$$+ \quad \Delta_{00}((L_i, a_i, b_i), (L_j, a_j, b_j)), T\big) \tag{6}$$

where the threshold T, and α (Equation 4) are parameters of the algorithm. Since the size of the histograms can be huge, we work on a scaled-down version of the images.

All the above measures search for high-level semantic correlation between images. In contrast, many image similarity measures which examine lower level features were also proposed, and it is important to analyze how well "simpler" methods compare with human resizing perception as well. The MPEG-7 standard [MPEG-7 2002; Manjunath et al. 2001] gathers several well-defined descriptors for visual content similarity, which are widely incorporated in Content-Based Image Retrieval (CBIR) systems. Such descriptors can be used for retargeting analysis as well, as the length of their representation is fixed regardless of the image size. These descriptors do not explicitly search for correspondence between the images, yet are still able to capture differences in local features such as color or gradients, which can show useful for quantifying retargeting results. We focus on two such descriptors. Their computation is fairly straightforward, and a quick review of their extraction is detailed next.

Edge Histogram [EH]. The Edge Histogram Descriptor [Manjunath et al. 2001] captures the spatial distribution of edges in the image. To represent localized edge distributions, the given image is subdivided into 4×4 sub-images, each of which is examined for 5 different edge orientations: vertical, horizontal, two diagonals, and isotropic (non-directional). For each sub-image, a normalized 5-bin histogram is obtained by classifying apparent edges to these five categories. The descriptor is then defined to be the combination of these histograms, which results in $4 \times 4 \times 5 = 80$ length description (and this length is fixed over all image sizes). The image intensity component (Y component in YUV colorspace) is used for edge extraction. Commonly, L_1 is used to measure distance between two EHDs. We therefore define EH$(S, T) = \|$EHD$(S) -$ EHD$(T)\|_1$.

Color Layout Descriptor [CL] . The Color Layout Descriptor [Kasutani and Yamada 2001] represents the spatial distribution of color in the image, and is similar in nature to JPEG compression. Again, the image is first partitioned into 64 non-overlapping blocks in an 8×8 grid. The average color in YUV colorspace is used as the color representative for each block. Each component of these

Algorithm	Parameter	Description
BDS	$patchSize = 7$	Size of square patches
	$nLevels = 4$	Number of pyramid levels
	$\alpha = 0.5$	Equation 1
	$D = \|\|$L*a*b$\|\|_2$	Patch distance.
BDW	$patchSizes$={8,16,32,64}	Size of square patches
	$D = \|\|$L*a*b$\|\|_2$	Patch distance
	D_{aggr} = mean	Distance aggregation
SIFTflow	$patchSizes = 8$	Size of square patches
	$\sigma = 300$	Equation 3
	$\alpha = 2$	Equation 3
	$d = 40$	Equation 3
	$nLevels = 4$	Number of pyramid levels
	$nIterations = 60$	Number of BP iterations
PatchMatch	$nIterations = 5$	Number of iterations
	$\alpha = 0.5$	decreasing factor
EMD	$T = 10$	Equation 6
	$\alpha = 0.5$	Equation 4

Table 1: *Selected parameter sets for each metric, and the PatchMatch algorithm. Note that not all parameters are explicitly referred to in this paper, but are given here for completeness. We refer to the corresponding publications for further details on the parameters and implementation. The notation $||L*a*b||_2$ abbreviates L_2-norm in L*a*b color space. In general, we found that distances in L*a*b colorspace produce better results than RGB distances. It was also clear that using multiple image (or patch) scales produces better results than using single scale.*

derived colors is then transformed by an 8×8 DCT, and the first few low-frequency coefficients of each channel are zigzag-scanned and quantized to form the descriptor. Typically, 6 Y coefficients are used together with 3 coefficients of both U and V channels, resulting in a descriptor of length 12. The color layout distance between two CLDs, $\text{CLD}(S) = (Y, U, V), \text{CLD}(T) = (Y', U', V')$ is defined as a weighted distance between their corresponding coefficients:

$$CL = \sqrt{\sum_{i \in Y} \alpha_i (Y_i - Y_i')^2} + \sqrt{\sum_{i \in U} \beta_i (U_i - U_i')^2} + \sqrt{\sum_{i \in V} \gamma_i (V_i - V_i')^2} \tag{7}$$

where Y_i, U_i, V_i denote the ith coefficient of each channel, and α, β and γ are weights, decreased along the coefficient scan order.

The data we collected in this experiment is the distance of every retargeted result to its source image under each one of the above measures. It should be noted that tuning the objective metrics is a laborious task. First, most of the algorithms contain numerous parameters which span a large configuration space that should be searched in a principled manner. Second, these algorithms are typically computationally expensive, and between 5 and 32 hours were required to calculate the results of a single metric on our entire dataset on a 2 quad-core 16 GB computer. We therefore choose to rely mainly on parameter settings proposed and used by the methods' authors, and applied some basic parameter optimization to improve the agreement between the objective measure rankings and the results of the subjective test (agreement is measured by Kendall's τ distance described in the next section). We investigated the difference in performance between different color spaces, image scales, patch sizes and regularization parameters. For the low-level descriptors, we do not perform further parameter optimization, as those measures have already undergone extensive evaluation (albeit not for a retargeting application). We leave further experiments with such descriptor for future work. Table 1 summarizes the best parameter settings for each algorithm, consequently used in our analysis.

References

BARNES, C., SHECHTMAN, E., FINKELSTEIN, A., AND GOLDMAN, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM TOG 28*, 3.

DONG, W., ZHOU, N., PAUL, J.-C., AND ZHANG, X. 2009. Optimized image resizing using seam carving and scaling. *ACM TOG 28*, 5, 1–10.

GUO, C., MA, Q., AND ZHANG, L. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *CVPR '08*.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11, 1254–1259.

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.

KRÄHENBÜHL, P., LANG, M., HORNUNG, A., AND GROSS, M. 2009. A system for retargeting of streaming video. *ACM TOG 28*, 5.

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.

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

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'09*.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2008. Improved seam carving for video retargeting. *ACM Trans. Graph. 27*, 3.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2009. Multi-operator media retargeting. *ACM Trans. Graph. 28*, 3.

SIMAKOV, D., CASPI, Y., SHECHTMAN, E., AND IRANI, M. 2008. Summarizing visual data using bidirectional similarity. In *CVPR '08*.

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 '07*.