A Perceptually Motivated Online Benchmark for Image Matting

Technical report corresponding to the CVPR'09 paper
TR-188-2-2009-03

Christoph Rhemann[1]*, Carsten Rother[2], Jue Wang[3], Margrit Gelautz[1], Pushmeet Kohli[2], Pamela Rott[1]

[1]Vienna University of Technology Vienna, Austria
[2]Microsoft Research Cambridge Cambridge, UK
[3]Adobe Systems Seattle, USA

Abstract

The availability of quantitative online benchmarks for low-level vision tasks such as stereo and optical flow has led to significant progress in the respective fields. This paper introduces such a benchmark for image matting. There are three key factors for a successful benchmarking system: (a) a challenging, high-quality ground truth test set; (b) an online evaluation repository that is dynamically updated with new results; (c) perceptually motivated error functions. Our new benchmark strives to meet all three criteria.

We evaluated several matting methods with our benchmark and show that their performance varies depending on the error function. Also, our challenging test set reveals problems of existing algorithms, not reflected in previously reported results. We hope that our effort will lead to considerable progress in the field of image matting, and welcome the reader to visit our benchmark at www.alphamatting.com.

1. Introduction

It is well known that the introduction of quantitative benchmarks for low-level vision problems, such as stereo [18] or optical-flow [2], has led to a considerable performance boost in the respective fields. To assure continuous progress, related recent work has focused on providing such benchmarks on the web, which enables the research community to add new results as they arise. This allows people from academia and industry to keep track, analyze and compare recently proposed work in these areas.

Unfortunately, no such standard benchmark has been developed so far for the task of image matting. Image Matting is a computer vision problem which aims to extract an object from its background by recovering correctly the opacity and corresponding foreground color of each pixel. Matting has a number of important applications which have led to numerous different algorithms over the past few years. Since a review of previous work in this area is out of the scope of this paper, we refer the reader to the recent survey of Wang et al. [22]. As many approaches to matting exist, a quantitative benchmark for these methods becomes vital to reveal their strengths and weaknesses, thus providing the ground for novel research directions.

The major goal of this work is to provide such a benchmark for image matting on top of a dataset with corresponding high quality ground truth. Unfortunately, recently proposed ground truth datasets [14, 23, 16] cannot be used straightaway for this task, since they have serious flaws. For instance the data in [14] is considerably affected by noise and the reference solutions in [23] are biased towards some matting algorithms. Although the dataset in [16] is of very high quality, it would greatly benefit from test images showing natural scenes (as opposed to the currently used artificial backgrounds) as well as from a larger diversity of image properties (e.g. images with a larger depth of field). Therefore, we augmented the high-quality matting database of [16] with so far missing images from natural scenes that feature e.g. different focus settings and translucent objects. This joint dataset largely reflects the challenges inherent to real images and provides our basis for the comparison of matting algorithms.

Another issue addressed in this paper is that none of the previously proposed datasets has emerged as an accepted standard. As a consequence, comparisons in subsequent work were conducted on different data, lowering their informative value. This is presumably due to the lack of an appropriate online benchmark system that allows other researchers to include novel results. Thus we establish a dynamic online benchmark that provides all data and scripts that enable the research community to complement our evaluation with new results. This will bring researchers in the favorable position to interactively analyze recent work which will hopefully inspire further research.

Our third contribution is to improve on the evaluation

*This work was supported in part by Microsoft Research Cambridge through its PhD Scholarship Programme and a travel sponsorship.

methodology for image matting that has been previously tied to simple pixel-wise error measures that do not always correlate to the visual quality as perceived by humans. Thus we go beyond these evaluation methodologies and develop quantitative error measures that are based on subjective human perception. More specifically, we concentrate on two properties of alpha mattes that considerably affect the visual quality of matting results, namely the connectivity of the foreground object and the preservation of gradients in the alpha matte. We develop error functions that estimate the compliance of these properties and validate that our measures are correlated to human perception in a user study. Our work is related to research in other areas of computer vision, where perceptual distance measures were developed for e.g. image segmentation [15, 5] or color constancy [7].

Experimental results show that our dataset is challenging and pronounces strength and weaknesses of image matting algorithms that were not apparent in previous evaluations. We found that the performance of algorithms varies under our perceptually motivated error measures that are based on the connectivity and gradient of the alpha matte. This motivates to strive for more complex perceptual error functions that combine these measures. However, we believe that this is a very challenging task, since we found indication that the visual perception of errors is ambiguous among humans.

The remainder of this paper is organized as follows. In sec. 2 we discuss the construction of our ground truth dataset and analyze its properties. We explain the design of our online benchmark in sec. 3 and derive our perceptually motivated error functions in sec. 4. Finally, we evaluate and discuss the performance of matting algorithms in sec. 5.

2. Database

Ideally, a ground truth dataset for image matting should feature several important properties. Firstly, the data should cover a variety of conditions found in real-world images such as color ambiguity, different focus settings, or high-resolution data. Secondly, the data should be challenging in order to further push the limits of current methods, and thirdly the data has to be paired with high quality ground truth alpha mattes to allow for a fair comparison. We strive to construct a dataset that has all these properties.

To obtain ground truth information for real-world images one could follow the approach in [23] where existing matting methods were applied to natural images and their results were manually combined to a reference solution. We applied this approach to several challenging natural images, but found the resulting alpha matte to be of low quality. Further we argue that such a dataset would be biased towards the algorithms that were used to construct the ground truth.

Since there seems to be no reasonable chance to derive alpha mattes with sufficient quality from real-world imagery we decided to capture high-quality ground truth mattes in a restricted studio environment by triangulation [19]. Our set of 8 images is considerably more challenging than previously used data and depicts natural (indoor) scenes that comprise of a variety of challenges one faces in real-world images, like different focus settings (see sec. 2.2). To derive an even more complete set of 35 images we augmented our data with the database proposed in [16]. Finally, we split up this set into 8 test and 27 training images (see sec. 3).

2.1. Data capture

To obtain a composite that can serve as test image for evaluation purposes we built up a natural scene that was then photographed with a foreground object. To derive a high-quality ground truth alpha matte for this composite, we carefully placed a monitor (Apple Cinema 30" HD) between the object and the scene, without moving neither the object nor the camera (all subsequent shots have to be perfectly aligned with the composition). We displayed four single-colored backgrounds (i.e. black, red, green and blue) on the monitor that were photographed with the foreground object. After capturing the object in front of the screen, the object was removed to photograph the plain backgrounds as well. This allowed us to extract a ground truth matte by triangulation [19]. In [16] the same setup was used but image compositions were simply obtained by photographing the objects in front of a monitor, which showed natural background images.

Following [16], all images were shot in unprocessed RAW format with a professional DSLR camera (Canon 1D MarkIII with a Canon 28-105mm zoom lens) at a resolution of 10.1 Megapixels with constant camera settings. To avoid camera shake, we locked the mirror of the camera (hence the shutter is the only moving part inside the camera) and used a remote control to trigger the shutter. This enabled us to take images that are registered to each other with sub-pixel accuracy. For computing the alpha matte, the RAW image data was transformed into RGB color images without gamma correction (linear gamma) in order to avoid the introduction of noise in dark areas. Finally, the images were cropped at a bounding box that was casually drawn around the foreground objects, resulting in test scenes with an average size of about 6 Megapixels.

To assure that our newly recorded ground truth mattes, as well as those of [16], are indeed of high quality, we evaluated their noise level. For this purpose, we manually marked pixels which have an alpha value of exactly 1 (i.e. truly foreground) and then computed the number of pixels in this region with an alpha value lower than 0.97. For our new data, 3.4% (0.3% for the data in [16]) of the pixels are below this threshold. This is a very good value, compared to the data in [14], where we found on average 26.7% of true foreground pixels with an alpha value below 0.97.

2.2. Image properties

Our images exhibit many characteristics of real-world images, like highly textured backgrounds, different depth of fields, as well as color ambiguity. We included a range of foreground objects that have different properties such as hard and soft boundaries, translucency or different boundary lengths and topologies (e.g. a tree with many holes).

Our dataset is challenging and exhibits various levels of difficulty. On our data the mean squared error (normalized over the number of pixels with unknown alpha values) computed using the algorithms of [23] and [14] (averaged over the algorithms) varies between 0.3 and 21.8 with an average value of 4.2. This is considerably larger than the average error rates we computed with the same procedure on the datasets of [23] and [14] which are 1.1 and 0.9, respectively.

2.3. User input

Image matting is a severely ill posed problem and therefore user interaction is necessary to solve it. The most common form of user interaction is the trimap interface, where the user manually partitions the image into foreground, background and unknown regions. Transparency values are then computed for the unknown regions only. Some matting algorithms are also capable to work on very sparse trimaps, commonly denoted as scribbles. However, scribbles are subject to an even higher variation of inputs compared to trimaps and are often only used to derive a more accurate trimap [11, 1, 16].

Therefore, we decided to simulate the user input by a set of three different trimaps for each test image. Two of them where generated automatically by dilating the unknown region of the high-resolution ground truth trimap by 22 and 44 pixels respectively. To account for more natural user input we also included a hand drawn trimap for every test case. These were generated by an experienced user given a paint tool with a set of three brushes (i.e. unknown, foreground and background) and flood filling capability. The user was imposed a time constraint of 2 minutes per image, which we found sufficient to create a reasonable trimap for all images.

Although most matting algorithms accept trimap input, we plan to extend our benchmark with matting results that were generated with other forms of user interaction or in a completely automatic way (e.g. [14] supports a component picking interface and a completely unsupervised mode).

3. Online Benchmark System

An important reason that has led to the success of recently proposed benchmarks in computer vision is that they have been made freely available on the web. Inspired by [18, 2] we designed an online benchmark that is accessible free of charge at www.alphamatting.com. Like other online benchmarks, a major advantage of our repository is that it can be dynamically updated with novel datasets or error measures, if needed in the future. We provide all scripts and data necessary to allow other researchers to submit new results. We hope that this will encourage many researchers to participate in the competition. A screenshot of our online benchmark is shown in fig. 1.

Selecting a representative test set. A comprehensive benchmark for matting algorithms should be carried out on a dataset that covers a large variation of different scenarios that are encountered in practical matting applications. Since we invite other researchers to submit their results to our benchmark, a very large dataset is unreasonable, especially when people process high-resolution images with unoptimized research code. For example, assuming an average computation time of two minutes per image, computing results for our dataset of 35 images on 3 different trimaps requires more than 3 hours. Hence, we need a dataset that is as small as possible but still largely maintains the same variations as the full set.

Therefore, we decided to split up our database into a test and training set. The test set comprises of 8 images for which the ground truth alpha mattes are hidden from the public, in order to largely prevent excessive parameter tuning. The remaining 27 images serve as training dataset with publicly available ground truth. We hope that this set will be used by other people for parameter learning.

To select a representative test set from our full database we applied the following strategy. Firstly, we manually assigned all images to four categories depending on the amount of transparencies in their respective ground truth matte. Then we computed error rates (mean squared error) for all images with a variety of matting algorithms (i.e. [23, 14, 9, 20, 6, 8]). From each of the four categories, we selected those two images that were most challenging for the algorithms (i.e. images with a large average error and diverging quality of results).

To confirm that we have a chosen a well-balanced subset, we compare the performance of matting algorithms on our subset against their performance on the average subset. Therefore, we computed the average ranking of the 6 aforementioned algorithms over all possible subsets of 8 images. Indeed this ranking is identical to the one obtained on our particular subset. Furthermore, we computed the average correlation of rankings obtained from every possible subset of 8 images with the rankings on the full set, which gives a value of 0.91. This is very close to the correlation value for our subset which is 0.87.

We further decided to provide our datasets in two different resolutions (i.e. full resolution 6 Megapixels and downscaled, where the longest image side is 800 pixels), since most current matting algorithms are not capable of processing high-resolution images. In this work we restricted our evaluation (sec. 5) to low-resolution data. However, in the

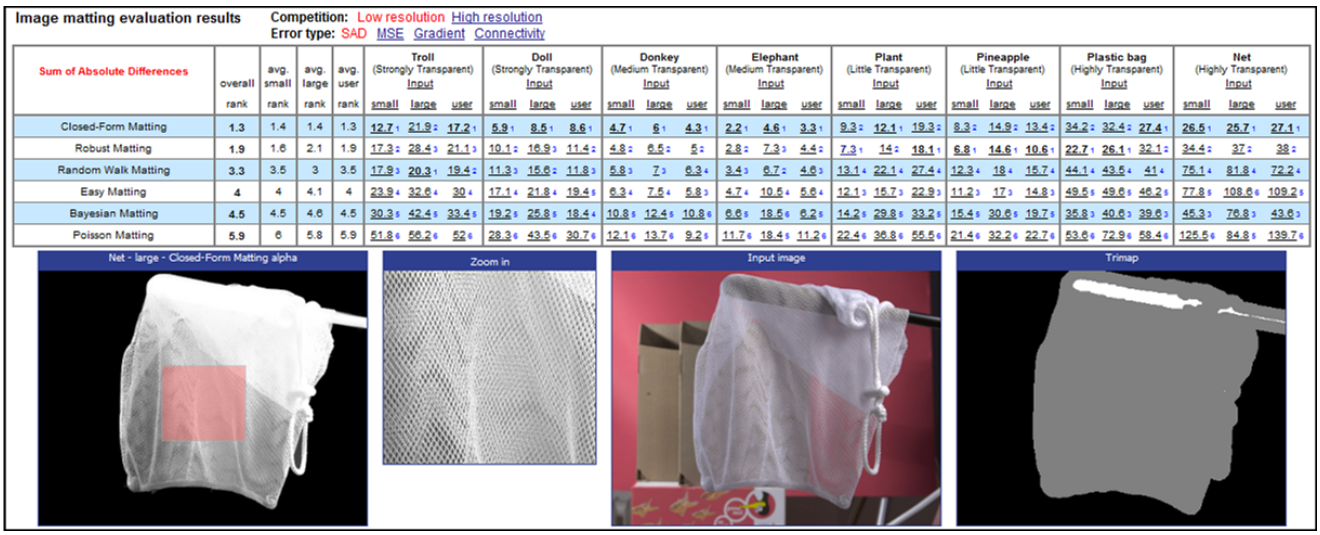

Figure 1. **Online benchmark.** A screenshot of our online evaluation table. The values in each cell correspond to the error generated by a specific method (rows) on a test image (columns). Moving the mouse over a specific error value shows images of the corresponding alpha matte (leftmost image). To allow for a better inspection of the result, a zoom-in of the alpha values in the red box is shown next to it. The zoomed-in area can be easily changed by moving this box. Further, we show the corresponding input image and trimap.

future we plan to supply the online benchmark with high-resolution images for those algorithms that can handle them.

4. Perceptually Motivated Error Measures

In order to quantitatively evaluate the performance of matting algorithms, their outputs (i.e. alpha mattes) have to be compared to the ground truth using an error metric. In previous work, simple metrics like the sum of absolute differences (SAD) or the mean squared error (MSE) have been used for this task. While these measures provide a good basis for comparison, they are not always correlated to the visual quality as perceived by a human observer. An example is depicted in fig. 2, which shows two image compositions where the SAD/MSE error is not correlated to the visual quality. This motivates to study error metrics that are better suited for a perceptual comparison of matting methods.

Clearly, the development of perceptually driven distance measures depends on the target application and thus we will focus on the commonly used application scenario of compositing the extracted foreground object onto a new background (cut & paste). To further reduce the complexity, we will restrict ourselves to pasting onto a homogeneously colored backing, which is an important application in the media industry (e.g. creating images for magazine covers).

Human observers judge the visual quality of image compositions by perceiving and weighing the different types of errors that appear in these images. This judgment depends on many different factors such as the color and texture of the resulting composite as well as the structure of the alpha matte. Ideally, one should learn a single visual error function over image patches that takes all these degrees of freedom into account. However, there are two problems with this approach. Firstly, image patches that are big enough to preserve the context of the depicted scene, e.g. of size 100x100 pixels, have an exponential number of potential colors and alpha values. Secondly, given the identical patch of an image composite, people might disagree on the visual error, hence a multi-modal error function is needed. For instance, 12% of the participants in our study prefered fig. 2 (d) over fig. 2 (c), while 88% decided the other way round. To widely circumvent these challenges, we concentrate on developing perceptual error functions for two specific error

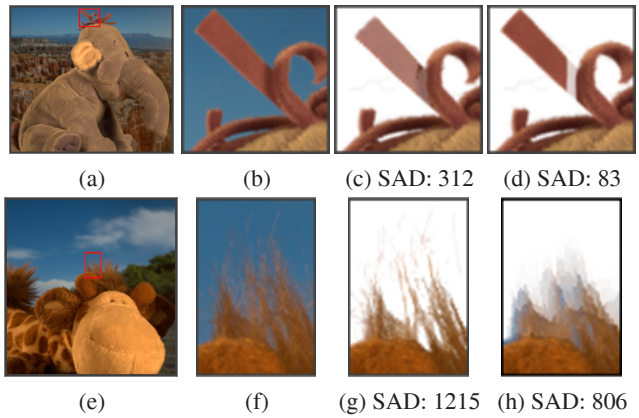

Figure 2. **Motivation for perceptual error measures.** Two images (a; e) were cropped to give the images shown in (b; f). Matting methods have been applied to generate new compositions (c-d; g-h). In both cases, the average user ranking was exactly opposite to the error computed by SAD or MSE. The top row (c-d) shows an example of our connectivity set, and the bottom row (g-h) an example of our gradient set.

categories where previously used error metrics, like SAD, largely disagree with humans. In an explorative pre-study with 4 subjects (3 males and 1 female) two error categories emerged that seem to considerably degrade the visual quality of image composites: (i) *connectivity* errors, which are a result of disconnected foreground objects, for example a disconnected piece of hair floating in the air; (ii) *gradient* errors, which are due to oversmoothing or erroneous discontinuities in the alpha matte (i.e. the gradient in the alpha matte diverges from the ground truth). Examples for each of these categories are depicted in fig. 2.

In the remainder of this section, we first derive the visual quality of image compositions in a user study (sec. 4.1). In sec. 4.2, we design perceptual distance measures and show that their correlation to the visual quality is superior, compared to previously used error measures, like SAD.

4.1. User study

The main goal of our user study is to infer the visual quality for image compositions from human observers in the presence of connectivity and gradient artifacts.

Data. We performed our psychophysical experiments on two sets of image compositions, each of them afflicted solely by connectivity or gradient artifacts. To construct these sets, we applied a variety of matting algorithms on the input images of our ground truth database and created composites by pasting the extracted foreground object onto homogeneously colored backgrounds. We then carefully selected crops of these compositions that mainly exhibited either connectivity or gradient artifacts. The size of the crops was chosen such that are small enough to isolate these error categories, but big enough to provide the user with sufficient contextual information to judge about its quality. In our pre-study, we found that crops with a size of about 100x100 pixels are a good tradeoff between these two factors.

Compositions created from the same image crop (but with different matting algorithms) were arranged into a single test case. Fig. 3 shows an example. To increase the number of composites per test case, we also included artificial images that we generated by interpolating some composites towards their ground truth. Note that by including these interpolations, the results of this study become more applicable to the output of future matting methods with higher quality results. From this pool of test cases, we have chosen only those whose composites could be easily sorted according to their quality (no ambiguities) and where we expected traditional error measures (e.g. SAD) to diverge from the human perception. For the study we used a total number of 20 test cases (10 for each error category), each test case comprising of 6 image compositions.

Study procedure. The study was carried out with 17 participants (8 males and 9 females) whose ages ranged from 24 to 67 years, with an average age of 36. The study aimed

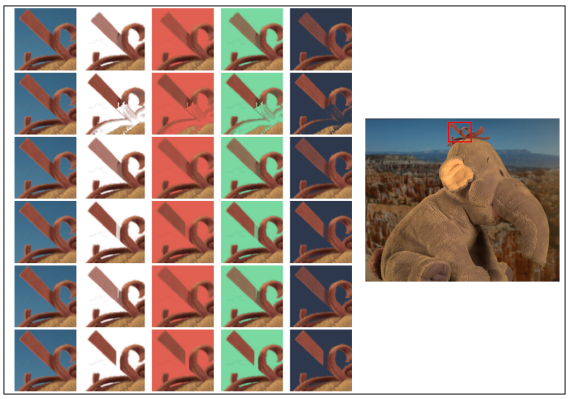

Figure 3. **Example test case of our study.** Explanation in text.

to derive an ordering of the compositions associated with each test case, from the judgment of the participants. Such an ordering can be obtained by means of absolute (on a discrete scale) or relative rankings. We preferred to derive relative rankings, since they have been shown to significantly raise the agreement between users in the context of web page ranking [4]. Relative rankings can be obtained by a sequence of pairwise comparisons (the user selects one out of a pair of images) or by sorting the compositions at a glance. In our pre-study we observed that the participants preferred to rank the compositions at a glance and therefore decided for the following experimental setup shown in fig. 3.

For each test case, the subjects were shown the associated 6 compositions in a list that they could interactively sort by moving the images on the screen (fig. 3(left)). Each list element shows the original image crop (left) together with composition on 4 homogeneously colored backgrounds (i.e. white and shades of red, green and blue). To provide the user with more contextual information, we also displayed the corresponding uncropped image (fig. 3(right)). For every participant, the compositions in each test case were shown in random order. This was done to overcome any bias of subjects against any particular initial position of the list of images.

Prior to the study the participants were told that they will be presented crops of photomontages that had been generated by inserting objects, extracted from a photograph, onto a single-colored background. Then we instructed the subjects to rank the results according to how realistic the image compositions appeared. The users were given the opportunity to indicate cases where two or more compositions could not be distinguished because they have the same quality. To reveal further details about the decision making process of the users we also recorded their verbal feedback.

4.2. Analysis of results

To obtain generalizable results, the study was evaluated with respect to the ranking of the "average user". In the av-

erage scores we accounted for image pairs that could not be clearly ranked (i.e. pairs where the average ranks differed by less than 0.2) by assigning them to the same score (14% and 8% of pairs in the gradient and connectivity set were affected). To validate that an analysis on the average observer basis is valid, we first analyze the variability of the user judgments with respect to the average rankings. Then, we examine to which extent several distance measures are correlated to these average scores. Since the distance measures give absolute error values, we converted them to relative rankings beforehand. To measure the similarity between two rankings we utilized the Kendall's τ measure [12] which is commonly used in statistics for comparing the correlation of ordinal random variables [10].

Agreement of observers. The correlation of the individual participants (averaged over all test cases and users) with the average user ranking was 0.90 and 0.87 for the connectivity and gradient test set, respectively. These are reasonably high values compared to the zero coefficient that would be given to a random ranking. However, the remaining variation in the user judgments implies that even for the identical image composition, that shows only a single class of artifacts, people disagreed on the visual error. This suggests that there is inherent ambiguity in the perception of errors and a single visual error function for image matting may not exist. Note that ambiguity in the perception of errors does not mean that there is no single global optimum (ground truth) for the alpha matte.

Error measures. Our perceptual error measures are:

- **Gradient.** We tried a number of different gradient measures, including the commonly used angular error between the gradient vectors, but found the following measure to work best. The difference between the gradients of the computed alpha matte α and its ground truth α^* is defined as $\sum_i (\nabla\alpha_i - \nabla\alpha_i^*)^q$, where $\nabla\alpha_i$ and $\nabla\alpha_i^*$ are the normalized gradients of the alpha mattes at pixel i that we computed by convolving the mattes with first-order Gaussian derivative filters with variance σ.

- **Connectivity.** A considerable amount of work has been devoted to the problem of measuring connectivity [17, 21]. Following recent work in this area [3], we define the degree of connectedness by means of connectivity in binary threshold images computed from the grayscale alpha matte.

In detail, we define the connectivity error of an alpha matte α with its corresponding ground truth α^* as $\sum_i (\varphi(\alpha_i, \Omega) - \varphi(\alpha_i^*, \Omega))^p$, where φ measures the degree of connectivity for pixel i with transparency α_i to a source region Ω. Consider fig. 4, which illustrates the intensity function of a row of pixels in an alpha matte. The source region Ω is defined by the largest connected region where both the alpha matte as well as its ground truth are completely opaque (illustrated by the red line in fig. 4). The degree of connectivity is based on the distance $d_i = \alpha_i - l_i$,

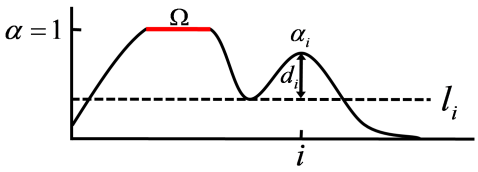

Figure 4. **Connectivity error.** See explanation in the text.

where l_i is the maximum threshold level where pixel i is 4-connected to Ω (dashed line in fig. 4). A pixel is said to be fully connected if $l_i = \alpha_i$. Finally, the degree of connectivity φ for pixel i is defined as

$$\varphi(\alpha_i, \Omega) = 1 - (\lambda_i \cdot \delta(d_i \geq \theta) \cdot d_i). \qquad (1)$$

This means that a pixel is fully connected if $\varphi = 1$ and completely disconnected if $\varphi = 0$. The δ function enforces that small variations in d_i below θ are neglected. We further weight d_i at disconnected pixels with their average distance λ_i to the source region: $\lambda_i = \frac{1}{|K|} \sum_{k \in K} dist_k(i)$, where K is the set of discretized alpha values in the range between l_i and α_i. The function $dist_k$ gives the normalized euclidean distance of i to the closest pixel that is connected to Ω at threshold level k. The intuition behind this is that unconnected parts that are further away from the connected region are visually more distracting.

Unfortunately, computing the connectivity under this metric is computationally rather expensive, since it requires to evaluate function $dist_k$ at each threshold level k. To make the computation tractable in our online evaluation system, we use a slightly modified version of this metric, which neglects the distance of unconnected islands to the connected region. This was done by simply setting λ_i in eq. (1) to a constant value of 1.

Agreement of error measures. The agreement of our error measures on the gradient test set (first row of table 1) shows that the correlation of SAD and MSE with the average human observer is rather low (0.45 and 0.51). Our connectivity measure performs similarly with a correlation of 0.47. The correlation for our computationally less expensive connectivity measure (shown in brackets in table 1) is 0.41. As expected our gradient measure outperforms all of them with a correlation of 0.75.

Analysis on the connectivity set (second row of table 1) shows that SAD and MSE exhibit an even lower correlation than on the gradient set (0.28 and 0.34) and also our gradient error (0.40) is not capable to capture errors in the connectivity. As expected our measure for connectivity performs well with a correlation coefficient of 0.75. Interestingly, our modified connectivity metric which neglects the distance of disconnected islands performs even slightly better with a correlation of 0.77.

Data	Grad.	Conn.	MSE	SAD	User consent
Grad.	**0.75**	0.47 (0.41)	0.51	0.45	0.87
Conn.	0.40	**0.75** (0.77)	0.34	0.28	0.90

Table 1. **Error measure correlations.** The correlation coefficients of four error measures for the connectivity and gradient set. Correlations of the modified connectivity metric that we use for online evaluation are shown in brackets.

4.3. Choice of parameters

We decided to choose the values for four important parameters of our error measures according to their robustness and correlation with the user scores. The robustness of error measures with respect to noise in the data is a test commonly used in information retrieval [24]. We distorted the alpha mattes with Gaussian noise (zero mean and variance ranging from 0.001 to 0.005) and ranked them using our new perceptual error measures. We then computed the correlation coefficients between these rankings and the ones derived on undistorted data. We repeated this K times (we found $K = 200$ sufficiently large) and used the average correlation coefficient as robustness score.

Let us consider fig. 5 (left) which shows the robustness of our gradient measure for different values of the parameter σ, which is the variance of the Gaussian derivative filters used to compute the gradients. We can see that for $\sigma = 0.2$ (blue curve), the robustness drops off quickly with increasing noise level. This is not surprising since a low σ makes the estimation of the gradient more sensitive to noise. For larger values of σ (1.4 and 3) the robustness is constantly high. Clearly, the choice of a parameter does not only depend on the robustness, but also on the correlation to the user scores (fig. 5 (right)). We can see that although a large value of $\sigma = 3$ (green curve in fig. 5 (left)) makes the measure robust to noise, the correlation of the gradient measure is rather low for this value. Thus we limited the parameters to a range where the error measures exhibit a robustness score of at least 0.9 and a correlation that is at worst 10% lower than its maximum value (avg. over all noise levels). Therefore a good choice is $\sigma \in \{1.2, .., 2.0\}$, where our measure is robust and highly correlated to the user scores.

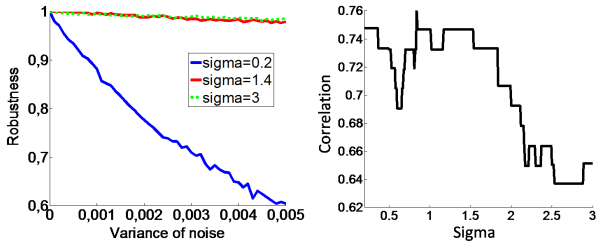

Figure 5. **Robustness of parameters.** See the text for explanation.

Accordingly, we can limit the remaining parameters of our error measures to the range, $q \in \{1, .., 3\}$, $\theta \in \{0.13, .., 0.25\}$ and $p \in \{1, .., 2\}$. Finally we select the number in each range which gives the maximum correlation (i.e. $\sigma = 1.4$, $q = 2$, $\theta = 0.15$ and $p = 1$). Clearly, our approach for parameter selection assumes that the user rankings are invariant to small noise in the alpha mattes. Note, if we were to have a larger training dataset we may circumvent this procedure completely and train these parameters directly, however, we leave this for future work.

5. Experiments

As a basis for evaluation, we have compared 6 matting methods that mostly represent the current state-of-the-art, namely *Bayesian matting* [6], *Closed-form matting* [13], *Easy matting* [9], *Poisson matting* [20], *Random walk matting* [8] and *Robust matting* [23].[1] For all algorithms, we used the implementations of the respective authors, except for [20] which we implemented ourselves. To offer a fair comparison we set the parameters for all algorithms to the values reported in the respective papers.

Performance. We evaluated all of the above mentioned algorithms on our 8 test images, using three different trimaps as inputs (see sec. 2.3). We computed the accuracy of the resulting alpha matte with respect to four error measures defined in sec. 4.2 (i.e. SAD, MSE, gradient and connectivity error). Each test case (image and trimap) gives a ranking of all algorithms. This rank, averaged over all test cases, is shown in table 2.

Method	SAD	MSE	Grad.	Conn.
Closed-form [13]	**1.3**	**1.4**	**1.5**	2.0 (1.8)
Robust matting [23]	1.9	1.8	1.7	3.4 (3)
Random walk [8]	3.3	3.2	3.5	**1.3 (1.4)**
Easy matting [9]	4.0	4.4	4.2	3.7 (4)
Bayesian matting [6]	4.5	4.3	4.3	5.0 (5.1)
Poisson matting [20]	5.9	5.9	6.0	5.6 (5.1)

Table 2. **Evaluation.** The table reports the overall ranks of the different algorithms with respect to four error measures. These ranks were obtained by averaging the ranks over all test cases i.e. all test image-trimap input pairs. Errors obtained with the modified connectivity metric that we use for online evaluation are shown in brackets.

When analyzing the results with respect to the SAD and MSE error measure, we observe that Closed-form matting and Robust matting outperform the other methods. We also notice that the performance of Robust matting, Bayesian matting and Easy matting is lower than what was reported in previous evaluations [22, 23, 16]. The main difference of these approaches to their competitors is that they use a data term in their objective function, which is derived from

[1] Wherever possible, we provide code (or links to it) for these methods on the evaluation website. Due to licensing issues we cannot provide code for [23, 6, 9].

global color models of true fore- and background regions. These data terms typically require to set a fair amount of free parameters. Hence, a potential over-fitting of these parameters to their respective test data may lead to a lower performance on our unseen data. Note that the test datasets for these methods were mostly composed of images with smooth backgrounds, whereas our dataset contains examples of highly textured backgrounds. A detailed inspection of these data terms shows that they are fairly sensitive to the exact placement of the trimap (i.e. true fore- and background regions). This sensitivity can introduce large artifacts in the alpha matte. Pure propagation based approaches, like Closed-form matting and Random walk matting, seem to suffer less from this problem. An exception is the propagation based Poisson matting algorithm that performed constantly worse than its competitors, since its assumption of smooth fore- and background colors is rarely met on our dataset.

On the other hand, visual inspection of the results shows that methods that model the fore- and background colors can sometimes overcome the color ambiguity problem. For instance, Closed-form matting (which does not have a global color model) tends to over-smooth holes in the foreground and shortens fine structures like hair. These structures were sometimes better captured by methods which have a global color model e.g. Robust matting. However, on average the drawbacks of the color model based methods prevailed their advantages, at least on our test set.

Performance on gradient error. Analyzing the scores with respect to the gradient error we see that Closed-form and Robust matting perform almost on par and the gap to Random walk matting increases slightly. This is mainly because Closed-form and Robust matting can better preserve the gradient of the alpha matte in regions like hair. In contrast, Random Walk matting tends to oversmooth the alpha matte which is penalized by our gradient measure.

Performance on connectivity error. When considering the rankings based on our connectivity error we see that the Random walk algorithm is clearly the best performer. This is not surprising, since alpha mattes generated by Random walk are perfectly connected, i.e. obtain a value of $\varphi = 1$ (eqn. 1) for each pixel. However, this does not mean that the connectivity error (which is the difference of the connectivity of the alpha matte with its ground truth) is zero, since the ground truth is usually not perfectly connected. However, the Random walk algorithm shows quite large errors under the other metrics, which motivates new research for future matting techniques that recover accurate alpha mattes while preserving the connectivity of the foreground object.

6. Conclusions and Future Work

We have presented a new benchmark for the evaluation of image matting algorithms that is freely available on the web at *www.alphamatting.com*. The evaluation of state-of-the-art matting algorithms on our challenging dataset reveals failures on images containing highly textured backgrounds, and images where the fore- and background cannot be differentiated on the basis of color alone. An important contribution of our work was the proposal and validation of perceptually motivated error measures based on the connectivity and gradient of the alpha matte. To the best of our knowledge, this is the first study which validates error measures for alpha matting using the composition quality as perceived by humans. We hope that our work will encourage researchers to develop new matting algorithms that pay more attention to visually important features such as connectivity.

Future work could concentrate on establishing more complex perceptual measures that take into account other factors such as color and texture of the image. Such an error function is highly desirable since it could be used by machine learning methods as a loss function for alpha matting. However, more research is needed, since results of our user study indicate that the visual perception of errors is ambiguous and thus a multi-modal function might be needed.

Acknowledgments

We thank Anat Levin and Guillermo Sapiro for helpful discussions and Daniel Scharstein for granting us to use the graphic design of the *middlebury evaluation website*.

References

[1] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*, 2007.

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

[3] U. Braga-Neto and J. Goutsias. Grayscale level connectivity: theory and applications. *TIP*, 2004.

[4] B. Carterette, P. Bennett, D. Chickering, and S. Dumais. Here or there. In *ECIR*, 2008.

[5] A. Cavallaro, E. Drelie-Gelasca, and T. Ebrahimi. Objective evaluation of segmentation quality using spatio-temporal context. In *ICIP*, 2002.

[6] Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A Bayesian approach to digital matting. In *CVPR'01*.

[7] A. Gijsenij, T. Gevers, and M. Lucassen. A perceptual comparison of distance measures for color constancy algorithms. In *ECCV*, 2008.

[8] L. Grady, T. Schiwietz, S. Aharon, and R. Westermann. Random walks for interactive alpha-matting. In *VIIP'05*.

[9] Y. Guan, W. Chen, X. Liang, Z. Ding, and Q. Peng. Easy matting: A stroke based approach for continuous image matting. In *Eurographics*, 2006.

[10] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2007.

[11] O. Juan and R. Keriven. Trimap segmentation for fast and user-friendly alpha matting. In *VLSM*, 2005.

[12] M. Kendall. *Rank Correlation Methods*. Hafner, 1955.

[13] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, 2006.

[14] A. Levin, A. Rav-Acha, and D. Lischinski. Spectral matting. In *CVPR*, 2007.

[15] B. Peng and O. Veksler. Parameter selection for graph cut based image segmentation. In *BMVC*, 2008.

[16] C. Rhemann, C. Rother, A. Rav-Acha, and T. Sharp. High resolution matting via interactive trimap segmentation. In *CVPR*, 2008.

[17] A. Rosenfeld. On connectivity properties of grayscale pictures. *Pattern Recognition*, 1983.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.

[19] A. Smith and J. Blinn. Blue screen matting. *SIGGRAPH 96*.

[20] J. Sun, J. Jia, C. Tang, and H. Shum. Poisson matting. *SIGGRAPH*, 2004.

[21] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *PAMI*, 1991.

[22] J. Wang and M. Cohen. Image and video matting: A survey. *Foundations/ Trends Comp. Graphics and Vision*, 2007.

[23] J. Wang and M. F. Cohen. Optimized color sampling for robust matting. In *CVPR*, 2007.

[24] Y. Zhou and W. Croft. Ranking robustness: a novel framework to predict query performance. In *CIKM*, 2006.