Confidence-Driven Image Co-matting

Linbo Wang\textsuperscript{a}, Tianchen Xia\textsuperscript{a}, Yanwen Guo\textsuperscript{a,}\textsuperscript{*}, Ligang Liu\textsuperscript{b}, Jue Wang\textsuperscript{c}

\textsuperscript{a} State Key Lab for Novel Software Technology, Nanjing University, China PR
\textsuperscript{b} School of Mathematical Sciences, University of Science and Technology, China PR
\textsuperscript{c} Adobe Research

Abstract

Single image matting, the task of estimating accurate foreground opacity from a given image, is a severely ill-posed and challenging problem. Inspired by recent advances in image co-segmentation, in this paper, we present a novel framework for a new task called co-matting, which aims to simultaneously extract alpha mattes in multiple images that contain slightly-deformed instances of the same foreground object against different backgrounds. Our system first generates trimaps for input images using co-segmentation, and an initial alpha matte for each image using single image matting. Each alpha matte is then locally evaluated using a novel matting confidence metric learned from a training data set. In the co-matting step, we first align the foreground object instances using appearance and geometric features, then apply a global optimization on all input images to jointly improve their alpha mattes, which allows high confidence local regions to guide their corresponding low confidence ones in other images to achieve more accurate mattes all together. Experimental results show that this co-matting framework can achieve noticeably higher quality results on an image stack than applying state-of-the-art single image matting techniques individually on each image.

Keywords: Image matting; image co-segmentation

1. Introduction

Image matting aims at recovering accurate foreground opacity from natural images. Specifically, given an input image $I$, it estimates the foreground opacity map $\alpha$ as well as a foreground and background image $F$ and $B$, according to the linear compositing equation:

$$I = \alpha F + (1 - \alpha)B,$$

where $\alpha$ is also called the alpha matte, whose per-pixel values are constrained to be in $[0, 1]$. Given its wide range of applications in image editing and film production, image matting has been extensively studied in the past few decades. However, the problem remains a formidable challenge due to its severely ill-posed nature: $F, B$ and $\alpha$ in Eqn. 1 are all unknown. To make the problem tractable, existing matting approaches usually require additional user input such as a trimap, which identifies pixels as definitely foreground (F), definitely background (B) and unknown (U).

Existing image matting algorithms mainly focus on extracting a matte from a single image. They can roughly be classified into three categories, sampling-based, propagation-based and hybrid-model-based approaches [1]. Sampling-based methods sample known F and B color pairs and select best ones among them for estimating the alpha value of an unknown pixel. Propagation-based approaches establish affinities between neighboring pixels to enforce local smoothness. Hybrid algorithms combine color sampling and local affinity together in a single objective function and solve the alpha matte using optimization. More details of these approaches are given in Sec. 2.

Although tremendous progress has been made in recent years, state-of-the-art single matting approaches still fail often on real-world images. The main reason is that the underlying assumptions made in matting methods are often violated in complicated real-world examples. Specifically, for images with overlapping F and B color distributions, sampling-based methods cannot identify the right foreground and background color samples for some unknown pixels, leading to severe errors in the alpha matte. Propagation-based methods, on the other hand, may perform poorly due to the violation of...
Figure 1: Overview of our system. Given two input images and their corresponding trimaps (indicated by green lines) in (a), we first generate an initial matte for each image in (b) using single image matting. Region-wise matting confidence shown in (c) is evaluated for each initial matte, and inter-image unknown region registration shown in (d) is established. All the information is then incorporated in a global optimization procedure to derive refined mattes of both images shown in (e).

local color line model assumption when the foreground or the background contains textures with high gradient edges. Correcting errors in single image matting requires the user to specify more accurate trimaps, which is a tedious process.

In this paper, we show that by considering multiple images that contain the same foreground object together (see Fig. 1a), one can get more accurate matting results than applying single image matting to each image individually. We call this new framework co-matting, which is inspired by recent advances in image co-segmentation. There are two important assumptions made in our approach. Firstly, we assume that different images contain instances of the same foreground object, and each instance is allowed to have some small deformations, but not large topology changes. Secondly, the input images are photographed against distinct backgrounds, which leads to different levels of difficulty for matting. Images meeting such requirements commonly arise in real world, e.g. multiple inconsecutive frames in a movie as shown in Fig. 1a.

With the above assumptions, our co-matting system first evaluates local matting confidence for initial alpha mattes (Fig. 1c), which are generated by applying a single image matting method on individual images (Fig. 1b). The metric itself is learned from a dataset based on multiple image features, including both image-specific features and matting-algorithm-dependent ones. We then align the foreground instances across different images using a two-step approach (Fig. 1d). Finally, we perform joint matting on all images in a new global optimization scheme, which simultaneously refines all alpha mattes (Fig. 1e). The optimization is set up in such a way that erroneous areas in one initial alpha matte, produced by the single image matting algorithm, are automatically improved with the aid of their counterparts in other images. Our system also allows the user to further improve the results by refining the trimap on one image, which will automatically improve the mattes on other images as well, through the optimization process. This allows the user to achieve high quality matting results on all images with minimal user input.

2. Related Work

2.1. Single image matting

Here we only briefly review some recent image matting approaches. For a more comprehensive survey, we refer the readers to [1].

Robust matting [2] pioneers the modern non-parametric sampling methods for image matting. For each unknown pixel, it picks best F and B sample pairs from an initial sample set according to color fitness estimated with Eqn. 1. Recent sampling-based methods follow this framework but seek to expand the sampling space in a more accurate and efficient way. Rheinmann et al. [3] collect nearby samples in geodesic distance instead of Euclidean distance. Shared matting [4] shoots rays in several directions and samples known boundary pixels on the rays. Best sample pairs are further propagated among neighboring pixels. He et al. [5] construct
a sample space consisting of all known boundary pixels and apply randomized patchmatch [6] to efficiently locate best samples for each unknown pixel. All these methods may fail in the presence of color ambiguity, which occurs when the linear combination of a false sample pair accidentally well explains an unknown pixel.

Propagation-based approaches compute the matte by solving global cost functions, which encode matting affinities among pixels. Various assumptions are made to analytically construct the affinity matrix. Closed-form matting [7] assumes the inherent F and B colors in a local window subject to a 4D color line model. Singaraju et al. [8] propose two simpler alternative models: Point-Point and Line-Point color models for special degenerated cases. Zheng et al. [9] assume the alpha value of a pixel is linearly dependent on those of its neighbors, and extend the derived algorithm to non-linear cases. Lee et al. [10] impose a linear model upon alpha values of pixels with similar colors based on the nonlocal principle. It is further extended to video matting in [11] and material matting in [12]. An inherent limitation of all these methods is that since the alpha matte is estimated in a propagation manner, small errors could spread and accumulate to form large errors.

Hybrid methods [2, 3, 4, 5] combine the above two approaches, assuming that sampling errors can be eliminated under smoothness constraints, while affinity-based propagation error can be suppressed by accurate color sampling. The recent work by Chen et al. [13] integrates sampling-based priors with both local and non-local affinities together to achieve the best performance on the benchmark dataset [14]. However, when dealing with real-world image with highly complex appearance, both sampling priors and affinity terms are error-prone, and combining them together does not help much.

2.2. Multi-image matting

A few systems have been proposed for dealing with multiple images in very specific hardware settings. Smith and Blinn [15] tackle the classical blue screen matting problem, which aims to recover the foreground matte from multiple images where the same object is shot against multiple known uniformly-colored backgrounds. Sun et al. [16] propose a method to estimate matte from well-aligned flash/non-flash image pairs. Joshi et al. [17] present a video matting system using a camera array. These systems are only applicable with very restricted settings and cannot be easily generalized to handle a wide range of real-world examples.

2.3. Image co-segmentation

Our work is largely inspired by image co-segmentation, which jointly performs Figure-Ground segmentation of an image pair with the same foreground and different backgrounds, in the absence of user guidance. It was first studied by Rother et al. [18] and has gained considerable attention afterwards [19, 20, 21, 22, 23]. More recently, the concept has been generalized to 3D shape segmentation problem [24, 25].

3. Our Co-matting Approach

3.1. Pre-processing

Image matting algorithms usually require the user to provide rough constraints for known foreground and background regions. Two kinds of user input are commonly used: trimaps and sparse scribbles. Comparing with the latter, a well-defined trimap can better constrain the solution space and lead to higher quality matting results. Therefore, our co-matting framework starts with trimap generation.

Manually specifying an accurate trimap is a tedious task for users. A viable alternative is to first partition the input image into foreground and background regions using binary segmentation, and then create a narrow band around the foreground boundary as the unknown region by applying morphological operations. Following this routine, we first use an existing co-segmentation technique [21] to perform joint binary image segmentation, and interactive scribbled-based segmentation refinement is only involved when the result is not satisfactory. After binary segmentation is done in each image, we then create a uniform unknown region using morphological operations. All unknown regions have the same width. In case that the foreground object contains both fuzzy and near-solid edges, we allow the user to touch up the unknown region using paint brushes on one image, and use an optical-flow-guided propagation method similar to the one proposed in a previous video matting system [26] to propagate the user modified trimap shape to other images.

After trimaps are generated, we use the Global Sampling matting [5] to generate an initial alpha matte for each input image. The initial alpha mattes will then be improved by our co-matting optimization approach.

3.2. The Optimization Formulation

There are two main goals we endeavor to achieve with the proposed co-matting framework. Firstly, it should allow high confidence local regions where local alpha mattes are accurate in one image to help improve
intra- and inter-image affinity, as well as an initial alpha matte computed by single image matting, constrained by constant prior weight (red line), matting confidence (green line) and sampling confidence (blue line). The arrow indicates how high confidence pixels guide the low confidence ones.

the matting result of their counterparts in other images. Secondly, we do not want low confidence regions where the alpha mattes could be erroneous to jeopardize high confidence regions in other images in the co-matting process. To achieve both goals, we present a new optimization framework by extending existing single image matting, constrained by constant prior weight (red line), matting confidence (green line) and sampling confidence (blue line). The arrow indicates how high confidence pixels guide the low confidence ones.

Figure 2: The graph model of the co-matting framework (assuming two input images). The alpha value of each pixel is determined by intra- and inter-image affinity, as well as an initial alpha matte computed by single image matting, constrained by constant prior weight (red line), matting confidence (green line) and sampling confidence (blue line). The arrow indicates how high confidence pixels guide the low confidence ones.

4. Estimating Local Matting Confidence

With initial alpha matte acquired, we turn our efforts on evaluating the matting "confidence", which describes how well a matting approach can do in a given image region. Essentially, the matting confidence directly correlates with the accuracy of the computed alpha matte. Notice that the definition of confidence here differs from that in previous sampling based approaches [2, 4, 5]. The latter is used to indicate how well the selected color samples satisfy the linear constraint in Eqn. 1 with the current unknown pixel. The matting confidence estimation is critical in our solution, as it predicts high-quality alpha regions and determines how matting information "flows" among different images in the co-matting framework.

\[
\alpha = \arg\min_{\alpha} [\lambda^T L^M \alpha + \lambda_r \alpha^T L^R \alpha + \lambda_c (\alpha - \tilde{\alpha})^T \Lambda (\alpha - \tilde{\alpha})],
\]

where \(\lambda^T L^M \alpha\) enforces local spatial matte smoothness of each image, with \(L^M\) constructed by stacking matting Laplacian matrices [7] along its diagonal. The second term encodes matte coherence across different images, based on the registration results among them. Specifically, \(L^R\) is a registration Laplacian defined as \(L^R = I - W^R\), where \(I\) is identity matrix and \(W^R\) is defined as:

\[
W^R_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ are registered;} \\
0 & \text{otherwise.} 
\end{cases}
\]

The third term in Eqn. 2 conditionally restrains the final alpha values to be consistent with initially estimated alpha values \(\tilde{\alpha}\). \(\Lambda\) is a diagonal matrix whose diagonal values set different weights for difference pixels:

\[
\Lambda_{ii} = \begin{cases} 
C & \text{if } i \text{ is in } F \text{ or } B; \\
\frac{C}{f_i} & \text{if } f(\alpha_i) > T_f; \\
f_i & \text{otherwise.} 
\end{cases}
\]

Finally, the two weights \(\lambda_r\) and \(\lambda_c\) in Eqn. 2 control the strength of the last two terms in Eqn. 2, and they are set at 100 and 0.1 in our experiments, respectively. The solution of the energy minimization problem can be obtained by solving a linear system with conjugate gradient method.

The above co-matting framework can also be viewed in a graph model perspective as shown in Fig. 2. It takes into account of two important factors: (1) the initial local matting quality in each image; and (2) the inter-image correlation among different images. Estimating local matting confidence allows us to identify "good" local regions that we can leverage to improve other regions. A large affinity is defined between each registered inter-image pixel pair (not necessarily dense), providing the linking-bridges that good alpha estimation flow along among different images. As a result, initial high confidence alpha values hold steady while low confidence ones are refined. In the following sections, we will describe how the two key ingredients, matting confidence (Sec. 4) and inter-image affinity (Sec. 5), are computed in our system.
Evaluating matting confidence is a challenging task, and as far as we know, no previous attempt has been made on this subject. We tackle the problem by learning an effective metric based on multiple features, which are designated to be consistent with observations. Before introducing the features and the learning process, we give a rough explanation about the motivation behind the feature choices.

In our approach, two types of features are investigated. The first category accounts for background complexity, which is a long-standing challenge for all existing matting techniques. To describe the background complexity, several region-wise features, including color, gradient and regularity are extracted. The second category of features are derived from performance analysis of existing matting techniques. Sampling confidence and matte discrepancy between different matting techniques are mainly explored for feature representation. Comparing with features in the first category, which only correlates with the content of the input image, these features are more specific to the limitations of existing matting algorithms. Combining features in two categories together, we are able to learn a metric that accurately predicts the matting quality a specific matting algorithm can achieve on a given image.

4.1. Granularity of feature extraction

To extract effective features for measuring matting confidence, the first and foremost task is to determine the granularity of feature extraction. In general, there are three levels of features: image-, region- and pixel-level. First of all, extracting features on the whole image level is not desirable in our framework. This is because that the final matting confidence is expected to be, rather than globally uniform, spatially-varying in different local regions, so that regions of different images can complement each other during the co-matting process. Furthermore, compared with pixel-wise features, region-wise ones tend to be more reliable and less sensitive to image noise. We thus choose to use region-wise features for measuring matting confidence.

A key question for computing region-wise features is how to partition the unknown region in a matting-confidence uniform manner. Considering the confidence is largely dependent on background appearance complexity of unknown region, which is difficult to evaluate under occlusion of foreground, we segment the unknown region based on two observations: (1) background appearance tends to be locally smooth; and (2) the band of the unknown region is usually narrow. Specifically, the partitioning process starts with applying Graph-based images segmentation approach [27] on the background region of the trimap, which divides the background region into sub-regions of uniform appearance. Thereafter, as shown in the example in Fig. 3, all intersection points of neighboring background sub-regions on the background boundary of the trimap are located, and lined to their spatial nearest neighbors on the foreground boundary. These lines, together with the original trimap boundaries, segment the unknown regions into multiple sub-regions. Pixels in each sub-region are expected to be of similar background appearance, hence uniform matting confidence. Furthermore, since there are certain other reasons accounting for matting confidence, our system also allows users to adjust the partitions if necessary. With unknown region partitioned, features extracted are described below.

4.2. Background-complexity-based features

Measuring image complexity has been extensively studied in the literature [28, 29]. The unique difficulty our approach faces, however, is that the unknown sub-regions are heavily occluded by the foreground object, thus directly computing features from them is inapplicable. Driven by the same observations behind our unknown region partitioning strategy, we choose to first extract complexity features for background sub-regions adjacent to the unknown region. The feature values are then propagated to the unknown sub-regions. Given a background region $B_i$ with pixel number $N_{B_i}$, the features we extract are:

- Color complexity (CC). This feature is measured by the normalized RGB color histogram entropy of the region. Let $H$ denote the normalized color histogram vector with bin number of $N_{ch}$ (we set as $4^*4^*4 = 256$, 4 bins for each color channel), the feature value is computed as

$$F_{B_i}^{CC} = \sum_{l=1}^{N_{ch}} H_l \log_2(H_l).$$
Intuitively, a smaller feature value indicates that the region contains more color species and thus more complex scene appearance.

- **Gradient complexity (GC).** It is defined as the average gradient magnitude of the region. Let $M$ denote the gradient magnitude vector of the region, it is computed as:

$$F_{GC}^B = \frac{1}{N_B} \sum_{k=1}^{N_B} M_k.$$ 

This feature reflects the gradient variation inside the region. A large value indicates severe gradient variation, which is usually caused by highly-textured backgrounds.

- **Regularity entropy (RE).** It is computed as the average regularity entropy of the local neighborhood around each pixel. Specifically, at each pixel location $k$, the gray value of a small window with size $S \times S$ is vector-quantized into $N_r$ bins, denoted as $V_k^i$, $k = 1, ..., N_r$. We set $S = 9$ and $N_r = 10$ in our experiments. Entropy of all vectors are then calculated and averaged as

$$F_{RE}^B = \frac{1}{N_B} \sum_{k=1}^{N_B} \sum_{i=1}^{N_r} V_k^i \log_2(V_k^i).$$

This feature measures the repeatability of local patterns in local regions. Intuitively, the more chaotic the region appears, the smaller the feature value becomes.

- **Region size (RZ).** It is inspired by the designation of Graph-based segmentation approach [27] in our unknown region partitioning strategy, which produces segmentation results by comparing inter-and intra-region difference. Large regions produced by this approach are usually of relatively small appearance variation, while small regions contain larger variations. Therefore, it may be safe to assume an inverse correlation between the region size and complexity.

Once we compute the feature vectors for background sub-regions, denoted as $F^B$, we propagate them into the adjacent unknown sub-regions. Specifically, the feature value $F_T$ of the unknown sub-region $T$ is computed as:

$$F_T = \frac{\sum_j N_j \frac{\text{length}(L_{B,T_j})}{\text{length}(L_{B,T})}}{N_B},$$

where $\text{length}(L_{B,T_j})$ denotes the length of the intersection line $L_{B,T_j}$ between sub-regions $B$ and $T_j$, and $T$ is the union of $T_j$, while $N_B$ is the number of background sub-regions.

### 4.3. Matting-algorithm-dependent features

As stated earlier, state-of-the-art matting approaches employ two fundamental techniques: (1) sampling known foreground and background colors for an unknown pixel for alpha estimation (color sampling); and (2) defining local matting affinities to enforce local matte smoothness (matting affinities). Recent approaches show that these two techniques augment each other and often lead to better results that neither of them can achieve alone. Nevertheless, no theoretical analysis has been made so far about under what circumstance, how much improvement can be achieved. Our experiments show that for an unknown pixel, if the two estimated alpha values, one from color sampling and the other from matting affinities, differ significantly from each other, the final alpha values usually contain large error. Therefore, computing the matte difference produced by these two distinct procedures may shed some light on matting confidence evaluation.

In addition, sampling-based matting techniques generally involve a sampling confidence evaluation process. As stated earlier, sampling confidence and the matting confidence we pursue here are not equivalent concepts. However, as a key piece of sampling based approach, it also contributes to matting confidence in a heuristic way. For this reason, we also incorporate it in our feature recipe. Moreover, color ambiguity on foreground and background may impact the matte quality of an unknown sub-region, which is also reported as limitation in both Closed-form matting [7] and Global Sampling matting [5]. Therefore, we further extract the following three algorithm-dependent features for matting confidence estimation. Note that different from complexity-based features, only the unknown region information are involved in feature derivation here.

- **Absolute matte difference (AMD).** Apart from Global Sampling matting, we also apply Closed-form matting [7] on each image to compute another matte. We then compute the average absolute matte difference for each unknown sub-region as a feature.

- **Average sampling confidence (ASC).** As in [5], sampling confidence for each unknown pixel is derived during the matting process. Feature value for each region is then defined by averaging the confidence values of each pixel it contains.
• Boundary color ambiguity (BCA). For each unknown sub-region, we collect known foreground and background colors along its inner and outer boundaries, and model them using Multivariate Gaussian distributions. KL divergence distance between the two models is then computed as a feature.

4.4. Learning a metric

Combining the features described above, we derive a matting confidence metric using statistic regression methods. After investigating several commonly-used models (e.g. Linear Regression, Support Vector Regression, etc), we choose Random Forest Regression [30] as the learning model given its good performance in our experiments.

Random Forest Regression is one of the well-known nonlinear regression models. It constructs an ensemble of classification and regression tree (CART) and predicts the target response as average response of all CARTs. During training, two kinds of randomness are introduced to avoid overfitting and to improve prediction accuracy. First, each tree is built on a set of samples randomly selected with replacement from the training data (bootstrap sampling). Second, during tree growing, splitting of each node is tested on a subset of randomly selected attributes (features). An important advantage is that bootstrap sampling leaves about 36% of the training data untouched, known as oob(out-of-bag) data. The oob data can be used to estimate the generalization error of each tree and help adjust parameters during tree building.

In our case, for training, we randomly collect 100 background images from internet and blend them with the foreground images using their corresponding ground-truth alpha mattes provided by the matting evaluation benchmark [14]. Global Sampling matting is then applied on all the synthesized images to estimate alpha mattes, and features are then extracted as articulated above. Thereafter, the regression model is established by fitting all the features to the ground-truth matting confidence, defined as:

\[
\exp\Big(\frac{|\alpha - \tilde{\alpha}|}{|\alpha - \tilde{\alpha}| - 1}\Big), \tag{5}
\]

where \(\alpha\) and \(\tilde{\alpha}\) are the ground-truth and estimated alpha mattes inside each unknown sub-region, respectively. We set the number of decision trees as 200 in our experiment.

5. Inter-image Unknown Region Alignment

Another important technical component in the comatting framework is to build correspondence among the unknown regions across different images. Directly aligning the unknown regions is difficult, since in these regions the foreground object is roughly the same, but the backgrounds are significantly different, thus directly computing image features in them for matching using conventional image registration methods is error-prone. Furthermore, if both foreground and background regions are locally smooth, matching ambiguity is inevitable if we only consider local features.

To avoid the difficulties mentioned above, we propose a two-phase, coarse-to-fine solution for aligning the unknown regions. It first identifies one image as a pivot to perform registration with others, which is the one that has the highest average matting confidence evaluated previously. We then apply SIFT-flow [31] between the pivot image and all other images for foreground alignment (i.e., applying it to the F regions in the trimaps). SIFT-flow combines densely-sampled SIFT feature with optical flow to estimate the pixel displacement field between two images. Compared with pure SIFT matching, it encourages uniform displacement in neighboring pixels. Compared with conventional optical flow, SIFT features achieve a higher degree of robustness by effectively characterizing view-invariant and brightness-independent image structures across images. After this step, most pixels near the foreground boundary are usually aligned well.

Next, we seek to align pixels in the unknown regions. As stated earlier, directly using the appearance features inside the unknown regions for matching are inapplicable due to the difference in background appearance. We thus assume that the registration filed between two images vary smoothly from the inner foreground region to the unknown region. With the assumption, a global non-rigid transformation between an unknown region pair is derived by performing a robust point matching algorithm [32] with thin-plate-spline (RPM-TPS) on two point sets, each consisting of pixels on the boundary between the foreground and the unknown region in each image. Specially, The PRM-TPS algorithm fits the non-rigid transformation with a thin-plate-spline function and develops an energy function encoding the fitting error as well as anchor points correspondence assignment. For optimization, it explores a EM style solution by alternating the process of transformation fitting and correspondence updating. For more detail, we refer the readers to [32].
6. Experimental Results

To verify the effectiveness of our system, we first report the quantitative evaluation results of the proposed matting confidence metric, then present co-matting results and comparisons on multiple examples as well as average time overhead in each experiment.

6.1. Evaluating the matting confidence metric

Given the multiple features proposed, we would like to analyze the informativeness of each one and their combinations. To this end, we conduct cross-validation tests on the dataset described in Sec. 4.4. Specifically, we first extract all 7 features from the training dataset and partition the features into 5 dividends. We then perform model evaluation 5 times, by selecting one of the dividends for testing and all others for training each time. It takes around 4.5 hours to extract \( \tilde{10}K \) features from all 1700 images (100 background image blended with 17 foreground ones drawn from the matting evaluation benchmark [14]). Once the features are generated, the process of cross-validation is quite fast, taking about 3.5 minutes.

Average SAD between the predicted confidence values and the ground-truth ones are reported in Fig. 4. In addition to the results obtained by all features combined, we also report results by using each individual feature only. The results show that, out of all the individual features, Absolute Matte Difference (AMD) performs the best, while the other features works comparatively well in predicting the matting confidence. Unsurprisingly, combining all the features together outperforms each individual feature.

6.2. Co-matting results

We test the proposed approach on a variety of examples, some are photographed by ourselves and others are collected from clips of movies and video dataset used in previous video matting algorithms. To demonstrate its superior to single image matting, we also compare our results with those of Closed-form matting [7] and Global Sampling matting [5], which are regarded as state-of-the-art propagation-based and sampling-based approaches, respectively.

Throughout the experiments, our co-matting approach successfully improves the results generated by initial single image matting. Fig. 5 shows an example containing two video frames extracted from the film “Life of Pi”. It shows that Closed-form matting [7] tends to produce over-smoothed, blurry matte regions where high contrast details are lost (see red arrows in the figure). On the other hand, Global Sampling matting [5] produces noisy results when foreground and background colors are ambiguous (see blue arrows in the figure). In comparison, our co-matting algorithm generates better quality mattes on both input images by leveraging the complementary nature of local regions between the two images.

Fig. 6 shows the results of two video frames from an example proposed in [33]. The two input images share very similar background appearance, except for a few local regions marked by the color rectangles. The main challenge of matting in these regions lies in the structural changes of the background, which leads to inconsistent results for previous single image matting methods. Our approach successfully remedies the flaws therein. Particularly, an interesting scenario occurs in the second image, where a nail on the wood building arises, neighboring the hair of the woman as indicated by red arrow. The nail succeeds to cheat the two single image matting approaches but fails to do so in our system thanks to the reference in the other image.

Our system is capable of dealing with more than two images at the same time. In Fig. 7, we show such an example by extracting 3 inconsecutive frames from a video clip. As stated in Sec. 5, to avoid repeatedly aligning unknown regions, the 2nd image is automatically selected as the pivot image for alignment. As shown in the figure, the resulting mattes produced by the proposed co-matting approach consistently meet our expectations. Regions of low initial matte quality (marked by solid color rectangles) are co-improved by other higher quality regions across images.

6.3. Running time

Table 1 summarizes the average time overhead of each experiment above. Image size and average number of pixels in the unknown regions of each group are provided as well. The whole algorithm is implemented using MATLAB and C++ combination on a laptop with Intel Core i5 2.67GHz CPU and 4GB memory.

Overall, the co-matting framework is more time-intensive compared with single image matting approaches. As such, the main time overhead concentrates
on the two sub-tasks, feature extraction and unknown region alignment. The former sub-task takes most of its time to compute AMD features, which involves running two single image matting algorithms, Closed-form matting and Global Sampling matting. The final optimization time (we use CHOLMOD [34] to solve the linear system) of Fig. 5 & 6 are both less than Closed-form matting, thanks to the reuse of matting Laplacian calculated previously in constructing the final linear system. Nevertheless, as the optimization time grows non-linearly with the number of unknown pixels, it exceeds Closed-form matting in the case of Fig. 7. Besides, in practice, we simplify the final linear system by excluding most of the known pixels, which are distant from unknown regions in each image and only correlate with known pixels in the corresponding matting Laplacian matrix. This trick brings no harm to the optimization results in principle and in practice, but largely saving the memory required and running time.

In terms of user assistance, it takes 3, 4 and 10 user scribbles for co-segmentation refinement of images in Fig. 5, 6 and 7 respectively. Unknown region partition rectification costs further 0, 2 and 4 times of user intervention. Although it seems quite a few of user assistance involved in the entire process, we have to emphasize that the pre-processing for creating trimaps is required by most existing image matting techniques. In general, it takes more user efforts to perform binary segmentation one by one, which is also the motivation behind co-segmentation techniques.
7. Conclusion and Future Work

In this paper, we have presented a framework for the new task of image co-matting, referring to jointly extracting alpha mattes from multiple images with the same foreground object (slight deformation allowed) on different backgrounds. A global energy function encoding single image matting confidence and inter-image affinities is proposed. The matting confidence is evaluated by learning a metric based on multiple visual features, while inter-image affinities are determined by unknown region alignment with SIFT flow and non-rigid matching. Experimental results show that our system can improve the matting quality by jointly considering multiple images.

The performance of our system is limited by the quality of initial mattes. If for the same local region all matting results across different image are bad, then our system cannot improve this local region much in this case. Also, our system cannot deal with large foreground deformation, as registration becomes a much harder problem. Therefore, our future work consists of investigating how to derive high quality initial mattes as well as a more robust unknown region registration strategy. We also plan to provide an interactive tool for image co-matting so the user can provide additional constraints in an efficient way for improving the final results.

References

[19] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, icoseg: In-
Figure 7: Comparison results of three images of a man before screen