Video Matting via Sparse and Low-Rank Representation

Dongqing Zou, Xiaowu Chen*, Guangying Cao, Xiaogang Wang
State Key Laboratory of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University, Beijing, China

Abstract

We introduce a novel method of video matting via sparse and low-rank representation. Previous matting methods [10, 9] introduced a nonlocal prior to estimate the alpha matte and have achieved impressive results on some data. However, on one hand, searching inadequate or excessive samples may miss good samples or introduce noise; on the other hand, it is difficult to construct consistent nonlocal structures for pixels with similar features, yielding spatially and temporally inconsistent video mattes. In this paper, we proposed a novel video matting method to achieve spatially and temporally consistent matting result. Toward this end, a sparse and low-rank representation model is introduced to pursue consistent nonlocal structures for pixels with similar features. The sparse representation is used to adaptively select best samples and accurately construct the nonlocal structures for all pixels, while the low-rank representation is used to globally ensure consistent nonlocal structures for pixels with similar features. The two representations are combined to generate consistent video mattes. Experimental results show that our method has achieved high quality results in a variety of challenging examples featuring illumination changes, feature ambiguity, topology changes, transparency variation, dis-occlusion, fast motion and motion blur.

1. Introduction

Video matting is to accurately extract a moving foreground matte from an input video while avoiding spatial and temporal artifacts. It is a fundamental and important computer vision problem with many applications, including hair modeling [7], dehazing [19] and so on. In the past few years, various video matting methods [14, 3, 23, 25] have been presented and achieved impressive matting results. Despite of much progress on video matting, it is still very challenging to achieve temporal consistency due to topology variation, motion blur, transparency changing and dis-occlusion.

Most video matting methods use optical flow to generate temporally consistent video mattes. However, optical flow can not guarantee to estimate an accurate motion for videos with complex scenes, resulting in temporal artifacts in video matting. Methods [10, 9] introduce a nonlocal prior to estimate the alpha matte. However, on one hand, searching inadequate or excessive samples may miss good samples or introduce noise; on the other hand, it is difficult to construct consistent nonlocal structures for pixels with similar features, yielding spatially and temporally inconsistent video mattes.

In fact, the nonlocal prior proposed in [10, 25] implies that nonlocal pixels with similar features are generated from the same subspace. Thus these pixels can be represented by several bases or atoms according to the sparse and low-rank representations [35, 8, 39]. Accordingly, if we can discover some subspaces that well represent the foreground and background of all frames, and build the relationships between pixels within the same subspace, the spatial and temporal relationships between pixels would be obtained.

According to the analysis above, it is reasonable to assume that pixels from the same object in different frames are drawn from one identical low-rank feature subspace, and all pixels in several successive frames lie on a union of multiple subspaces. This assumption can be justified by natural statistic and observations of videos. Therefore, if each pixel can be represented as a linear combination of atoms, we can pursue a low-rank and sparse representation for all pixels. With the sparse constraint, each pixel in the video will be only represented with several related atoms, which in theory is consistent with the principle of nonlocal matting methods [10, 9]. With the low rank constraint, pixels with similar features in the same frame are represented with the same atoms in a dictionary, thus spatial consistency is achieved. Moreover, under this constraint, pixels with similar features from successive frames are represented with the same atoms too. Low-rank constraint contributes to ensuring temporally consistent mattes.

In this paper, we propose a novel video matting method via sparse and low-rank representation in this paper. With

*corresponding author (email: chen@buaa.edu.cn)

some sparse inputs on some key frames, we first learn a dictionary which consists of two sub-dictionaries. These two sub-dictionaries describe the contents of known foreground and background regions in key frames, respectively. With the learned dictionary, we then represent all pixels in the input video while pursuing a low-rank and sparse representation to obtain a coefficient matrix. Finally, coupled with multi-frame Laplacian used for enhancing the local smoothness of alpha values, the alpha matte of each frame is solved.

The key contributions of this work include: 1) A novel video matting method via sparse and low-rank representation is proposed. Our method achieves spatial and temporal consistency and overcomes the matte artifacts caused by topology changing, feature ambiguity and motion variation. 2) A novel dictionary learning algorithm is proposed to well represent the foreground and the background regions in the target video, which contributes to improving the matting accuracy. We demonstrate the superior performance of our method on standard databases by comparing with state-of-the-art methods.

2. Related Work

In this section, we review only the most relevant works to ours. A more comprehensive survey on image and video matting can be found in [30].

Sparse Representation. In the past few years, the sparse representation has been applied to the problem domain of image processing, such as image super-resolution [37], image and video denoising and inpainting [27], cross-style image synthesis [33]. Sparse representations have also been applied to face recognition [36], image background modeling [6], and image classification [26]. More related works can be found in the comprehensive surveys [35] and [17]. Recently, Jubin *et al*. [20] proposed a sparse coding image matting method, in which the sum of the sparse codes of foreground pixels is regarded as the estimate of the alpha matte. Different from this matting method [20], our method further constraints that pixels with similar features should have similar alpha values through low-rank representation. Besides, instead of directly using all known pixels as a dictionary for alpha matting, we propose to learn a discriminative dictionary to better represent the pixels in unknown regions.

Low-Rank Representation. Comparing to sparse representation, low-rank representation has a better performance in discovering global structures of data. The low-rank representation can reveal the relationships of the samples: the within-cluster affinities are dense while the between-cluster affinities are all zeros. The low-rank representation has been applied to many applications of image processing including image denoising [34], face recognition [8], classification [39] and so on [28]. To the best of our knowledge, no work, however, has applied the low-rank representation to solve video matting problems, we adopt the low-rank representation for video matting for the first time.

Video Matting. The effectiveness of existing video matting methods is dependent on accurate optical flow or special hardware systems. Chuang et al. [14] interpolated the trimaps across the video by using forward and backward optical flow. [22, 4] applied the optical flow to generate the trimaps for each frame according to the trimaps of key frames. Eisemann et al. [16] proposed a spectral video matting method which warps matting components using optical flow. Lee et al. [23] extended the robust matting [31] into a temporally coherent video matting method by using optical flow to define an anisotropic kernel. Wang et al. [32] proposed a co-matting method which propagates a trimap to other images by using the optical flow. Affinity motion was introduced in [15, 25] to obtain temporally consistent video mattes. Different from optical flow based methods, hardware-assisted systems [21] focus on automatically generating trimaps for all video frames. Different from these methods, we apply the sparse and low-rank representation to construct nonlocal structures for pixels to solve video mattes. Our method is also related to some image matting methods such as KNN matting [10], Learning based image matting [12] and Closed-form matting [24].

3. Sparse and Low-Rank Constraints

3.1. Sparsity on Matting

Chen et al. [10] proposed a nonlocal smooth prior guided image matting method, which estimates the alpha value of each pixel by preserving the nonlocal structure of each pixel. Later, Chen et al. [9] proposed a KNN matting which capitalizes on matching K nonlocal neighborhoods for solving alpha matte. The basic idea of these two nonlocal prior guided image matting methods is to search K samples $\{x_i\}_{i=1}^{K}$ to represent pixel j with a set of weights $\{w_i\}_i^K$. The estimated alpha value of pixel j is calculated as $\alpha_j = \sum_i^K w_i \alpha_i$, and $\alpha_i = 1$ if pixel i is in foreground regions, 0 otherwise. Obviously, the selected K nearest samples for each pixel are sparse in the image, so the nonlocal prior to image matting can be equally transformed into a sparse representation problem. That is, the pixels in known foreground and background regions of an image $\mathbf{X}$ can be treated as the dictionary $\mathbf{D}$, a corresponding sparse code matrix $\mathbf{W}$ which subjects to $\mathbf{X}$=$\mathbf{DW}$ can be calculated. As a result, the alpha value of each pixel j in unknown regions can be estimated according to the corresponding sparse codes $\mathbf{W}_j \in \mathbf{W}$, namely, $\alpha_j = \sum_i \mathbf{W}_j^i \delta(\mathbf{D}_i)$, $\delta(\mathbf{D}_i) = 1$ if $\mathbf{D}_i$ represents foreground, 0 otherwise.

Specifically, let $\mathbf{X}_i$ represent the image i in RGBXY space, if the appropriate dictionary $\mathbf{D}$ representing known

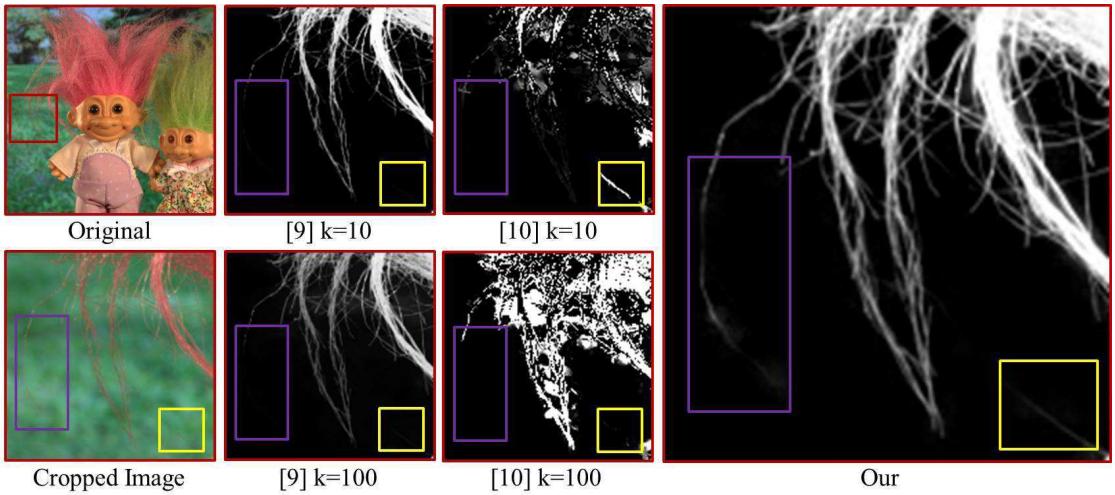

Figure 1. It is difficult to get good nonlocal structures by using previous methods [9, 10]. In the first row, the matting results are obtained with recommended parameter K by these two methods. In the second row, a large K is required to capture the more information by [10]. The last column shows the result by our method.

pixels is available, the KNN searching problem in matting can be defined as such a sparse representation problem,

$$\arg\min_{\mathbf{W}_i} ||\mathbf{X}_i - \mathbf{D}\mathbf{W}_i||_0 + ||\mathbf{W}_i||_0, \tag{1}$$

where $||\bullet||_0$ denotes zero-norm, which is used to count the number of non-zero entries in representation matrix $\mathbf{W}_i$. The atoms in $\mathbf{D}$ corresponding to non-zero weights in $\mathbf{W}_i$ are the expected samples. The sum of those weights corresponding to foreground atoms is the estimated alpha matte.

The sparsity constraint can benefit the sample selection for matting. Previous methods [10, 9] fixed the number of nearest neighbors during samples searching, which may result in bad mattes for some images. As shown in Figure 1, for methods [10, 9], inadequate neighbors will result in incorrect alpha values for some pixels, while excessive neighbors will introduce noise. In comparison, by applying the sparse representation, best samples which could reconstruct each pixel can be selected for alpha estimation. Moreover, appropriate number of samples for each pixel can be automatically computed, and the number of samples for different pixels can be different, which helps to remove noise while improving accuracy of alpha estimation, yielding good results.

3.2. Low-Rankness on Matting

In *image matting*, a good matting result expects that pixels with similar features have similar alpha values. According to the nonlcoal image matting methods [10, 9], pixels with similar features are expected to be represented by K similar neighbors, so they should have similar nonlocal structures in a feature space. Such nonlocal structure constraints in sparse representation require pixels with similar features to have similar representations over the learned dictionary. As a result, the representation matrix $\mathbf{W}_i$ in Eq. (1) is expected to be low-rank. The sparse and low-rank constraint is defined as,

$$\arg\min_{\mathbf{W}_i} ||\mathbf{X}_i - \mathbf{D}\mathbf{W}_i||_0 + ||\mathbf{W}_i||_0 + ||\mathbf{W}_i||_*, \tag{2}$$

where the $||\bullet||_*$ denotes the matrix nuclear norm, which is used to find lowest possible rank of a matrix.

Analogously, we can get similar conclusion on *video matting*. Since frames in a video shot usually describe the same scene, all frames $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ in a video shot lie on low-dimensional subspaces [35], too. Therefore, the concatenation of corresponding representation matrixes $\{\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n\}$ over the dictionary $\mathbf{D}$ is expected to be low-rank. The representation matrix $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n\}$ can be obtained by minimizing

$$\begin{aligned} &\min \sum_i^n (||\mathbf{X}_i - \mathbf{D}\mathbf{W}_i||_0 + ||\mathbf{W}_i||_0) + ||\mathbf{W}||_*, \\ &\forall p, q, (w_i)_{p,q} \in \mathbf{W}_i, \quad s.t. \quad (w_i)_{p,q} >= 0. \end{aligned} \tag{3}$$

where the $(w_i)_{p,q}$ represents the response of pixel q in i_{th} frame over p_{th} atom in dictionary $\mathbf{D}$, and the non-negative constraint on $(w_i)_{p,q}$ is set to avoid generating negative alpha values. Let t denote the number of atoms in learned dictionary $\mathbf{D}$, n represent the number of frames and m represent the number of pixels in a frame, the $\mathbf{W}$ in Equation

3 is:

$$\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n\} = \begin{vmatrix} (w_1)_{1,1} & \cdots & (w_i)_{1,q} & \cdots & (w_n)_{1,m} \\ \vdots & \ddots & \vdots & & \vdots \\ (w_1)_{p,1} & \cdots & (w_i)_{p,q} & \cdots & (w_n)_{p,m} \\ \vdots & & \vdots & \ddots & \vdots \\ (w_1)_{t,1} & \cdots & (w_i)_{t,q} & \cdots & (w_n)_{t,m} \end{vmatrix}$$

3.3. Optimization

To solve the Equation 3, we first convert it into the following equivalent problem:

$$\begin{aligned} &\min \sum_{i}^{n}(||W_i||_1 + \lambda||E_i||_1) + \gamma||W||_* \\ &s.t. \quad X_i = DS_i + E_i; \quad W_i = J_i; \\ &\quad W_i = S_i; \quad W_i = T_i, \quad T_i >= 0. \end{aligned} \tag{4}$$

where the parameters J_i, S_i, T_i are auxiliary variables used to solve this equation. Specifically, this energy function can be solved with the inexact augmented Lagrange multiplier method or alternating direction method (ADM) [29], which equivalents to minimize the following augmented Lagrange function:

$$\begin{aligned} &\min(\gamma||W||_* + \sum_{i}^{n}(||J_i||_1 + \lambda||E_i||_1) + \sum_{i}^{n}(\langle A_i, W_i - J_i\rangle \\ &+ \langle Y_i, X_i - DS_i - E_i\rangle + \langle V_i, W_i - S_i\rangle + \langle U_i, W_i - T_i\rangle \\ &+ \frac{\mu}{2}||X_i - DS_i - E_i||_F^2 + \frac{\mu}{2}||W_i - J_i||_F^2 \\ &+ \frac{\mu}{2}||W_i - S_i||_F^2 + \frac{\mu}{2}||W_i - T_i||_F^2)) \end{aligned}$$

where $A_1, ..., A_n, Y_1, ..., Y_n, V_1, ..., V_n, U_1, ..., U_n$ are Lagrange multipliers, and $\mu > 0$ is a penalty parameter. The inexact ALM method for this equation is outlined in Algorithm 1. Note that the sub-problems of the algorithm are convex and they all have closed-form solutions.

4. Video Matting

Since the obtained representation matrix **W** encodes the spatially and temporally consistent nonlocal structures for all pixels, the nonlocal structures hold for alpha values of all pixels according to [10, 25]. To get the representation matrix of the input video, we first learn a dictionary, which consists of two sub-dictionaries, from known foreground and background regions in key frames. With the obtained representation matrix, we construct the nonlocal relationships between alpha values to enhance temporal consistency. Finally, we extend the matting Laplacian to multi-frame matting Laplacian to enhance the local smoothness of alpha values.

Algorithm 1 Optimization of problem (4) by ADM.

Input: Data $\{X_i\}$, dictionary D, parameters λ and γ.
Initialize: $A = U = V = Y = 0, S = T = J = 0,\ \mu = 10^{-6}$.
while not converged **do**

1. Fix the others and update $J_1, ..., J_n$ by

$$J_i = \arg\min_{J_i} \frac{1}{\mu}||J_i||_1 + \frac{1}{2}||J_i - (W_i + \frac{A_i}{\mu})||_F^2.$$

2. Fix the others and update $S_1, ..., S_n$ by

$$S_i = (D^T D + I)^{-1}(D^T(X_i - E_i) + W_i + \frac{(D^T Y_i + V_i)}{\mu}).$$

3. Fix the others and update $T_1, ..., T_n$ by

$$T_i = W_i + \frac{U_i}{\mu}, T_i = max(T_i, 0).$$

4. Fix the others and update W by

$$W = \arg\min_{W} \frac{\gamma}{2\mu}||W||_* + \frac{1}{2}||W - M||_F^2.$$

where M is a matrix formed as follows:

$$M = [F_1, F_2, ..., F_n],$$

in which $F_i = \frac{1}{3}(J_i + S_i + T_i - \frac{(A_i + V_i + U_i)}{\mu})$.
5. Fix the others and update the $E_1, ..., E_n$ by

$$E_i = \arg\min_{E_i} \frac{\lambda}{\mu}||E_i||_1 + \frac{1}{2}||E_i - (X_i - DS_i + \frac{Y_i}{\mu})||_F^2.$$

6. Update the multipliers

$$\begin{aligned} A_i &= A_i + \mu(W_i - J_i), \\ Y_i &= Y_i + \mu(X_i - DS_i - E_i), \\ V_i &= V_i + \mu(W_i - S_i), \\ U_i &= U_i + \mu(W_i - T_i). \end{aligned}$$

7. Update μ by $\mu = \min(1.1\mu, 10^{10})$.
(ρ=1.9 in all experiments).
8. Check the convergence condition: $X_i - DS_i - E_i \to 0, W_i - J_i \to 0, W_i - S_i \to 0$ and $W_i - T_i \to 0$.
end while
return W.

4.1. Discriminative Dictionary Learning

A dictionary **D**, which consists of two sub-dictionaries $\{\mathbf{D}_f, \mathbf{D}_b\}$, is first learned from the users labeled key frames

to represent the target video. Besides $\mathbf{D}$ is required to have powerful reconstruction ability, a good dictionary $\mathbf{D}$ should have powerful discriminative ability. The discriminative ability of the dictionary $\mathbf{D}$ means that the corresponding sub-dictionary has good representation ability to the associated class while poor representation ability for other classes. Accordingly, the within-class weights in the representation matrix should be non-zero while the between-class weights should be all zeros. The rationality of this assumption for $\mathbf{D}$ attributes to the fact that most alpha values of pixels are 1 in foreground, while 0 in background. According to the fact that the alpha value of a pixel is proportional to the sum of corresponding coding coefficients over $\mathbf{D}$, a dictionary $\mathbf{D}$ with powerful discriminative capability is expected to learned to achieve a better matting result.

Let $\mathbf{X}_f$ and $\mathbf{X}_b$ denote the pixels from foreground and background in selected key frames, respectively. Let $\mathbf{Z}_f = \{\mathbf{Z}_f^f, \mathbf{Z}_f^b\}$ denote the coefficient matrix representing $\mathbf{X}_f$ over $\mathbf{D}$, and $\mathbf{Z}_b = \{\mathbf{Z}_b^f, \mathbf{Z}_b^b\}$ denote the coefficient matrix representing $\mathbf{X}_b$ over $\mathbf{D}$. $\{\mathbf{Z}_i^j | i, j = f, b\}$ is the coding coefficient of $\mathbf{X}_i$ over the sub-dictionary $\mathbf{D}_j$. The discriminative dictionary learning model is defined as,

$$\min_{(D,Z_i)} \sum_i (||\mathbf{X}_i - \mathbf{D}\mathbf{Z}_i||_F^2 + ||\mathbf{X}_i - \mathbf{D}_i\mathbf{Z}_i^i||_F^2 + \sum_{j \neq i} ||\mathbf{D}_j\mathbf{Z}_i^j||_F^2), \quad (5)$$

where the first two terms are reconstruction terms, which expects that the $\mathbf{X}_i$ should be well represented by the dictionary $\mathbf{D}$ and corresponding sub-dictionary $\mathbf{D}_i$. The last term constrains that the coefficients $\mathbf{Z}_i^j$ of $\mathbf{X}_i$ over sub-dictionary $\mathbf{D}_j$ should approach to zero, such that $||\mathbf{D}_j\mathbf{Z}_i^j||_F^2$ is small. The Eq. (5) can be solved by using the quadratic programming algorithm [38].

Figure 2 shows a comparison between using and without using the discrimination constraint for matting. The matte $\mathbf{I}_\alpha = \mathbf{Z}_f * \mathbf{1} + \mathbf{Z}_b * \mathbf{0}$. Figure 2 (b) is the matting result with the represent matrix of the learned dictionary without using discrimination constraint, wherein the alpha values of some pixels in background regions are incorrectly estimated. This is because some pixels in background are represented by some atoms in foreground subdictionary $\mathbf{D}_f$, which results in lots of positive corresponding sparse codes in $\mathbf{Z}_b^f$, yielding incorrect alpha values. In contrast, by considering the discrimination constraint during dictionary learning, our method can correctly distinguish the foreground from the background, leading to an accurate alpha matte, as shown in Figure 2 (c).

4.2. Temporally matting

Nonlocal Structure. Given the dictionary $\mathbf{D} = \{\mathbf{D}_f, \mathbf{D}_b\}$ with discriminative ability, we construct the nonlocal low-rank and sparse relationships between pixels for video matting. Specifically, given a video with n frames $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$, we obtain the nonlocal low-rank and sparse

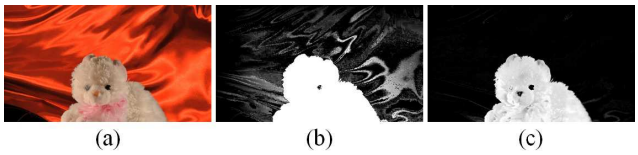

Figure 2. Comparison between using and without using discrimination constraint for dictionary learning. (a) is the target scene, (b) is the matting result with representation matrix of the learned dictionary without using discrimination constraint, (c) is the result obtained by using our discriminative dictionary.

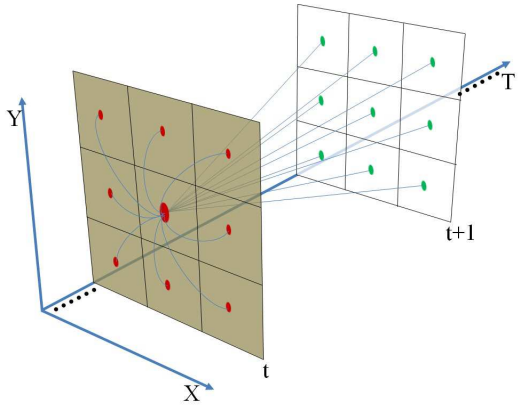

Figure 3. Multiframe local matting Laplacian. The 2m×2m matting Laplacian encodes the relationships across successive two frames to enhance local smoothness.

relationship $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n\}$ between all pixels by using Equation 3. The nonlocal relationships are used to measure the affinities for all alpha values of corresponding pixels, which is defined as,

$$\min \sum_i^n \sum_j^m (\alpha_{ij} - \alpha_D \mathbf{w}_{ij})^2, \quad (6)$$

where α_{ij} represents the alpha value of the pixel j in ith frame, m denotes the number of pixels in a frame, and $\alpha_D = \{\alpha_f, \alpha_b\}$ represents the alpha values of all atoms in dictionary $\mathbf{D}$. $\alpha_f = \mathbf{1}$ for the corresponding atoms in foreground sub-dictionary, and $\alpha_b = \mathbf{0}$ for the all atoms in background sub-dictionary. $\mathbf{w}_{ij} = [(w_i)_{1,j}, ..., (w_i)_{t,j}]^T$ is the jth column weights vector in $\mathbf{W}_i \subset \mathbf{W}$.

Multi-frame Local Laplacian. As pointed out by Chen et al. [11], nonlocal prior alone for image matting will fail in capturing local structures of semitransparent objects, resulting in spatial incoherent matting results. By combining the nonlocal prior with the local Laplacian, good results will be obtained. We thus are inspired to extend the image matting Laplacian to multi-frame Laplacian to complement the nonlocal structure for video matting. Specifically, following the principles in previous works [25, 13], we assume that the color line model for a local 3×3 window also holds for a $3 \times 3 \times 2$ cube formed with pixels in two consecutive

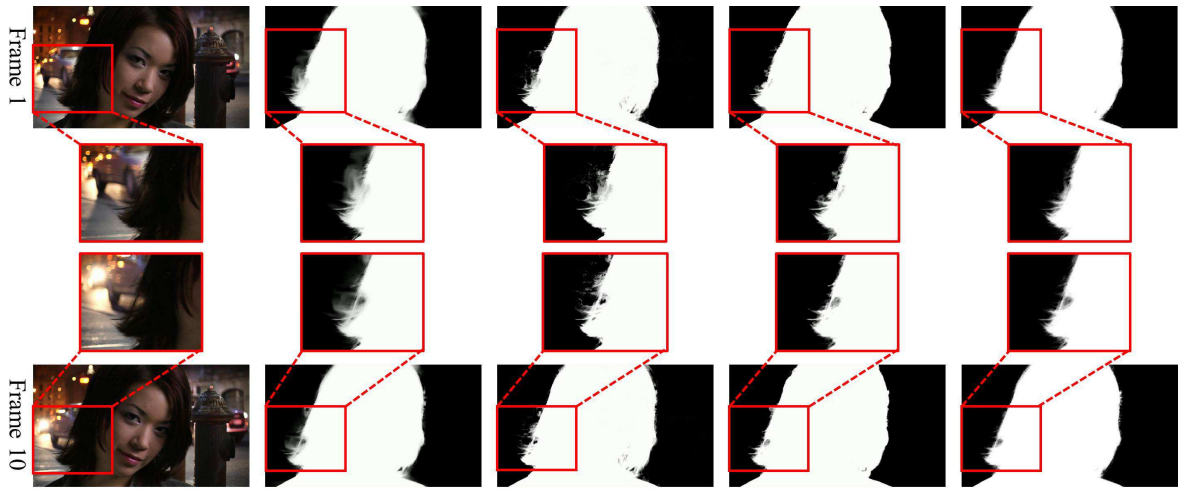

Figure 4. Comparisons with [5, 9, 24] on the video *city*. This example demonstrates that our matting method can handle illumination changes and feature ambiguity.

frames, as shown in Figure 3. Therefore, for the pixel i and j in a cube c_k, the multi-frame local Laplacian W_{ij}^{mlap} is defined as,

$$W_{ij}^{mlap} = \delta \sum_{k}^{(i,j)\in c_k} \frac{1 + (C_i - \mu_k)\left(\sum_k + \frac{\epsilon}{18} I\right)^{-1}(C_j - \mu_k)}{18}. \quad (7)$$

Here, the parameter δ controls the strength of the local smoothness. μ_k and $\sum_k$ represent the color mean and variance in each cube. ϵ is a regularization coefficient which is set to 10^{-5}. C_i is the feature of pixel i. Our cube size is fixed as $3 \times 3 \times 2$ for all examples.

Closed-form Solution. Pixels with known alpha values $\{g_i\}$ from the trimap and dictionary **D** are first collected to form a subset $\mathbb{S}$. The energy function for solving alpha values for the input video is defined as:

$$\begin{aligned} E = \lambda \sum_{i\in\mathbb{S}} (\alpha_i - g_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{m} (\alpha_{ij} - \alpha_D \mathbf{w}_{ij})^2 \\ + \sum_{i=1}^{n}\sum_{j=1}^{m} \Big(\sum_{k\in N_j} W_{jk}^{mlap} (\alpha_{ij} - \alpha_k) \Big)^2 \end{aligned} \quad (8)$$

where the set N_j is the set of neighbors of the pixel j, including neighboring pixels in $3\times3\times2$ cube. Equation 8 can be further rewritten into a matrix form as:

$$E = (\alpha - G)^T \Lambda (\alpha - G) + \alpha^T L \alpha, \quad (9)$$

in which

$$L = \begin{bmatrix} L_D & -\mathbf{W} \\ -\mathbf{W}^T & L_u \end{bmatrix} \quad (10)$$

Here, **W** is the representation matrix for the video, and $L_D = \mathbf{W} * \mathbf{W}^T$. L_u is a block diagonal matrix consisting of multi-frame Laplacian matrixes for all frames, namely, $L_u = diag(L_u^1; ...; L_u^n)$. The matrix L is symmetric and can be solved with the Nystrom method [18].

The Equation 10 is a quadratic function about α, which can be minimized by solving the linear equation in closed-form solution:

$$(\Lambda + L)\,\alpha = \Lambda G. \quad (11)$$

In fact, instead of solving the alpha values for all frames, we can reduce to solve only two successive frames at a time. In this way, the alpha mattes of next coming frames are solved progressively along the time axis until all frames are processed. The 2-frame affinity matrix is effective to ensure temporal information, since both forward and backward affinities are taken into consideration when defining L_u. Moreover, comparing to solving all frames, a matrix of size $2m \times 2m$ is built, where m is the total number of pixels to be processed in one frame, which will drastically reduce the running time and memory consumption.

5. Experiment

We demonstrate our method on various videos and compare it to state-of-the-art methods to show the effectiveness achieved by our video matting method. Due to space limitations, we are only able to show selected results in the paper as a demonstration. More results can be found in the supplementary material.

We compare our method to KNN matting [9], Close-formed matting [24], and Video Snapcut [5] to demonstrate the performance of our method. For fairly comparing the matting performance, we use the professional datasets for video matting [1] and image matting [2] to perform comparison. Each test video in this dataset is coupled with cor-

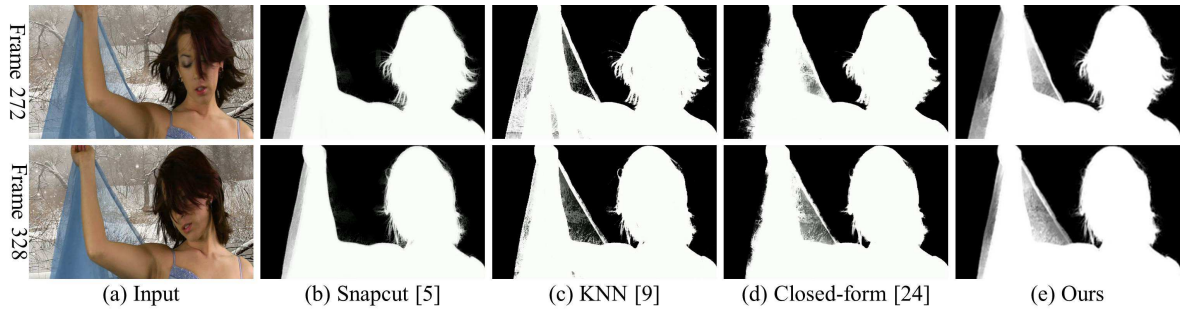

Figure 5. Comparisons with [5, 9, 24] on the video *snow*. This example demonstrates that our method can handle transparency variation.

responding temporally consistent trimaps (the groundtruth is not publicly available).

5.1. Qualitative Comparison

Illumination changes and feature ambiguity. As shown in Figure 4 (a), when the illumination is changing over the sequence, low contrast and feature ambiguous regions around objects will arise, feature based nearest neighbors searching will result in an inaccurate estimation of alpha mattes for the foreground, as the results shown in Figure 4 (c) produced by [9]. Local smoothness priors alone results in blurry and unclear boundaries, as the results shown in Figure 4 (b) (d) generated by [5, 24]. Figure 4 (e) by our method shows that sparse and low-rank constraints are effective in constructing consistent and good nonlocal structures, thus producing better results in the presence of illumination changes and feature ambiguity.

Transparency variation. Transparency variation of the foreground will result in large variation of feature values for the corresponding pixels in different frames, yielding incorrect mattes and temporal incoherence in video matting, as the results by [5, 9, 24] shown in Figure 5. In contrast, our method is robust to the transparency variation and achieves an accurate and temporally consistent video matting result.

Dis-occlusion and Changing topology. Figure 6 demonstrates that our method is able to handle dis-occlusion via obtaining best nonlocal atoms from learned dictionary. In comparison, video snapcut [5] fails in distinguishing background from the foreground when topology changes, resulting in a bad video matte, as the matting result on frame 137 shown in Figure 6.

Shape changes. Figure 7 demonstrates that our method can handle shape changes via sparse and low-rank constraints. KNN matting [9] produces a spatially inconsistent result in which alpha value of some pixels in foreground are estimated incorrectly, Snapcut [5] generates temporally inconsistent mattes, as shown in Figure 7 (b)(c), respectively. Our method generates a spatially and temporally consistent video matte, as shown in Figure 7 (d).

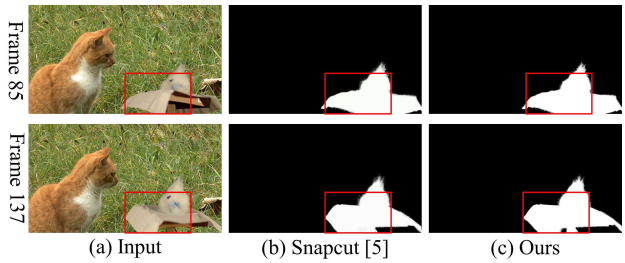

Figure 6. Comparisons with [5] on the video *slava*. This example demonstrates that our method can handle dis-occlusion and topology changes.

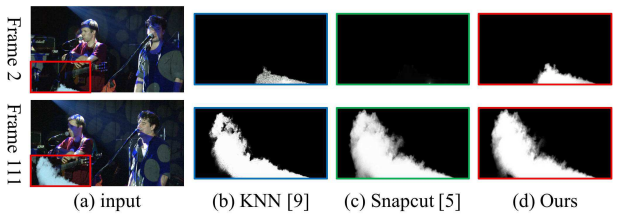

Figure 7. Comparisons with [9, 5] on the video *concert*. This example demonstrates that our method can handle shape changes.

Fast motion and motion blur. Figure 8 shows an example to demonstrate the ability of our method on handling the foreground with fast motion and motion blur. In this example, the woman repeatedly and briskly swings her arm up and down, and thus generating large motion blur. The ability of KNN matting [9] degrades so greatly for this case that it is hard to extract the regions between foreground and background. Our method works well in this complex situation.

Sparse inputs. Our method can generate spatially and temporally consistent matting results with limited user interactions. As shown in Figure 9, we only treat the first frame as the only keyframe with a sparse trimap, and run our method automatically on the other frames without any user interaction. Our method generates the temporally consistent results while the Snapcut [5] fails to accurately extract the foreground object with the sparse inputs.

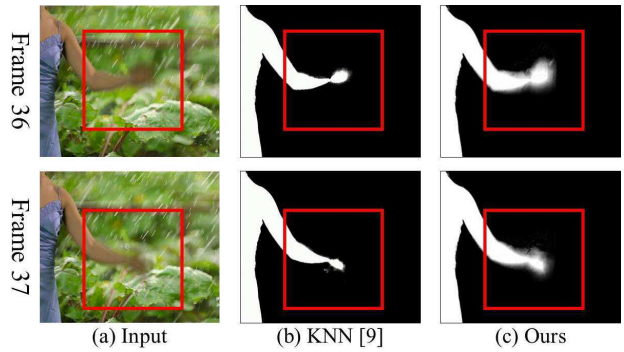

Figure 8. Comparisons with [9] on the video $rain$. This example demonstrates that our method can handle fast motion and motion blur.

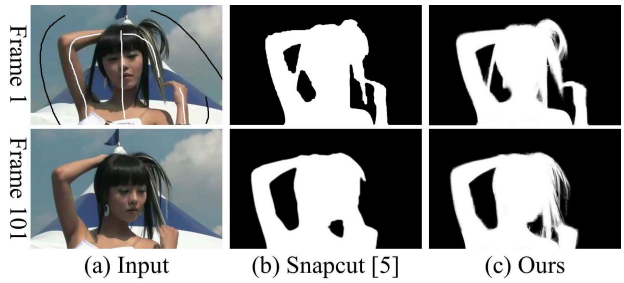

Figure 9. Comparisons with [5] when using sparse inputs. Only strokes on the first frame are input and our method gets a better matting result.

5.2. Quantitative Comparison

We perform quantitative comparison to evaluate temporal coherence, by measuring differences in alpha values between successive frames according to [23]. The measure of the difference $dif(i)$ for the i_{th} pixel in t_{th} frame is defined as,

$$dif(i) = \frac{\alpha_i(t+1) - \alpha_i(t)}{I_i(t+1) - I_i(t)},$$

where the $\alpha_i(t)$ represents the alpha value of i_{th} pixel in t_{th} frame, and its RGB color feature is denoted by $I_i(t+1)$.

Table 1 shows the comparisons between our matting method and the KNN matting [9], video snapcut [5], and Closed-Form matting [24] on five videos. These five videos are $city$, $concert$, $snow$, $slava$, and $rain$. Obviously, our method generates more coherent results on each video than previous methods. Here, due to space limited, we only show the average alpha difference $\sum_{t=1}^{n}\sum_{i=1}^{m} dif(i)/(m*n)$ of the whole video. The difference of alpha values frame by frame can be found in our supplementary file.

We also use the benchmark database for image matting to evaluate the performance of our method [2]. Our method ranks first according to the measurement of gradient error, ranks fifth and sixth according to connectivity error and MSE, respectively, and ranks tenth according to the SAD.

	KNN [9]	Snapcut[5]	Closed-form[24]	Ours
city	0.2677	0.1466	0.1507	**0.1391**
concert	0.0440	0.0345	0.0337	**0.0243**
snow	0.4362	0.3972	0.8724	**0.2578**
slava	0.1422	0.1374	0.1335	**0.0947**
rain	0.6627	0.8626	0.9538	**0.2560**

Table 1. Comparisons of error rates of different methods on five videos.

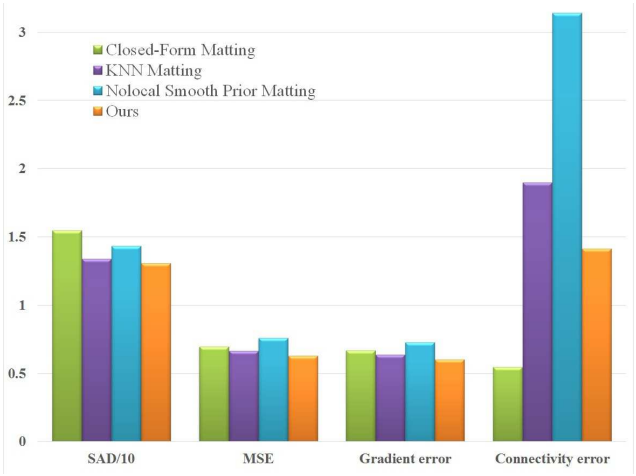

Figure 10. Quantitative evaluation of our method and the methods of KNN matting [9], Closed-form matting [24] and nonlocal smooth prior guided image matting [10].

Our method is promising to get better performance by using some pre-optimization or post-optimization, as some previous methods did. We compare the quantitative error of our work with some related image matting methods (including KNN mating [9], Closed-form matting [24] and nonlocal smooth prior guided image matting [10]) in Figure 10. Our method generates smallest errors.

6. Conclusion

In this paper, we proposed a novel video matting method via sparse and low-rank representation. We are the first to apply sparse and low-rank representation to construct consistent non-local structures for pixels from different frames. Our method generates spatially and temporally consistent video matting results, and alleviated the interactions for users. Comparisons on standard benchmark databases show that our work outperforms state-of-the-art methods.

7. Acknowledgement

We thank the reviewers for their valuable feedback. This work is supported in part by grants from NSFC (61325011) & (61532003), 863 Program (2013AA013801), and SRFDP (20131102130002).

References

[1] http://videomatting.com/. 6

[2] http://alphamatting.com/. 6, 8

[3] N. Apostoloff and A. W. Fitzgibbon. Bayesian video matting using learnt image priors. In *CVPR*, pages 407–414, 2004. 1

[4] X. Bai, J. Wang, and D. Simons. Towards temporally-coherent video matting. In *Computer Vision/Computer Graphics Collaboration Techniques*, pages 63–74. Springer, 2011. 2

[5] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: Robust video object cutout using localized classifiers. *ACM Trans. Graph.*, 28(3):70:1–70:11, 2009. 6, 7, 8

[6] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *ECCV*, pages 155–168, Berlin, Heidelberg, 2008. 2

[7] M. Chai, L. Wang, Y. Weng, X. Jin, and K. Zhou. Dynamic hair manipulation in images and videos. *ACM Trans. Graph.*, 32(4):75:1–75:8, 2013. 1

[8] C. Chen, C. Wei, and Y. F. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *CVPR*, pages 2618–2625, 2012. 1, 2

[9] Q. Chen, D. Li, and C.-K. Tang. Knn matting. In *CVPR*, pages 869–876, june 2012. 1, 2, 3, 6, 7, 8

[10] X. Chen, D. Zou, Q. Zhao, and P. Tan. Manifold preserving edit propagation. *ACM Trans. Graph.*, 31(6):132:1–132:7, 2012. 1, 2, 3, 4, 8

[11] X. Chen, D. Zou, S. Z. Zhou, Q. Zhao, and P. Tan. Image matting with local and nonlocal smooth priors. In *CVPR*, pages 1902–1907, 2013. 5

[12] I. Choi, S. Kim, M. Brown, and Y.-W. Tai. A learning-based approach to reduce jpeg artifacts in image matting. In *ICCV*, pages 2880–2887, 2013. 2

[13] I. Choi, M. Lee, and Y.-W. Tai. Video matting using multi-frame nonlocal matting laplacian. In *ECCV*, pages 540–553, 2012. 5

[14] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. *ACM Trans. Graph.*, 21(3):243–248, 2002. 1, 2

[15] D. Corrigan, S. Robinson, and A. Kokaram. Video matting using motion extended grabcut. pages 1–9, 2008. 2

[16] M. Eisemann, J. Wolf, and M. A. Magnor. Spectral video matting. In *VMV*, pages 121–126, 2009. 2

[17] M. Elad, M. A. T. Figueiredo, and Y. Ma. On the Role of Sparse and Redundant Representations in Image Processing. *Proc. the IEEE*, 98(6):972 –982, 2010. 2

[18] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *PAMI*, 26(2):214–225, 2004. 6

[19] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2011. 1

[20] J. Johnson, D. Rajan, and H. Cholakkal. Sparse codes as alpha matte. In *BMVC*, pages 245–253, 2014. 2

[21] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. *ACM Trans. Graph.*, 25(3):779–786, July 2006. 2

[22] J. Ju, J. Wang, Y. Liu, H. Wang, and Q. Dai. A progressive tri-level segmentation approach for topology-change-aware video matting. *CGF*, 32(7):245–253, 2013. 2

[23] S.-Y. Lee, J.-C. Yoon, and I.-K. Lee. Temporally coherent video matting. *Graphical Models*, 72(3):25–33, 2010. 1, 2, 8

[24] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, pages 61–68, 2006. 2, 6, 7, 8

[25] D. Li, Q. Chen, and C.-K. Tang. Motion-aware knn laplacian for video matting. In *ICCV*, pages 3599–3606, 2013. 1, 2, 4, 5

[26] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, pages 1–8, 2008. 2

[27] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008. 2

[28] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, pages 853–860, 2012. 2

[29] Y. Shen, Z. Wen, and Y. Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. Technical Report UILU-ENG-09-2215, 2011. 4

[30] J. Wang and M. F. Cohen. Image and video matting: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(2):97–175, Jan. 2007. 2

[31] J. Wang and M. F. Cohen. Optimized color sampling for robust matting. In *CVPR*, pages 1–8, 2007. 2

[32] L. Wang, T. Xia, Y. Guo, L. Liu, and J. Wang. Confidence-driven image co-matting. *Computers & Graphics*, 38:131–139, 2014. 2

[33] S. Wang, L. Zhang, L. Y., and Q. Pan. Semi-coupled dictionary learning with applications in image super-resolution and photo-sketch synthesis. In *CVPR*, pages 2216–2233, 2012. 2

[34] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems*, pages 2080–2088. 2009. 2

[35] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proc. the IEEE*, 98(6):1031–1044, 2010. 1, 2, 3

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, Feb. 2009. 2

[37] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *Trans. Img. Proc.*, 19(11):2861–2873, 2010. 2

[38] M. Yang, Z. Lei, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. In *ICIP*, pages 1601–1604, 2010. 5

[39] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In *CVPR*, pages 1673–1680, 2011. 1, 2