Spatiotemporal Texture Reconstruction for Dynamic Objects Using a Single RGB-D Camera

Hyomin Kim, Jungeon Kim, Hyeonseo Nam, Jaesik Park, and Seungyong Lee

POSTECH

Abstract
This paper presents an effective method for generating a spatiotemporal (time-varying) texture map for a dynamic object using a single RGB-D camera. The input of our framework is a 3D template model and an RGB-D image sequence. Since there are invisible areas of the object at a frame in a single-camera setup, textures of such areas need to be borrowed from other frames. We formulate the problem as an MRF optimization and define cost functions to reconstruct a plausible spatiotemporal texture for a dynamic object. Experimental results demonstrate that our spatiotemporal textures can reproduce the active appearances of captured objects better than approaches using a single texture map.

CCS Concepts
• Computing methodologies → Texturing;

1. Introduction
3D reconstruction methods using RGB-D images have been developed for static scenes [NIH+11, NZIS13, DNZ+17] and dynamic objects [NFS15, IZN+16, DKS+16]. The appearance of the reconstructed 3D models, which are often represented as color meshes, is crucial in realistic content creation for Augmented Reality (AR) and Virtual Reality (VR) applications. Subsequent works have been proposed to improve the color quality of reconstructed static scenes [GWO+10, ZK14, JJKL16, FYY+18, LGY19] and dynamic objects [OERF+16, DCC+18, GLD+19]. For static scenes, most of the approaches generate a global texture atlas [WMG14, JJKL16, FYY+18, LGY19], and the main tasks consist of two parts: selecting the best texture patches from input color images and relieving visual artifacts such as texture misalignments and illumination differences when the selected texture patches are combined.

While single texture map is widely used to represent a static scene, it is not preferred for a dynamic object due to appearance changes over time. A multi-camera system could facilitate creating
a spatiotemporal (time-varying) texture map of a dynamic object, but such setup may not be affordable for ordinary users. As an alternative, recovering a spatiotemporal texture using a single RGB-D camera can be a practical solution.

Previous reconstruction methods using a single RGB-D camera maintain and update voxel colors in the canonical frame [NFS15, IZN*16, GXY*17, YGX*17, YZG*18]. However, this voxel color based approach cannot properly represent detailed time-varying appearances, e.g., cloth wrinkles and facial expression changes for capturing humans. Kim et al. [KKPL19] proposed a multi-scale optimization method to reconstruct a global texture map of a dynamic model under a single RGB-D camera setup. Still, the global texture map is static and cannot represent time-varying appearances of a dynamic object.

In this work, we propose a novel framework to generate a spatiotemporal texture map for a dynamic template model using a single RGB-D camera. Given a template model that is dynamically aligned with input RGB-D images, our main objective is to produce a spatiotemporal texture map that provides a time-varying texture of the model. Since a single camera can obtain only partial observations at each frame, we bring color information from other frames to complete a spatiotemporal texture map. We formulate a Markov random field (MRF) energy, where each node in the MRF indicates a triangular face of the dynamic template model at a frame. By minimizing the energy, for every frame, an optimal image frame to be used for texture mapping each face can be determined. Recovering high-quality textures is another goal of this work because it is crucial for free-viewpoint video generation of a dynamic model. Due to imperfect motion tracking, texture drifts may happen when input color images are mapped onto the dynamic template model. To resolve the problem, we find optimal texture coordinates for the template model on all input color frames. As a result, our framework can handle time-varying appearances of dynamic objects better than the previous approach [KKPL19] based on a single static texture map.

The key contributions are summarized as follows:

- A novel framework to generate a stream of texture maps for a dynamic template model from color images captured by a single RGB-D camera.
- An effective MRF formulation including selective weighting to compose high-quality time-varying texture maps from single view observations.
- An accelerated approach to optimize texture coordinates for resolving texture drifts on a dynamic template model.
- Quantitative evaluation and user study to measure the quality of reconstructed time-varying texture maps.

2. Related Work
2.1. Dynamic Object Reconstruction

3D reconstruction systems for dynamic objects can be classified into two categories: template-based and templateless. A representative templateless work is DynamicFusion [NFS15], which shows real-time non-rigid reconstruction using a single depth camera. VolumeDeform [IZN*16] considers a correspondence term made from SIFT feature matching to improve motion tracking. BodyFusion [YGX*17], which specializes in human body, exploits human skeleton information for surface tracking and fusion.

Template-based non-rigid reconstruction assumes that a template model for the target object is given, and deforms the template model to be aligned with input data. Most template-based methods [LAGP09, ZNI*14, GXW*15, GXW*18] are based on embedded deformation (ED) graph model [SSP07], which is also utilized in templateless works [LGV*13, DTF*15, DDK*16, DDF*17]. High-end reconstruction systems using a number of cameras can capture a complete mesh of the target object at each timestamp, so they commonly use template tracking between keyframes for compressing redundant data [CCS15, PKC*17, GLD*19].

In this paper, we use a template-based method [GXW*15, GXW*18] to track the motion of a dynamic object, while any method can be used if it can generate a sequence of meshes with a fixed connectivity conforming a RGB-D image sequence.

2.2. Appearances of Reconstructed Models

Static scenes When scanning a static scene with a single RGB-D camera, color information obtained by volumetric fusion tends to be blurry due to inaccurate camera pose estimation. Gal et al. [GWO*10] optimize projective mapping from each face of a mesh onto one of the color images by performing combinatorial optimization on image labels and adjusting 2D coordinates of projected triangles. Zhou and Kolkin [ZK14] jointly optimize camera poses, non-rigid correction functions, and vertex colors to increase the photometric consistency. Jeon et al. [JKL16] optimize texture coordinates projected on input color images to resolve misalignments of textures. Bi et al. [BKR17] propose patch-based optimization to align images against large misalignments between frames. Fu et al. [FYX18] and Li et al. [LYG19] improve the color of a reconstructed mesh by finding an optimal image label for each face. Fu et al. [FYX20] jointly optimize texture and geometry for better correction of textures. Huang et al. [HTD*20] exploit a deep learning approach to reconstruct realistic textures.

Non-rigid objects with multiple cameras Non-rigid object reconstruction has received attention in the context of performance capture using multi-view camera setup [CCS*15, PKC*17, DDK*16, OERF*16]. Fusion4D [DDK*16] is a real-time performance capture system using multiple cameras, and it shows the capability to reconstruct challenging non-rigid scenes. However, it uses volumetrically fused vertex colors to represent the appearance, resulting in blurred colors of the reconstructed model. To resolve the problem, Holoportation [OERF*16] proposes several heuristics, including color majority voting. Montage4D [DCC*18] considers misregistration and occlusion seams to obtain smooth texture weight fields. Offline multi-view performance capture systems [CCS*15, GLD*19] generate a high-quality texture atlas at each timestamp using mesh parameterization, and temporally coherent atlas parameterization [PKC*17] was proposed to increase the compression ratio of a time-varying texture map.

Non-rigid scenes with a single camera Kim et al. [KKPL19] produces a global texture map for a dynamic object by optimizing...
texture coordinates on the input color images in a single RGB-D camera setup. Pandey et al. [PTY19] propose semi-parametric learning to synthesize images from arbitrary novel viewpoints. PFu [SHN19] combines learning-based 3D reconstruction, volumetric deformation, and light-weight bundle adjustment to reconstruct a clothed human model in a few seconds from a single RGB-D stream. These methods produce either a single texture map for the reconstructed dynamic model [KKPL19, SHN19] or temporally independent blurry color images for novel viewpoints [PTY19]. They all cannot generate a sequence of textures varying with motions of the reconstructed model over time. Our framework aims to produce a visually plausible spatiotemporal texture for a dynamic object with a single RGB-D camera.

3. Problem Definition

3.1. Spatiotemporal Texture Volume

Given a triangular mesh $\mathcal{M}$ of the template model, we assume that a non-rigid deformation function $\Psi_t$ is given for each time frame $t$, satisfying $\Psi_t(\mathcal{M}) = \mathcal{M}_t$, where $\mathcal{M}_t$ denotes the deformed template mesh fitted to the depth map $D_t$. Our goal is to produce a spatiotemporal (time-varying) texture map $\mathcal{T}$ for the dynamic template mesh. By embedding $\mathcal{M}$ onto a 2D plane using a parameterization method [Mic11], each triangle face $f$ of $\mathcal{M}$ is mapped onto a triangle in the 2D texture domain. A spatiotemporal texture map can be represented as stacked texture maps (or spatiotemporal texture volume), whose size is $W \times W \times T$, where $W$ is the side length of a rectangular texture map and $T$ is the number of input frames. We assume that our mesh $\mathcal{M}_t$ keeps a fixed topology in the deformation sequence, and the position of face $f$ in the 2D texture domain remains the same regardless of time $t$. Only the texture information of face $f$ changes with time $t$ to depict dynamically varying object appearances.

3.2. Objectives

To determine the texture information for face $f$ of the deformed mesh $\mathcal{M}_t$ at time $t$ in a texture volume, we need to solve two problems: color image selection and texture coordinate determination. We should first determine which color image in the input will be used for extracting the texture information. A viable option is to use the color image $C_t$ at time $t$. However, depending on the camera motion, face $f$ may not be visible in $C_t$, and in that case, texture should be borrowed from another color image $C_{t'}$ where the face is visible. Moreover, when there are multiple such images, the quality of input images should be considered, as some images can provide sharper textures than others.

Once we have selected the color image $C_{t'}$, we need to determine the mapped position of $f$ in $C_{t'}$. Since we already have the deformed mesh $\mathcal{M}_{t'}$ at time $t'$, the mapping of $f$ onto $C_{t'}$ can be obtained by projecting face $f$ in $\mathcal{M}_{t'}$ onto $C_{t'}$. However, due to the errors in camera tracking and non-rigid registration, mappings of $f$ onto different input color images may not be completely aligned. If we use such mappings for texture sampling, the temporal texture volume would contain little jitters in the part corresponding to $f$.

An overall pipeline of our system is shown in Figure 2. Starting from RGB-D images and a deforming model, we determine the aligned texture coordinates using global texture optimization (Section 4). Then, to select color images for each face $f$ at all timestamps, we utilize MRF optimization (Section 5). To reduce redundant calculation, we pre-compute and tabulate the measurements that are required to construct the cost function before MRF optimization. Finally, we build a spatiotemporal texture volume and refine textures by conducting post-processing.

4. Spatiotemporal Texture Coordinate Optimization

Camera tracking and calibration errors induce misalignments among textures taken from different frames. Non-rigid registration
errors incur additional misalignments in the case of dynamic objects. Kim et al. [KKPL19] proposed an efficient framework for resolving texture misalignments for a dynamic object. For a template mesh $M$, the framework optimizes the texture coordinates $\xi$ of mesh vertices on the input color images $C$ so that sub-textures of each face extracted from different frames match each other. The optimization energy is defined to calculate the photometric inconsistency among textures taken from different frames:

$$E(\xi, P) = \sum |C(\xi) - P|^2,$$

where $P$ denotes proxy colors computed by averaging $C(\xi)$. $\xi$ and $P$ are solved using alternating optimization, where $\xi$ is initialized with the projected positions $\Phi$ of the vertices of deformed meshes on corresponding color images.

To obtain texture coordinates with aligned textures among frames, we adopt the framework of Kim et al. [KKPL19] with some modification. Their approach selects keyframes to avoid using blurry color images (due to abrupt motions) in the image sequence and considers only those keyframes in the optimization. In contrast, we need to build a spatiotemporal texture volume that contains every timestamp, so optimization should involve all frames. We modify the framework to reduce computational overhead while preserving the alignment quality.

Our key observation is that the displacements of texture coordinates determined by the optimization process smoothly change among adjacent frames. In Figure 3, red dots indicate initially projected texture coordinates $\Phi(V)$ of vertices $V$, and the opposite endpoints of blue lines indicate optimized texture coordinates $\xi(V)$. Based on the observation, we regularly sample the original input frames and use the sampled frames for texture coordinate optimization. Texture coordinates of the non-sampled frames are determined by linear interpolation of the displacements at the sampled frames.

In the optimization process, we need to compute the proxy colors $P$ from the sampled image colors. To avoid invalid textures (e.g., background colors), we segment foreground and background on the contrast, we need to build a spatiotemporal texture volume that contains every image frame for each face $f$ of the dynamic template. In our experiment, where we sampled every frame, we chose keyframes to avoid using invalid textures. Their approach selects keyframes to avoid using invalid textures. However, this approach produces a static texture map that does not reflect dynamic appearance changes.

To solve the labeling problem, we use MRF (Markov Random Field) [GG84, Bes86] optimization. The label $l_{f,t}$ for face $f$ at time $t$ is a node in the MRF graph, and there are two kinds of edges in the graph. A spatial edge connects two nodes $l_{f,t}$ and $l_{f',t}$ if faces $f$ and $f'$ are adjacent to each other in the template mesh. A temporal edge connects two nodes $l_{f,t}$ and $l_{f,t+1}$ in the temporal domain. To determine the value of $l_{f,t}$ on the MRF graph, we minimize the following cost function.

### 5.1. Cost Function

Our cost function consists of data term $E_d$ and smoothness term $E_s$ as follows,

$$E(L) = \sum_{f \in F} \sum_{l \in L} (E_d(l_{f,t}) + \lambda \sum_{(f',t) \in \mathcal{N}_{f,t}} E_s(l_{f,t}, l_{f',t+1})).$$

where $F$ is the set of faces in the template mesh and $\mathcal{N}_{f,t}$ is the set of adjacent nodes of $l_{f,t}$ in the MRF graph.

#### 5.1.1. Data Term

In many cases, a natural choice for the label of $l_{f,t}$ would be frame $t$. However, if face $f$ is not visible in $C_t$, i.e., $t \notin \mathcal{L}_f$, we cannot choose $t$ for $l_{f,t}$. In addition, even in the case that $t \in \mathcal{L}_f$, $C_t$ could be blurry and some other $C_{t'} \notin \mathcal{L}_f$ could be a better choice. To find a plausible label for $l_{f,t}$, in either case, we assume that similar shapes would have similar textures too. The data term consists of two terms $E_{\text{qual}}$ and $E_{\text{geo}}$:

$$E_{\text{data}}(l_{f,t}) = E_{\text{qual}}(l_{f,t}) + E_{\text{geo}}(l_{f,t}).$$

$E_{\text{qual}}$ is designed to avoid assigning low-quality texture, such as blurred region. $E_{\text{qual}}$ is defined as follows:

$$E_{\text{qual}}(l_{f,t}) = \chi_{\text{blur}}(l_{f,t} - \xi(f, t_{f,t} + 1)) + \chi_{\text{blur}}(1 - \xi(f, t_{f,t} + 1)).$$
where $\chi_0(a)$ is a step function whose value is 1 if $a \geq \theta$, and 0 otherwise. The first term estimates the blurriness of face $f$ by measuring the changes of the optimized texture coordinates $\xi$ of vertices among consecutive frames, $l_f$ and $l_{f+1}$. The second term uses the dot product of the face normal $n$ and the camera’s viewing direction $c$. $E_{\text{geo}}$ prefers to assign an image label if it provides sharp texture and avoids texture distortion due to slanted camera angle. By taking a step function $\chi_0$ in $E_{\text{geo}}$, no penalty is imposed once the image quality is above the threshold, keeping the plausible label candidates as much as possible.

$E_{\text{geo}}$ ensures to select a timestamp with a similar shape. The geometry term measures global similarity $E_{\text{geo}}$ and local similarity $E_{\text{loc}}$ between shapes:

$$E_{\text{geo}}(l_f) = \omega_0 E_{\text{geo}}(l_f) + (1 - \omega_0) E_{\text{loc}}(l_f).$$ (5)

$E_{\text{geo}}$ considers overall geometric similarity between frames and is defined as follows:

$$E_{\text{geo}}(l_f) = \sum_p \mathcal{R}_{l_f-l_{f+1}}(p) \Phi(K(M_t), C_{l_f}) \cup \Phi(M_{f+1}, C_{l_{f+1}}).$$ (6)

where $p$ denotes an image pixel, and $\mathcal{R}_{l_f-l_{f+1}}(p) = \min (1, |\mathcal{T}_K(M_t)(p) - \mathcal{T}_M(p)|)$ is the clipped difference between two depth values $\mathcal{T}_K(M_t)$ and $\mathcal{T}_M$, that are rendered from $K(M_t)$ and $M_{f+1}$, respectively. Here, $K \in \mathbb{R}^{4 \times 4}$ denotes a 6-DoF rigid transformation that aligns $M_t$ to $M_{f+1}$.

$E_{\text{geo}}$ is devised to find an image label $l_f$ such that the deformed template meshes $M_t$ and $M_{f+1}$ are geometrically similar. It measures the sum of depth differences $\mathcal{R}$ over the union area of the projected meshes on image $C_{l_f}$ and $\Phi(K(M_t), C_{l_f}) \cup \Phi(M_{f+1}, C_{l_{f+1}})$, where $p$ in Eq. (6) is a pixel in the union area on $C_{l_f}$. We obtain $K$ with RANSAC based rigid pose estimation that utilizes coordinate pairs of the same vertices in $M_t$ and $M_{f+1}$.

Even though we found a geometrically similar frame $l_f$, with $E_{\text{geo}}$, local shapes around face $f$ do not necessarily match among $M_t$ and $M_{f+1}$. Local similarity $E_{\text{loc}}$ guides shape search using local geometric information and is defined as $E_{\text{loc}} = -\frac{1}{2} \sum \tau_j \cdot \tau_j$, where $(j, f')$ denotes one of three vertex pairs from face $f$ in $M_t$ and $M_{f'}$. We utilize SHOT [TSDS10] to obtain a descriptor $\tau$ of the local geometry calculated for each vertex. SHOT originally uses the Euclidean distance to find the neighborhood set. Since we have a template mesh, we exploit the geodesic distance to obtain the local surface feature. To reduce the effect of noise, we apply median filter to SHOT descriptor values of vertices.

### 5.1.2. Smoothness Term

The smoothness term is defined for each edge of the MRF graph. Its role is to reduce seams in the final texture map and to control texture changes over time. To this end, we define spatial smoothness $E_{\text{spa}}$ in texture map domain and temporal smoothness $E_{\text{temp}}$ over time axis of the spatiotemporal texture volume. That is,

$$E_s(l_f, l_{f'}, l_{f''}) = \omega_0 E_{\text{spa}}(l_f, l_{f'}) + \omega_0 E_{\text{temp}}(l_f, l_{f'-l_{f''}}).$$ (7)

**Figure 4:** Texture dynamism analysis. (a) texture map obtained by partial observation at timestamp 11, (b) SSIM map between timestamps 11 and 12 (darker is smaller), (c) partial texture map at timestamp 12. Blue and sky-blue boxes are on clothes whose appearances are frequently changing due to wrinkles, while green and yellow boxes are on consistent texture regions. Note that the texture under the blue box is not shown in (a) and (b), as it is invisible at timestamps 11 and 12. (d) The red curve corresponding to the red box on the mouth in (c) shows small SSIM values only at certain timestamps (e.g., opening and closing the mouth).

$E_{\text{spa}}$ is defined as follows:

$$E_{\text{spa}}(l_f, l_{f'}) = \frac{1}{|\text{edge}|} \sum_{p,p' \in \text{edge}} \left( |C_{l_f}(p) - C_{l_{f'}}(p')| + |G_{l_f}(p) - G_{l_{f'}}(p')| \right).$$ (8)

where $\text{edge}$ is a shared edge between two faces $f$ and $f'$, and $(p, p')$ are the corresponding image pixels when the edge is mapped onto $C_{l_f}$ and $C_{l_{f'}}$. If the nodes $l_f$ and $l_{f'}$ have different image labels, color information should match along the shared edge to prevent a seam in the final texture map. In addition, we use the differences of gradient images, $G_{l_f}$ and $G_{l_{f'}}$, to match the color changes around the edge between $f$ and $f'$. $E_{\text{temp}}$ is defined as follows:

$$E_{\text{temp}}(l_f, l_{f'}, l_{f''}) = \frac{1}{|\text{face}|} \sum_{p,p' \in \text{face}} |C_{l_f}(p) - C_{l_{f'}}(p')|,$$ (9)

where $\text{face}$ denotes the projection of face $f$ onto $C_{l_f}$ and $C_{l_{f'}}$, and $(p, p')$ denotes corresponding image pixels on the projected faces. $E_{\text{temp}}$ measures the intensity differences and promotes soft transition of the texture assigned to $f$ over time shift $t \rightarrow t'$.

### 5.2. Selective Weighting for Dynamic Textures

Textures on a dynamic object may have different dynamism depending on their positions. Some parts can have dynamic appearance changes with object motions, e.g., cloth wrinkles under human motions, while other parts may have almost no changes of
textures over time, e.g., naked forearms of the human body. We verified this property with an experiment. After we have obtained optimized texture coordinates for a deforming mesh sequence, we can build an initial incomplete spatiotemporal texture volume by simply copying the visible parts in the input images onto the corresponding texture maps (Figure 4a). Then, we compute the SSIM map between adjacent frames in the texture volume on the overlapping areas (Figure 4b). Figure 4d shows changes in the SSIM values over time for the marked parts in Figure 4c. In the blue and cyan colored regions, SSIM fluctuates along time, meaning that these parts have versatile textures. Such parts need to be guaranteed for dynamism in the final spatiotemporal texture, while smooth texture changes would be fine for other parts.

To reflect this property in the MRF optimization, we modify the smoothness terms on MRF edges. We ignore a temporal MRF edge if the labels of the linked nodes exhibit a lower SSIM value, i.e., less correlation. In our temporal selective weighting scheme, the following new temporal smoothness term replaces $E_{\text{temp}}(l_f,t_f,t'_f)$ in Eq. (7):

$$
E_{\text{temp}}(l_f,t_f,t'_f) \leftarrow E_{\text{temp}}(l_f,t_f,t'_f) \cdot \chi_0(\min\{S(l_f,t_f),S(l_f,t'_f)\}),
$$

(10)

where $S(l_f,t)$ is the average SSIM value of the pixels inside face $f$ computed among consecutive frames. $\chi_0(a)$ is a step function whose value is 1 if $a \geq \theta_0$, and 0 otherwise. This scheme encourages the dynamism of textures over time when there is less correlation among frames.

In a similar way, we update the spatial smoothness term $E_{\text{spa}}(l_f,t_f)$ in Eq. (7) as follows:

$$
E_{\text{spa}}(l_f,t_f) \leftarrow E_{\text{spa}}(l_f,t_f) \left(2 - \frac{\chi_0(S_m(f,k)) + \chi_0(S_m(f',k))}{2}ight),
$$

(11)

where $S_m(f,k) = \min_{k \in C_f} S(f,k)$. If there is a label with a small SSIM value in $C_f$, the part has a risk of temporal discontinuity due to the updated temporal smoothness term in Eq. (10). Therefore, in that case, we enforce the spatial smoothness term to impose spatial coherence when temporal discontinuity happens. In Figure 4c, the red box is likely to exhibit temporal changes, e.g., mouth opening. Our spatial selective weighting scheme increases the spatial smoothness term for the red box to secure spatially coherent changes of texture.

5.3. Labeling and Post Processing

Since our MRF graph is large to solve, we divide the problem and solve it using alternating optimization. The idea of fixing a subset and optimizing the others is known as block coordinate descent (BCD). In [CK14, TWW∗16], the BCD technique is applied to uniform and non-uniform MRF topology. Our MRF graph consists of spatial edges connecting adjacent mesh faces at a single frame and temporal edges connecting corresponding mesh faces at consecutive frames. As a result, our MRF topology is spatially non-uniform but temporally uniform. We subdivide the frames into two sets: even-numbered and odd-numbered. We first conduct optimization for even-numbered frames without temporal smoothness terms. After that, we optimize for odd-numbered frames while fixing the labels of even-numbered frames, where temporal smoothness terms become unary terms. We iterate alternating optimization among even- and odd-numbered frames until convergence. To optimize a single set MRF graph (even or odd), we use an open-source library released by Daniel et al. [TWW∗16].

After labeling is done, we have a frame label per face at each timestamp. Then we fill the spatiotemporal texture atlas volume by copying texture patches from the labeled frames. When there are drastic brightness changes among different frames, spatial seams may happen on label boundaries (Figure 5a). To handle such cases, we apply gradient processing [PKCH18] to texture slices in the volume, resolving spatial seams (Figure 5c).

6. Results

6.1. Experiment Details

Setup All datasets shown in this paper were recorded by an Azure Kinect DK [Mic20] with 3072p resolution for color images, and 640 × 576 for depth images. All experiments were conducted on a workstation equipped with an Intel i7-7700K 4.2GHz CPU, 64GB RAM, and NVIDIA Titan RTX GPU. Our RGB-D sequences used for experiments contain 200 to 500 frames. The size of a texture slice in a spatiotemporal texture volume is 4000 × 4000. We set $\theta_0 = 0.9$, $\omega_0 = 10$, $\omega_2 = 2$, $\theta_t = 20$, $\theta_3 = 0.3$, $\theta_4 = 0.95$ for all examples. Source code for our method can be found at https://github.com/posgraph/coupe.Sptiotemporal-Texture/.

<table>
<thead>
<tr>
<th>Scene</th>
<th># Frame</th>
<th>Tex Opt (min)</th>
<th>Labeling (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>265</td>
<td>11.0</td>
<td>23.0</td>
</tr>
<tr>
<td>2</td>
<td>210</td>
<td>9.3</td>
<td>17.9</td>
</tr>
<tr>
<td>3</td>
<td>372</td>
<td>8.8</td>
<td>20.0</td>
</tr>
<tr>
<td>4</td>
<td>247</td>
<td>5.6</td>
<td>23.9</td>
</tr>
<tr>
<td>5</td>
<td>484</td>
<td>19.2</td>
<td>56.0</td>
</tr>
<tr>
<td>6</td>
<td>210</td>
<td>8.6</td>
<td>25.1</td>
</tr>
</tbody>
</table>

Table 1: Computation times. ‘Tex Opt’ means global texture coordinate optimization. The input images of scenes 3 and 4 are of lower resolutions than those of other scenes.

Timing The global texture coordinate optimization takes about 9 ~ 20 minutes for a dynamic object. Solving MRF on average takes about 5 ~ 7 seconds per frame. The color correction takes about 20 seconds per frame, but this step can be parallelized. Timing details are shown in Table 1.
Dynamic object reconstruction
We reconstruct a template model by rotating a color camera around the motionless object and using Agisoft Metashape† for 3D reconstruction. Then we apply mesh simplification [HCC06] to the reconstructed template model and reduce the number of triangle faces to about 10~30k. To obtain a deformation sequence of the template mesh $\mathcal{M}$ that matches the input depth stream, we adopt a non-rigid registration scheme based on an embedded deformation (ED) graph [SSP07]. The scheme parameterizes a non-rigid deformation field using a set of 3D affine transforms. It estimates the deformation field to fit mesh $\mathcal{M}$ into an input depth map $D_t$ at time $t$ using an $l_0$ motion regularizer [GXW15].

6.2. Analysis
Texture coordinate optimization
To accelerate the texture coordinate optimization process, we regularly sample frames and interpolate the texture coordinate displacements computed from the samples. Table 2 shows that our approach achieves almost the same accuracy compared to the original one using all frames. Using fewer sample frames reduces more computation, and we sample every four frames for the time-accuracy tradeoff.

<table>
<thead>
<tr>
<th>sampling</th>
<th>1/2</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>error ratio</td>
<td>1.0063</td>
<td>1.02220</td>
<td>1.02309</td>
<td>1.0427</td>
</tr>
</tbody>
</table>

Table 2: Error ratio of our modified texture coordinate optimization depending on the sampling ratio, where $1/n$ means sampling every $n$ frames. We used the value of Eq. (1) to measure the error of a texture coordinate optimization result. The error ratio was computed by dividing the error of our sampling approach by that of the original approach, taking the average of all datasets.

Global & local geometric similarity
Our method utilizes global and local geometric similarities to assign the most reasonable frame labels to the faces of the dynamic template mesh. In Figure 7a, back parts that do not show obvious geometric patterns are not distinguishable by local similarity, and global similarity helps to assign proper frame labels. On the other hand, as shown in Figure 7b, local similarity helps to find appropriate textures if local geometry changes should be considered to borrow natural appearances from other frames.

Selective weighting
Figure 8 shows that spatial selective weighting can help preventing abnormal seams. Figure 6 compares different temporal weighting schemes. Our temporal selective weighting makes the mouth motion consistent (Figure 6a), and incurs more dynamism on the back part by dropping the smoothness constraint (Figure 6b).

6.3. Quantitative Evaluation
Coverage
We computed the amount of textures generated by our method. Naive projection can cover 29.5% ~ 39.6% of the mesh.
6.4. Qualitative Comparison

In Figure 14a, we compare our results with a texture-mapped model using single images. The single image-based texture has black regions since some parts are invisible at a single frame, while our texture map covers the whole surface of the object. Compared to global texture map [KKPL19], our spatiotemporal texture is more appropriate to convey temporal contents (Figure 14b). In particular, wrinkles on the clothes are expressed better. Additional results are shown in Figures 15 and 16.

6.5. User Study

For a spatiotemporal texture reconstructed using our method, texture quality of originally visible regions can be measured by comparing with the corresponding regions in the input images, as we did in Section 6.3. However, for regions invisible in the input images, there are no clear ground truths. Even if an additional camera is set up to capture those invisible regions, the captured textures are not necessarily the only solutions, as various textures can be applied to a single shape, e.g., eyes may be open or closed with the same head posture.

To evaluate the quality of reconstructed textures for originally invisible regions without ground truths, we conducted a user study using Amazon Mechanical Turk (AMT). We generated a set of triple images using input images, dynamic model rendered with spatiotemporal texture, and dynamic model rendered with global texture [KKPL19]. We picked the triple images to be as similar as possible to each other, while removing the backgrounds of input images. Figure 10a shows an example. We prepared 40 image triples (120 images) and asked to evaluate whether the given images are realistic or not in two ways: One task is to choose a more realistic image among a pair of images. The other is to score how realistic the given images are on a scale of 1-10. We set up 25 workers whose hit approval rates are over 98% to answer a single question. As a result, we collected total 3,000 pairwise comparison results and 3,000 scoring results.

The preference matrix in Table 4 summarizing the pairwise comparison results shows that the input images are clearly preferred to rendered images using reconstructed textures. This result would be inevitable as geometric reconstruction errors have introduced unnaturalness on top of texture reconstruction artifacts. The errors are mainly caused by the template deformation method, which may not completely reconstruct geometric details. For example, in Figure 10a, the sleeve of T-shirt is tightly attached to the arm in the reconstructed mesh. Still, Table 4 shows our spatiotemporal textures dominate global textures, where only texture reconstruction qual-

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Methods</th>
<th>Scene 1</th>
<th>Scene 2</th>
<th>Scene 3</th>
<th>Scene 4</th>
<th>Scene 5</th>
<th>Scene 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ours</td>
<td>28.0900</td>
<td>30.6023</td>
<td>28.2425</td>
<td>28.0716</td>
<td>23.4724</td>
<td>27.3731</td>
</tr>
<tr>
<td></td>
<td>Ours (1/3)</td>
<td>28.8495</td>
<td>29.9831</td>
<td>28.4498</td>
<td>27.6200</td>
<td>23.1869</td>
<td>27.2290</td>
</tr>
<tr>
<td></td>
<td>Ours (1/2)</td>
<td>28.8779</td>
<td>30.0269</td>
<td>28.5851</td>
<td>27.8902</td>
<td>23.2763</td>
<td>27.2566</td>
</tr>
<tr>
<td>Blurriness</td>
<td>Input images</td>
<td>0.7937</td>
<td>0.8004</td>
<td>0.8096</td>
<td>0.9095</td>
<td>0.8220</td>
<td>0.8440</td>
</tr>
<tr>
<td></td>
<td>Kim et al. [KKPL19]</td>
<td>0.7971</td>
<td>0.8006</td>
<td>0.8865</td>
<td>0.9101</td>
<td>0.8384</td>
<td>0.8514</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.7896</td>
<td>0.7912</td>
<td>0.8282</td>
<td>0.9065</td>
<td>0.8104</td>
<td>0.8425</td>
</tr>
</tbody>
</table>

Table 3: Texture reconstruction errors and blurriness. PSNR and blurriness were measured excluding the background. For blurriness, smaller means sharper. In the top rows, average PSNR was computed using all frames. In the middle rows, 1/n in parentheses means sampling every n frames, and average PSNR was computing using 1/6 sample frames that were not involved in texture reconstruction.

Figure 8: Without (a) and with (b) spatial selective weighting. With spatial selective weighing, coherent textures are obtained in regions where textures may change dynamically.
Figure 9: Texture reconstruction quality. (a) the bluriness of input and reconstructed textures as well as average vertex motions of frames. (b) (c) input (top) and rendered (bottom) images. The reconstructed texture is apparent in (b), although the input is blurry due to fast motion. In (c), the reconstructed texture preserves the sharpness of the input.

Figure 10: User study. The image triple in (a) consists of an input image with background removed, a rendered dynamic mesh using spatiotemporal texture, a rendered dynamic mesh using global texture [KKPL19], from left to right. In (b), from left to right, the averages are 6.798, 6.053, and 5.084, and the standard deviations are 2.4197, 2.3138, and 2.3653.

Table 4: Preference matrix from the user study. The value of each row refers to the number of times the image on that row is preferred when compared with the image on a particular column.

<table>
<thead>
<tr>
<th></th>
<th>Raw image</th>
<th>[KKPL19]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>[KKPL19]</td>
<td>137</td>
<td>-</td>
<td>190</td>
</tr>
<tr>
<td>Ours</td>
<td>182</td>
<td>810</td>
<td>-</td>
</tr>
</tbody>
</table>

6.6. Comparison with Learning-based Methods

As mentioned in Section 2, there are learning-based methods to reconstruct textures of objects. However, there are important differences between our method and those methods. Firstly, our method uses a template mesh and an image sequence from a single camera as the input, whereas the learning-based methods use a single image [SHN*19, PTY*19] or multi-view images [SHN*19]. Although it is unfair, we tested the authors’ implementation of [SHN*19] on single images of our dataset (Figure 11). The generated textures on the originally invisible areas are clearly worse than our results. Note that multi-view images that can be used in [SHN*19] should be captured by multiple cameras at the same time. Then the setting differs from our case of a single camera, making comparison using multi-view images inappropriate. Secondly, due to memory limitation, the networks of [SHN*19, PTY*19] cannot fully utilize the textures of our 4K dataset. This would lead to poor quality compared to our results. Thirdly, the learning-based methods target only humans, while our method does not have the limitation, as demonstrated in the top two examples of Figure 15. Finally, our method reconstructs a completed texture volume that can be readily used for generating consistent novel views in a traditional rendering pipeline, but synthesized novel views from learning-based methods are not guaranteed to be consistent.

6.7. Discussions and Limitations

Rich textures Our method would work for rich textures as far as our texture coordinate optimization works properly. Texture coordinate optimization may fail in an extreme case. For example, when the texture is densely repeated, the misalignment in geometric reconstruction could be larger than the texture repetition period. Then, the texture correspondence among different frames could be wrongly optimized.

Computation time Our approach takes quite some time and it would be hard to become real-time. On the other hand, texture reconstruction for a model is needed to be done only once, and our method produces high-quality results that can be readily used in a conventional rendering pipeline.
In Figure 12, a directional light moves continuously in the input images (see the movement of shadow). However, the reconstructed textures in (e) on invisible regions show sudden brightness changes.

Figure 12: Variant lighting example. A directional light moves continuously in the input images (see the movement of shadow). However, the reconstructed textures in (e) on invisible regions show sudden brightness changes.

Dependency on geometry processing If non-rigid registration of dynamic geometry fails completely, our texture reconstruction fails accordingly. However, the main focus of this paper is to reconstruct high-quality textures, and our method would be benefited from the advances of single-view non-rigid registration techniques. In our approach, texture coordinate optimization helps handle possible misalignments from non-rigid registration of various motions. For fast motions, as shown in Figure 9, our labeling data term could filter out blurry textures, replacing them with sharper ones.

Fixed topology As template-based non-rigid models have been steadily researched until recently [LYLG18, YDXZ20], our method assumes that each model in the motion sequence has the same fixed topology. As a result, our method may not handle extreme shape changes, such as taking off clothes or looser clothes.

Variant lights and relighting Our method may not produce plausible texture brightness changes in a variant lighting environment. In Figure 12, a directional light is continuously moving when the input images are captured. However, the reconstructed textures on invisible regions show sudden brightness changes at some frames. As a result, sometimes temporal flickering can be found on original invisible regions. Another limitation of our method is that lighting is not separated from textures but included in the restored textures. This limitation would hinder relighting of the textured objects, but could be resolved by performing intrinsic decomposition on the input images.

Unseen textures Our method cannot generate textures that are not contained in the input images. Unnatural textures could be produced for regions invisible at all frames (Figure 13).

Figure 13: Failure case of an unseen region. The red boxed region is never visible from any input color image, and the recovered texture is unnatural.

There are several ways to improve our approach. First, to conduct the labeling process more efficiently, considering a specific period like an approach by Liao [LH13] would be a viable option. Second, our framework separates the 3D registration part and texture map generation part. We expect better results if the geometry and texture are optimized jointly [FYLX20].

Acknowledgements

This work was supported by the Ministry of Science and ICT, Korea, through IITP grant (IITP-2015-0-00174), NRF grant (NRF-2017M3C4A7066317), and Artificial Intelligence Graduate School Program (POSTECH, 2019-0-01906).

References


Figure 14: Qualitative comparison. (a) (left) When single views are used for dynamic texture mapping, there are blank regions since the object is not fully observable at a single view. (right) Our approach reconstructs textures on the whole area since it produces a spatiotemporal volume containing every frame. Besides, our method produces sharp textures, even if there is a blurred part in an input image. (b) rendered images of a dynamic model using global texture map [KKPL19] (left) and our spatiotemporal texture (right). Our spatiotemporal texture reproduces wrinkles with reasonable directions.

Figure 15: Additional results. For each example, the top row shows the results from the global texture mapping based approach. The bottom row shows our results. Note that every rendered image was captured from the opposite side of the actual camera’s position in the frame.
Figure 16: Additional results. (left) snapshots of texture mapped geometry. (right) rear and front views of dynamic geometries at two timestamps. Each image trio shows the underlying geometry, global texture map-based result, and our result.
Kim et al. / Spatiotemporal Texture Reconstruction for Dynamic Objects Using a Single RGB-D Camera


