Separating Texture and Illumination for Single-Shot Structured Light Reconstruction

Minh Vo
mvpo@cs.cmu.edu

Srinivasa G. Narasimhan
srinivas@cs.cmu.edu

Yaser Sheikh
yaser@cs.cmu.edu

Robotics Institute, Carnegie Mellon University, USA

Abstract

Active illumination based methods have a trade-off between acquisition time and resolution of the estimated 3D shapes. Multi-shot approaches can generate dense reconstructions but require stationary scenes. In contrast, single-shot methods are applicable to dynamic objects but can only estimate sparse reconstructions and are sensitive to surface texture. In this work, we develop a single-shot approach to produce dense reconstructions of highly textured objects. The key to our approach is an image decomposition scheme that can recover the illumination and the texture images from their mixed appearance. Despite the complex appearances of the illuminated textured regions, our method can accurately compute per pixel warps from the illumination pattern and the texture template to the observed image. The texture template is obtained by interleaving the projection sequence with an all-white pattern. Our estimated warping functions are reliable even with infrequent interleaved projection. Thus, we obtain detailed shape reconstruction and dense motion tracking of the textured surfaces. We validate the approach on synthetic and real data containing subtle non-rigid surface deformations.

1. Introduction

Structured light based shape reconstruction algorithms are divided into two categories: multi-shot [4, 6, 15] and single-shot methods [7, 14, 17]. Multi-shot methods can estimate per-pixel depth map for a wide range of objects using temporal coding of the illumination patterns but require the scene to be stationary during image acquisition. Single-shot methods can work for dynamic objects by decoding the spatial structure embedded in the illumination pattern but generate low spatial resolution reconstructions and are susceptible to the high frequency texture on the object surface (see Figure 1 for examples).

In this work, we present a single-shot structured light system that can estimate high spatial and temporal resolution depth even for highly textured objects. The proposed method is single-shot in the sense that it does not use the illuminated images in the previous frames to temporally decode the illumination pattern. Our system consists of one camera and one projector. The projector generates high frequency illumination patterns needed to estimate dense 3D shape. We decompose the observed mixed appearance images into the surface texture and the projected illumination pattern. Because of this decomposition, high resolution 3D shape of the object in the textured regions can be reliably estimated. In addition, we obtain dense tracking inside the texture region in the presence of the illumination pattern.

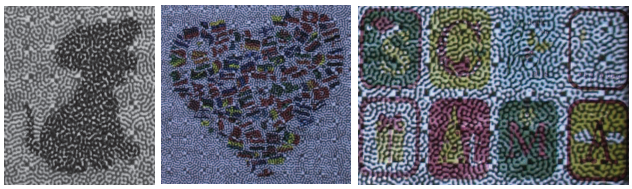

Figure 1. Conventional single-shot structured light systems often fail for highly textured objects. The mixture of albedo variations and high frequency illumination patterns makes it difficult to establish reliable and dense camera-projector correspondences.

To achieve the illumination-texture separation, we develop an optimization framework that estimates warps of both the illumination pattern and a reference texture template to compose the observed image. The texture template is obtained by infrequently interleaving the projection sequence with an all-white pattern. The warping functions are computed starting from a sparse set of correspondences between the camera and the projector. The results are greedily propagated into textured areas where spatial correspondences cannot be directly estimated. Figure 1 shows several types of surface texture that our method can handle.

Since the method computes warps to a template, it does not exhibit drift over time. Moreover, unlike conventional single-shot structured light systems whose performances degrade as surface texture frequency increases, our method

http://www.cs.cmu.edu/ ILIM/projects/IL/TextIllumSep

achieves better decomposition accuracy with higher frequency texture. Finally, despite being presented for single-shot approaches, this method can also be used in conjunction with multi-shot systems by spatially modulating the illumination patterns with a random pattern [18]. We demonstrate dense and accurate decomposition and reconstruction results on both synthetic and real data with non-rigidly deforming objects.

2. Related Work

Active illumination has been used to estimate shape estimation problem with major focus on designing coded patterns that are robust to occlusion, depth discontinuity, and albedo variations [11]. A conventional structured light system has to choose appropriate illumination patterns depending on its temporal or spatial resolution requirements. Since multi-shot methods can robustly generate high spatial resolution but require stationary scenes, motion compensation schemes have been developed to handle slowly moving objects [16]. Another common approach is to interleave the patterns for structure estimation with patterns optimized for computing motion [8].

On the other hand, single-shot methods sacrifice spatial resolution for high temporal resolution reconstruction. However, because of the spatial coding strategy, generally these methods cannot deal with textured objects and they are forced to rely on specific light patterns for different types of surface texture. Koninckx et al. [9] mitigate the problem by introducing a feedback loop that changes patterns according to the error in the decoding process. Yet, they rely on heuristic rules to set the color codes depending on different textured surfaces. In principle, this method treats the surface texture as a nuisance and designs illumination patterns robust to the texture. Conversely, our method considers the texture as an additional source of information that needs to be recovered along with the 3D shapes. To the best of our knowledge, this is the first work to explicitly separate high frequency texture and illumination patterns in the context of structured light system.

Another solution for single-shot methods to handle object with both textured and textureless regions is to employ multi-view stereo systems. These systems treat the illumination pattern as surface texture to assist the matching [19]. However, since this method requires at least two cameras in addition to a projector in between, baseline between cameras is larger which makes correspondence estimation hard.

Our image decomposition method is similar in spirit to intrinsic image estimation [5, 12]. These works make smoothness assumptions about the environment light (the sun, the sky, or indoor lights) and estimate the reflectance and shading images from images captured by a single camera. Conversely, our work decomposes the high frequency illumination patterns and texture from the observed mixed appearance image. Additionally, our method is specifically designed for single-shot structured light systems that consist of one camera and one projector.

3. Texture-Illumination Decomposition

Consider an object being illuminated by a projector. Similar to intrinsic image estimation, the brightness $I(x, y)$ at location (x, y) in the observed image is modeled as a multiplication of a texture image $I_T(x, y)$ and the illumination image $I_L(x, y)$ at that location:

$$I(x, y) = I_T(x, y)I_L(x, y). \tag{1}$$

The texture image I_T is the image observed if the projector illuminates an all-white pattern on the object. The illumination image I_L is the incident lighting pattern. Equation 1 is ill-posed because it has two unknowns $I_T(x, y)$ and $I_L(x, y)$ in one equation. To handle this under-constrained problem, we require additional knowledge of the reference templates for the illumination and texture source and a proper per-pixel initialization of the unknowns. Since the illumination image is a projection of the known projector pattern, this pattern serves as one of our references. The reference texture template can be obtained by interleaving the projection sequence with a white pattern. The initialization problem is solved by the greedy correspondence growing algorithm [3]. We describe the method in detail below.

3.1. Mathematical Formulation of the Objective

Figure 2 shows a sequence of images of an object being illuminated by the projector. Initially, the projector illuminates the scene with an all-white pattern so that the texture template T is observed. The appearance of this template I_T changes according to the movement of the object. Because image deformation is high-dimensional and non-linear, analytic forms that describe consistent deformation behavior over the entire image do not exist. Thus, we locally model this distortion by a warping function f and employ constant gain a_T and offset b_T to approximate for the intensity changes between the two images due to changes in surface normals, light directions and ambient illumination:

$$I_T(x, y) = a_T T(f(x, y)) + b_T, \tag{2}$$

where $f(x, y)$ maps the coordinate of point (x, y) in the observed image to its corresponding location in the texture template T. Similarly, the illumination image I_L is the projection of light pattern L on the object and hence, is related to each other by a set of local warping functions g. We assume the projector has been photometrically calibrated and adopt a linear model to relate the brightness of the pure illumination image to the projecting pattern:

$$I_L(x, y) = a_L L(g(x, y)) + b_L, \tag{3}$$

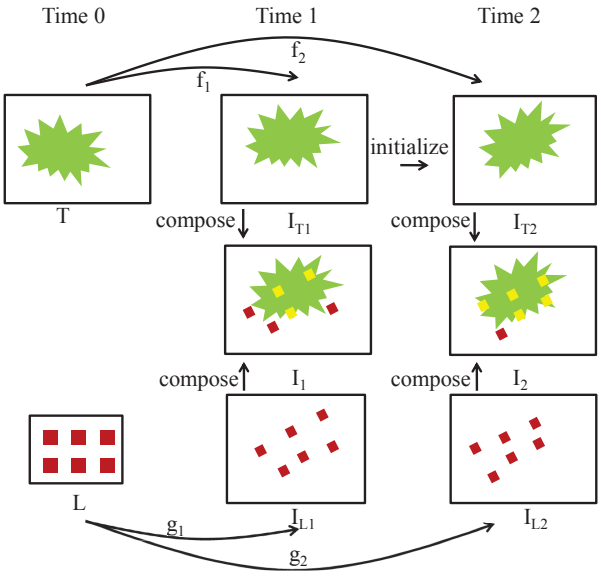

Figure 2. Texture and illumination image decomposition. The green texture region in the template T is warped to the pure texture image I_{Ti} by the function f_i. The red squares from the illumination pattern L are mapped to the pure illumination image I_{Li} by the function g_i. Properly composing the illuminated red squares in I_{Li} with the green region in I_{Ti} turns them into yellow squares, which constitute the mixture image I_i observed at time i. Notice that spatially adjacent squares are likely to have similar warping parameters. By using the estimated warping parameters of the previous frame to initialize warps between the observed frame to its reference templates L and T, the warping function can be reliably estimated even with infrequent interleaving.

where $g(x, y)$ relates the point (x, y) to its matched in the illumination image L and a_L, b_L are constant gain and offset to compensate for the brightness difference between the projector pattern and the observed illumination.

To increase the robustness of the warping functions to large deformation, we set them to be locally affine and minimize the following cost function for their shape parameters and the photometric compensation coefficients:

$$\sum_{k=-N}^{N} \sum_{l=-N}^{N} [I_T(x_k, y_l) I_L(x_k, y_l) - I(x_k, y_l)]^2, \quad (4)$$

over a patch of size $(2N + 1) \times (2N + 1)$ centered at point (x_0, y_0), where we want to decompose. These warping functions are given as:

$$f(x_k, y_l) = \begin{bmatrix} x_0 + p_0 + (p_2 + 1)k + p_3 l \\ y_0 + p_1 + p_4 k + (p_5 + 1)l \end{bmatrix},$$
$$g(x_k, y_l) = \begin{bmatrix} x_0 + q_0 d_x + (q_1 + 1)k + q_2 l \\ y_0 + q_0 d_y + q_3 k + (q_4 + 1)l \end{bmatrix}, \quad (5)$$

where (d_x, d_y) is the normalized vector representing the direction of the epipolar line in the projector image, and $q_{0..4}, p_{0..5}$ are the affine warp parameters. Similar to stereo matching, we simplify the parameterization of g by constraining it to lie on the camera-projector epipolar line. Equation 4 is optimized at every patch in the textured regions using the Gauss-Newton method.

3.2. Pixel-wise Initialization

Due to perspective distortion in the illumination image I_L and the large deviation from the texture template T for fast moving objects, good initial guesses for the warping parameters are required to optimize equation 4. Even with random patterns [4], the repetitive nature of the spatial neighbor coding illumination pattern exacerbates the local minimum problem. We solve the initialization problem as follows: start from the texture boundary regions and greedily propagate the results to the interior texture area. We employ the three steps to initialize points at the boundary.

Step 1: Compute dense matching between camera and projector using a greedy correspondence growing algorithm [3]. This greedy growing strategy and the use of a random illumination pattern allows us to establish dense correspondences everywhere except for the textured surface regions. Here, the texture boundary is naturally defined as places where spatial correspondences are not obtainable.

Step 2: For a pixel that is close to the textured region boundary, we exhaustively search in its local neighbor for patches on the illumination pattern and texture template that minimizes the cost defined in Equation 4. Depending on the motion of the objects, the search range of pixels close to the texture boundary is set a priory. Because of the deformation between patches in the templates and the ones in the pure texture and illumination images, patches in the templates are pre-warped before being used to examine their contribution to the cost function. The warping parameters of the illumination pattern are initialized from its spatial neighbor computed in step 1. Those of the texture templates are set to either its temporal neighbor if available or to the zero deformation state.

Step 3: Refine the best locations of the illumination and texture patches by optimizing the cost function in Equation 4 using a standard Gauss-Newton method.

Owing to the use of the texture warping parameters estimated in the previous frame as initialization, our warping functions can robustly warp observed patch to its reference template that is temporally far away. Hence, only infrequent projection of the interleaving white frame is needed and the obtained results have high temporal resolution. This strategy is analogous to the approach of Tian and Narasimhan et al. [13] who use less distorted patches to estimate globally optimal sets of warping parameters.

In spite of the greedy correspondence growing strategy, erroneous guesses can not propagate long as examined and thresholded by the cost function in Equation 4. Hence, our

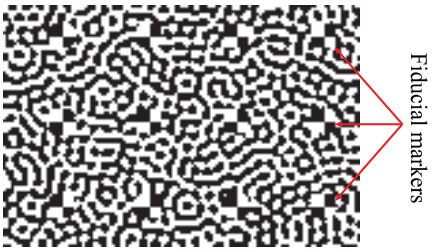

Figure 3. Part of the illumination pattern. The fiducial markers are embedded into the random pattern to provide a sparse set of correspondences between the camera and the projector.

method avoid both the global ambiguity of the illumination pattern and being stuck in regions where occlusion, surface discontinuity, or severe foreshortening occurs. Since only a few seed points are needed initially, the good correspondences can be quickly propagated to non-decomposable regions in the earlier frames.

4. Results

We validate the performance of our approach on both synthetic and real cloth sequences containing a range of texture frequencies. The non-rigidity of cloth makes dense decomposition and shape reconstruction challenging. We show the results for the sequences in which the all-white pattern is projected once in 30 frames. For all of our experiments, this interleaving interval gives good trade-off between the temporal resolution of the results and the speeds of the moving objects. We also fix the patch size to be 19×19. The 3D shapes are estimated by triangulating the correspondences obtained after the decomposition. We split the region of interest into sub-regions and independently execute them in parallel to take advantage of the multi-core architecture of modern computer. Currently, our algorithm can decompose on average 4112 points every second on a Quad-core i7 CPU (3.6 GHz).

4.1. Illumination pattern

Figure 3 shows our static bandpass random binary illumination pattern [4]. The size of the speckle in this pattern can be tuned to provide suitable contrast for illuminating objects of different size. Fiducial checkerboard markers are uniformly seeded at every 32 pixels inside this pattern to provide set of sparse spatial correspondences. These correspondences are computed using template matching along epipolar lines. Because the distance between these markers is usually magnified in the camera image, these markers do not cause ambiguities in the propagation process.

4.2. Synthetic Data

Our synthetic cloth composed of 64,000 vertices is generated using the OpenCloth engine [10] and can deform in subtle and non-rigid ways. The camera and projector resolution are set to 1920×1080 and 1280×800, respectively.

Figure 4 shows the decomposition result on the synthetic bear cloth sequence. We intentionally have the texture of the bears to be very similar to the illumination pattern. This extreme case severely violates the assumptions of any methods powered by independent component analysis or smoothness prior. Thus, such methods are not applicable. Conversely, our method recovers the pure texture and illumination images that well resemble the groundtruth. The small noticeable defect in frame 2 does not expand and is fixed in frame 18. Hence, there is little-to-no drift in the estimated warping functions.

Figure 5 shows the decomposed images, the 3D shape as well as the flow from the observed image to the reference template obtained from texture warping functions. Without our separation, the correlation score between the projector patch and its corresponding patch in the observed image is very low in the textured regions. The depths estimated in these regions are eliminated, which results in large holes in the 3D shape. By separating the pure texture and illumination images, the proposed method successfully estimates the 3D shape with the exceptions of places where severe foreshortening occurs.

To quantify the decomposition error, we compare the correspondences obtained after the decomposition with those estimated on the synthesized pure illumination and texture images. The pure illumination image is created by projecting the illumination pattern onto a textureless cloth. Texture image is obtained by texturemapping the cloth with binarized Perlin-noise pattern generated by error-diffusion dithering and projecting white light on the cloth. This binarized random pattern is known to give accurate tracking results [1, 8]. We compute the spatial correspondences between camera and projector on the pure illumination image and the temporal correspondences on the pure texture image using patch-based tracking methods [2]. These correspondences are the best possible estimation at a given frame and hence serve as ground truth.

We define our error metric as the normalized error in decomposing texture and illumination images:

$$\frac{1}{N}\sqrt{\left(\frac{\widehat{x_i}-x_i}{W}\right)^2+\left(\frac{\widehat{y_i}-y_i}{H}\right)^2}, \tag{6}$$

where N is the number of points inside the texture regions, $(\widehat{x_i},\widehat{y_i})$, and (x_i,y_i) are the ground truth and the estimated correspondence locations, respectively. Depending on whether the error of the illumination or texture is being evaluated, (W,H) could be either the resolution of the projector or camera image.

Figure 6 shows the performance of our method on the bear and flowery cloth sequences. The bear cloth has higher frequency texture than the flowery cloth. With longer in-

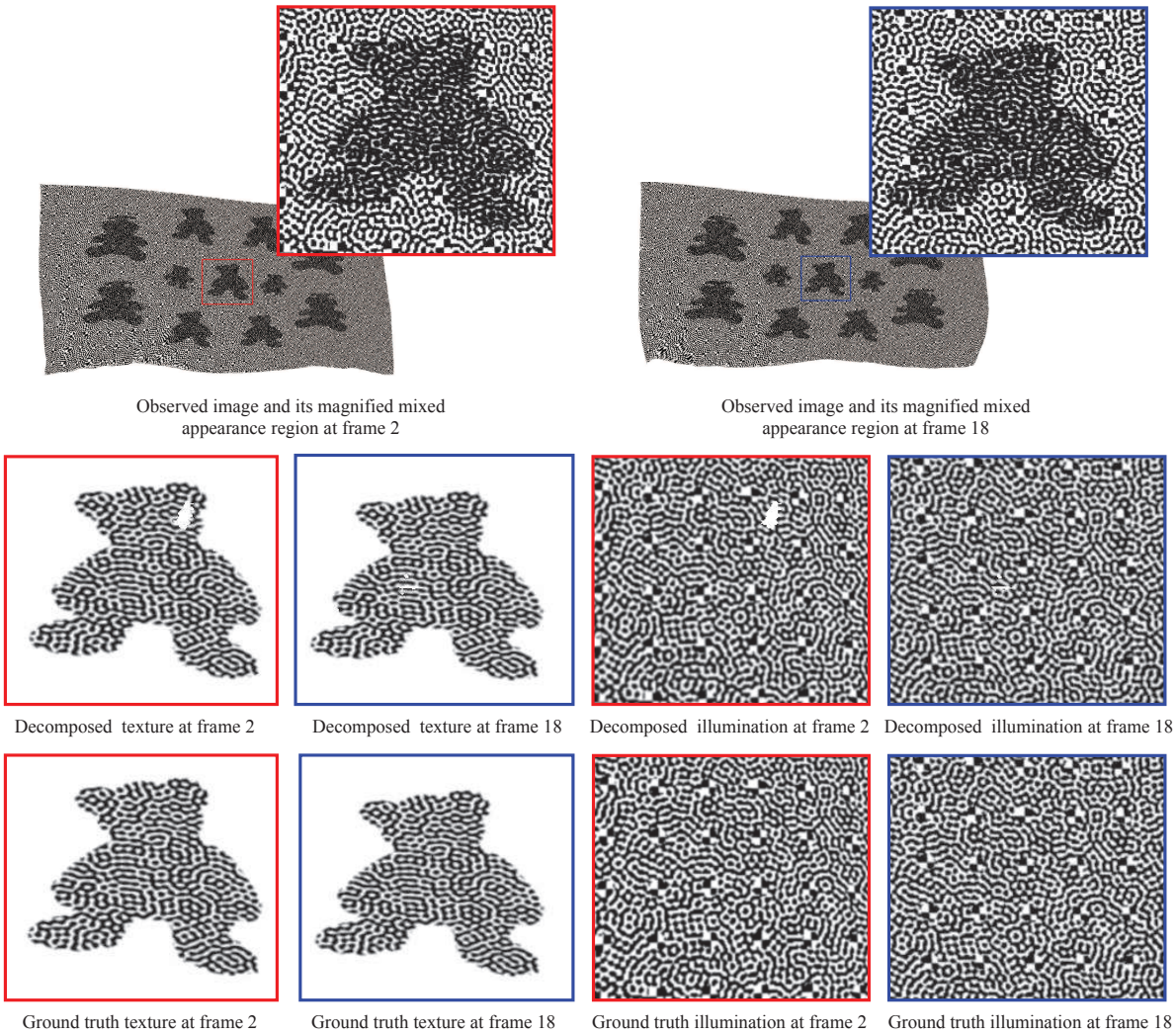

Figure 4. Decomposition of the texture and illumination images of the synthetic bear cloth sequences. The interleaving pattern is projected every 30 frames. The same region of the bear in the two images is magnified to show how its intricate appearance changes over time when it is illuminated by the projector. Because of cloth folding, bear appears smaller in the vertical direction. Visually, the illumination images experiences higher deformation than the texture images. The noticeable defect in the decomposition at frame 2 has been fixed in frame 18.

terleaving sequences, larger deformation in the image degrades the algorithm performance. Yet, no explosion in the decomposed texture and illumination error is observed. The texture decomposition error and the fraction of correspondences estimated for flowery sequence are not as good as for the bear sequence. This indicates that the algorithm performs better with higher frequency texture. The explanation for this phenomenon is similar to the optical flow problem: tracking highly textured surfaces suffers from less drift. Notice that because of the coherent structure of the bear, the algorithm fails when the entire texture boundary cannot be estimated reliably. This results in the abrupt drop in the fraction of estimated correspondences. Nevertheless, the initial illumination decomposition error for the flowery sequence remains competitive to the bear sequence until influence of the decomposed texture error overwhelms the overall result.

4.3. Real Data

We conduct several real experiments with different cloth deformation and different texture frequencies. For all of our experiments, the scenes are illuminated using a 1280×800 DLP View Sonic projector and the images are acquired by the Canon XH-G1s HD 1920×1080 camera operating at 30fps. The camera and projector are calibrated using the method of Vo et al [15]. As in the synthetic dataset, we project an all-white frame every 30 frames. The results are presented without any post-processing.

Figure 7 shows our results on the dogs and flag sequences. Despite the simple image formation model, our approach can handle complex appearances of real world

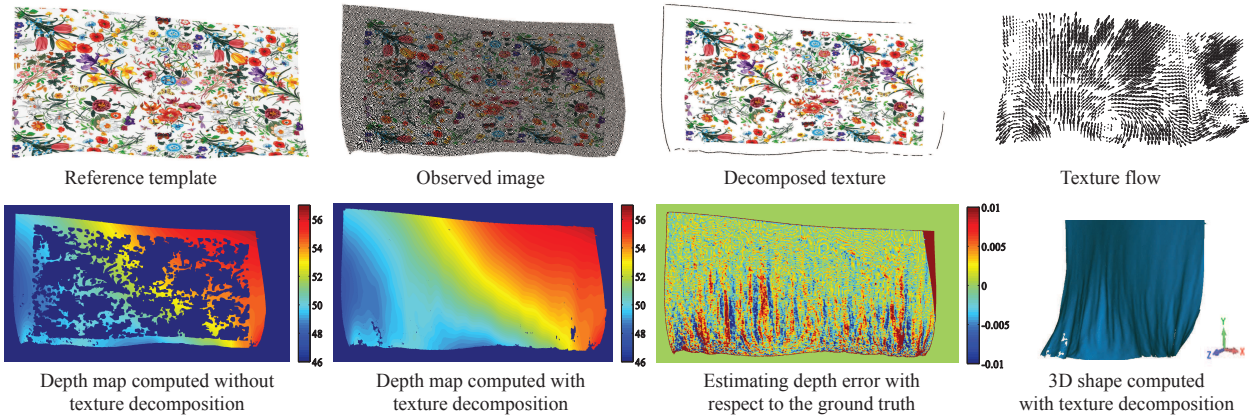

Figure 5. Decomposition of the texture and illumination images from a synthetic flowery cloth sequence with interleaving pattern projected every 30 frames. The depth map estimated after removing the texture not only shows its completeness over depth map obtained without texture removal but its high accuracy with respect to ground truth depth. The displacement of the texture regions with respect to the reference template is also estimated from the decomposition results.

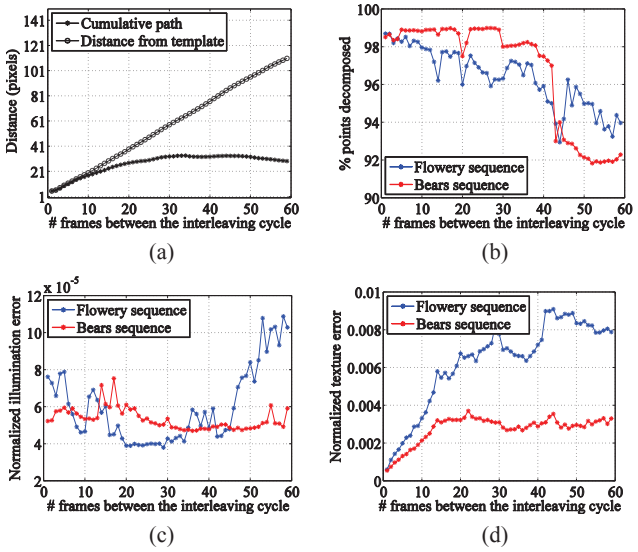

Figure 6. The accuracy and robustness of the decomposition with respect to the interleaving period. **(a)** The median distance from the current frame to the reference template and the accumulative median displacement over frames **(b)** Percentage of camera-projector correspondences obtained in the textured regions **(c)** Normalized illumination error **(d)** Normalized texture error. The error is evaluated only for points inside the textured regions and is computed by Equation 6.

textured objects illuminated by the projector. We believe this is due to the local block decomposition strategy which is robust to global lighting variation. As shown in the flag sequence, while the texture decomposition could be incomplete, especially for low frequency textured objects, the quality of the decomposed illumination is less affected by the texture frequency. Due to the relatively large size of the textured region, the 3D shape of this folding cloth cannot

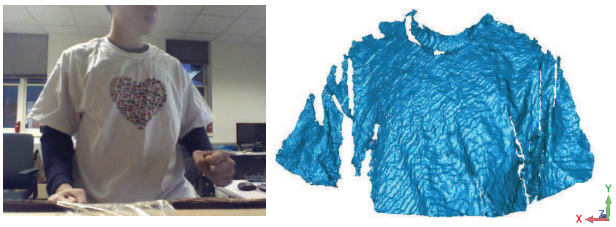

Figure 8. 3D shape from the Kinect. Note that smoothing filters has been applied to generate the mesh from raw point cloud data.

be obtained by hole-filling or interpolation algorithms. Despite the mixed appearance, the texture flow estimated with respect to the reference template is also obtained.

Figure 8 shows the 3D shape obtained from the Kinect sensor. For fair comparison with the performance of the proposed method shown in the T-shirt sequence (see Figure 7), the same subject is standing at a similar distance to the sensor. Visually, the quality 3D shape from our method outperforms that of the Kinect. It is noteworthy that unless smoothing filter is applied to raw Kinect results, the mesh generation fails as the surface normal computed from raw point cloud is noisy.

4.4. Texture Flow vs. Illumination Flow

Besides the motion of the textured regions, there is also an apparent motion due to the illumination pattern. As illustrated in Figure 9, the flow direction of the illumination pattern and the texture are remarkably different from each other. Unlike the motion flow which presents the movement of points on the object surface, the observed illumination is the projection of the ray emanating from the light source and hence, its flow field must move only on the epipolar line of the camera-projector system. Furthermore, this

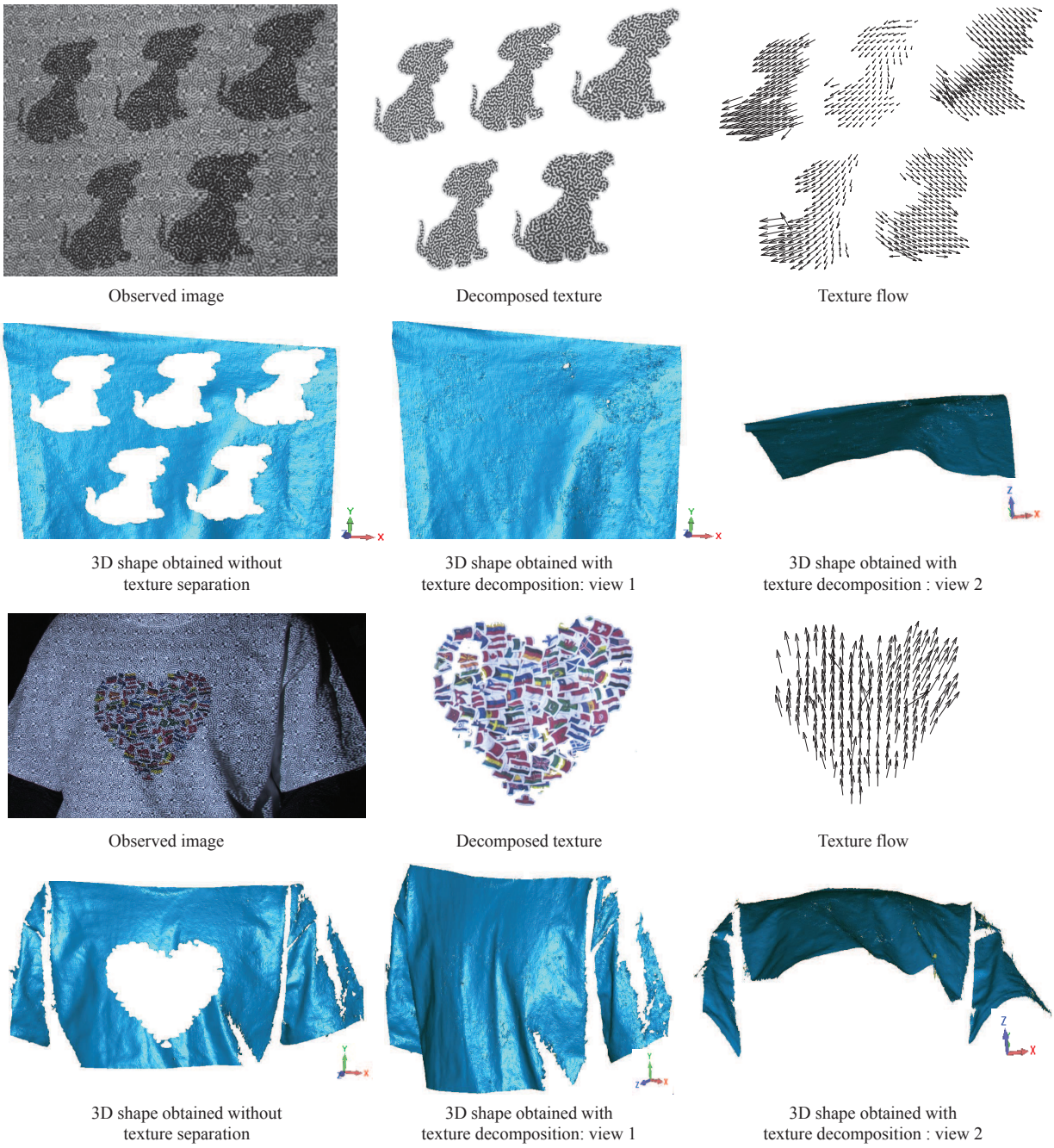

Figure 7. Decomposition of the texture and illumination images from real cloth sequences. The interleaving pattern is projected every 30 frames. After the decomposing process, the 3D shape can be obtained in the textured regions. No post-processing is applied. Applying hole-filling methods on the 3D shape estimated without texture decomposing cannot yield appealing results because of the relatively large textured regions. Despite the mixed appearance, the texture flow faithfully shows how the cloth is moving.

flow field encodes the changes in depths of rays emanating from the light source that hit the object. Because establishing spatial correspondences between camera and projector is much more difficult than estimating temporal correspondences, especially in the wide baseline scenario, any structured light system can gain benefit from the temporal coherency of the illumination flow. Future work will investigate different approaches to incorporate illumination flow constraint into structured light reconstruction algorithms.

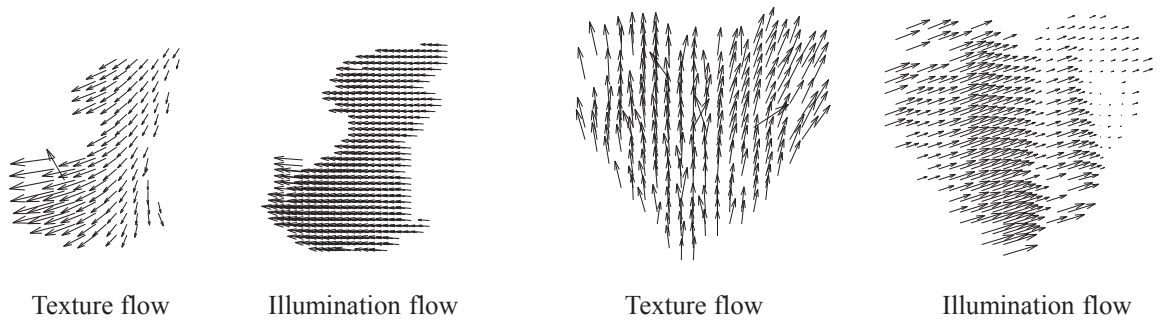

Figure 9. Observed motion flow and illumination flow when a moving textured object is illuminated by point light source. While the directions of texture flow follow the movement of the object, the directions of the illumination flow are constrained to be on the camera-projector epipolar line. The figure is best viewed in the electronic version.

5. Discussion

While we only show the results for a single deforming object, our algorithm is applicable to general scenes containing multiple objects. As long as the seed points, i.e. correspondences established in low frequency textured regions, are available on the object, these correspondences can propagate to the entire object. Nevertheless, because of the nature of the patch decomposition approach, our algorithm cannot handle well the textured regions at the occluding boundary.

For objects with completely high frequency texture, the proposed algorithm will fail. Yet, such cases are rare and most objects have a mixture of low and high frequency texture regions (see Figure 5). While our approach may also fail for objects with low frequency texture, the camera-projector correspondence can be established to reconstruct 3D shape. Moreoever, these correspondences can also be exploited to separate the texture out of the observed image.

Since the interleaving sequence is dependent on the motion of the object, the interleaving period has to be adapted to different applications. Nevertheless, in an era 60-fps consumer grade camera, there is no need to interleave every frame, especially for daily human activity.

Acknowledgement: This research was supported in parts by an ONR Grant N00014-11-1-0295, a NSF Grant IIS-1317749, and a NSF Grant No. 1353120.

References

[1] B. Atcheson, W. Heidrich, and I. Ihrke. An evaluation of optical flow algorithms for background oriented schlieren imaging. *Experiments in fluids*, 2009.

[2] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 2004.

[3] J. Cech, J. Sanchez-Riera, and R. Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, 2011.

[4] V. Couture, N. Martin, and S. Roy. Unstructured light scanning to overcome interreflections. In *ICCV*, 2011.

[5] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *PAMI*, 2006.

[6] M. Gupta and S. K. Nayar. Micro phase shifting. In *CVPR*, 2012.

[7] H. Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi. Dynamic scene shape reconstruction using a single structured light pattern. In *CVPR*, 2008.

[8] S. Konig and S. Gumhold. Image-based motion compensation for structured light scanning of dynamic surfaces. *International Journal of Intelligent Systems Technologies and Applications*, 2008.

[9] T. P. Koninckx, A. Griesser, and L. Van Gool. Real-time range scanning of deformable surfaces by adaptively coded structured light. In *3DIM*, 2003.

[10] M. Movania. Opencloth https://code.google.com/p/opencloth/, 2011.

[11] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado. A state of the art in structured light patterns for surface profilometry. *Pattern recognition*, 2010.

[12] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *PAMI*, 2005.

[13] Y. Tian and S. G. Narasimhan. Globally optimal estimation of nonrigid image distortion. *IJCV*, 2012.

[14] A. O. Ulusoy, F. Calakli, and G. Taubin. One-shot scanning using de bruijn spaced grids. In *ICCV Workshops*, 2009.

[15] M. Vo, Z. Wang, B. Pan, and T. Pan. Hyper-accurate flexible calibration technique for fringe-projection-based three-dimensional imaging. *OpEx*, 2012.

[16] T. Weise, B. Leibe, and L. Van Gool. Fast 3d scanning with automatic motion compensation. In *CVPR*, 2007.

[17] S. Yamazaki, A. Nukada, and M. Mochimaru. Hamming color code for dense and robust one-shot 3d scanning. In *BMVC*, 2011.

[18] Z. Yang, Z. Xiong, Y. Zhang, J. Wang, and F. Wu. Depth acquisition from density modulated binary patterns. In *CVPR*, 2013.

[19] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, 2003.