Structured-Light Stereo:
Comparative Analysis and Integration of Structured-Light and Active Stereo for Measuring Dynamic Shape

Wonkwi Jang, Changsoo Je, Yongduek Seo, and Sang Wook Lee

Department of Media Technology
Department of Electronic Engineering
Sogang University, Shinsu-dong 1, Mapo-gu, Seoul 121-742, Republic of Korea
Office: +82-2-711-8916 Fax.: +82-2-706-4216
E-mail addresses: wkjang@sogang.ac.kr (W. Jang), vision@sogang.ac.kr (C. Je),
yndk@sogang.ac.kr (Y. Seo), slee@sogang.ac.kr (S. W. Lee)

Abstract
In computer vision, two major active range imaging methods have been frequently employed for rapid and efficient shape recovery: (a) conventional active stereo vision and (b) conventional structured-light vision. This paper presents a comparative analysis and an integration of the two active approaches, namely, a structured-light stereo approach for the acquisition of dynamic shape. We firstly investigate the strengths and weaknesses of the two approaches in terms of accuracy, computational cost, field of view, depth of field, and color sensitivity. Based on this analysis, we propose a novel integrated method, the structured-light stereo, to recover dynamic shapes from a wider view with less occlusion by taking most of the benefits of the two approaches. The main idea is as follows. We first build a system composed of two cameras and a single projector (just a basic setup for conventional active stereo), and the projector projects a single “one-shot” color-stripe pattern. The next step is to estimate reliable correspondences between each camera and the projector via an accurate and efficient pattern decoding technique, and some remaining unresolved regions are explored by a stereo matching technique, which is less sensitive to object surface colors and defocus due to the projector’s short depth of field, to estimate additional correspondences. We demonstrate the efficacy of the integrated method through experimental results.
Keywords
Active stereo; structured light; range imaging; color stripe; permutation; comparative analysis.

1. Introduction
Dynamic shape recovery based on stereo vision [1] has been actively researched as an important subject in computer vision. One of the essential issues in stereo vision is to determine the correspondences between the two views. When an object or a scene is distinctively textured, we can obtain dense range images via existing passive stereo matching techniques. However, objects with a lack of textures, such as human skin, bring in visual ambiguities, so it is hard to estimate the correspondences, and the resulting resolution of the range images could be too low to adequately represent the original shapes. A way to alleviate this issue is to employ additional lights such as projectors, which can generate virtual textures onto object surfaces. Using the object’s appearance as well as the virtually created textures considerably resolves the correspondence problem in most cases. This approach can be divided largely into the conventional active stereo (AS), which uses arbitrary illumination (or patterns), and the conventional structured light (SL), which uses coded illumination.

In AS, random dot [2], line or stripe [3], or sinusoid [4,5] patterns are firstly projected onto objects for reducing visual ambiguities, and then pixel-wise correspondences are estimated by the stereo matching techniques of passive stereo. As this approach is largely based on the passive stereo approach except for easily generateable projection patterns, little independent research has been done to the best of our knowledge. On the other hand, since SL requires well-designed patterns for encoding spatial lighting information at each pattern element, extensive studies have been undertaken [6,7,8,9,10,11,12,13]. Some studies have been carried out focusing on scene properties [14] and the scattering environment [15]. We can classify SL patterns largely into two types: temporally coded patterns and spatially coded patterns, depending on the coding types. Temporally coded patterns use more than one pattern image to label each visual element distinctively. For example, binary-encoded stripe patterns [16] require $\log_2 N$ pattern images to encode $N$ unique labels. Although this approach can provide high-resolution range images, it requires high-synchronization equipment in order to capture rapidly moving objects since multiple patterns are necessary for a stationary shape recovery. Some of recent spatially coded patterns resolve the problem with the use of spatial neighborhoods [17,18,19], and this makes it possible to design “one-shot” patterns for dynamic shape recovery. However, when spatial neighborhoods are unidentifiable due to
occlusions or depth discontinuities, this approach may fail to decode the correct labels. Continuous patterns such as color or gray sinusoidal patterns [20,21,22,23,24,25] can theoretically have many pattern elements. However, most of them are not one-shot patterns, and they are sensitive either to noise and object color (low-frequency patterns) or to object geometry (high-frequency patterns produce unwrapping problem). Consequently, each approach has its own benefits and drawbacks, but we do not consider temporally coded patterns in this paper due to the expensive cost of the high-synchronization equipment. Accordingly, we further consider spatially coded patterns only.

The AS and SL approaches have been taken into account independently for dynamic shape recovery so far. Even for static shape recovery, only a few integrations of AS and SL have been proposed [26,27]. In this paper, we investigate the strengths and weaknesses of AS and SL through various criteria for rapid high-resolution range imaging. Based on this analysis, a novel integrated method, the structured-light stereo is proposed, which takes advantage of the merits of both of the two approaches for recovering dynamic shapes from a wider view with less occlusion.

The rest of this paper is organized as follows. Section 2 presents the comparative analysis of AS and SL, and Section 3 proposes the integrated method, the structured-light stereo. We demonstrate the efficacy of the integrated method through experimental results in Section 4. Finally, in Section 5, we conclude the work with a discussion of future work.

2. Comparative Analysis of AS and SL

Prior to the comparative analysis, we select representative techniques for the AS and SL approaches, to make a proper integration for dynamic shape recovery.

2.1. Representative Technique Selection of AS and SL

Our interest is to obtain high-resolution range images for dynamic scenes with commercially available equipment. Thus, one-shot patterns, which do not require synchronization, are appropriate. Moreover, high-frequency sinusoidal patterns that have unwrapping problems are inappropriate [22]. Most modern one-shot patterns are based on color [17,18,19] and geometry [28]. As a color-based technique, the authors of [17,18] proposed a spatially coded pattern using color-stripe permutations, which encodes the position of each stripe uniquely with its spatial neighborhoods (spatially-windowed uniqueness). As a geometry-based technique, the approach in [28] uses a simple grid pattern and the calibrated geometry between a camera and a projector. Instead of explicitly encoding
the pattern element positions as in [17,18,19], this technique estimates which grid intersections on an acquired image correspond to each of the vertical and horizontal lines on the pattern with the help of the calibrated geometry. The geometry-based technique has less sensitivity to depth discontinuities and occlusions due to little use of the spatial neighborhoods, which are required for the former color-based techniques, but has faults in that the resulting solution can fall in a local minimum, and measuring the confidence is not trivial. However, both of the approaches have similar characteristics in terms of accuracy, field of view, projector’s depth of field sensitivity, and color sensitivity, which are criteria that are more important for the comparison, so it does not matter which one is preferred. Accordingly, we select the color-based technique [17] as the representative of the SL approach. In AS, the stereo matching techniques of passive stereo give similar results for highly textured scenes [29], contrary to SL, which requires paying attention to the design of patterns. We can thus select any techniques having low cost in AS and, accordingly, choose template matching such as the sum of squared differences (SSD), sum of absolute differences (SAD), and normalized cross-correlations (NCC). In general, SSD penalizes large differences to a greater degree compared to SAD. NCC is insensitive to the linear scale in illumination [30], and has a higher complexity compared to SSD, which we use in this work. Consequently, we selected two techniques: SL using multiple color-stripes and AS using template matching as the representatives of SL and AS, respectively. When the techniques are not general enough to represent these approaches, we give additional comments as necessary.

2.2. Comparative Analysis

The AS and SL approaches are compared in terms of accuracy, computational cost, field of view (FoV), projector’s depth of field (DoF), and color sensitivity. Note that each approach is now represented as the selected technique with some exceptions, and the pattern employed for AS is the same as the pattern for SL because the pattern effectively produces spatial uniqueness in both approaches, similar to other patterns [2,3,4,5]. Firstly, the accuracy of each approach can be evaluated by how exact the obtainable correspondences are. In the case of SL, since encoding and decoding are performed on stripes, stripe-wise correspondences are obtained. Unfortunately, each stripe should usually have a greater than one pixel-width in the camera image [17], since subpixel-width stripes cannot be properly extracted (the lower bound is one pixel), and the stripe width spatially varies with geometry (there should be broader stripes). This means that when stripe-wise correspondences are estimated, the pixels
on the same stripe have the same correspondence, which causes relatively low resolution and accuracy. Although subpixel localization can be performed in both AS and SL, their basic accuracy resolutions are based on the pixel size or stripe-width. In the case of AS, on the other hand, most matching techniques [29] have at least pixel-wise accuracy in distinctively textured regions. For instance, in the case of template matching, if the window size is larger than the pixels of multiple stripes that guarantee uniqueness, then accurate pixel-wise correspondences are obtained with no consideration of issues such as depth continuities and occlusions. Fig. 1 depicts the accuracy difference between the two approaches. As a consequence, AS typically attains higher resolution than SL does, but with the locally-smoothness assumption on object surfaces, the difference is visually ignorable. Moreover, there exist some schemes of sub-stripe localization [31].

Fig. 1. Comparison of accuracy resolutions and unit windows of SL and AS in disparity computation: (A) image of a statue onto which a color-stripe pattern [17] is projected, and (B) an enlarged view of the left box representing the accuracy resolutions and unit windows of AS and SL.

Recovering dynamic shapes requires a high computational cost for handling a number of stereo images. In particular, stereo matching techniques vary from simple techniques requiring low cost, such as template matching, to modern optimization-based techniques requiring high cost such as, graph cuts [32]. Let us consider the lower-cost technique, template matching. In active stereo, the whole area of each camera image will mostly contain sufficient illumination texture, and template matching should be performed for whole image
pixels unlike in passive stereo of a less-textured scene. Therefore, the required time complexity is $\Theta(mnpqS)$ where $m \times n$ is the image size, $p \times q$ is the template window size, $S$ is the search range, and $\Theta$ is the big theta. Basically the search range is two-dimensional (e.g. $u \times v$), and can be simplified to a one-dimensional disparity range, $D$, after image rectification, resulting in $\Theta(mnpqD)$. The computational complexity can be decreased by using an efficient multi-resolution matching approach based on a multi-level pyramid such as in [33]. The complexity is decreased to $\Theta(mnpq\log D)$ or $\Theta(mnpqh)$, where $h$ is the number of levels of the pyramid. For example, based on [33], $D = 3 \times 2^{h-1}$. The template window size and the disparity range commonly depend on the image size ($M \equiv mn$). By assuming $\Theta(pq) = \Theta(M)$ and $\Theta(D) = \Theta(M^{0.5})$, we get $\Theta(M^2\log M)$, since $\Theta(\log(M^{0.5})) = \Theta(\log M)$. It can be noted that the computational complexity can be further reduced using dynamic programming [34,35]. However, dynamic programming relies on the monotonicity assumption, which can be readily violated by occlusions. In this discussion on computational complexity, we intend to allow typical occlusions for both of AS and SL, and thus exclude the reduction by the monotonicity assumption. In case of SL, on the other hand, the correspondences come directly from the pattern decoding process based on the color-stripe sequence, and the time complexity is just $\Theta(mn) = \Theta(M)$, which is also the same in the case of multi-layer decoding [18] since each pixel is still decoded only once. Therefore, the computational cost of AS is $\Theta(M\log M)$ times more expensive compared to that of SL.

![Fig. 2. Projector-camera configurations and the FoVs: (A) SL, (B) AS, and (C) the integrated method (structured-light stereo).](image)

Range imaging from AS and SL uses two views, so only the common regions from the two views are employed for range imaging. The two approaches thus roughly have similar fields...
of view (FoVs). However, AS is better than passive stereo (PS) only in the intersection of the three FoVs of the two cameras and a projector. Therefore, the active FoV of AS is smaller than that of SL. Contrary to SL, which requires a camera and a projector as the minimal system configuration, AS is equipped with two cameras and a projector, but utilizes only two views of the cameras explicitly, since the projector basically plays the role of generating virtual textures. This implies that, with the AS system configuration, we can also potentially employ the projector view to additionally apply the SL technique without additional equipment. Fig. 2 demonstrates the FoV of the three cases in similar system configurations in which all views look at the statue. The solid lines represent the view of each camera, and the dotted lines represent the view of the projector. The dark blue curves on the statue represent the recoverable shape regions among the views. We note that if SL techniques could be run on the AS system configuration, or if the AS and SL approaches could be integrated, it would be possible to recover object shapes from a wider view with less occlusion (see Section 3).

![Fig. 2](image)

Commercially available projectors have short depths of field (DoFs) since they are mainly used for projections onto planar screens. Therefore, it is infeasible to focus on multiple objects spaced over a large depth interval or an object having a large amount of depth. This makes the identification of color stripes on defocus-blurred regions difficult. In case of AS,
however, explicit decoding of the color stripes is not required, and accordingly, it is less sensitive to defocus blur compared to the SL approach as long as the imaged surfaces have sufficient distinctiveness.

We conducted an experiment to quantitatively compare AS and SL. Fig. 3(a) represents the experimental configuration. As detailed techniques (including parameters) and hardware types will be given in Sections 3 and 4, respectively, we now focus on an analysis of the experimental results. Firstly, the projector is set to focus on the bottom region of the white dotted box on the slanted plane, and then projects the pattern onto the plane. Fig. 3(b) shows the acquired image from the left camera, which is the region of the white dotted box in Fig. 3(a). Although the depth of the box region is only 20cm, the upper region of the box is slightly out of focus, as shown in Fig. 3(c), which is the enlarged region of the white box in Fig. 3(b). It is thus hard to identify the color stripes on those regions due to the blended colors. Fig. 3(d) and 3(e) represent the color-stripe segmentation result and the decoded stripe IDs, respectively. Since the projected pattern is substantially defocused in the upper part of the camera image, the segmentation and decoding results have many errors around the upper part. As we can see in Fig. 3(e), only about 50% is properly decoded. On the other hand, we can obtain most correspondences of those regions via stereo matching of AS represented in Fig. 3(f), which shows less sensitivity to the projector’s DoF. Unfortunately, it is a physically difficult problem to achieve both a long DoF and an abundant amount of light at the same time, so defocus sensitivity is a significant criterion.

![Fig. 4. Result of a highly contrasted and strongly colored object: (A) camera image, (B) result of SL, and (C) result of AS.](image)

Color sensitivity is also an important issue. When a colored stripe pattern is projected onto object surfaces having high contrasts or strong colors, it is hard to separate each stripe, so pattern decoding is not possible in those regions. Contrary to this approach, stereo matching
techniques in AS do not discriminate between the projected pattern image and the appearance of objects, and instead utilize both types of information at the same time. Thus, if only the distinctiveness is preserved, AS is able to extract correct correspondences. Fig. 4 shows the results of AS and SL with a highly contrasted and strongly colored object surface.

As a supplement, the AS approach requires more preprocessing. One assumption with stereo vision is that the color responses of the two sensors are similar. However, each sensor has its own characteristics, and more significantly, the same sensors cannot ensure the exact same responses. To handle this problem, photometric calibration is required. In the case of SL, pattern images are already given, and the number of colors is relatively small, so it is easy to correct the pattern colors offline. However, in case of AS, one of the stereo images is always corrected online, so the approach requires more time to process. In addition, there are two additional matters in the AS approach. The first one is the image rectification, which is required for reducing the searching time, and the second one is the disparity range, which limits the range of the search when matching to reduce the searching time. On the other hand, some existing efficient SL approaches such as those in [17,18] guarantee the global uniqueness of the subpatterns. Table 1 summarizes the comparative analysis of the two approaches discussed so far.

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>SL</th>
<th>Structured-light stereo (SLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic accuracy</td>
<td>pixel</td>
<td>stripe width</td>
<td>pixel / stripe</td>
</tr>
<tr>
<td>resolution</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational</td>
<td>( \Theta(M^2 \log M) )</td>
<td>( \Theta(M) )</td>
<td>( \Theta(M + M^2 \log M) )</td>
</tr>
<tr>
<td>time complexity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field of view</td>
<td>one intersection</td>
<td>two intersections</td>
<td>three intersections</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>defocus, object color</td>
<td>image rectification, disparity range</td>
<td>image rectification, disparity range</td>
</tr>
<tr>
<td>Schemes to reduce</td>
<td>image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the searching cost</td>
<td>rectification, disparity range</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This section presents the Structured-Light Stereo, a novel integrated method based on the comparative analysis by taking advantage of most of the benefits of both approaches to recover dynamic shapes from a wider view with less occlusion. The integration of AS and SL has rarely been studied. Scharstein and Szeliski [26] proposed an integrated method based on binary-encoded stripe patterns [16], and employed it to obtain the ground-truth disparities for comparing the existing stereo matching techniques of passive stereo. Wang et al. [27] proposed an integrated method based on phase shifting. Although their methods [26, 27] provide reliable and accurate shapes for static scenes, the computational cost and the consideration of real-time capturing for dynamic shape recovery, which are main issues in our work, were not deliberated. Unlike existing integrations [26, 27], our integrated method employs a single one-shot color-stripe pattern to acquire dynamic shapes. We provide an overview of our integration method in Section 3.1, and give a detailed description of the method in Section 3.2.

3.1. Overview of Structured-Light Stereo

We integrate the two approaches based on the AS system configuration which is comprised of a (uncalibrated) projector and two (calibrated) cameras, as shown in Fig. 2. The pattern we are employing is the one-shot spatially coded pattern [17] used in the above comparative analysis. The SL approach explicitly utilizes the pattern information by decoding it, whereas the AS approach implicitly utilizes the pattern information as virtual textures.

Through the comparative analysis, we confirmed that the computational cost of the SL technique is significantly lower than the AS one, so we first decode the acquired stereo images in order to estimate very reliable correspondences efficiently. Most uncertain regions that were not decoded due to the sensitivity to defocus and object surface colors are then resolved by a following stereo matching technique. The remaining processes are to geometrically calibrate the projector, and to recover depths from the estimated disparity maps (see Section 3.2). Note that since we exploit information from all (three) views explicitly, it is possible to recover shapes from a wider view having less occlusion than each of AS and SL. The fourth column of Table 1 summarizes the advantages of the proposed method compared to the AS and SL approaches. In addition to the broader field of view, the method improves the accuracy by preferentially selecting reliable parts of the acquired disparity maps from the three intersections (see Section 3.2 for the details).

3.2. Structured-Light Stereo Details
Fig. 5 represents the flowchart of the proposed structured-light stereo. Each box shows an independent process, and each arrow represents the relation among processes and the data being transferred. A detailed explanation of each process follows.

Geometric and photometric calibrations are required before reconstructing scenes represented as black boxes in Fig. 5. The geometric calibration is to estimate the (intrinsic/extrinsic) parameters of the stereo camera, and the parameters are used for the reconstruction from three views, and for stereo rectification. To achieve this, we employed the method presented in [36]. Although the projector is required to be calibrated as well, we will consider this later. The photometric calibration is to estimate the transformation parameters to make the color responses of the stereo cameras consistent (including the projector). We employed the method in [37] to deal with color coupling, nonlinearities, and biases in the transformations.

After the calibrations, we project the pattern image onto the target objects, and capture stereo images. Each stereo image pair is then preprocessed as shown in the blue boxes in Fig. 5. The first one is image rectification, which reduces the searching domain from 2D to 1D,
and the second one is color transformation that transforms the color responses of a camera to the reference one using the parameters estimated during the photometric calibration.

We can now estimate correspondences among the views as shown in the green boxes in Fig. 5. As already discussed, we first take advantage of the SL approach. When every two adjacent stripes are forced to have different colors for facilitating the identification of the stripes, \(N(N-1)^{k-1}\) unique color-stripe sequences with the length \(k\) can be generated from \(N\) different colors [17,38]. For instance, if we take \(N = 3\) (RGB) and \(k = 7\), then we can obtain 192 \((= 3 \times (3-1)^{7-1})\) unique labels, which are proper for the usual projector resolutions (width or height of 1,280 or less pixels). If we take \(N = 3\) (RGB) and \(k = 9\), then we can obtain 768 \((= 3 \times (3-1)^{9-1})\) unique labels, which are even sufficient for Full-HD (1920 × 1080) projectors with a stripe width of two or three projector-pixels.

If the numbers \(N\) and \(k\) are fixed for pattern encoding and decoding, every pixel of the acquired images is labeled with a stripe ID even if some stripes are not identifiable due to factors such as depth discontinuities. Therefore, in order to obtain reliable correspondences, we decode the color-stripe sequences with a decoding length, \(k_d = k_e + \alpha\), where \(k_e\) is the length used in the encoding, and \(\alpha\) is a positive integer. Actually, we use a modified version of an initial single layer in multi-layer decoding [18]. Since the number of colors is three, and every two consecutive stripe-colors should be different, the number of detectable colors of each additional stripe is two. Hence, the error rate with \(\alpha\) is given as \(E_\alpha = 2^{-\alpha}E_0\), where \(E_0\) is the error rate in decoding with the length \(k_e\). When \(\alpha = 3\), for example, the error rate goes down to \(E_\alpha = 0.125E_0\). Through the expanded length of decoding, this approach is able to obtain very reliable correspondences, but it also increases the extent of the regions that are not decoded since they do not include sufficiently-long-and-correct color sequences. This is, however, not a serious issue in the structured-light stereo since the following stereo matching can considerably resolve the problem.

When we perform pattern decoding on acquired stereo images with the above strategy, we obtain two highly reliable disparity sets, \(S_L\) and \(S_R\), which represent the projector-camera correspondences of the left image region \((L)\), and those of the right image region \((R)\), respectively. From this, we can infer the common region, \(C_S \equiv S_L \cap S_R\) without expensive matching techniques since each stripe is uniquely labeled, and the pattern is shared in two views. We can also obtain the regions that belong to the left view only and the right view only, which are denoted as \(C_L \equiv S_L - C_S\) and \(C_R \equiv S_R - C_S\), respectively. This process is called
disparity map separation, and the results are used especially in the following disparity range estimation and in projector calibration.

Decoded regions in the decoding result are limited due to the projector’s short DoF and strong surface color and reflectance. We thus find an additional disparity set, $C_A$ from the intersection of the unresolved regions in $L$ and $R$, $(L−S_L) \cap (R−S_R)$ using a stereo matching technique. Fig. 6 shows a schematic illustration of the spatial relationship between the acquired disparities, $C_S$, $C_A$, $C_L$, and $C_R$. Visual ambiguities are considerably reduced by the projected pattern, so we use a simple template matching technique, SSD, with a $19 \times 7$ window size and a disparity search range $\lambda$. The template window is not square, since the pattern of vertical color-stripes has a high-frequency horizontal variation of intensities, and therefore the pattern-projected scene image will have more effective intensity profiles horizontally than vertically. Note that it is assumed that the spatial variation in object color is considerably less rapid than that of pattern color in its encoding direction. This assumption is employed in most of the conventional SL methods, since without this assumption, reliable recovery of the original pattern colors might be infeasible. For example, if the spatial variation in object color is similar to that of the light pattern, we could not distinguish whether any local variation of color in the camera image is caused by the light pattern or by the object color. In addition, the pattern guarantees uniqueness with nine stripes, so the window width should be $18 \sim 27$ as most stripes have a width of two or three pixels in camera images. In matching, searching the whole solution space demands a high
computational cost, so it is desirable to automatically limit the search range $\lambda$ depending on the object distance from the cameras. The decoding result in SL can be used to determine the search range. Given the correspondences in $C_S$, we can achieve it by finding the maximum disparity in $C_S$, and regard it as the lower bound of the search range. Although, through the matching process, we can obtain additional correspondences, the result is not reliable such as for occluded or shadowed regions. We thus check the left-right consistency, which is not as reliable as the pattern decoding, which provides theoretical confidence, and then remove the inconsistent data.

As we have the camera parameters and correspondences in $C_S$, $C_A$, $C_L$, and $C_R$, it is possible to estimate the depths in the $C_S$ and $C_A$ regions, whereas depth estimation from $C_L$ and $C_R$ is infeasible since we have no projector (intrinsic/extrinsic) parameters. We thus first estimate the 3D coordinates corresponding to the 2D coordinates in the $C_S$ and $C_A$ regions via triangulation with the given camera parameters.

Through the geometric calibration process, we estimated the parameters of the stereo camera, whereas we disregarded the calibration of the projector. This is due to the difficulty in the acquisition of the projector view as an active sensor. In addition, since projectors have relatively short DoFs, it can be desirable to adopt projectors that have an auto-focusing function to recover object shapes regardless of the initial position depending on the applications. Therefore, we need a self-calibration technique for the projector, but the estimation of the parameters of the projector is possible without the self-calibration, as we know the 2D pattern coordinates in $C_S$ and the corresponding 3D coordinates. When the 3D to 2D data is given, it is simple to estimate the $2 \times 4$ projector projection matrix using singular value decomposition with no consideration of outliers because the correspondences in $C_S$ are highly reliable with the error rate, $E_\alpha = 2^{-\alpha}E_0$. With the projection parameters, it is now possible to estimate the remaining depths in $C_L$ and $C_R$. Finally, the depths estimated in $C_S$, $C_A$, $C_L$, and $C_R$ are merged into one in a single 3D frame without the registration problem as in [26]. These processes are represented as red boxes in Fig. 5.

4. Experimental Results

We have conducted experiments for demonstrating the performance of the proposed method. The experimental result for a static scene and that for a dynamic scene are presented in Sections 4.1 and 4.2, respectively.
Fig. 7. Experimental results of the multiple objects. (A and B) preprocessed captured stereo image, (C) color segmented image of (B), (D) decoded stripe IDs of (B), (E) disparity map obtained from stereo matching, (F) depth image obtained by the structured-light stereo from the viewpoint of the right camera. Some additional discontinuity artifacts are produced by view generation.
4.1. Static Scene: Colored Multiple Objects

The experimental system for the static scene is composed of two Sony XC-003 640 × 480 3CCD cameras with 16mm lenses and an Epson EMP-7700 1024 × 768 LCD projector. In this experiment, the RGB one-shot pattern with $k = 9$ is projected into the scene, and therefore $N = 768$ as discussed in Section 3.2. The static scene experiment is to recover the shape of multiple colored objects that are separated from each other by depth to verify that regions which are out of focus and strongly colored are able to be recovered from a wider view with less occlusion compared to each of AS and SL.

We positioned six objects, a yellow waterdrop toy (close middle), a donut-like yellow cushion (left), a white handbag (right), a white toy dog on a box (center), a red toy hippo (distant left), and a white teddy bear (distant right), on a table as shown in Fig. 7(a) and (b). The projector is set to focus on the white doll, and projects the color-stripe permutation pattern onto the scene. From the acquired stereo image, we first decode each image. Fig. 7(d) shows the pattern decoding result of the right image, and we can see that the object shapes located on the left side were not recovered due to the projector’s short DoF and object surface colors. Moreover, the box below the white doll was also not recovered because of the extremely high reflectance of the object surface. Fig. 7(c) represents the color segmentation result, which immediately affects the pattern decoding, and shows the cause of the partially unsuccessful pattern decoding. Although the pattern is robust to system noises, nonlinearities, and object surface colors, it is difficult to deal with complex situations such as those in our experiment. However, in contrast to pattern decoding, stereo matching can handle those scenes with less sensitivity as shown in Fig. 7(e). The result obtained by the proposed method is presented in Fig. 7(f), and this view is generated with respect to the right camera. Although most depths are recovered from AS, the right side of the handbag and the occluded region which is to the right side of the white doll were recovered from SL as representative examples. It is also noted that although stereo matching has less sensitivity, there are some regions which were recovered only by pattern decoding (not by stereo matching) such as in the case that the template window size is unable to ensure uniqueness.

4.2. Dynamic Scene: Fine Facial Expression

The experimental system for the dynamic scene is composed of two BASLER A504kc 1280 × 1024 CMOS cameras with 50 mm lenses and an InFocus LP650 1024 × 768 DLP projector. In this experiment, the RGB one-shot pattern with $k = 7$ is projected into the scene, and thus $N = 192$ as discussed in Section 3.2. The dynamic scene experiment is to recover the
dynamic shape of a finely expressed human face. The projector alternately projects two images, the RGB one-shot pattern image and a white image. By synchronizing the projector and cameras at 60 fps, we sequentially capture the face images corresponding to each projection from the two camera viewpoints. The maximum feasible frame rate is 80 fps in the synchronized projector-camera system. Fig. 8 shows the experimental setup for acquiring the fine facial expression.

Fig. 9 shows the experimental results of the fine facial expression. Fig. 9(A) is the camera image with white illumination for acquiring the texture color, and (B) is an image of the RGB one-shot pattern projection for dynamic shape acquisition. Fig. 9(C) shows the stripe segmentation result from (B), and (D) shows the depth result by the proposed method. We use the images with white illumination for the facial texture color, and render the acquired results. Fig. 9(E) is a point-rendered view, (F) and (G) are two polygon-rendered views, and (H) and (I) are polygon-rendered views after interpolation as a postprocessing. Fig. 10 shows the results from four selected frames. Note that the differences between frames are quite apparent even though the facial expression is very fine.

5. Conclusion and Future Work

This paper presented a comparative analysis of AS and SL for dynamic scene recovery in terms of accuracy, computational cost, field of view, sensitivities to defocus and object color. Based on the analysis, we also proposed an integrated method, the structured-light stereo, by taking most of the benefits of the two approaches for recovering wider-view dynamic shapes with less occlusion, and demonstrated the performance of the proposed method through experimental results.

We paid considerable attention to the selection of each representative of AS and SL, but it is infeasible to represent a class of techniques with a single specific technique since, for example, the color sensitivity is different depending on the specific SL technique used. Therefore it is required to research more general and in-depth comparative analyses as future work. Since the proposed SLS employs the SL technique that is based on a single projector, the SLS may have some limitations in case of using multiple projectors. When the structured-light patterns projected by multiple projectors, are superposed, extracting the original patterns is nontrivial. Thus the SLS presented in this paper, should be further integrated with the multi-projector SL method [39], to establish a multi-projector SLS technique.
Fig. 8. Experimental setup for recovering dynamic shape and texture. A projector and two cameras are synchronized to sequentially capture the scene images corresponding to each projection image from the two camera viewpoints.

A) B) C) D) E)

F) G) H) I)

Fig. 9. Experimental results of fine facial expression. (A) white illumination for texture acquisition, (B) RGB one-shot pattern projection for dynamic shape acquisition, (C) stripe segmentation result, (D) depth result by the proposed method, (E) point-rendered view, (F and G) polygon-rendered views, (H and I) polygon-rendered views after postprocessing.
Acknowledgement

We thank the editor and three reviewers for their valuable comments and suggestions. This research was supported by the Converging Research Center Program funded by the Ministry of Education, Science and Technology (2012K001342) and by Basic Science Research

Fig. 10. Experimental results from selected frames of fine facial expression. (A) white illumination for texture acquisition, (B) RGB one-shot pattern projection for dynamic shape acquisition, (C) polygon-rendered view after postprocessing.
Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A2009461).

References


