Fusion of Active and Passive Sensors for Fast 3D Capture

Qingxiong Yang, Kar-Han Tan, Bruce Culbertson, John Apostolopoulos

HP Laboratories
HPL-2010-97

Keyword(s):

3D, Depth, Sensors, Time of flight, Stereo, Multiview, real time, remote collaboration, human interaction

Abstract:

We envision a conference room of the future where depth sensing systems are able to capture the 3D position and pose of users, and enable users to interact with digital media and contents being shown on immersive displays. The key technical barrier is that current depth sensing systems are noisy, inaccurate, and unreliable. It is well understood that passive stereo fails in non-textured, featureless portions of a scene. Active sensors on the other hand are more accurate in these regions and tend to be noisy in highly textured regions. We propose a way to synergistically combine the two to create a state-of-the-art depth sensing system which runs in near real time. In contrast the only known previous method for fusion is slow and fails to take advantage of the complementary nature of the two types of sensors.

External Posting Date: August 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: August 21, 2010 [Fulltext]
To be presented at IEEE International Workshop on Multimedia Signal Processing 2010, Saint-Malo, France. October 4, 2010

Fusion of Active and Passive Sensors for Fast 3D Capture

Qingxiong Yang* Kar-Han Tan† Bruce Culbertson† John Apostolopoulos†
*University of Illinois at Urbana Champaign †Hewlett-Packard Laboratories
http://vision.ai.uiuc.edu/~qyang6/ http://www.hpl.hp.com/people/kar-han_tan/

***Abstract*—We envision a conference room of the future where depth sensing systems are able to capture the 3D position and pose of users, and enable users to interact with digital media and contents being shown on immersive displays. The key technical barrier is that current depth sensing systems are noisy, inaccurate, and unreliable. It is well understood that passive stereo fails in non-textured, featureless portions of a scene. Active sensors on the other hand are more accurate in these regions and tend to be noisy in highly textured regions. We propose a way to synergistically combine the two to create a state-of-the-art depth sensing system which runs in near real time. In contrast the only known previous method for fusion is slow and fails to take advantage of the complementary nature of the two types of sensors.**

I. Introduction

In the future, users in a conference room will be able to naturally interact with large, immersive display walls without needing to use dedicated controllers or wear special markers. Depth sensing systems will be able to capture in real time the 3D shape, pose, and positions of users in the room and use that information to enable highly engaging experiences for seamless collaboration, data visualization, or 3D entertainment. Depth sensing is therefore a key enabling technology that new products can leverage to deliver enhanced functionalities that provide differentiation in the marketplace.

At the same time, depth sensing is also one of the hardest fundamental challenges of computer vision. 3D reconstruction using passive stereo with multiple cameras is a well-studied problem and to date they still will not work reliably in many cases. Non-textured and featureless regions of a scene are particularly challenging for stereo as there simply is insufficient visual information for establishing correspondence across multiple cameras. The possible solution is to propagate information from textured pixels to the non-textured pixels. Most of the community's efforts are focused on this problem which is named *disparity optimization* in [1]. A number of excellent optimization methods have been proposed, and the state-of-the-art methods are either based on belief propagation (BP) [2], [3] or graph cuts (GC) [4]. Both BP and GC are formulated in an energy-minimization framework [5], where the objective is to find a disparity solution that minimizes a global energy function. Nevertheless, these methods are known to be quite fragile in practice and slow.

MMSP'10, October 4-6, 2010, Saint-Malo, France. 978-1-4244-8112-5/10/$26.00 2010 IEEE

Alternative, laser range scanners can provide extremely accurate and dense 3D measurement over a large working volume [6], [7], [8], [9], [10], [11]. However, most of these high-quality scanners measure a single point at a time, limiting their applications to static environments only.

Recently, a new class of active depth sensing system based on the time-of-flight (TOF) principle is emerging: [12], [13], [14], [15], [16]. The TOF principle is similar to that of LIDAR scanners with the advantage that whole scene is captured at the same time. A typical TOF system features a near-infrared (NIR) pulse illumination component, as well as an image sensor with a fast gating mechanism. Based on the known speed of light, the system coordinates the timing of NIR pulse wave emissions from the illuminator with the gating of the image sensor, so that the signal reflected from within a desired depth range is captured exclusively. The amount of pulse signal collected for each pixel corresponds to where within the depth range the pulse was reflected from, and can thus be used to calculate the distance to a corresponding point on the captured subject.

TOF sensors are generally of low resolution: 320×240 or less. Most sensor fusion approaches using TOF sensors aim at enhancing the resolution of depth maps by combining it with a single passive camera. By assuming that depth discontinuities always corresponds to color discontinuities, Yang *et. al.* [17] used a high-resolution color image to reduce the depth ambiguities after up-sampling from a low-resolution (64×48) depth image acquired from Canesta sensors [14] via joint bilateral filtering. Kopf [18] also employed joint bilateral filter during a number of image operations, e.g., tone mapping, colorization, stereo matching, to improve up-sampling results. The main difference between these two methods is that in [18], joint bilateral filter is applied to the depth image directly, which will oversmooth depth discontinuities. [17] instead uses a voting scheme, which better preserves the depth discontinuities and is more robust to the noisy. However, the complexity of [17] is much higher than [18], since the number of depth hypotheses used for voting is generally large (60 in [17]). TOF sensors are able to sense depth even in non-textured regions regardless of the low resolution. In fact they perform poorly on heavily textured surfaces for which stereo excels. This offers hope that we may finally be able to build reliable depth sensing systems. The advantage and disadvantage of the active and passive sensors are listed in Table I. As can be seen, active and passive sensors complement each other.

TABLE I
ADVANTAGE AND DISADVANTAGE OF ACTIVE AND PASSIVE SENSORS. THE ENTRIES IN BOLD CORRESPOND TO THE ADVANTAGES.

Sensor	Resolution	Highly-textured region	Non-textured region
Active	low	Non-robust	**Robust**
Passive	**High**	**Robust**	Non-robust

The only prior art we are aware of that fuses depth estimation from TOF sensor and stereo sensor is [19]. Our method is superior as [19] does not take into account the complementary nature of the two kinds of depth sensors and would weight data from both subsystems equally regardless of whether the scene is textured or non-textured. While they did not report running times in their paper, since they employ belief propagation, a very expensive operation, with the same camera and disparity search resolution we expect their algorithm to require 30 seconds per frame for a 3-view setup and 15 seconds per frame for a 2-view setup. In contrast, our system runs at 8 fps and correctly takes into account the complementary nature of the sensors.

II. ALGORITHM OVERVIEW

We propose fusing the two kinds of sensors in a synergistic fashion, relying on passive stereo in highly textured regions while using data from active depth sensors in featureless regions. A photo of our experiment setup is shown in Fig. 1 (a) and a flow chart of our algorithm is shown in Fig. 1 (b).

One key challenge we faced is that data from the active sensors are of a much lower resolution than those from typical RGB cameras used in the passive subsystem. To address this problem we developed an algorithm for up-sampling a depth map at real time based on hierarchical bilateral filtering in Sec. III-B.

For synergistic fusion, we make use of a TOF signal strength image provided by the active depth sensor which indicates the signal strength received at each sensor pixel. This allows us to compute a TOF sensor confidence map. A stereo confidence map is also computed based on local image features. The two confidence maps are then incorporated into the cost function used to populate the 3D volume created by a plane-sweeping stereo matching algorithm.

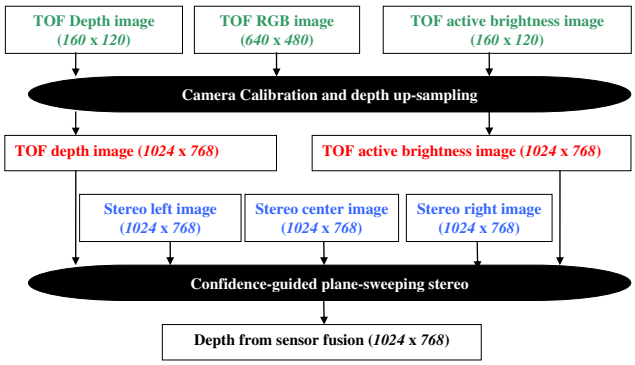
III. FAST SENSOR FUSION

In this section, we present a method to combine TOF and passive sensors to create a state-of-the-art depth sensing system which runs in near real time. Our method is separated into two steps as shown in Fig. 1 (b): camera calibration and depth up-sampling (Sec. III-A and III-B) and confidence-guided plane-sweeping stereo matching (Sec. III-C).

A. Camera calibration

Our depth sensing system combines a TOF sensor with three stereo cameras. TOF sensor can produce a depth image and a TOF signal strength image of 160×120 resolution with an operational range up to about four meters. The TOF sensor

(a) Experiment setup.



(b) Algorithm overview.

Fig. 1. System overview. (a) is a photo of our depth sensing system and (b) is the proposed framework. The inputs of our sensor fusion system are a low-resolution (160×120) depth image, a low-resolution (160×120) TOF signal strength image and a high-resolution (640×480) RGB image from the TOF depth sensor (shown in green) and three high-resolution (1024×768) images from passive stereo sensors (shown in blue). The output of the system is a high-resolution (1024) depth image. Our system can be separated into two steps. First, we up-sample the low-resolution images captured from the TOF depth sensor and then register them with the passive stereo sensors. Second, we compute a depth image from the images captured from the passive stereo sensors and the registered depth and TOF signal strength images from TOF depth sensor.

can also produce a 640×480 RGB image. The captured depth image and TOF signal strength image can be mapped to the RGB image via a lookup table using the depth image. We thus captured a sequence of images of a calibration pattern simultaneously from the TOF sensor and the stereo sensors, and then use the calibration toolbox [21] to compute the calibration/intrinsic matrix, radial distortion coefficients, and projection matrix of each sensor as shown in Fig. 4.

B. Real Time Depth Image Up-Sampling

We next up-sample the depth and TOF signal strength images captured from TOF depth sensor. We have previously proofed that by enforcing the depth edges to be consistent with the color edges, joint bilateral filter can be used to up-sample the depth image up to $100\times$ resolution [17]. However, it is too slow for real-time applications. In this section, we present a hierarchical method for real-time depth image up-sampling.

Let R be the vector of all depth hypotheses, D_T^i be the up-

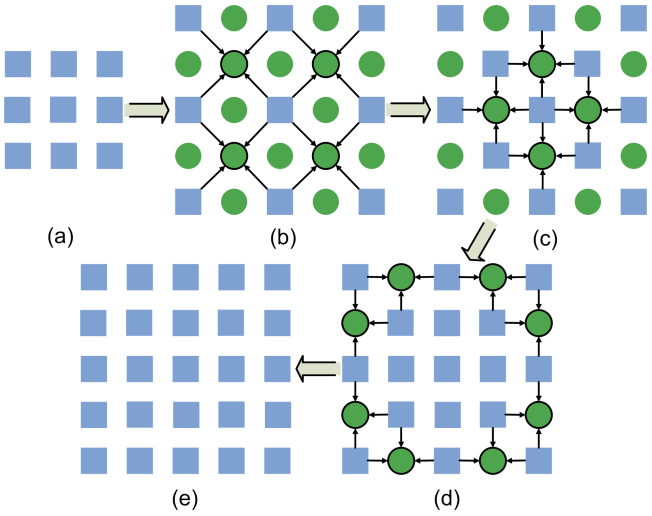


Fig. 2. Hierarchical up-sampling. (a) is the depth image in coarse scale, (b) is the nearest up-sampling result of (a). The unconfident pixels are marked as green circles, and blue squares are confident pixels with depth values sampling from (a). The depth values of unconfident (un-sampled) pixels with four confident neighbors are estimated as shown in (b) and then marked confident in (c). (c) also shows that the depth values of the other unconfident pixels except the unconfident edge pixels are estimated based on the updated depth images. These unconfident pixels are also marked as confident after estimation as shown in (d). The unconfident edge pixels are eliminated in the last step. (e) shows that every pixel in the image are marked as confident.

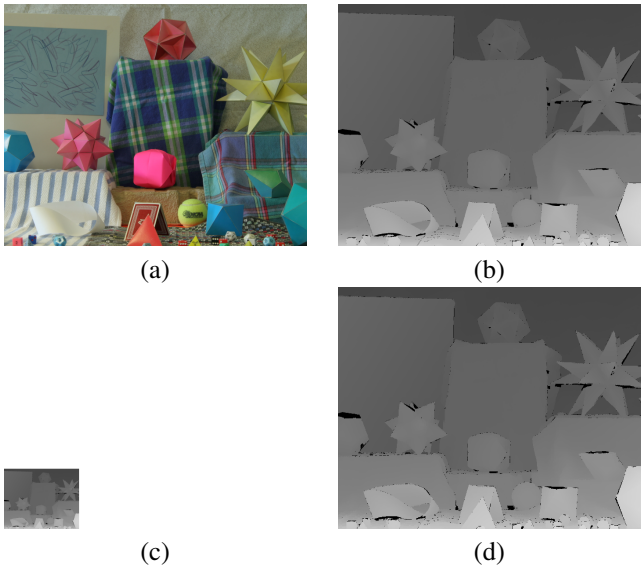


Fig. 3. Evaluation using Middlebury data set [20]. (a) High resolution color image; (b) Ground-truth disparity image; (c) Low resolution disparity image; (d) Up-sampled high resolution disparity image.

sampled depth image using nearest neighbor method at scale i, $C_T^i = \{0, 1\}$ as a binary confidence map associated with D_T^i, I_T^i as the registered color image, p as a pixel in D_T^i, and q as another pixel in the neighborhood $N(p)$ of p, the joint bilateral up-sampling problem can then be expressed as:

$$D_T^{i,new}(p) = \underset{z \in R}{\operatorname{argmin}} \frac{1}{\sum_{q \in N(p)} F(p,q) \cdot G(I_T^i(p), I_T^i(q)) \cdot C_T^i(q)} \cdot \sum_{q \in N(p)} F(p,q) \cdot G(I_T^i(p), I_T^i(q)) \cdot C_T^i(q) \cdot |z - D_T^i(q)|, \quad (1)$$

where $F(p,q)$ and $G(I_T^i(p), I_T^i(q))$ are the spatial and range weighting function of the joint bilateral filter, respectively, $|z - D_T^i(q)|$ is the penalty cost value for assigning depth hypothesis z to pixel q and $C_T^i(q) = 0$ identifies un-sampled pixels. The updated depth image $D_T^{i,new}$ is then up-sampled using nearest neighbor method and fed to the next scale for further refinement. An example of our hierarchical up-sampling method on a 3×3 depth image is shown in Fig. 2.

Figure 3 presents the visual evaluation of our up-sampling method on the Middlebury data set [20]. We down-sampled a high-resolution disparity image (treated as ground truth, showing in Figure 3 (a)) obtained using structured light scanning to a low resolution disparity image (Figure 3 (c)), and then used our method to up-sampled the low resolution disparity image to its original resolution based on the low resolution disparity image and the high resolution color image (Figure 3 (b)), and then compared the up-sampled disparity image (Figure 3 (d)) and the original high-resolution disparity image (Figure 3 (b)). Visually, there is little difference between Figure 3 (b) and (d). Note that the black pixels in (b) are invalid pixels where structured light scanning fails.

We use this hierarchical method to up-sample the registered depth and TOF signal strength image. Finally, the 2D image points of the up-sampled depth and TOF signal strength image are projected as 3D points using the up-sampled depth image and then captured back by the central stereo sensor. Specifically, let a pixel location in the up-sampled depth image be represented as a homogeneous 3-vector $\mathbf{p} = [\mathbf{x}, \mathbf{y}, 1]^T$, the z-depth value of $\mathbf{p}$ be $z_{\mathbf{p}}$ (obtained from the TOF depth image), the calibration/intrinsic matrix of the TOF sensor be $\mathbf{K}_T$, and projection/extrinsic matrix be $\mathbf{P}_T = [\mathbf{R}_T | (-\mathbf{R}_T \mathbf{t}_T)]$, the 3D point $\mathbf{q}$ corresponding to the 2D pixel $\mathbf{p}$ can then be computed as follows

$$\mathbf{q} = \mathbf{R}_T^{-1}(\mathbf{K}_T^{-1} z_{\mathbf{p}} \mathbf{p}) + \mathbf{t}_T. \quad (2)$$

Let the calibration/intrinsic matrix of the central stereo camera be $\mathbf{K}_C$ and the projection/extrinsic matrix be $\mathbf{P}_C$, the corresponding 2D pixel location in the stereo camera be $\mathbf{p}_C = [\mathbf{x}_C, \mathbf{y}_C, 1]$, then

$$\mathbf{p}_C \sim K_C \mathbf{P}_C \mathbf{q} = K_C \mathbf{P}_C (\mathbf{R}_T^{-1}(\mathbf{K}_T^{-1} z_{\mathbf{p}} \mathbf{p}) + \mathbf{t}_T). \quad (3)$$

Using Eqn. (2) and (3), we can map every pixel in the depth and TOF signal strength image to the stereo camera. However, due to occlusions, source pixels do not map directly to a single pixel in the destination space. In this case, we keep only the pixel which is closest to the center of central stereo sensor. Also, there will be "holes" in these destination

images. We simply need to set the active brightness values to zeros for these pixels. Zero active brightness value means that the confidence of the TOF sensor is zero, thus we can use the result from stereo matching to fill in these "holes". The resolution of the obtained depth and TOF signal strength image is the same as the resolution of the stereo sensor: 1024×768. This step is summarized in Fig. 5.

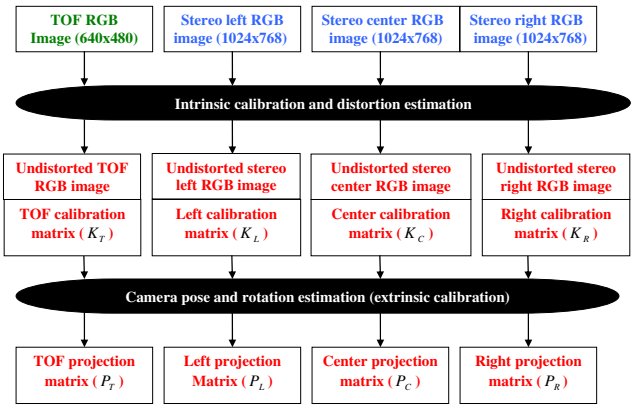


Fig. 4. Camera calibration.

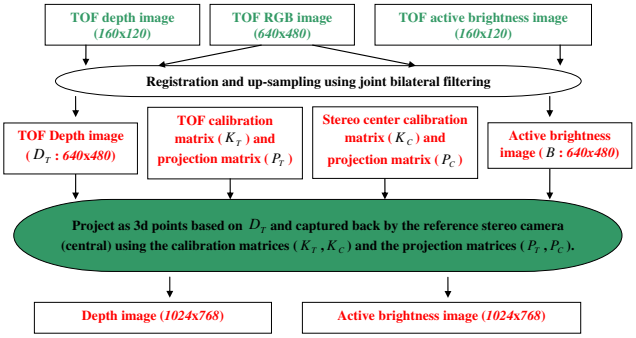


Fig. 5. Up-sample the depth and TOF signal strength images captured from the TOF sensor.

The up-sampled TOF signal strength image provided by the TOF depth sensor indicates the signal strength received at each sensor pixel and allows us to compute a confidence map for the depth image obtained from the TOF depth sensor. Let the TOF signal strength image be B and noise be Gaussian distributed, a confidence value at each pixel location $\mathbf{p}$ can be computed based on B using a Gaussian function:

$$C_T(\mathbf{p}) = exp(-\frac{b^2}{2B(\mathbf{p})^2}), \tag{4}$$

where b is a constant used to control the shape of the confidence function.

C. Confidence-Guided Plane-Sweeping Stereo

We finally integrate the contribution of the passive and TOF sensors via confidence-guided plane-sweeping stereo. Plane-sweeping stereo tests a family of plane hypotheses and records for each pixel in a reference view the best plane using some dissimilarity measure. The algorithm works with any number of cameras, and images need not be rectified. In our experiments, the z direction of the reference (central) camera is used. For each plane $\mathbf{d}$, the images captured from the left and right stereo sensors are projected on the current plane $\mathbf{d}$, and rendered in the reference review (central) as $I_L^{\mathbf{d}}$ and $I_R^{\mathbf{d}}$. Let I_C be the image captured by the reference stereo sensor, the matching cost for each pixel location $\mathbf{p}$ at the reference image is then computed for the left and right stereo sensor as:

$$M_L(\mathbf{p},\mathbf{d}) = ||I_L^{\mathbf{d}}(\mathbf{p}) - I_C(\mathbf{p})||, \tag{5}$$
$$M_R(\mathbf{p},\mathbf{d}) = ||I_R^{\mathbf{d}}(\mathbf{p}) - I_C(\mathbf{p})||. \tag{6}$$

The coarse left and right matching cost $M_L(\mathbf{p},\mathbf{d})$ and $M_R(\mathbf{p},\mathbf{d})$ are very noisy, either local or global optimization methods can be applied to them separately for de-noising, e.g., box filter, bilateral filter, joint bilateral filter, symmetric joint bilateral filter, loopy belief propagation, graph cuts. In our implementation, we used box filter which is the fastest.

The matching cost for each pixel is compared at each plane hypothesis, and the smaller one is selected as correct:

$$M(\mathbf{p},\mathbf{d}) = \min(M_L(\mathbf{p},\mathbf{d}), M_R(\mathbf{p},\mathbf{d})). \tag{7}$$

A stereo confidence map can then be computed based on local image features. Specifically, assume that the cost is perturbed by Gaussian noise and $\mathbf{d}_S(\mathbf{p}) = \arg\min_{\mathbf{d}} M(\mathbf{p},\mathbf{d})$ is the plane corresponding to the lowest matching cost and is the correct depth at pixel location $\mathbf{p}$, we wish to estimate the likelihood that $\mathbf{d}_S(\mathbf{p})$ does not have the lowest cost after the cost is perturbed. This likelihood is proportional to $exp(-\frac{(M(\mathbf{p},\mathbf{d})-M(\mathbf{p},\mathbf{d}_S(\mathbf{p})))^2}{2\sigma_S^2})$ for some σ_S that depends on the strength of the noise. A stereo confidence value at each pixel location $\mathbf{p}$ can then computed as the inverse of the sum of these probabilities for all possible depths:

$$C_S(\mathbf{p}) = (\sum_{\mathbf{d}\neq\mathbf{d}_S(\mathbf{p})} exp(-\frac{(M(\mathbf{p},\mathbf{d}) - M(\mathbf{p},\mathbf{d}_S(\mathbf{p})))^2}{2\sigma_S^2}))^{-1}. \tag{8}$$

Let the depth image up-sampled from the TOF depth sensor be D_T, we combine the stereo sensors and TOF depth sensor by using D_T, the stereo sensor confidence C_S and TOF sensor confidence C_T to update the matching cost:

$$M^F(\mathbf{p},\mathbf{d}) = (1 - W(\mathbf{p}))M(\mathbf{p},\mathbf{d}) + W(\mathbf{p})\min((\mathbf{d} - D_T(\mathbf{p}))^2, \eta), \tag{9}$$
$$W(\mathbf{p}) = \frac{(1 - C_S(\mathbf{p}))C_T(\mathbf{p})}{(1 - C_T(\mathbf{p}))C_S(\mathbf{p}) + (1 - C_S(\mathbf{p}))C_T(\mathbf{p})}, \tag{10}$$

where η is a constant to reject outliers. Let $1 - C_T(\mathbf{p})$ in Eqn. (10) be the stereo confidence from the TOF sensor and $1 - C_S(\mathbf{p})$ be the TOF confidence from stereo sensor, $(1 - C_S(\mathbf{p}))C_T(\mathbf{p})$ is then the fused TOF sensor confidence, $(1 - C_T(\mathbf{p}))C_S(\mathbf{p})$ the fused stereo sensor confidence, and $W(\mathbf{p})$ in Eqn. (10) the normalized fused TOF sensor confidence. The behavior of $W(\mathbf{p})$ with respect to the original TOF sensor confidence $C_T(\mathbf{p})$ and the original stereo sensor confidence $C_S(\mathbf{p})$ is presented in Fig. 6. Note that $W(\mathbf{p}) = 0.5$ when the

confidence obtained from the stereo sensor and TOF depth sensor is the same: $C_S(\mathbf{p}) = C_T(\mathbf{p})$.

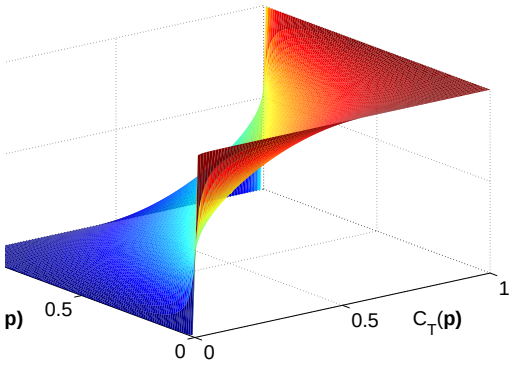


Fig. 6. Behavior of $W(\mathbf{p})$ (Eqn. 10) with respect to the original TOF sensor confidence $C_T(\mathbf{p})$ and the original stereo sensor confidence $C_S(\mathbf{p})$.

Depth values at each pixel location $\mathbf{p}$ are computed by selecting the plane that corresponds to the minimum matching cost:

$$D(\mathbf{p}) = \arg\min_{\mathbf{d}} M^F(\mathbf{p}, \mathbf{d}). \tag{11}$$

Finally, to retain sub-pixel accuracy from the TOF depth sensor, we can assume that the matching cost function is a polynomial function, and sub-pixel accuracy can be obtained by polynomial interpolation [17]. We also use the real-time joint bilateral filtering method presented in [22] to smoothen the obtained depth map.

IV. Experiment

We implemented the algorithm with a GPU acceleration, and some results can be seen in Figure 7. The red and green boxes show where the active sensor fails, and the blue and cyan boxes indicated where passive stereo fails. Our algorithm is able to produce depth maps that are visually superior to those produced by the individual sensor subsystems. We used 1024×768 stereo cameras, and for 48 levels of stereo disparity our algorithm is able to output depth maps at around 8 frame per second.

V. Conclusion and Future Work

We have presented a fast depth sensing system which takes advantage of the complementary nature of passive stereo sensor and active depth sensor in the paper. Our GPU implementation on a NVIDIA Geforce 9800 GTX GPU shows that our system is able to output depth maps at around 8 frame per second for 1024×768 stereo cameras and 48 levels of stereo disparity. An additional contribution of the paper is a real-time depth image up-sampling method which is very useful as the current available active depth sensors are all of very low resolution. An unsolved problem in our system is that it is invalid for large black regions as both the active depth sensor and stereo sensor fail in this case.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1/2/3, pp. 7–42, April-June 2002.

[2] W. T. Freeman, E. Pasztor, and O. T. Carmichael, "Learning low-level vision," *IJCV*, vol. 40, no. 1, pp. 25–47, 2000.

[3] J. Sun, N. Zheng, and H. Y. Shum, "Stereo matching using belief propagation," *PAMI*, vol. 25, no. 7, pp. 787–800, 2003.

[4] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[5] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *PAMI*, vol. 8, no. 4, pp. 413–242, 1986.

[6] J. Batlle, E. Mouaddib, and J. Salvi, "Recent progress in coded structured light as a technique to solve the correspondence problem: A survey," *Pattern Recognition*, vol. 31, no. 7, pp. 963–982, 1998.

[7] P. Besl, *Active Optical Range Imaging Sensors, in Advances in Machine Vision, chapter 1*, 1989, pp. 1–63.

[8] R. Jarvis, "A perspective on range finding techniques for computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 122–139, 1983.

[9] D. Poussart and D. Laurendeau, *3-D Sensing for Industrial Computer Vision, in Advances in Machine Vision, chapter 3*, 1989, pp. 122–159.

[10] J. Salvi, J. Pagès, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[11] T. C. Strand, "Optical three-dimensional sensing for machine vision," *Optical Engineering*, vol. 24, no. 1, pp. 33–40, 1985.

[12] "3dv systems, z-cam," http://www.3dvsystems.com/home/index.html.

[13] S. R. S.-. MESA Imaging AG, "The swiss center for electronics and microtechnology," http://www.csem.ch/fs/imaging.htm.

[14] "Canesta inc, canestavision^{TM} electronic perception development kit," http://www.canesta.com/html/$development_kits$.htm.

[15] "Pmd technologies, pmd s3," http://www.pmdtec.com/.

[16] "Fotonic, fotonic-b70," http://www.fotonic.com/content/News-And-Press/Default.aspx.

[17] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images." in *CVPR*, 2007.

[18] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)*, vol. 26, no. 3, p. to appear, 2007.

[19] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *CVPR*, 2008.

[20] "Middlebury stereo test bed," http://vision.middlebury.edu/stereo/data/.

[21] J. Bouguet, "Matlab camera calibration toolbox," in *Caltech Technical Report*, 2000.

[22] Q. Yang, K.-H. Tan, and N. Ahuja, "Real-time o(1) bilateral filtering," in *CVPR*, 2009, pp. 557–564.

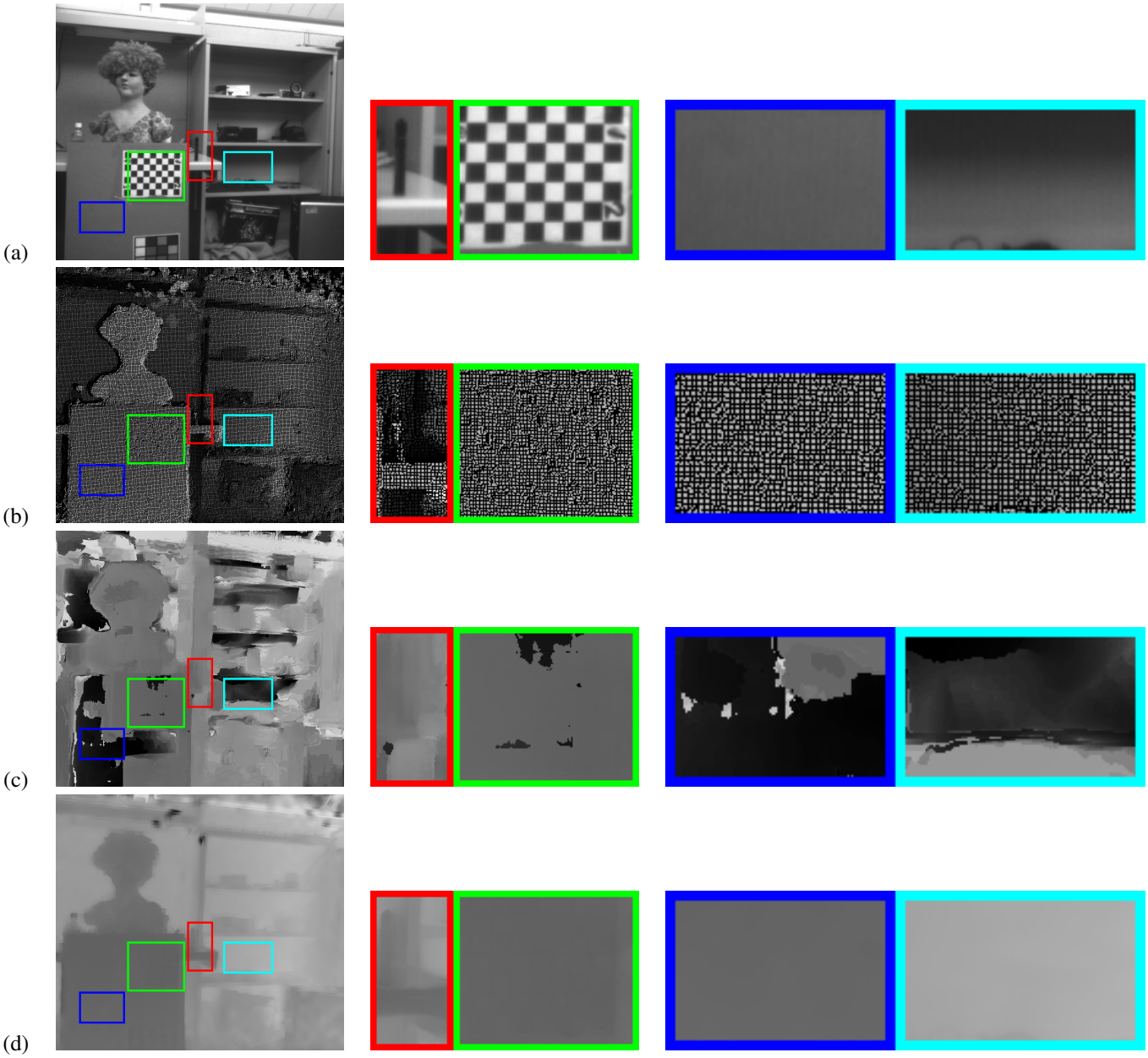


Fig. 7. Visual evaluation of the proposed method. From top to bottom: (a) Reference camera image; (b) Depth map obtained from TOF sensor; (c) Depth map computed from stereo matching; (d)Fused depth map. The red box in (b) shows that active sensor is invalid for thin-structured objects because it is of low resolution. The red box in (d) shows the improvement after sensor fusion. The irregular holes (in black color) inside the green box in (b) are due to the incorrect depth values introduced by the texture variance inside the green box in (a). If the depth values are correct, the holes will be regular as shown in the blue and cyan boxes in (b). The corresponding depth values computed from stereo vision are presented in the green box in (c), which shows that stereo matching is more robust in this situation and can be used to improve the performance of the active sensor as presented in the green box in (d). The uneven depth values in the blue and cyan boxes in (c) shows that stereo matching is very fragile for low-textured regions. However, the depth values obtained from the active sensor is accurate as shown in the blue and cyan boxes in (b). The fused depth values are presented in the blue and cyan boxes in (d). As can be seen, sensor fusion does greatly improve the performance of stereo matching on low-textured regions.