2012 Second Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization & Transmission

Sensor Fusion for Depth Estimation, including TOF and Thermal Sensors

Jeroen van Baar, Paul Beardsley
Disney Research Zurich
Zurich, Switzerland
jeroen@disneyresearch.com

Marc Pollefeys, Markus Gross
Institute for Visual Computing
ETH Zurich
Zurich, Switzerland

***Abstract*—This paper describes the computation of depth maps for a high-quality reference camera augmented by a set of satellite sensors. The satellite sensors include support cameras, a TOF (time-of-flight) sensor, and a thermal camera, all rigidly attached to the reference camera. There is extensive previous work on computing depth maps with stereo alone, and high-quality results have been achieved. However it has proved difficult to achieve good results for cases such as textureless areas, or similar fore- and background colors. We show that with our proposed sensor fusion we can achieve high quality results.**

The paper makes two contributions. The first is a method for combining TOF data with multi-camera data that includes reasoning about occlusions, to produce an improved depth estimate near depth discontinuities. The second contribution is to show the benefit of thermal sensing as a segmentation prior. Thermal cameras were formerly high-cost devices but are now available at the same cost as machine vision cameras. This work demonstrates their advantages, particularly for scenes including humans.

***Keywords*-Computer vision; Image processing; Sensor fusion**

I. Introduction

This paper describes a high quality reference camera augmented with satellite sensors of different modalities. Figure 1 shows our experimental setup.[1]

The goal is to obtain high-quality depth maps. Existing work has shown the advantage of active sensors for measuring 3D information, including TOF sensors and structured illumination. The Kinect has been a landmark device in bringing 3D sensing to the mass market, using a camera plus an IR illumination source that projects a speckle pattern onto the scene. But there are limitations with both approaches - TOF response falls off on an oblique surfaces, as found for example near the occluding contour of a curved object. While structured illumination delivers sparse information, not pixel-dense measurement. Both approaches suffer when the surface is poorly reflective.

The first contribution of this paper is to demonstrate how TOF data is combined with data from multiple cameras to overcome the problem of fall-off and produce a high-quality depth measurement at an occluding contour. The depth map is actually computed for the (high-resolution) reference camera, supported by the lower resolution TOF sensor and support cameras. The method utilizes support cameras on both sides of the TOF sensor in order to do explicit reasoning about occlusions.

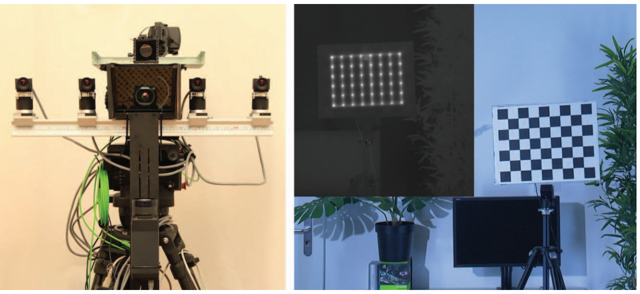

Fig. 1. **Left** The high quality reference camera is at center (obscured by a beam splitter). The four satellite cameras are arranged approximately along a horizontal line with the reference camera. The depth camera is at the top. The beam splitter passes visible light to the reference camera and reflects thermal radiation to the thermal camera at the bottom. **Right** Resistors are accurately mounted below the surface and in registration with the corners of the checkerboard pattern for easy calibration of the thermal and reference camera.

Our second contribution is to demonstrate the advantage of sensor fusion including thermal sensing. Thermal cameras detect emitted heat and aid segmentation by detecting thermal gradients between humans and the background. Human skin, even when covered by some layers of clothing, typically gives a reasonably strong thermal gradient with the background. A beam splitter is used to capture registered images for the reference camera and the thermal camera. Beam splitting with two visible light cameras has a disadvantage that only half the light enters each camera. But beam splitting for a visible light camera and a thermal camera using a K-glass beam splitter results in most of the incident visible light entering the visible light camera, and most of the thermal radiation entering the thermal camera.

Replicating the high quality reference camera to create a (homogenous) multi-camera system may prove very costly. Instead we aim to exploit current hardware advances. The prize of lesser quality cameras continues to fall. In addition those types of cameras continue to miniaturize, for example, small USB board cameras have a weight of 30g and dimensions of a

[1] For a practical system, we would envisage a more compact arrangement of sensors around the reference camera, but the focus of the current work is on the algorithms not the hardware.

978-0-7695-4873-9/12 $26.00 © 2012 IEEE
DOI 10.1109/3DIMPVT.2012.69

couple of cm. New low-end thermal cameras are a similar size and price to machine vision cameras [1]. Building a multiple sensor modalities single system is thus becoming practical. Our experimental setup in Figure 1 is merely a prototype. It is feasible to eventually envisage a compact cinematographic camera augmented with satellite sensors with little impact on its normal handling and workflow.

A. Paper Organization

The remainder of the paper is organized as follows: in Section II we discuss related work; a detailed technical description is given in Section III; results are presented in Section IV followed by a discussion and conclusion (Section V).

II. Related Work

Sensor fusion has received a lot of attention in recent years, especially in combining visual data with Time-of-Flight (ToF) depth data [2], [3], [4]. We similarly exploit ToF data in combination with visual data.

Guan et al. [4] use a combination of cameras and ToF depth sensors to reconstruct 3D objects. Given their setup they can only reconstruct single objects within some bounded volume. We are interested in depth for general (indoor) scenes, typically involving multiple dynamic subjects. Scanning approaches using IR-based depth sensors can reconstruct objects [5] and environments [6], [7]. These approaches are based on a dense sampling with the depth sensor, rather than using data at a single point in time, and therefore do not permit dynamic objects or scenes. Tola et. al [8] capture 3D video using a ToF sensor and registered cameras, by generating a 3D mesh based on the depth values from the depth sensor. The images from the camera are then used as textures for the 3D mesh. This approach does not consider the resolution of the registered cameras. ToF depth sensors have been used together with stereo for enhancing depth maps [9] and also computing alpha mattes [10]. Good results are demonstrated, however the scenes either have only very small volume, or else can the foreground subject can be segmented reliably from the background. We aim our fusion to general scenes with multiple dynamic subjects instead. In addition we show that we can reconstruct depth maps at HD-resolution using a fusion approach.

For an overview of depth (disparity) methods and multi-view stereo see [11], [12]. High quality results can be obtained using global methods based on formulating the depth reconstruction as a Markov Random Field. See [13], [14] for an evaluation of the various methods. The drawback of global methods is their performance, since the problem size depends on the number pixels in the images. Determining the correct parameters to obtain the best results could thus be time consuming. To avoid the running times of global methods, Larsen et al. [15] propose an iterative approach as an approximation to Belief Propagation. They back project the current depth hypothesis onto the cameras in their multi-view setup. We adopt a similar approach in our local method, and show that we can obtain high quality depth maps from fusion of the various modalities. Our method can obtain results which are comparable to global methods in under a minute.

The fact that depth and color discontinuities are often correlated has often been exploited. In [16] the authors propose to oversegment images of video sequences and compute depth maps with consistent object boundaries. Others exploit color discontinuities for weighting in global optimization methods to avoid smoothing over depth boundaries, for example [17]. All these color based approaches fail in areas with lack of (color) contrast, such as textureless areas, but also similarly colored fore- and background objects. In addition to ToF data we therefore also propose to exploit thermal IR.

Fusion with thermal IR has been used in previous work. In [18] the authors propose to exploit thermal IR for tracking humans in video using so-called spatiograms. Their goal is the reliable tracking of occurrences rather than accurate segment boundaries. We exploit thermal IR in the context of computing depth maps for general scenes.

III. Reconstructing Depth Maps

A. Calibration

The intrinsics and extrinsics of the reference and satellite cameras are calibrated using a standard approach [19]. The reference camera and the depth camera are calibrated together using the method described in [20].

A beam splitter is used to capture registered images from the reference camera and the thermal camera. Calibration is done using a custom calibration target with a conventional checkerboard pattern that has heated resistors on the vertices. This enables detection of corresponding features in the reference image and the thermal image. The beam splitter is first physically adjusted to bring the corresponding features into coarse registration, then a homography is computed to produce an accurate registration between the images, following the method of [21].

In addition to geometric calibration, we also perform a color calibration to transform the color spaces of the satellite cameras to that of the reference camera using the method of [22].

B. Depth Maps from Sensor Fusion

This section describes the computation of depth maps for the high quality reference camera. The problem is to obtain a high quality depth map by fusing (a) multi-view stereo (MVS) data from the reference camera plus satellite cameras, with (b) low resolution depth data from the depth camera. We also exploit the thermal signal as we will explain in this section.

The depth computation contains the following steps:

1) Compute the initial data cost for the MVS by sweeping a depth plane through a discretized depth volume [23].
2) Fuse the MVS stereo data with ToF depth data, including re-projection onto the satellite cameras.
3) Perform plane fitting to reduce noise and improve final result.

We compute depth maps based on a discretization of the depth volume into a set of depth layers. For each pixel of the reference image we determine the depth plane with largest support and assign the depth to this pixel. We define support as the maximum of the combined contribution of MVS stereo, ToF depth and re-projection onto the satellite images. The support per pixel can then be formulated as:

$$E_p(x) = w_{st} \cdot E_{p,st} + w_{ToF} \cdot E_{p,ToF} + w_{re} \cdot E_{p,re}, \quad (1)$$

where st stands for stereo, and re for reprojection on the satellite cameras.

MVS Stereo: To determine the contribution from MVS stereo, a plane is swept through the depth volume, and the images from the satellite images are projected onto this plane at each depth. We then determine the agreement between the pixel of the reference image with pixels of the satellite images using normalized cross correlation (NCC). One reason for choosing NCC stems from the fact that the we use heterogenous cameras as the reference and satellite cameras. NCC provides robustness given the different resolutions and field of views of the cameras. The adaptive NCC proposed by Heo et al. [24] provides further robustness. The NCC is computed both for a log-space image as well as an RGB image. Compared to *standard* NCC, the correlation is normalized according to a bilateral weighting. This provides better localization of intensity discontinuities. The contribution E_{stereo} is then finally computed as $1 - max(NCC, 0)$.

Occlusion Reasoning: We have two satellite cameras on each side of the reference camera. This allows initial occlusion reasoning by comparing the NCC results between the two left cameras, and the two right cameras. If their contributions differ by more than some threshold NCC_{thresh} we discard the NCC value that is furthest from 1. In our examples we use $NCC_{thresh} = 0.75$.

ToF Depth: For a pixel in the reference camera, the contribution of the ToF camera is determined by the difference of the depth plane D_i from discretization and the depth value reported by the ToF camera at the corresponding pixel. As described in [20], a 3D point P is first transformed to the ToF camera's coordinate space, and then projected onto its image plane. We then look up the associated depth value D_{ToF}. The contribution for the ToF depth data is then determined according to:

$$E_{ToF} = \exp(s \cdot min(|D_i - D_{ToF}|, \tau)/\sigma_{ToF}) \quad (2)$$

with $s = 1/\tau$. In our case the ToF camera has a range of 5 meter, and we obtained good results with $\tau = 0.1m$.

Reprojection onto Satellite Cameras: The contribution from stereo is computed based only on the satellite cameras. Because we combine with ToF depth in the fusion step, we further aim to improve the support for each pixel at a particular depth by reprojecting the associated 3D point onto the satellite cameras. The depth camera is a time-of-flight sensor that measures phase shifts of reflected modulated IR illumination. As discussed earlier, it is less reliable near depth discontinuities where the IR signal strikes an oblique surface so that there is a reduction in the signal reflected back at the sensor. This is an important step to improve the overall quality of the depth map.

Given a point P representing a pixel in the reference image on the current depth plane, we reproject P onto the satellite camera image planes using the projection matrices computed during calibration of the setup. We compare the color of the associated pixel in the satellite camera, with that of the pixel in the reference image. The final contribution $E_{p,re}$ to equation 1 is then:

$$E_{p,re} = \exp(-\left(\sum_{i=0}^{n}\left(\frac{|RGB_{ref} - RGB_{sat}|}{3}\right)\right)/\sigma_{re}) \quad (3)$$

Occlusion Reasoning: Additional occlusion reasoning can be performed again during this step. We compare the color similarity between the left and right satellite cameras. If the difference is larger than some threshold we only consider the contribution from the lesser of the two. Equation 3 then becomes:

$$E_{p,re} = \begin{cases} \exp(-L/\sigma_{re}), & \text{if } L - R > \tau \\ \exp(-R/\sigma_{re}), & \text{if } R - L > \tau \\ \exp(-(L+R)/\sigma_{re}), & \text{otherwise} \end{cases} \quad (4)$$

with L and R corresponding to equation 3 for the left and right cameras.

C. Plane Fitting for Improving Depth Estimates

The initial depth map can be obtained by evaluating equation 1 and apply the winner-take-all strategy of selecting the depth with largest support. The resulting depth map may be very noisy. To reduce this noise, we perform an iterative plane fitting approach. We emphasize that this is not a global method. Rather, we perform an iterative, somewhat conservative, plane fitting approach which aims to retain some details obtained by evaluating equation 1 at the expense of allowing some noise to remain in the depth map.

We first segment the reference image into regions using the method described in [25]. We extended this method to incorporate the thermal signal in addition to the color information. By also considering the thermal signal we aim to correctly segment along object boundaries in areas with similar fore- and background colors. For each segment we then perform the following steps:

1) Pick four pixels and fit least squares plane to the corresponding depth values.
2) Determine inliers and outliers for this plane hypothesis.
3) For the outliers, re-estimate the depths according to the plane hypothesis.
4) Reproject 3D points associated with the re-estimated depths onto the satellite cameras.
5) Determine if re-estimated depths are acceptable.

If a scene is well approximated by planar segments, plane fitting approaches can achieve very high quality results. In general however scenes may have cluttered, or detailed objects in the background which are not well approximated by planes. Our approach is therefore to apply plane fitting conservatively.

We sample the depth values in a segment according to a RANSAC [26] approach. We discard plane hypotheses which are slanted more than 45° degrees as these are very likely wrong estimates. We also omit planes which do not have enough inliers. This means that for some segments we cannot reliably estimate planes and we leave the initial depth values as is.

For the cases when a segment is well approximated by a plane, we compute the depths for the pixels classified as outliers according to the fitted plane. We then reproject the corresponding 3D points onto the satellite to check if the newly estimated depths are consistent with the color information from the satellite images. If the depths are not consistent with the color information we discard the plane.

Occlusion Reasoning: We again take into account that we may be dealing with an occlusion region. We separately check for color consistency for the left and right satellite cameras. However in addition we also check if the pixel in a satellite image is occluded by comparing the depth value. For each segment which is well represented by a plane, we store the depth value at the reprojected locations in the satellite images. We overwrite the current depth value whenever a smaller depth value reprojects to a particular satellite image pixel. We can either iterate this approach for a fixed number of iterations, or until the number of segments that are updated is below some threshold.

D. Smoothing

The final step is to smooth the computed depth maps using the trilateral filtering described in [27]. We trivially extend the trilateral filter to incorporate the thermal signal as well. In addition we also incorporate the dis-/similarity between neighboring segments in the filter. As we will discuss in the results section, this smoothing step is especially effective if the depth discontinuities have been accurately estimated. We will show that for some challenging example scenes our sensor fusion can obtain such accurate estimates.

E. Comparative Analysis

The ideal way to evaluate the system would be to compare the results with ground-truth. Laser scanners offer millimeter accuracy 3D measurements, but are unable to handle dynamic scenes because of the scanning time of the laser. Thus accurate ground-truth is not readily available.

As an alternative, we computed depth maps using a stereo pair of two multi-spectral cameras that are capable of recording in RGB and IR, plus several Kinects to project IR speckle pattern onto a scene. Conventional stereo matching is carried out on the images, with the IR speckle providing useful texture for to disambiguate the matching on homogeneous surfaces. Color gradients from the registered RGB images can be used to impose first order smoothing priors. This approach has its own limitations, but it does provide a basis for showing how our results compare against another method which is expected to provide high-quality depth. We present a comparison in the next section.

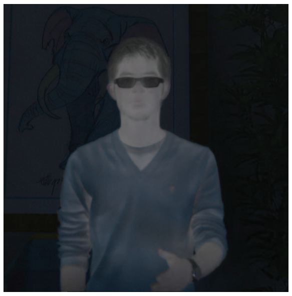

Fig. 3. Thermal image superimposed on the reference camera color image. The thermal signal is high for humans, especially for exposed skin, compared to many inanimate objects.

IV. Results

A. Hardware

Figure 1 shows the camera rig. The high quality camera is a 3-CMOS Sony PWM-350K capturing 720p HD; the satellite cameras are Point Grey Grasshoppers, resolution 1600x1200; the depth camera is a Swiss Ranger SR4000, resolution 176x144; the thermal IR camera is a FLIR SC-645, resolution 640x480. All devices capture at 25fps and are synced using an external synchronization trigger.[2] Example images captured with the rig at one time instant are shown in Figure 2. An example of the registration between the reference RGB image and the thermal image is shown in Figure 3, with the thermal image superimposed on the color image.

B. Fusion via Local Method

For all the results presented in this section we used 50 depth layers to discretize the depth volumes. In Figures 4 and 5 we present results obtained with the method discussed in Section III. Textureless areas which are difficult for stereo are reconstructed correctly in the fusion with the ToF depth. Exploiting discontinuities in the thermal signal, the hair and fingers of the subjects in Figures 4 are correctly reconstructed as well.

By exploiting the thermal signal, the segmentation can be correct when there is no color discontinuity present. The left column of Figure 6 shows a zoomed-in area from the color image (for the example from the left column of Figure 4). Along the boundary between the subjects' hair there is no distinct color difference. The thermal signal in the middle column shows a clear difference, and the segmentation correctly segments along the boundary as shown in the right column. The the depths are correctly reconstructed as a result (see Figure 4).

C. Comparative Analysis

We compare our local method to a global method which optimizes the depth reconstruction formulated as a MRF. Figure 7 shows the results obtained using Tree-reweighted

[2] The thermal camera is not triggered on each frame, but its stream is started on the external trigger and then runs synchronous at 25fps.

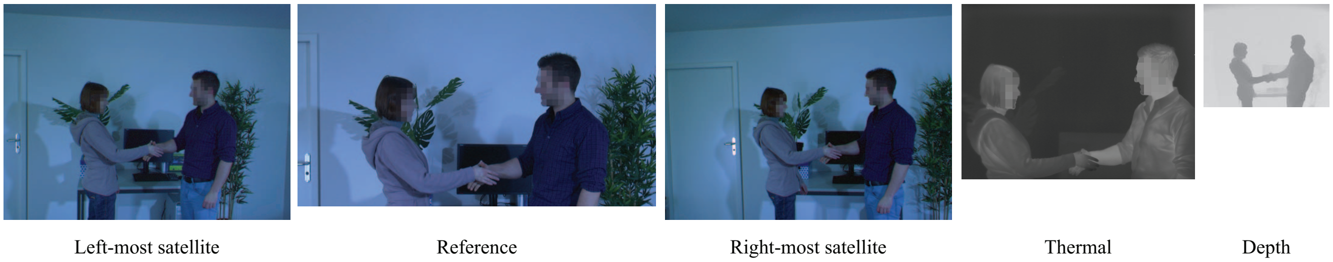

Fig. 2. Example images from different cameras at one time instant. **Top -** depth camera. **Center -** cinematographic camera. **Left and right -** the outermost satellite cameras. **Bottom -** thermal camera.

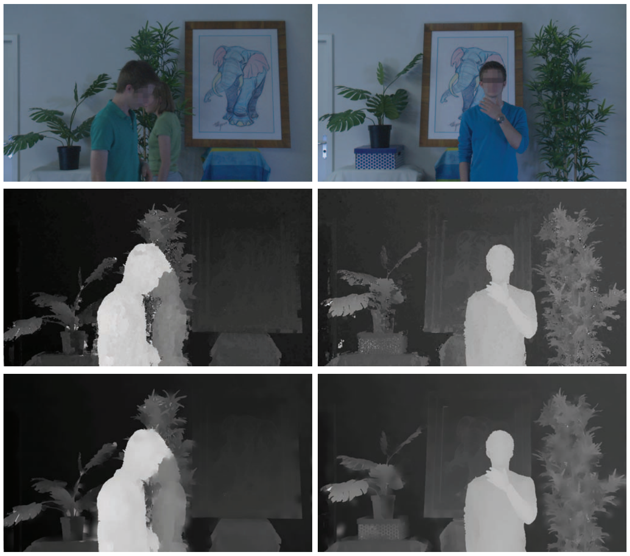

Fig. 4. **First row** Reference camera image. **Second row** Depth map before smoothing. **Third row** Depth after trilateral smoothing. In the first column the hair of the two subjects is reconstructed correctly. In the second column the hand, including fingers, is accurately reconstructed.

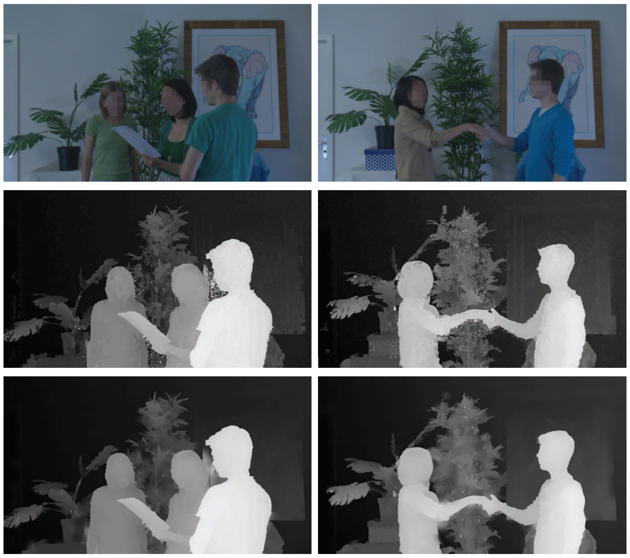

Fig. 5. **First row** Reference camera image. **Second row** Depth map before smoothing. **Third row** Depth after trilateral smoothing. Additional examples showing results for challenging scenes including textureless areas, and cluttered background (plants).

message passing (TRW) [28]. The left image shows the result when the thermal signal is incorporated as a smoothness prior. The right image shows the result without. The depth map is notably wrong in the region where the subjects' hair overlap.

As noted in the previous section, we simultaneously capture IR patterns projected with Kinects and capture the scene using multi-spectral cameras. We again emphasize that this is not a ground truth comparison. We reconstruct the scene using semi-global matching [29] on the IR images. We calibrate the multi-spectral cameras with respect to the rig which allows to reproject the resulting depth map onto the reference camera image, shown in the top row of Figure 8. The holes in the depth map are due to occlusion between the multi-spectral stereo cameras and reprojection onto the high resolution reference image. The quality of the depth obtained with our fusion approach for the foreground subject compares to depth obtained from stereo on the projected IR patterns. In the background the depths have different values. This is due to the fact that the ToF depth data contributes most in these textureless areas. The depth values reported by the ToF sensor at further distances are slightly different from real world measurements.

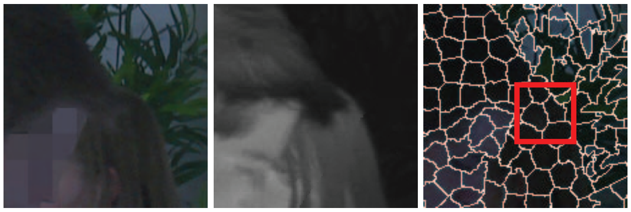

Fig. 6. **Left** Color image. **Middle** Thermal image. **Right** Segment boundaries. Although there is no clear color discontinuity between subjects' hair, the thermal image clearly shows a different signal. The segmentation is able to segment along the thermal boundary.

D. *Comparison of Modalities*

To give a better idea of the contribution of each modality, we omitted the reprojection onto the satellite cameras for the

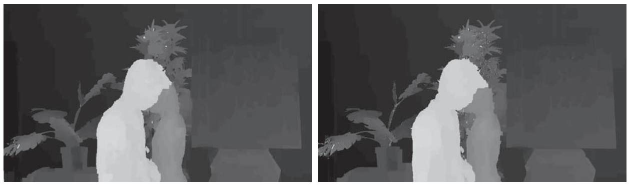

Fig. 7. Comparison with global MRF method (Tree-reweighted message passing) for computing depths. **Left** Without thermal segmentation prior. **Right** With thermal segmentation prior. For the latter case, the boundary for the foreground objects' hair is incorrectly reconstructed, due to the lack of color gradient.

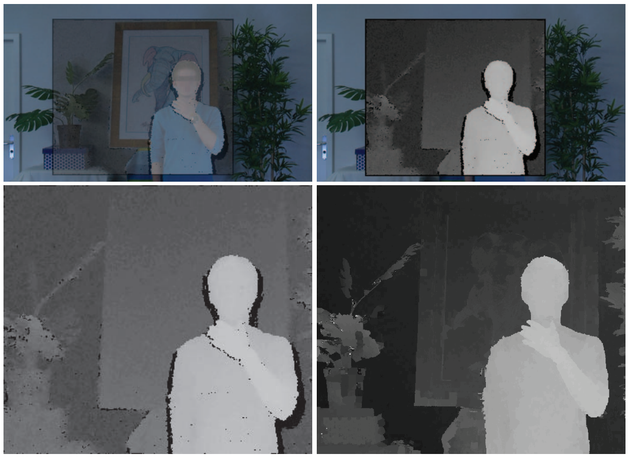

Fig. 8. Comparison with stereo from projecting IR speckle patterns. The depth map using IR patterns is obtained with additional multi-spectral cameras. The top row shows the reprojection onto the reference camera of our experimental rig. The bottom row compares the depth obtained by stereo from the IR patterns with our approach.

examples shown next. The depth maps were reconstructed using TRW. Figure 9 compares the results for stereo (RGB), stereo + thermal (RGB+T), stereo + depth (RGB+D) and stereo + depth + thermal (RGB+D+T), in rows two through five for several example scenes. The TRW parameters were fixed for all results. In the first column of Figure 9, (RGB) and (RGB+T) yield equivalent results. When the information from the depth camera is fused, the space between the two foreground subjects is reconstructed at the correct depth. Finally, in row four, (RGB+D+T) preserves the shape of the nose for the foremost subject. For the second column, when fusing the information from the depth camera, the paper leaflet is no longer being reconstructed compared to the (RGB) and (RGB+T) cases. This, together with the missing leaves for the plants in the background, demonstrates the problem of thin structures for the depth camera. In the third column of Figure 9, the hand shape is better preserved for both cases where thermal is considered in the smoothness prior. In the last column of Figure 9 the plant pot is reconstructed accurately along its boundary when thermal is considered.

V. Discussion and Conclusion

This paper describes a reference camera augmented with satellite sensors. We described a straightforward method to fuse the modalities and compute high quality depth maps for the reference camera. The performance for the results presented in the previous section is less than one minute per frame. Further optimization may reduce this to several seconds per frame.

Experimental results were shown for scenes with dynamic objects and background clutter. Large textureless areas such as background walls are handled by fusing depth from stereo with ToF depth. We also showed cases where surfaces of the same color overlapped at different depths. Of particular interest is the case where human subjects or body parts are overlapping - we showed that different subjects may have different thermal signatures, and therefore an occluding contour can be found even though the skin color is similar. We provided additional comparisons to show the contribution of each modality separately. Furthermore we compared our approach to some other techniques of computing depth maps, and showed that we obtain comparable results.

Future work will focus on exploiting the multiple modalities to generate temporally smooth depth for video sequences. Temporally smooth depth maps require accurate motion / flow estimation. As depth sensors continue to improve, the depth data can be directly exploited for computing accurate motion information.

Finally, we argued that the required satellite sensors for this system are becoming more compact and more low-cost. It is feasible to envisage a compact clip-on device that attaches to a cinematographic-quality reference camera, to enable robust and accurate computation of depth maps.

References

[1] ICI, "Centurion thermal camera," www.infraredcamerasinc.com, 2009.

[2] M. Lindner, M. Lambers, and A. Kolb, "Sub-pixel data fusion and edge-enhanced distance refinement for 2d/3d images," *Int. Journal of Intell. Systems Tech. and Appl.*, vol. 5, no. 3, 2008.

[3] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Pattern Anal. and Mach. Intell.*, vol. 31, no. 3, 2009.

[4] L. Guan, J.-S. Franco, and M. Pollefeys, "3d object reconstruction with heterogeneous sensor data," *3DPVT08*, 2008.

[5] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *IEEE Comp. Vision and Pattern Recog. Workshops (CVPRW)*, 2008.

[6] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS*. MIT Press, 2005.

[7] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *ACM Symp. on User Interf. Softw. and Tech. (UIST)*, 2011. [Online]. Available: http://doi.acm.org/10.1145/2047196.2047270

[8] E. Tola, C. Zhang, Q. Cai, and Z. Zhang, "Virtual view generation with a hybrid camera array," *EPFL technical report*.

[9] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," *IEEE Comp. Vision and Pattern Recog. (CVPR)*, 2008.

[10] J. Zhu, M. Liao, R. Yang, and Z. Pan, "Joint depth and alpha matte optimization via fusion of stereo and time-of-flight sensor," in *IEEE Comp. Vision and Pattern Recog. (CVPR)*, 2009.

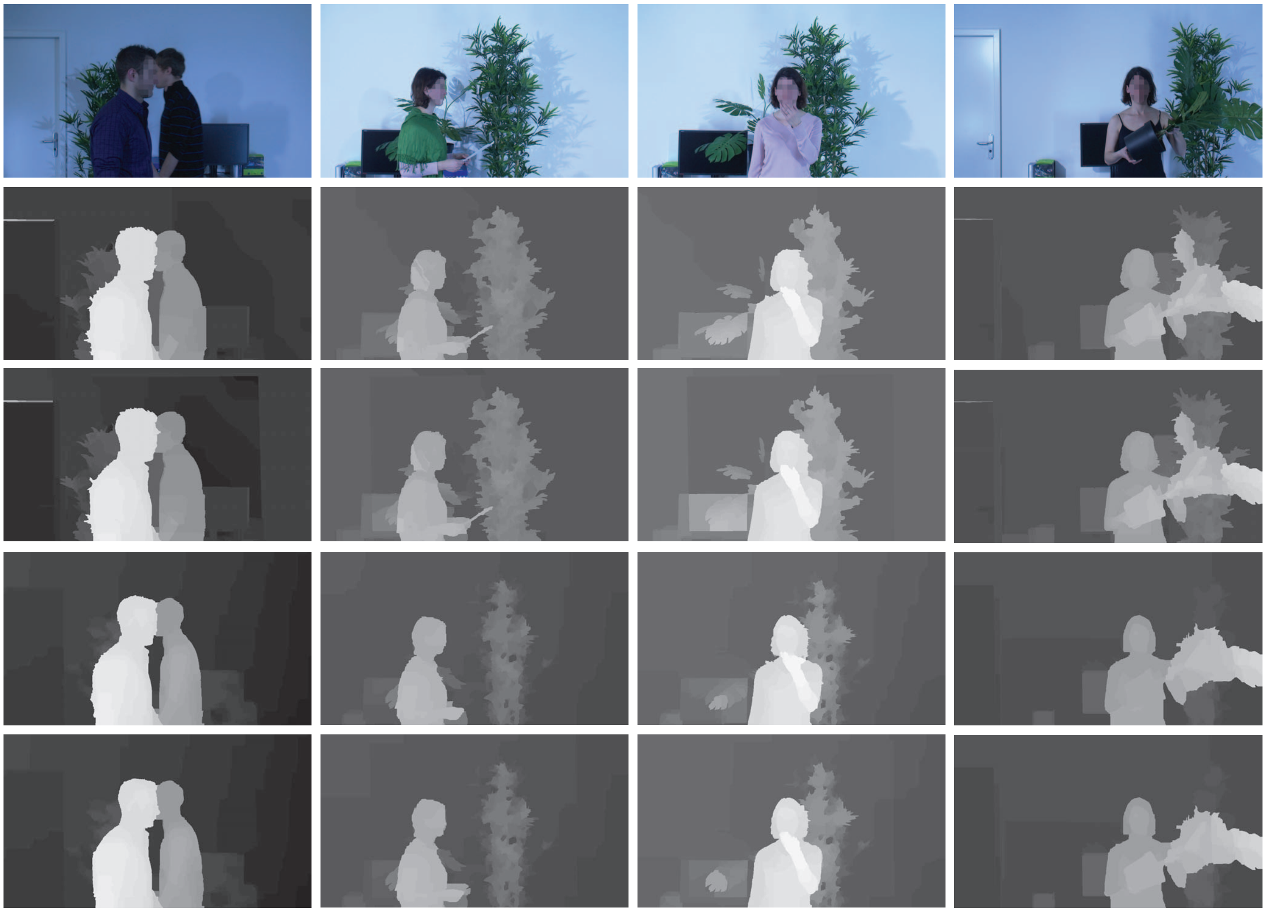

Fig. 9. Several example scenes. Comparison between stereo (2nd row), stereo + thermal (3rd row), stereo + depth (4th row), and finally stereo + depth + thermal (5th row). The first row shows the reference camera input images.

[11] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. Journal Comp. Vision*, vol. 47, no. 1, 2002.

[12] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *IEEE Comp. Vision and Pattern Recog. (CVPR)*, vol. 1, 2006.

[13] Miscellaneous, “”middlebury stereo evaluation”,” ”http://vision.middlebury.edu/stereo/eval/”, retrieved: June 2012.

[14] ——, “”middlebury multi-view stereo evaluation”,” ”http://vision.middlebury.edu/mview/eval/”, retrieved: June 2012.

[15] E. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs, “Simplified belief propagation for multiple view reconstruction,” *3D Data Process., Visual., and Transm. (3DPVT)*, 2006.

[16] C. L. Zitnick and S. B. Kang, “Stereo for image-based rendering using image over-segmentation,” *Int. Journal of Comp. Vision*, vol. 75, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1285519.1285520

[17] P. Felzenszwalb and D. Huttenlocher, “Efficient belief propagation for early vision,” *Int. Journal of Comp. Vision*, vol. 70, no. 1, 2006.

[18] C. O. Conaire, N. E. O’Connor, and A. F. Smeaton, “Thermo-visual feature fusion for object tracking using multiple spatiogram trackers,” *Mach. Vision Appl.*, vol. 19, September 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1416799.1416803

[19] Y.-Y. Bouquet, “Camera calibration toolbox,” 2012.

[20] Q. Yang, R. Yang, J. Davis, and D. Nister, “Spatial-depth super resolution for range images,” *IEEE Comp. Vision and Pattern Recog. (CVPR)*, 2007.

[21] C. Conaire, N. O’Connor, E. Cooke, and A. F. Smeaton, “Comparison of fusion methods for thermo-visual surveillance tracking,” *9th Int. Conf. on Information Fusion*, 2006.

[22] A. Ilie and G. Welch, “Ensuring color consistency across multiple cameras,” *IEEE Int. Conf. on Comp. Vision (ICCV)*, 2005.

[23] R. Yang and M. Pollefeys, “Multi-resolution real-time stereo on commodity graphics hardware,” *IEEE Comp. Vision and Pattern Recog. (CVPR)*, 2003.

[24] Y. S. Heo, K. M. Lee, and S. U. Lee, “Robust stereo matching using adaptive normalized cross-correlation,” *IEEE Pattern Anal. and Mach. Intell.*, vol. 33, no. 4, 2011. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2010.136

[25] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “Slic superpixels,” *Technical Report 149300 EPFL*, no. June, 2010.

[26] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Comm. of ACM*, vol. 24, no. 6, Jun. 1981. [Online]. Available: http://doi.acm.org/10.1145/358669.358692

[27] B. Smith, L. Zhang, and H. Jin, “Stereo matching with nonparametric smoothness priors in feature space,” *IEEE Comp. Vision and Pattern Recog. (CVPR)*, 2009.

[28] V. Kolmogorov, “Convergent tree-reweighted message passing for energy minimization,” *IEEE Pattern Anal. and Mach. Intell.*, vol. 28, no. 10, 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2006.200

[29] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, 2008. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.1166