Deep Learning for Confidence Information in Stereo and ToF Data Fusion

Gianluca Agresti, Ludovico Minto, Giulio Marin, Pietro Zanuttigh
Department of Information Engineering
University of Padova, Italy
{agrestig,mintolud,maringiu,zanuttigh}@dei.unipd.it

***Abstract*—This paper proposes a novel framework for the fusion of depth data produced by a Time-of-Flight (ToF) camera and a stereo vision system. The key problem of balancing between the two sources of information is solved by extracting confidence maps for both sources using deep learning. We introduce a novel synthetic dataset accurately representing the data acquired by the proposed setup and use it to train a Convolutional Neural Network architecture. The machine learning framework estimates the reliability of both data sources at each pixel location. The two depth fields are finally fused enforcing the local consistency of depth data taking into account the confidence information. Experimental results show that the proposed approach increases the accuracy of the depth estimation.**

I. Introduction

There exist many different devices and algorithms for real-time depth estimation including active lighting devices and passive systems exploiting only regular cameras. The first family includes structured light cameras and Time-of-Flight (ToF) sensors while the most notable example of the second family are the stereo setups. None of these solutions is completely satisfactory, active devices like Time-of-Flight and structured light cameras are able to robustly estimate the 3D geometry independently of the scene content but they have a limited spatial resolution, a high level of noise and a reduced accuracy on low reflective surfaces. Passive stereo vision systems, although widely used for the simple technology, have various limitations, in particular their accuracy strongly depends on the scene content and the acquisition is not very reliable on uniform or repetitive regions. On the other side, passive stereo vision systems have a high resolution and a limited amount of noise. The characteristics of the two approaches are complementary and the fusion of data from the two systems has been the subject of several research studies in the last years.

This paper proposes a depth estimation algorithm combining together stereo and ToF data. An effective solution for this task needs two fundamental tools: the estimation of the reliability of the data acquired by the two devices at each location and a fusion algorithm that exploits this information to properly combine the two data sources. The reliability of ToF data has traditionally being estimated by using noise models for these sensors [1]. ToF sensors are typically affected by various sources of noise. Shot noise can be estimated from the amplitude and intensity of the received signal, but, evaluating depth estimation issues specific of their working principles like the mixed pixels and the multi-path error is more challenging. In particular the latter, due to light rays scattered multiple times before reaching the sensor, is very difficult to be directly estimated and removed. In this work instead we use a deep learning framework to estimate confidence data for ToF information.

Stereo data confidence is typically estimated with different metrics based on the analysis of the shape of the cost function [2]. These metrics capture the effects of the local matching cost computation, but most recent stereo vision techniques exploit complex global optimization schemes whose behavior is not captured by standard metrics. For this reason, coherently with the approach used for ToF data, we exploit a deep learning framework also for stereo confidence information.

Finally we use an extended version of the Local Consistency (LC) framework able to exploit the confidence data [3], [4] to perform the fusion of the two data sources.

The proposed algorithm starts by reprojecting ToF data on the stereo camera viewpoint and upsamples the data to the spatial resolution of the stereo setup by using a combination of segmentation clues and bilateral filtering [3]. Then confidence information for ToF depth data is estimated. For this task we developed an ad-hoc Convolutional Neural Network (CNN) that takes in input multiple clues, i.e., the stereo and ToF disparities, the ToF amplitude and the difference between the left image and the right one warped according to disparity information, providing a hint of the stereo matching accuracy, and jointly estimates both stereo and ToF confidence measures.

It is customary that the training of CNNs requires a good amount of data with the corresponding ground truth information. At the time of writing there are no available datasets collecting these data and furthermore the acquisition of accurate ground truth data for real world 3D scenes is a challenging operation.

For this reason we rendered 55 different 3D synthetic scenes using $Blender$ [5] with examples of various acquisition issues including reflections and global illumination. Realistic stereo and ToF data have been simulated on the rendered scenes using LuxRender [6] and a simulator realized by Sony EuTEC starting from the simulator presented

by Meister et al. [7]. This dataset, that represents another contribution to this paper, has been used to train the CNN that proved to be able to accurately estimate a confidence measure for both stereo and ToF depth acquisitions even in challenging situations like mixed pixels on boundaries and multi-path artifacts affecting ToF estimations.

Finally, the two data sources are fused together. The proposed fusion algorithm has been derived from [4]. The framework extends the LC method [8] to combine the confidence measures of the data produced by the two devices. It computes a dense disparity map with subpixel precision by combining the two sources of information enforcing the local consistency of the measures weighted according to the computed confidence information.

The next section summarizes the related works in Section II. Then, Section III introduces the general architecture of the approach. Section IV describes the deep learning network used to compute confidence information. The fusion algorithm is described in Section V. The synthetic dataset is described in Section VI and the results are finally discussed in Section VII. Section VIII draws the conclusions.

II. Related Works

Stereo vision systems can estimate depth data from two standard images by exploiting the well known triangulation principle. A significant amount of research studies focused on this family of 3D data acquisition systems and a detailed review can be found in [9]. The depth estimation accuracy of these systems depends on many factors, including not only the specific matching algorithm used to estimate the disparity map but also the photometric content of the scene. In particular, the estimation is prone to errors in regions with fewer details, e.g. a planar wall with no texture, or repetitive patterns. For this reason it is important to estimate the reliability of the computed data. An exhaustive review about techniques for confidence estimation in stereo vision system can be found in [2]. Notice how the confidence information for stereo systems used to be computed with deterministic algorithms based on the analysis of the matching cost function and only recently deep learning approaches have been exploited for this task [10], [11], [12].

ToF cameras have also attracted the attention of the research community working on depth acquisition systems [1], [13], [14], [15], [16], [17], since they can acquire depth information in real-time and many low cost devices using ToF principles are currently available in the consumer market. Differently from stereo vision systems, ToF cameras can estimate accurately the depth values also in regions without texture or with repetitive patterns since they don't rely uniquely on the scene content for depth estimation. On the other side these devices have various limitations as the low resolution and high noise levels. Furthermore, they are affected by systematic errors as multi-path interference, mixed pixels and noisy estimation on low reflective regions. In [16] it is possible to find a detailed analysis of the various error sources and [17] focuses on the effects of the low reflectivity of the scene on the depth estimation.

ToF cameras and stereo vision systems are based on different depth estimation principles and they have complementary strenghts and weaknesses, therefore a fusion of the data acquired from the two sources can lead to a more reliable depth estimation. Different works on stereo-ToF depth fusion can be found in the literature, e.g., [18] and [1] present two complete reviews of the different approaches.

The combination of a ToF camera with a stereo vision system in order to estimate and then fuse two depth maps of the scene has been used in several works [17], [19], [20], [21]. In order to perform the fusion Zhu et Al. [22], [23], [24] used a MAP-MRF Bayesian formulation where a global energy function is optimized with belief propagation. Dal Mutto et Al. [25] also used a probabilistic formulation and computed the depth map with a ML local optimization. In a more recent version of the approach [26] a global MAP-MRF optimization scheme has been used in place of the ML optimization. Nair et Al. [27] instead used a variational fusion framework. An interesting contribution of this approach is the use of confidence measures for the ToF and stereo vision systems in order to control the process. Evangelidis et Al. [28] estimate the depth information by solving a set of local energy minimization problems. Another solution [3] consists in using a locally consistent framework [8] to fuse the two data sources. This approach has been improved in [4] by extending the LC framework driving the fusion process with the depth map confidences in order to take in account the different nature and reliability of the data sources. In this paper we started from the idea in [4], but instead of using heuristic cues to compute the confidence data we present a CNN framework for confidence estimation.

A strictly related problem is the fusion of the data delivered by a ToF camera with that from a single color camera [29], [30], [31], [32], [33], [34]. For this task different strategies have been proposed, including methods based on bilateral filtering [30], [31], edge-preserving schemes [33] and methods exploiting confidence information for ToF data [32].

III. Proposed Method

The considered acquisition system is made of a ToF camera and a stereo vision system each producing an estimation of depth data from the corresponding viewpoint. The goal of the proposed method is to provide a dense depth map from the point of view of one of the color cameras of the stereo setup.

We will assume that the two acquisition systems have been jointly calibrated (this can be done using ad-hoc techniques for this kind of setups, e.g., the one proposed in [25]). The algorithm includes four steps (see Figure 1):

1) The depth information acquired from the ToF sensor is firstly reprojected to the reference color camera viewpoint. Since ToF sensors have typically a lower

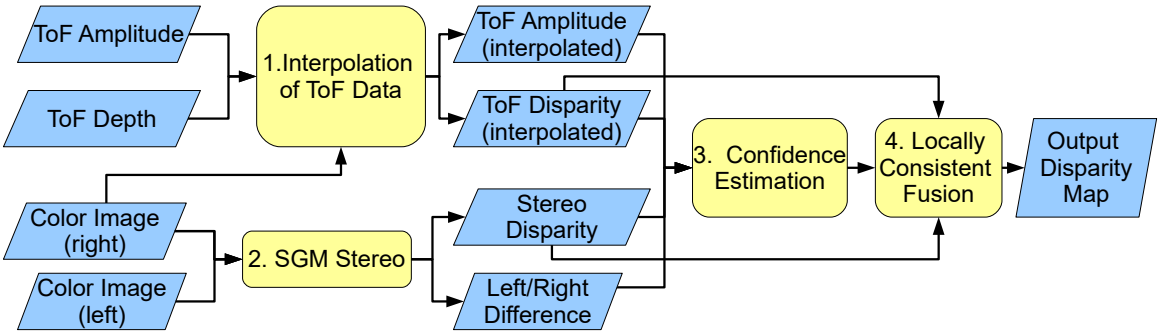


Fig. 1: Flowchart of the proposed approach.

resolution than color cameras, it is also necessary to interpolate the ToF data. For this task we used the approach proposed in [3]. More in detail the color image is firstly segmented using Mean-Shift clustering [35] and then an extended cross bilateral filter with an additional segmentation-based term besides the standard color and range ones is used to interpolate the data and to produce a high resolution depth map aligned with the color camera lattice. Since the fusion algorithm works in disparity space, the computed depth map is also converted to a disparity map with the well known inversely proportional relationship between the two quantities. For more details on the ToF reprojection and upsampling we refer the reader to [3].

2) A high resolution depth map is computed from the two color views by applying a stereo vision algorithm. The proposed approach is independent of the stereo algorithm used to compute the disparity map, however for the current implementation we used the Semi-Global Matching (SGM) algorithm [36]. This algorithm performs a 1D disparity optimization on multiple paths that minimizes an energy term made of point-wise or aggregated matching costs and a regularization term.
3) Confidence information for the stereo and ToF disparity maps are estimated using the Convolutional Neural Network architecture of Section IV.
4) The upsampled ToF data and the stereo disparity are finally fused together using an extended version of the LC technique [8] as described in Section V.

IV. Learning the Confidence with Deep Networks

We use a Convolutional Neural Network (CNN) model to estimate the confidence information that will be used in the fusion algorithm of Section V. The proposed CNN takes as input both ToF and stereo clues and outputs the confidence map for each of the two devices.

The input data are extracted from the ToF and stereo disparity maps, the ToF amplitude and the color images by applying a simple and fast preprocessing step. For any given scene i in the dataset the following data are considered:

- $D_{T,i}$ the ToF disparity map reprojected on the reference camera of the stereo vision system.
- $A_{T,i}$ the ToF amplitude image reprojected on the reference camera of the stereo vision system.
- $D_{S,i}$ the stereo disparity map.
- $I_{R,i}$ the right stereo image converted to greyscale.
- $I_{L',i}$ left stereo image converted to grayscale and reprojected on the right camera using the disparity computed by the stereo algorithm.

The first clue, $\Delta'_{LR,i}$, is extracted from the left and right greyscale images $I_{L',i}$ and $I_{R,i}$ in a two-step procedure. First, the absolute difference between their scaled versions is computed:

$$\Delta_{LR,i} = \left| \frac{I_{L,i}}{\mu_{L,i}} - \frac{I_{R,i}}{\mu_{R,i}} \right| \tag{1}$$

where the scaling factors $\mu_{L,i}$ and $\mu_{R,i}$ are the averages calculated on the left and right images respectively. The value returned by Eq. (1) is then divided by $\sigma_{\Delta_{LR}}$, the average of the standard deviations calculated for each $\Delta_{LR,j}$ for j varying across the scenes in the training set:

$$\Delta'_{LR,i} = \Delta_{LR,i} / \sigma_{\Delta_{LR}} \tag{2}$$

The other three clues $D'_{T,i}$, $D'_{S,i}$ and $A'_{T,i}$ are obtained straightforwardly from ToF and stereo disparities and ToF amplitude by applying a normalization similar to the one in Eq. (2), that is:

$$D'_{T,i} = D_{T,i} / \sigma_{D_T} \tag{3}$$

$$D'_{S,i} = D_{S,i} / \sigma_{D_S} \tag{4}$$

$$A'_{T,i} = A_{T,i} / \sigma_{A_T} \tag{5}$$

where σ_{D_T}, σ_{D_S} and σ_{A_T} are the average of the standard deviations calculated for each disparity or amplitude representation in the training set. Finally, the four clues $\Delta'_{LR,i}$, $D'_{T,i}$, $D'_{S,i}$ and $A'_{T,i}$ are concatenated in a four-channel input image which is fed to the CNN in order to produce

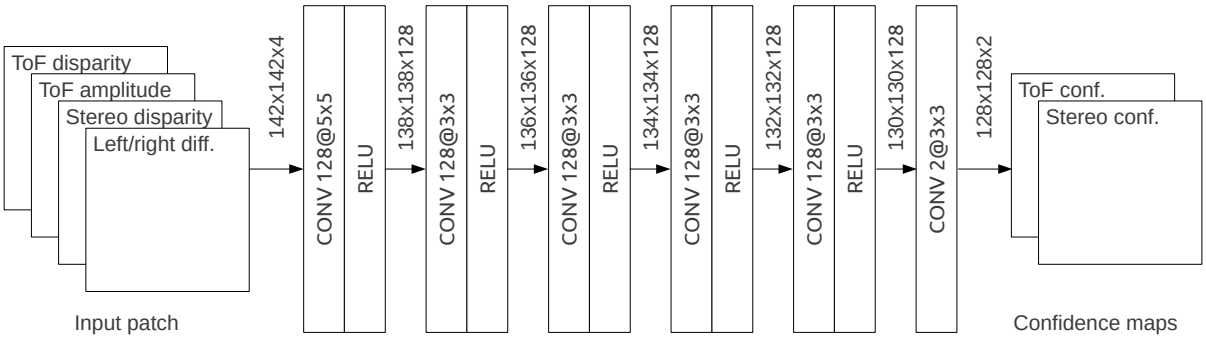


Fig. 2: Architecture of the proposed Convolutional Neural Network (CNN). The size of the outputs produced at the end of each convolutional layer is also reported for the case where a single 4-channel training patch of size 142×142 is fed as input to the CNN.

two confidence maps P_T and P_S for ToF and stereo data respectively.

The inference process is better explained by Figure 2 that shows the architecture of the proposed CNN. It contains a stack of six convolutional layers each followed by a ReLU non-linearity with the exception of the last one. The first five convolutional layers have 128 filters each, the first layer has a window size of 5×5 pixels while all others have a window size of 3×3 pixels. The last convolutional layer has only two filters, producing as output a two-channels image where the two channels contain the estimated ToF and stereo confidence respectively. Notice that, in order to produce an output with the same resolution of the input, no pooling layers have been used. At the same time, to cope with the boundary effect and size reduction introduced by the convolutions, each input image is padded by 7 pixels along their spatial dimensions, where the padded values are set to be equal to the values at the image boundary.

A. Training

A large set of training examples has been generated by randomly extracting a number of patches from each scene in the training set. Each patch has a size of 128×128 pixels (that becomes 142×142 after padding). During this process, a set of standard transformations have also been applied to augment the number of training examples and ease regression, namely rotation by ± 5 degrees, horizontal and vertical flipping. In our experiments, we extracted 30 patches from each of the 40 scenes included in the training set, and considering also their transformed versions at the same corresponding location we obtained a total of 6000 patches.

Both ToF and stereo ground truth confidence maps have been generated from the absolute error of the two disparity estimations against the disparity ground truth of the scene, that is available in the dataset. More specifically, each ground truth confidence map has been computed by first clipping all values above a given threshold in the corresponding disparity absolute error map, then dividing by the same threshold in order to obtain values in the range $[0, 1]$.

The network has been trained to minimize a loss function defined as the Mean Squared Error (MSE) between the estimated ToF and stereo confidence maps and their corresponding ground truth. To this aim, we employed a standard Stochastic Gradient Descent (SGD) optimization with momentum 0.9 and batch size 16. We started the training with an initial set of weight values derived with Xavier's procedure [37] and an initial learning rate of 10^{-7} subject to a constant decay by a coefficient of 0.9 every 10 epochs. The network has been implemented using the MatConvNet framework [38]. The training of the network took about 3 hours on a desktop PC with an Intel i7-4790 CPU and an NVIDIA Titan X (Pascal) GPU.

V. Fusion of Stereo and ToF Disparity

The final step is the fusion of the disparity maps from the upsampled ToF camera and the stereo vision system using the confidence information estimated by the deep learning framework. For each pixel location, we aim at combining different depth hypotheses from the two subsystems to obtain a more accurate depth estimation. For this task we used a modified version of the approach of [4]. This method is based on the Locally Consistent (LC) technique [8], a method firstly introduced for the refinement of stereo matching data. In [3] this method has been extended in order to be used with multiple disparity hypotheses as in the case of our setup. A further extension has been proposed in [4], that modifies the original formulation to account for the confidence measures and introduces sub-pixel precision.

The idea behind the method is to start by computing the plausibility of each valid depth measure at each point. The plausibility is a function of the color and spatial consistency of the data. Multiple plausibilities are then propagated to neighboring points. In the final step the overall plausibility is accumulated for each point and a winner-takes-all strategy is used to compute the optimal disparity value. For more details on this technique we refer the reader to [3], [4], [8]. Notice that in this work the

parameters of the method have been set to $\gamma_s = 8$ and $\gamma_c = \gamma_t = 4$.

The extension of the method proposed in [3] produces reasonable results, but has the limitation that assign the same weight to the two data sources without accounting for their reliability.

The further extension of the method proposed in [4] assigns different weights to the plausibilities depending on the confidence estimation for the considered depth acquisition system computed at the considered point:

$$\Omega'_f(d) = \sum_{g \in \mathcal{A}} \Big(P_T(g) \mathcal{P}_{f,g,T}(d) \; + \; P_S(g) \mathcal{P}_{f,g,S}(d) \Big) \quad (6)$$

where $\Omega'_f(d)$ is the plausibility at point f for depth hypothesis d, $\mathcal{P}_{f,g,T}(d)$ is the plausibility propagated by neighboring points g according to ToF data and $\mathcal{P}_{f,g,S}(d)$ is the one according to stereo data. Finally $P_T(g)$ is the ToF confidence value at location g and $P_S(g)$ the stereo confidence value. In [4] the confidence information comes from a deterministic algorithm based on the noise model for the ToF sensor and from the cost function analysis for the stereo system, while in the proposed approach the confidence is estimated with the CNN of Section IV.

VI. Synthetic Dataset

Another contribution of this paper is the new synthetic dataset that we will call $SYNTH3$. This dataset has been developed for machine learning applications and is split into two parts, a training set and a test set. The training set contains 40 scenes, among them the first 20 are unique scenes while the remaining ones are obtained from the first set by rendering them from different viewpoints. Notice that, even if the number of scenes is low if compared with the datasets used in other fields, it is still the largest dataset for stereo-ToF fusion currently available and each scene has different characteristics. The test set is composed by the data acquired from 15 unique scenes.

Each synthetic scene is realized by using the $Blender$ 3D rendering software [5]. The $Blender$ scenes have been downloaded from the $BlendSwap$ website [39]. We have appropriately modified the scenes and generated a dataset for Stereo-ToF data by rendering these scenes into virtual cameras.

The various scenes contain furnitures and objects of several shapes in different environments e.g., living rooms, kitchen rooms or offices. Furthermore, some outdoor locations with non-regular structure are also included in the dataset. In general, they appear realistic and suitable for the simulation of Stereo-ToF acquisition systems. The depth range across the scenes goes from about 50 cm to 10 m, providing a large range of measurements.

We have virtually placed in each scene a stereo system with characteristics resembling the ones of the ZED stereo camera [40] and a ToF camera with characteristics similar to a Microsoft Kinect v2 [1], [41]. The stereo system is composed by two Full-HD (1920×1080) color cameras with a baseline of 12 cm and the optical axes and image planes parallel to each other. The data acquired from these cameras are already rectified. Also the image plane and optical axis of the Kinect sensor are parallel to those of the ZED cameras and the Kinect sensor is placed under the right camera of the stereo system at a distance of 4 cm. The considered acquisition system is depicted in Figure 3 that shows the relative positions of the 2 cameras and ToF sensor. Table I summarizes the parameters of the acquisition system.

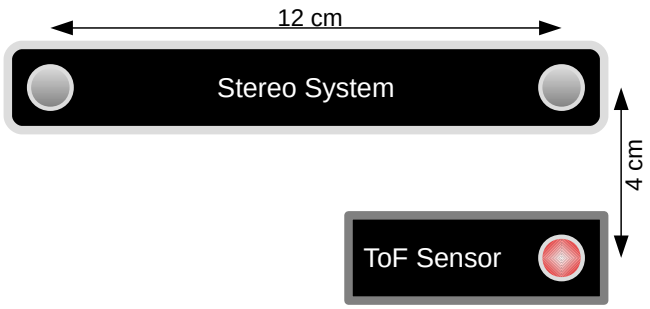


Fig. 3: Representation of the Stereo-ToF acquisition system. The Figure shows the relative position of the color cameras and ToF camera.

	Stereo setup	ToF camera
Resolution	1920×1080	512×424
Horizontal FOV	$69°$	$70°$
Focal length	$3.2\ mm$	$3.66\ mm$
Pixel size	$2.2\ \mu m$	$10\ \mu m$

TABLE I: Parameters of the stereo and ToF subsystems.

The dataset contains for each scene: 1) the 1920×1080 color image acquired by the left camera of the stereo system, 2) the 1920×1080 color image acquired by the right camera of the stereo system, 3) the depth map estimated by the ToF sensor on the synthetic scene and 4) the relative amplitude map of the ToF sensor.

The color images have been generated directly in $Blender$ using the $3D$ renderer $LuxRender$ [6]. The data captured by the Kinect ToF camera have been obtained by using the $ToF\text{-}Explorer$ simulator developed by Sony EuTEC. The $ToF\text{-}Explorer$ simulator was first designed according to the ToF simulator presented by Meister et al. in [7] that accurately simulate the data acquired by a real ToF camera including different sources of error as shot noise, thermal noise, read-out noise, lens effect, mixed pixels and the interference due to the global illumination (multi-path effect). The ToF simulator uses as input the scene information generated by $Blender$ and $LuxRender$.

Moreover, the dataset contains also the scene depth ground truth relative to the point of view of both the Kinect and the right color camera of the stereo system. To the best of our knowledge $SYNTH3$ is the first synthetic dataset

containing all the aforementioned data that can be used for deep learning applications. The dataset can be downloaded from *http://lttm.dei.unipd.it/paper_data/deepfusion*.

VII. Experimental Results

The proposed fusion algorithm has been trained on the training set and then evaluated on the test set of the $SYNTH3$ dataset. As pointed out in Section VI, the test set contains 15 different scenes. The thumbnails of the various scenes are shown in Figure 4, notice how they contain different challenging environments with different acquisition ranges, complex geometries and strong reflections. Also different materials, both textured and untextured have been used. The acquisition setup and the camera parameters are the same of the training set discussed in Section VI. Ground truth data have been computed by extracting the depth map from the $Blender$ rendering engine and then converting it to the disparity space. The algorithm takes in input the 512×424 depth and amplitude maps from the ToF sensor and the two 960×540 color images from the cameras. The color cameras resolution has been halved with respect to the original input data in the dataset. The output is computed on the point of view of the right camera at the same (higher) resolution of color data and it has been cropped to consider only on the region that is framed by all the three sensors.

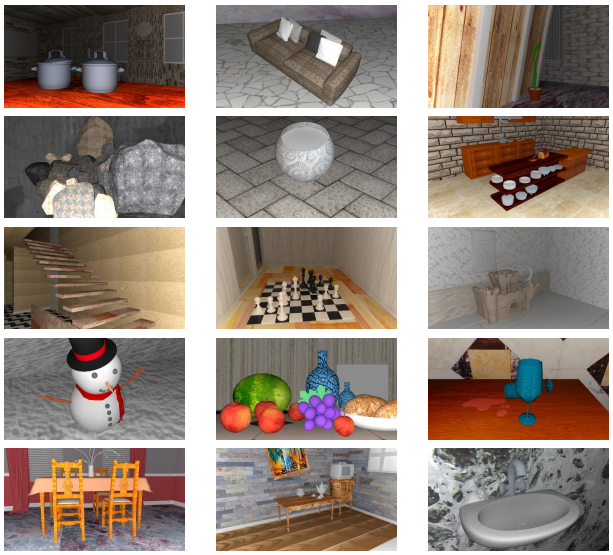
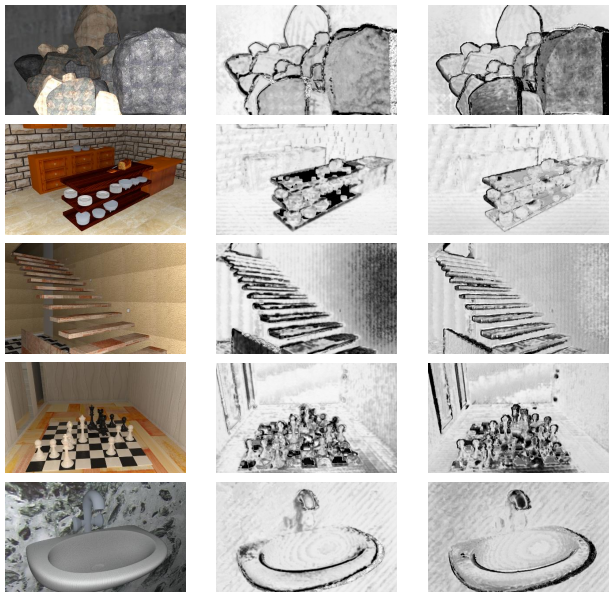


Fig. 4: Test set used for the evaluation of the performance of the proposed method. The figure shows the right camera color image for each scene in the dataset.

Before evaluating the performances of the fusion scheme we analyze the confidence information computed with the proposed CNN used to control the fusion process. Figure 5 shows the color image and the confidence maps for a few sample scenes.

The second column shows the ToF confidence, it is noticeable how the CNN is able to estimate the areas of larger error by assigning low confidence (darker pixels in the images). A first observation is that in most of the confidence maps it is possible to see how the error is larger in proximity of the edges. It is a well-known issue of ToF sensors due to the limited resolution and due to the mixed pixel effect. Furthermore, by looking for example at rows 2 and 4, it is visible how the CNN has also correctly learned that the ToF error is higher on dark surfaces due to the lower reflection (e.g., on the dark furniture in row 2 or on the black squares of the checkerboard in row 4, or on some of the rocks in row 1). The multi-path is more challenging to be detected, but row 4 shows how the confidence is lower on the wall edges or behind some stairs element in row 3. Concerning the stereo confidence the results are also good. Also in this case the limited accuracy on edges is detected and a low confidence is assigned. Furthermore, some surfaces with uniform or repetitive patterns have a lower confidence, e.g., some rocks in row 1.



a) Color view b) ToF conf. c) Stereo conf.

Fig. 5: Confidence information computed by the proposed deep learning architecture for some sample scenes: a) Color view; b) Estimated ToF confidence; c) Estimated stereo confidence. Brighter areas correspond to higher confidence values, while darker pixels to low confidence ones.

The computed confidence information is then used to drive the fusion process. Figure 6 shows the output of the proposed algorithm on some sample scenes. Column 1 and 2 show a color view of the scene and the ground truth disparity data. The up-sampled, filtered and reprojected ToF data are shown in column 3, while column 4 contains the corresponding error map. Columns 5 and 6 show the disparity estimated by the stereo vision algorithm and the corresponding error map. Concerning stereo data, for this work we used the OpenCV implementation of the SGM stereo algorithm with pointwise Birchfield-Tomasi metric, 8

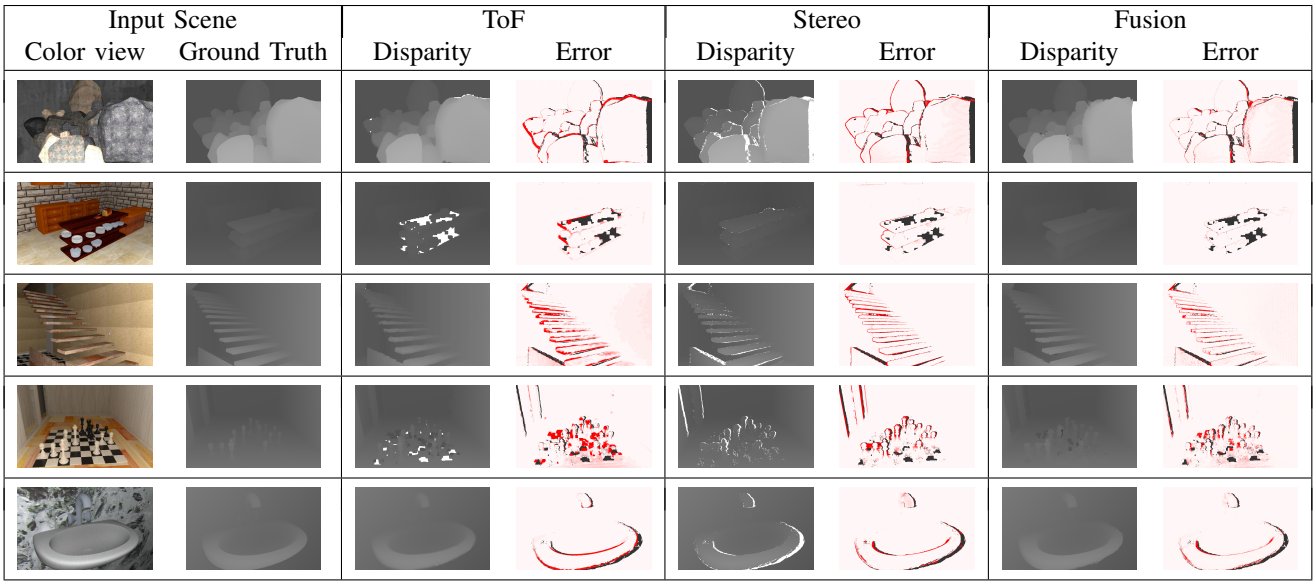


Fig. 6: Results of the proposed fusion framework on 5 sample scenes (one for each row). In error images, grey pixels correspond to points excluded since they are not valid on one of the disparity maps. The intensity of red pixels is proportional to the absolute error. (*Best viewed in color*).

paths for the optimization and a window size of 7×7 pixels. The fused disparity map and its relative error are shown in columns 7 and 8. Starting from ToF depth data, this is the more accurate of the two data sources, the filtered and interpolated data is quite accurate, even if there are issues in proximity of edges that are sometimes not too accurate. Also low-reflective surfaces like the black checkers in row 4 are very noisy and sometimes not acquired at all. The multi-path affects some regions like the steps of the stairs. Stereo based disparity maps usually have sharper edges but there are several artifacts due to the well-known limitations of this approach. The fusion algorithm reliably fuse the information coming from the two sensors providing a depth map with less artifacts on edges and free from the various problems of the stereo acquisition.

The numerical evaluation of the performances is shown in Table II and confirms the visual analysis. The table shows the RMS in disparity space averaged on all the 15 scenes. For a fair comparison, we considered as valid pixels for the results only the ones having a valid disparity value in all the compared disparity maps (stereo, ToF and fused disparities). By looking at the averaged RMS values, the ToF sensor has a high accuracy with a RMS of 2.19, smaller than the RMS of 3.73 of the stereo system. This is a challenging situation for fusion algorithms since it is difficult to improve the data from the best sensor without affecting it with errors from the less efficient one. However confidence data helps in this task and the proposed approach is able to obtain a RMS of 2.06 with a noticeable improvement with respect to both sensors. Comparison with state-of-the-art approaches is limited by the use of the new dataset and the lack of available implementations of the competing approaches. However, we compared our approach with the highly performing method of Marin et Al. [4]. This approach has a RMS of 2.07, higher than the one of the proposed method even if the gap is limited and the results comparable. The method of [4] outperforms most state-of-the-art approaches, so also the performances of the proposed method are expected to be competitive with the better performing schemes, with the advantage that the proposed approach does not involve heuristics.

Method	RMS Error
Interpolated ToF	2.19
SGM Stereo	3.73
Proposed Stereo-ToF Fusion	**2.06**
Marin et Al. [4]	2.07

TABLE II: RMS in disparity units with respect to the ground truth for the ToF and stereo data, the proposed method and [4]. The error has been computed only on non-occluded pixels for which a disparity value is available in all the methods.

VIII. Conclusions and Future Work

In this work we presented a scheme for the fusion of ToF and stereo data exploiting confidence information coming from a deep learning architecture. We created a novel synthetic dataset containing a realistic representation of the data acquired by the considered acquisition setup. A convolutional neural network trained on this dataset has been used to estimate the reliability of ToF and stereo data, obtaining reliable confidence maps that identify the most critical acquisition issues of both sub-systems.

The fusion of the two data sources has been performed using an extended version of the LC framework that combines the confidence information computed in the previous step and provides an accurate disparity estimation. The results show how the proposed algorithm properly combines the outputs of the two sensors providing on average a disparity map with higher accuracy with respect to each of the two sub-systems, also considering that ToF data have typically high accuracy.

Further research will be devoted to improve the deep learning architecture in order to obtain a more reliable confidence information. The use of deep learning architectures to directly compute the final output will be also explored. Finally we plan to extend the proposed dataset to better train the machine learning algorithms and to evaluate the method also on real data.

Acknowledgment.: We would like to thank the Computational Imaging Group at the Sony European Technology Center (EuTEC) for allowing us to use their $ToF\text{-}Explorer$ simulator. We also thank prof. Calvagno for his support and gratefully acknowledge NVIDIA Corporation for the donation of the GPUs used for this research.

REFERENCES

[1] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, *Time-of-Flight and Structured Light Depth Cameras: Technology and Applications*, 1st ed. Springer International Publishing, 2016. [Online]. Available: http://www.springer.com/book/9783319309712

[2] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[3] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo, "Locally consistent tof and stereo data fusion," in *Workshop on Consumer Depth Cameras for Computer Vision (ECCV Workshop)*. Springer, 2012, pp. 598–607.

[4] G. Marin, P. Zanuttigh, and S. Mattoccia, "Reliable fusion of tof and stereo depth driven by confidence measures," in *European Conference on Computer Vision*. Springer International Publishing, 2016, pp. 386–401.

[5] "Blender website," https://www.blender.org/, Accessed July 31st, 2017.

[6] "Luxrender website," http://www.luxrender.net, Accessed July 31st, 2017.

[7] S. Meister, R. Nair, and D. Kondermann, "Simulation of Time-of-Flight Sensors using Global Illumination," in *Vision, Modeling and Visualization*, M. Bronstein, J. Favre, and K. Hormann, Eds. The Eurographics Association, 2013.

[8] S. Mattoccia, "A locally global approach to stereo correspondence," in *Proc. of 3D Digital Imaging and Modeling (3DIM)*, October 2009.

[9] B. Tippetts, D. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *Journal of Real-Time Image Processing*, pp. 1–21, 2013.

[10] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure." in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[11] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map." in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[12] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] M. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras: Principles, Methods and Applications*, ser. SpringerBriefs in Computer Science. Springer, 2013.

[14] F. Remondino and D. Stoppa, Eds., *TOF Range-Imaging Cameras*. Springer, 2013.

[15] D. Piatti and F. Rinaudo, "Sr-4000 and camcube3.0 time of flight (tof) cameras: Tests and comparison," *Remote Sensing*, vol. 4, no. 4, pp. 1069–1089, 2012.

[16] T. Kahlmann and H. Ingensand, "Calibration and development for increased accuracy of 3d range imaging cameras," *Journal of Applied Geodesy*, vol. 2, pp. 1–11, 2008.

[17] S. A. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of stereo vision and time of flight imaging for improved 3d estimation," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, pp. 425–433, 2008.

[18] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, M. Eisemann, M. Magnor, and D. Kondermann, "A survey on time-of-flight stereo fusion," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ser. Lecture Notes in Computer Science, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. Springer Berlin Heidelberg, 2013, vol. 8200, pp. 105–127.

[19] K.-D. Kuhnert and M. Stommel, "Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction," in *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2006, pp. 4780 – 4785.

[20] A. Frick, F. Kellner, B. Bartczak, and R. Koch, "Generation of 3d-tv ldv-content with time-of-flight camera," in *Proceedings of the 3DTV Conference*, 2009.

[21] Y. M. Kim, C. Theobald, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and tof sensor fusion for dense 3d reconstruction," in *Proc. of 3D Digital Imaging and Modeling (3DIM)*, October 2009.

[22] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[23] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 899–909, 2010.

[24] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1400–1414, 2011.

[25] C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, "A probabilistic approach to ToF and stereo data fusion," in *Proc. of 3DPVT*, Paris, France, 2010.

[26] ——, "Probabilistic tof and stereo data fusion based on mixed pixels measurement models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2260–2272, 2015.

[27] R. Nair, F. Lenzen, S. Meister, H. Schaefer, C. Garbe, and D. Kondermann, "High accuracy tof and stereo sensor fusion at interactive rates," in *Proceedings of European Conference on Computer Vision Workshops (ECCVW)*, 2012.

[28] G. Evangelidis, M. Hansard, and R. Horaud, "Fusion of Range and Stereo Data for High-Resolution Scene-Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2178 – 2192, 2015.

[29] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *In Proc. of NIPS*. MIT Press, 2005, pp. 291–298.

[30] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[31] Q. Yang, N. Ahuja, R. Yang, K. Tan, J. Davis, B. Culbertson, J. Apostolopoulos, and G. Wang, "Fusion of median and bilateral filtering for range image upsampling," *IEEE Transactions on Image Processing*, 2013.

[32] S. Schwarz, M. Sjostrom, and R. Olsson, "Time-of-flight sensor fusion with depth measurement reliability weighting," in *Proceedings of the 3DTV Conference*, 2014, pp. 1–4.

[33] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, "A novel interpolation scheme for range data with side information," in *Proceedings of the European Conference on Visual Media Production (CVMP)*, 2009.

[34] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1141–1148.
[35] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 –619, 2002.
[36] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.
[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
[38] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.
[39] "Blend swap website," https://www.blendswap.com/, Accessed July 31st, 2017.
[40] "Zed stereo camera," https://www.stereolabs.com/, Accessed July 31st, 2017.
[41] J. Sell and P. O'Connor, "The xbox one system on a chip and kinect sensor," *IEEE Micro*, vol. 34, no. 2, pp. 44–53, 2014.