A Framework for Fast Low-Power Multi-Sensor 3D Scene Capture and Reconstruction

Aleksandra Chuchvara, Mihail Georgiev, Atanas Gotchev

Tampere University of Technology, Tampere, Finland
{aleksandra.chuchvara,mihail.georgiev,atanas.gotchev}@tut.fi

Abstract. We present a computational framework, which combines depth and colour (texture) modalities for 3D scene reconstruction. The scene depth is captured by a low-power photon mixture device (PMD) employing the time-of-flight principle while the colour (2D) data is captured by a high-resolution RGB sensor. Such 3D capture setting is instrumental in 3D face recognition tasks and more specifically in depth-guided image segmentation, 3D face reconstruction, pose modification and normalization, which are important pre-processing steps prior to feature extraction and recognition. The two captured modalities come with different spatial resolution and need to be aligned and fused so to form what is known as view-plus-depth or RGB-Z 3D scene representation. We discuss specifically the low-power operation mode of the system, where the depth data appears very noisy and needs to be effectively denoised before fusing with colour data. We propose using a modification of the non-local means (NLM) denoising approach, which in our framework operates on complex-valued data thus providing certain robustness against low-light capture conditions and adaptivity to the scene content. Further in our approach, we implement a bilateral filter on the range point-cloud data, ensuring very good starting point for the data fusion step. The latter is based on the iterative Richardson method, which is applied for efficient non-uniform to uniform resampling of the depth data using structural information from the colour data. We demonstrate a real-time implementation of the framework based on GPU, which yields a high-quality 3D scene reconstruction suitable for face normalization and recognition.

Keywords: ToF, 2D/3D, depth, fusion, denoising, NLM, face, ICP

1 Introduction

In the fields of pattern recognition, computer vision and biometrics, 2D and 3D scene capture and understanding are active areas of research due to the variety of practical applications such as biometric identification/authentication for security and access control purposes, behavioral and psychological analysis for various commercial applications, object tracking, and many others. Among visual scenes, scenes containing human faces are of particular interest and the task of face recognition has been thoroughly studied mainly by using two-dimensional (2D) imagery. One particular advantage of conventional face recognition systems based on 2D images is a fast and low-cost data acquisition. However, 2D facial images can vary strongly depending on

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

many factors such as viewpoint, head orientation, illumination, different facial expressions, even aging and makeup, which can significantly decrease the system performance. Thus, in most cases, it is necessary to maintain a canonical frontal facial pose and consistent illumination in order to achieve good recognition performance.

In order to overcome the above-mentioned limitations and improve face recognition accuracy, more and more approaches suggest utilizing 3D facial data, which can be acquired by dedicated range sensors. Systems utilizing structural information of the facial surface are less dependent to the pose and/or illumination changes, which mostly affecting 2D image based systems. Accurate depth sensing is a challenging task especially in presence of a real-time constraint. Recent advances in Time-of-Flight (ToF) depth sensing technologies made fast acquisition of 3D information about facial structure and motion a feasible task. ToF sensors are much less expensive and more compact than other traditional 3D imaging systems used for 3D model acquisition. ToF sensors can deliver range data at high frame rates enabling real-time interactive applications. ToF depth sensors acquire depth information in a form of perspective-fixed range images often referred to as 2.5D models, which provides valuable information for object detection and recognition and can greatly assist tasks of face segmentation, i.e. removal of non-facial data such as neck, torso and hair, and face normalization, i.e. alignment of the face data in a canonical position. These tasks are usually performed before the actual feature extraction and recognition take place.

A number of ToF imaging applications has been proposed in the fields of face detection and recognition [1-4], gesture recognition [5], and real-time segmentation and tracking [6, 7]. A survey on face recognition and 3D face recognition using depth sensors has been presented in [1]. A face recognition system based on ToF depth images has been proposed in [2]: as the performance of 3D face recognition is highly dependent upon the distance noise, the problem of low quality of the ToF data has been specifically addressed. A ToF face detection approach has been presented in [3], where range data yields a significant robustness and accuracy improvement of the face detector. Other systems tend to utilize multimodal approaches combining 2D and 3D features. A face recognition system using a combination of color, depth and thermal-IR data has been proposed in [4]. The system is calibrated and tested in order to select optimal sensor combination for various environmental conditions. In [5], a real-time 3D hand gesture interaction system based on ToF and RGB cameras has been presented: an improved hand detection algorithm utilizing ToF depth allows for recognition of complex 3D gestures. A framework for real-time segmentation and tracking fusing depth and color data has been proposed in [6], aimed at solving some common problems, such as fast motion, occlusions and tackling objects with similar color. In [7], a low-complexity real-time segmentation algorithm utilizing color and ToF depth information has been presented. The robust performance for the approach is based on simultaneous analysis of depth, color and motion information.

The major advantage of a ToF sensor compared to other depth estimation techniques is its ability to deliver entire depth map at a high frame rate and independently of textured surfaces and scene illumination. However, current ToF devices have certain technology limitations associated with their working principle, such as low sensor resolution (e.g. 200×200 pixels) compared to Full High-Definition (HD) (1920×1080) of color cameras, inaccuracies in depth measurements, and limited ability to capture

color information [8]. A solution is to combine a depth sensor with one or multiple 2D cameras responsible for color capture into a single multisensor system.

In this paper, we propose an end-to-end multi-sensor system combining a conventional RGB camera and a ToF range sensor with the purpose to perform real-time 3D scene sensing and reconstruction, which is instrumental for face recognition tasks, e.g. 3D face reconstruction, depth-guided segmentation, pose modification and normalization. Different combinations of high-resolution video cameras and low-resolution ToF sensors have been studied. The setups most related to our work, which utilize configuration with a single color view and a single ToF sensor, are described in [9-12]. A rather straightforward data fusion scheme has been implemented in [9]. The data fusion is implemented by mapping the ToF data as 3D point clouds and projecting them onto the color camera sensor plane, resulting in pixel-to-pixel projection alignment of color and depth maps. Subsequently, the color image is back-projected as a texture onto ToF sensor plane. Approaches described in [11] up-sample low-resolution depth maps by means of adaptive multi-lateral filtering (proposed in [13]) in a way that it prevents edge blurring in geometry data while smoothing it over continuous regions. With GPU implementation, this approach is shown to be feasible for real-time applications. An efficient method for the ToF and color data fusion, which generates the 3D scene on-the-fly utilizing FPGA hardware, has been presented in [10]. In [12] an efficient spatio-temporal filtering approach has been proposed that simultaneously denoises the depth video and increases its spatial resolution to match the color video.

2 System overview

In this section, an end-to-end 3D video system for real-time 3D scene sensing and reconstruction is presented (Fig. 1). The input data is generated by a multisensor setup combining a ToF depth sensor and a color camera. The two sensors are displaced and have different field of view and spatial resolution. Therefore, the distance data acquired from the low-resolution ToF sensor needs to be projectively aligned and fused with color data; a process, referred to as 2D/ToF data fusion. Computationally, this requires reprojection of the depth data from the low-resolution grid to world coordinates and then back on the higher-resolution grid of the color sensor followed by resampling to get the depth values at the grid points.

An accurate fusion of distance and color information requires reliable estimation of the relative positions of the cameras and their internal calibration parameters. A

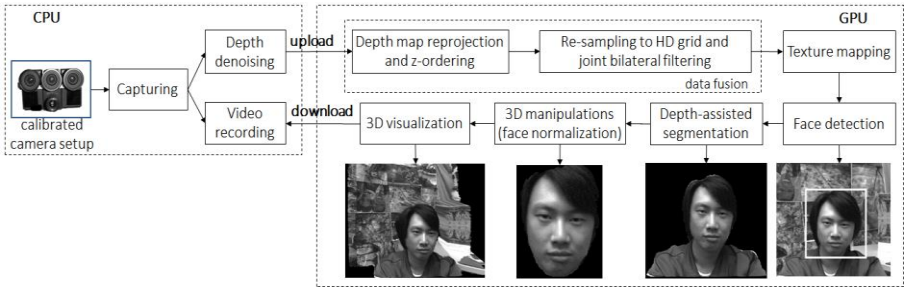

Fig. 1 Capturing, streaming and data processing pipeline

number of approaches have been published aimed at calibrating multiple color cameras with range sensors [14-16]. Calibration between depth and color sensors by a maximum likelihood solution utilizing a checkerboard similar to a stereo system calibration has been proposed in [14, 15]. Range-specific calibration techniques have been proposed in [14, 16]. First, the cameras are calibrated separately for the internal parameters, such as focal length and principal point coordinates, resulting in the camera calibration matrices $\boldsymbol{K}_{RGB}$ and $\boldsymbol{K}_{TOF}$. Second, the stereo calibration step provides the external camera parameters: a rotation matrix - $\boldsymbol{R}_{3x3}$ and a translation vector – $\boldsymbol{t}_{3x1}$, which form the relative transformation $\boldsymbol{RT}_{TOF\rightarrow RGB}=[\boldsymbol{R}|\boldsymbol{t}]$ between optical centers of the ToF camera and the RGB camera. Due to the fixed setup, these parameters have to be determined only once during the preliminary initialization stage in offline mode.

In our setup, a ToF camera based on the Photonic Mixer Device (PMD) principle is used [17]. To get distance data, a PMD sensor measures the phase-delay between an emitted wave and its reflected replica. A typical PMD consists of a beamer, an electronic light modulator and a sensor chip (e.g. CMOS or CCD). The beamer is made of an array of light-emitting diodes (LED) operating in near-infrared wavelengths (e.g. 850 nm). It radiates a point-source light of a continuously-modulated harmonic signal which illuminates the scene. The light reflected from object surfaces is sensed back by pixels of the sensor chip, which collects pixel charges for some interval denoted as *integration time*. For each pixel, the range data is estimated in relation to phase-delay between the sensed signal and the one of the light modulator. The phase-delay estimation is performed as a discrete cross-correlation of several successively captured samples taken between equal intervals during same modulation periods of fixed frequency. Denote the sample data as R_n (n=1, 2,…, N-1, $N\geq 4$). The amplitude A and phase φ of the signal are estimated from the sampled data, while the sensed distance D is proportional to the phase φ:

$$A=\frac{2}{N}\sum_{n=0}^{N-1}\left|R_n e^{-j2\pi\frac{n}{N}}\right|,\ \varphi=\arg\left(\sum_{n=0}^{N-1}R_n e^{-j2\pi\frac{n}{N}}\right),\ D\propto\frac{\varphi}{4\pi f}c_L, \tag{1}$$

where j is the imaginary unit, f is the frequency of the emitted signal and c_L is the speed of light through dry air (~298.109km/h).

Due to the operational principles of the ToF range sensor, significant amount of noise is present in the captured range data. The depth measurement noise is amplified during the fusion process and degrades the quality of the fused data. Thus, prior to the data fusion step, denoising of depth data should be performed in order to reduce the noise and remove outliers. The problem of denoising of ToF data has been addressed in a number of works [18-21]. Modern denoising approaches, such as edge-preserving bilateral filtering [22] and non-local filtering [23], have been modified to deal with ToF data [19, 21]. In [18], a range-adaptive bilateral filter has been proposed, by adjusting its size according to ToF amplitude measurements, since the noise level in distance measurements varies depending on the amplitude of the recorded signal [19]. In our work, we specifically consider the 2D/ToF fusion in the so-called low-sensing mode. In such a mode, the ToF sensor is restricted, e.g. by technological limitations, to operate in poor imaging conditions. These include low number of emitting diodes, or low power or short integration time. In such conditions, the noise becomes a dominant problem, which should be addressed by dedicated denoising methods [20, 21].

2.1 ToF denoising

Measurement accuracy of a ToF sensor is limited by the power of the emitted signal and depends on many factors, such as light intensity, different reflectivity of surfaces, distances to objects in a scene, etc. Erroneous range measurements [21] can be caused e.g. by multiple reflections to the sensed signals, sensing objects having low-reflectivity materials and colors, or surfaces of small incident.

When in low-powered sensing more, the ToF sensor is usually also with low spatial resolution, meeting technological limitations such as requirements for miniaturization of the beamer size and reducing the number of LED elements for cost-efficient hardware and embedding into portable devices. This leads to very noisy range images of a very low resolution. Degradations in the range data impede the projective mapping function in the 2D/3D fusion process. The case is illustrated in Fig. 2: one can observe that while the original noisy range data represents some scene structure, the fused output is fully degraded and useless. The process of surface based *z*-ordering becomes extremely unstable and no confidence of occluded and hidden data can be estimated (Fig. 2 3rd row). Due to the noise influence on the projected position, some areas of the rendered surface get artificially expanded and shadow some true areas. The effect impedes the non-uniform resampling at the stage of data fusion and also illustrates the importance of proper care of the range noise prior to fusion procedure.

We specifically address the noise reduction as a post-capture stage applied to low-sensed range data with the aim to achieve a 2D/ToF data fusion result with quality as if the ToF sensor was working in normal operating mode (Fig. 2). We have proposed a three-stage denoising technique: a raw data (system) denoising, point-cloud projection and denoising, and non-uniform resampling combined with depth refinement.

For the system denoising stage, we propose a technique based on the state-of-the art *non-local means* (NLM) denoising approach [24]. The general idea of NLM filtering is to find blocks (patches) similar to a reference block and to calculate a noise-free estimate of the central pixel of that reference block based on weighted average of the corresponding pixels in the similar blocks, where weights are proportional to the measured similarities.

In our approach, the signal components of the phase-delay and the amplitude of the sensed signal are regarded as components of a complex-valued random variable and processed together in a single step. The map for similarity search, denoted by U in our approach is chosen to be the pre-computed maps of $(A, \varphi) - (A_U, \varphi_U)$ given in Eq.(1) and pixel-wise combined into a complex-valued map, denoted by Z, while the modified NLM filter (NLM_{CLX})is given by:

$$\begin{aligned} &(\varphi_U, A_U) \to Z = A_U\left(e^{j\varphi_U}\right),\ Z \to A_U = |Z|,\ Z \to \varphi_U = \arg(Z) \\ &NLM_{CLX}[x] = \frac{1}{C_N(x)} \int_{\Omega} \exp\left(-\frac{G \times |Z(x+.) - Z(y+.)|(0)}{h^2}\right) Z(y)dy\,. \end{aligned} \tag{2}$$

In the equation C_N denotes a normalization factor, G is a Gaussian kernel, Ω is the search range, U is the pixel map, x is the index of filtered pixel, y is the running index of center pixels in similarity patches, h is a tuning filter parameter, and $\times\cdot(0)$ denotes a centered convolution operator, while $(+\ .)$ denotes the range of pixel indices of spatial neighborhood.

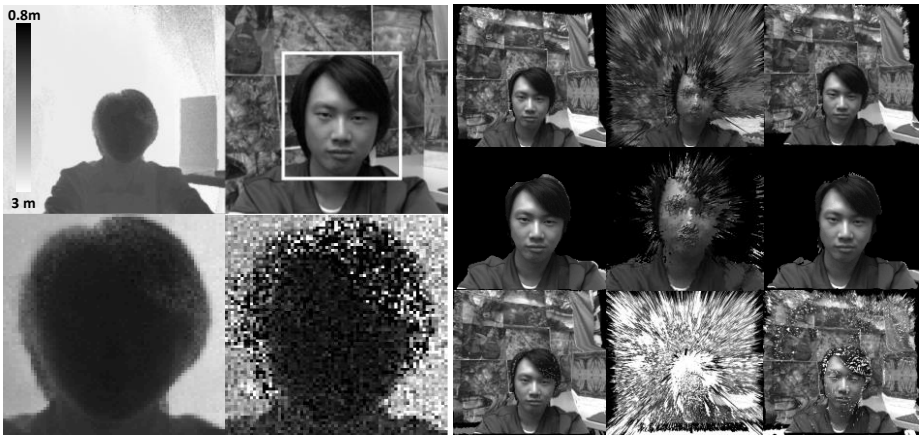

Fig. 2 Role of denoising. Left, (clockwise): ToF range image; detected face region; noisy (I_T=50μs) face region; GT face region. Right, in rows: fusion, segmentation and z-ordering; using (in columns) GT, noisy, denoised depth.

The complex-valued representation of the sensed signal facilitates the block search stage and leads to better filter adaptivity and similarity weighting with reduced computational complexity. The complex-domain filtering implies simultaneous filtering of all signal components in a single step, thus it provides additional feedback in the form of improved (noise-reduced) confidence parameter given by *A*, which can be utilized in iterative de-noising schemes. Such NLM_{CLX} filter can be easily extended to similarity search in temporal neighborhood of successively-captured frames, where temporal similarity is beneficial [28]. Such spatio-temporal filtering provides significant boost in denoising for small number of frames (e.g. 3) and for decreased size of the searched spatial neighborhood. The spatio-temporal search allows additional trade-off for using shorter integration times, and thus increases the number of frames which can be used. This provides an effective over-complete observation structure instrumental for effective denoising. Further modification is focused on speed. We have proposed a real-time version of NLM_{CLX} denoising filter – *S(implified)*NLM_{CLX}. It is only slightly inferior in terms of quality, however achieves an $O(1)$ complexity. The idea is to first apply a global pre-filtering in order to simplify the patch similarity search by utilizing summed area tables (SATs) and smart look-up table data fetching [28].

In contrast to other 2D/ToF fusion approaches, which suggest working in planar coordinates, our approach includes a second-stage denoising refinement data projected in the 3D world coordinate system (i.e. data point cloud), It makes use of the surface information presented in the point cloud of range measurements [20]. The property that surf mesh data could exist on the unique optical ray formed by corresponding range pixel and camera center can be used for a surface mesh denoising in the flavor of the techniques given in [13].

2.2 Resampling and data fusion

Depth and color data fusion refers to a process of depth data reprojection and color-aided resampling. As the two cameras in the setup cannot be placed at the same posi-

tion, their viewpoints are shifted relative to each other and also viewing directions may be slightly different. Thus, to align the depth information with the color image, every pixel of the depth image needs to be reprojected in the coordinate system of the color image. This transformation is usually performed in two steps, often referred to as 3D warping. First, each pixel of the ToF image $d(u, v)$ is converted into a corresponding 3D point in terms of a homogeneous coordinate – $(x, y, z, w)^T$:

$$x = u \cdot z / f_{TOF}, y = v \cdot z / f_{TOF}, z = f_{TOF} \cdot d(u,v) \Big/ \sqrt{f_{TOF}^2 + u^2 + v^2}, \; w = 1, \tag{3}$$

where f_{TOF} is the focal length of the ToF sensor. The second step is to project every 3D point on the image plane of the color camera:

$$(x,y,w)_{RGB} = \boldsymbol{K}_{RGB} \cdot \boldsymbol{RT}_{TOF \to RGB} \cdot (x,y,z,w)^T, \; (u,v)_{RGB} = (x,y)_{RGB} / w_{RGB}, \tag{4}$$

where $\boldsymbol{RT}_{TOF \to RGB}$ and $\boldsymbol{K}_{RGB}$ are calibration parameters of the system and $(u, v)_{RGB}$ are the image coordinates for a global point $(x, y, z, w)^T$ within the *RGB* image grid.

In order to obtain a dense depth map, the depth data should be upscaled to match the resolution of the color image. One particular problem is that the points projected to the color camera image grid do not fall to the integer positions and also are very sparse due to the low resolution of depth data. The unknown values at the regular positions of color image grid need to be estimated from sparsely scattered irregular data, imposing a non-uniform resampling problem. An accurate non-uniform to uniform resampling in real time is a challenging task and still an open-research problem [24]. It is important to utilize structural information presented in the high-resolution color data in the depth resampling process to properly align the multisensory data.

A non-uniform to uniform iterative resampling method has been proposed in [20]. The proposed method makes use of the color information by applying color-driven bilateral filter to the result of interpolated depth at each iteration (referred to as joint- or cross-bilateral filter [13]). The depth-refining resampling starts from an initial depth approximation on the high definition grid obtained by fitting Voronoi cells (V) around projected irregular depth samples followed by nearest-neighbor interpolation in the flavor of [25]. So-interpolated depth map d undergoes a joint-bilateral filtering (*JBF*) as defined in [13]. The role of the *JBF* is to smooth any blocky artifacts arising from the preceeding interpolation while aligning it with object edges of the color image. In order to estimate the error between the interpolated and the initial depth values, the values at the starting irregular locations are calculated by bilinear interpolation (L) of the refined depth d. The obtained error values are then interpolated again with V and added to d, and the procedure is repeated. By defining the depth map ob-

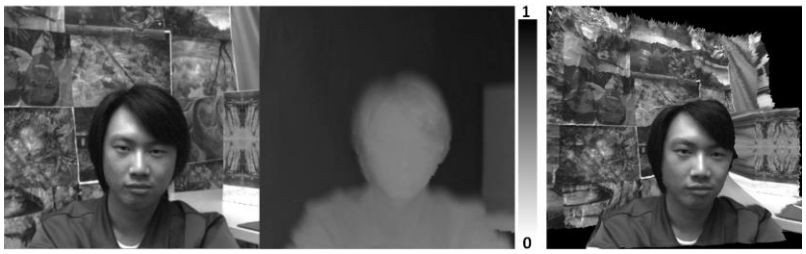

Fig. 3 Data fusion: view-plus-depth frame and 3D textured surface

tained after *i*-th iteration as d_i, initial depth values at irregular positions as *z*, and a relaxation parameter by λ, the whole iterative procedure can be formalized as follows:

$$d_{i+1} = JBF(d_i + \lambda V(z - L(d_i))). \tag{5}$$

The use of bilinear interpolation for depth values calculation at the irregular locations is motivated by the observation that depth maps can be modeled as piecewise-linear functions at local neighborhood and interpolators of higher degree are not beneficial.

The point-based depth reprojection comes along with one particular problem, namely z-ordering of projected points for detecting possible dis-occlusions. This requires an additional pre-processing of the projected data: ToF samples which are not visible from the color camera position are considered as "hidden points" and have to be filtered out before interpolation, in order to prevent the liking of hidden background information. A solution to the z-ordering problem is provided by our GPU-based rendering approach [26]. In order to generate a depth map corresponding to the viewpoint of the color camera, the algorithm makes use of a 3D mesh representation of the scene obtained using the ToF depth data. The depth data is represented as a triangulated surface mesh and then rendered as if observed from the point of view of the color camera. This constructs a depth map projectively aligned with the color image. During the rendering process, some depth testing is performed automatically and only the minimum per-pixel z-distance is stored in the depth buffer. Modern GPUs provide hardware support for triangulated non-uniform bi-linear resampling, which can be used for the fast resampling of the obtained depth map. However, the depth map obtained with bilinear interpolation needs to be further refined by a color-controlled filtering, e.g. *JBF*, so that color edges and depth discontinuities get aligned. The color image can be applied as a texture to the refined 3D depth surface. The textured surface can be then rotated, scaled and translated as needed to generate any arbitrary view (Fig. 3).

2.3 Depth-assisted segmentation

In a color plus depth imagery, a region containing a face can be detected by a cascade classifier mechanism, as in [27]. Then, the associated depth can be utilized for efficient face segmentation. An example of fast depth-assisted segmentation approach utilizing 2D color and depth information as well as motion information has been proposed in [7], where motion between two consecutive color and range frames is analyzed to locate preliminary region of interest within the scene. Then, a refinement algorithm delineates the segmented area. Object motion in a scene is detected by tracking pixel-wise color and depth changes between two consecutive frames. The temporal differences between both color and depth are mutually thresholded with certain value resulting in a region mask of detected motion. Initial foreground mask is estimated by applying a region growing algorithm, which uses so called "pixel seeds", which in our case are the ones detected in motion mask. The idea of the region growing algorithm is the following: a chosen seeding pixel is compared for similarity with the neighboring ones, and then added to the seeding region, thus growing it.

The foreground mask obtained with the growing algorithm can contain false inclusion of background areas due to errors in the depth map. This kind of errors can be

tackled by using more precise edge information from the color data, which results in improved foreground mask. To reduce boundary errors, so-called 'tri-map' is generated as follows: pixels inside the foreground mask are marked as 'certain foreground', the ones outside – as 'certain background', and pixels near the edges are marked as 'uncertain'. Then a *K-nn* search of the nearest pixels is performed in order to decide whether an uncertain pixel belongs to foreground or background by comparing its color to the certain foreground and certain background neighbors. The segmentation results, obtained using the described method, are illustrated in Fig. 2 Right.

2.4. Face normalization by Iterative Closest Point methods

A proper face alignment is viable for biometric applications involving facial data such as facial feature extraction, expression estimation, motion tracking and recognition. Usually such applications heavily rely on the use of trained classified data where certain face pose of limited misalignment variation was utilized for training. The process of face alignment to certain pose is referred in the state of the art literature as to face normalization [3, 4, 32, 33]. A rigid alignment for face normalization utilizing degrees of freedom (DoF) such as: angle rotations, translation shifts, and scaling can be obtained by utilizing so called Iterative Closest Points (ICP) algorithm [31].

Basically, the ICP algorithm is data registration applied on 3D data point clouds [30]. First, the algorithm takes a source of point cloud data as reference and target one as template. Then, for each point in the template, it locates the closest point in the source, and aims at minimizing the error between these points by applying a rigid transformation between the two meshes. This process is repeated until a threshold error is reached. The ICP solution may vary according to data selection [34], outlier filtering [31, 34], minimization constraints [29], or "closeness" metrics [34].

3 Performance validation on biometric pre-processing tasks

We illustrate the performance of our 3D capture and processing framework by experiments characteristic for typical biometric tasks such as face detection, tracking and recognition [2, 3, 4]. The first experiment demonstrates face projection alignment and fusion of color and depth data in the presence of noise. The second experiment demonstrates the performance of the system for ICP-based face normalization. The experimental equipment consists of custom designed 2D/ToF camera setup consisting of a Prosilica GE-1900C high-definition color camera and a PMDTech CamCube 2.0 ToF device mounted on a rig, where both cameras are vertically aligned with a baseline B = 6 cm. The scene represents a person frontally facing the cameras at a distance of 1.2 meters and sitting in front of a flat background situated in 2.5 meters.

3.1 2D/ToF fusion of face images

The face is detected utilizing a real-time modification of the Viola and Jones algorithm [27], [35] (Fig. 2). The detected face region U_F in the range map is used to quantify the denoising performance. The effect of noise in low-powered sensing envi-

ronment was simulated by changing the integration times of the ToF sensor for range $I_T \in [2000 \div 50]\mu s$, where the normal operating mode is corresponds to I_T=2000μs. To get ground truth data (GT), we have averaged 200 consecutively captured frames in normal operating mode. The low-sensing case is characterized by measured amplitude the reflected signal A<250 units [21]. For such amplitudes, it is expected that the error of the measured depth exceeds the one specified for normal operating mode [17]. The corresponding input low-sensed (and potentially wrong) depth pixels are counted as percentage of all available depth pixels and denoted as "BAD" (Table 1). Pixels having measurement error twice exceeding the corresponding GT range value after processing are considered uninformative and marked as "IMP". The noisy data has been processed by our $SNLM_{CLX}$ approach, working in real time. The denoising results are given in Table 1 and depicted in Fig. 4, where the comparison metrics are calculated as follows:

$$PSNR[dB] = 20\log_{10}\left(\frac{D_{MAX}}{MSE}\right),\ MSE = \frac{1}{S_U}\sum_{j=0}^{S_U-1}(U_F(j) - U_F'(j))^2, \tag{6}$$

where U_F and U_F' correspond to noisy input and denoised range output of face regions, S_U is number of pixels, and D_{MAX}=7.5m. The results demonstrate a robust denoising performance as the processed output is quite close to the ground truth data. Facial features such as filtrum, nose, and eyelids are apparently visible (c.f. Fig. 4). The denoising improvement is substantial and higher than 14 dB. As commented in [2], a denoising improvement of 12-14 dB ensures some 50-70 percent improvement in recognition when PCA or M(LDA) classifiers are used (see Table 1 in [2]).

Table 1. Denoising performance of proposed algorithms

$I_T[\mu s]$	2000	1000	800	500	400	200	100	80	50
BAD pixels, [%]	1	20	26	44	51	100	100	100	100
IMP pixels[%]	0	0	0	0	0	0.1	1.7	6.5	11.8
Noisy,[dB]	40.18	36.29	35.43	31.64	29.82	23.41	14.92	12.29	14.32
$SNLM_{CLX}$,[dB]	47.23	38.15	39.62	37.89	37.73	36.01	32.75	30.21	30.02

3.2 Face normalization

For the face normalization experiment we have implemented the classical ICP approach as presented in [29]. The following test was performed. A face with given GT

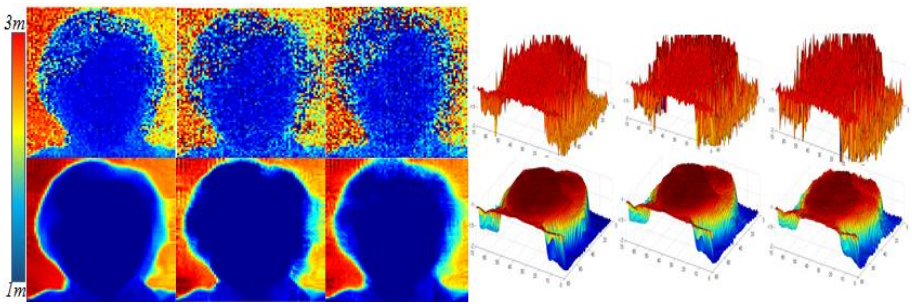

Fig. 4 Image and surface plots. Rows: noisy input and denoised output for I_T = 200, 100, 50 μs.

geometry was transformed in 6 degrees of freedom (DoF) consisting of 3D axial angles Pitch(φ), Yaw (θ), Roll(ψ) and vector of translation shifts $T = (x, y, z)$. The angles got arbitrary values in the (extreme) range [-30°, 30°] and the translation was fixed to a rather big shift of 40 cm. For the face normalization purposes, the proper estimation of T is considered insignificant [4], what is important in our case is to acquire data "*en face*". The results of face normalization by ICP are given in Table 2 and depicted in Fig. 5. Two metrics were utilized: the mean and the variance of misalignment error denoted as E_{MEAN} and E_{VAR} respectively. The results demonstrate the benefit of denoising pre-processing of low-sensed data. The results of E_{MEAN} and E_{VAR} show a substantial decrease of misalignment error (i.e. low E_{MEAN}) and improvement of robustness (i.e. low E_{VAR}). There is an interesting anomaly in the result. The face normalization performance of denoised results for less noisy input data (e.g. I_T=1000μs) is inferior to the ones for shorter integration times (e.g. I_T=50μs), thus noisier. This is explained by the adaptive mechanism of our filter. For the sensed data that is *expected* to have relatively small amount of noise according to higher values of A, our technique does not apply heavy filtering. However, in the case of facial data, low-sensing artifacts caused by multi-reflectivity path is observed in the areas of eye pupils, and hair [21]. They are presented in the depth modality but not in the amplitude one. Still, the obtained results for less noisy data show good robust performance for E_{MEAN}= ~1° and E_{VAR} = ~0°. An additional demonstration of ICP performance for face normalization is given in Fig. 6, where misaligned angle error is visually depicted for the extreme angle misalignments of 30°.

Table 2. Face normalization by ICP

	$I_T[\mu s]$	1000			200			100			50		
	DoF	φ	Θ	Ψ	Φ	θ	ψ	Φ	θ	ψ	φ	θ	Ψ
Noisy	E_{MEAN},[°]	0.4	1.1	1.2	1.9	3.3	2.5	4.1	4.1	4.2	4.4	4.3	4.5
	E_{VAR} ,[°]	0.1	1.3	0.1	2.2	7.4	5	9.5	13	10	14	19	12
Denoised	E_{MEAN} ,[°]	0.5	0.5	1.5	0.3	0.4	0.9	0.2	0.3	0.3	0.5	0.8	0.2
	E_{VAR} ,[°]	0.1	0.4	0	0	0.5	0	0	0.3	0	0	0.7	0

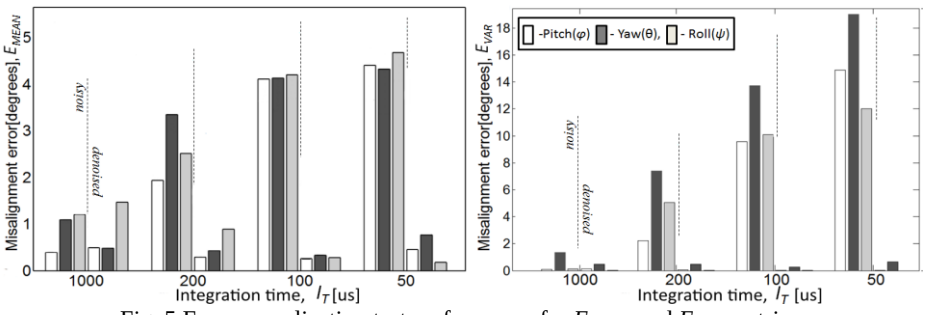

Fig. 5 Face normalization test performance for E_{MEAN} and E_{VAR} metrics

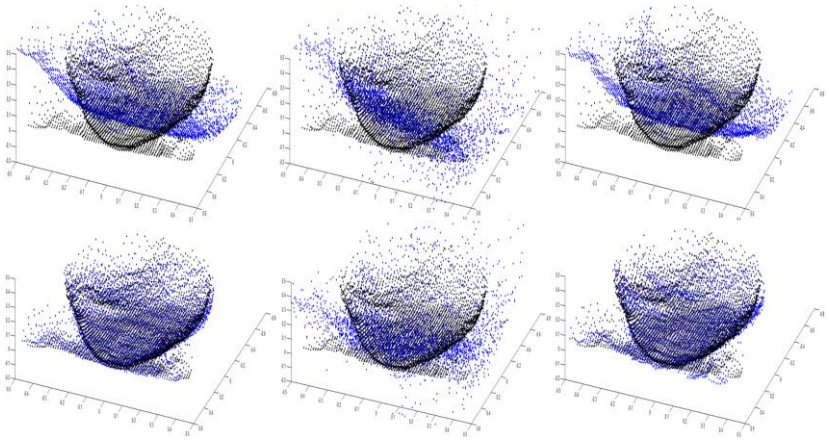

Fig. 6 Demonstration of ICP misalignment performance (columns): a) GT, b) noisy data – I_T=50μs, and c) denoised output; (rows): d) initial displacement – (φ, θ, ψ) = 30°, e) face normalized output by ICP

4 Conclusion

In this paper, we have presented a framework for 3D scene capture and reconstruction. It includes hardware and software systems aimed at efficient sensing and computing in real time. The hardware module includes an RGB color camera of high definition and a ToF active range sensor of rather low resolution. The two sensors are vertically aligned and properly jointly calibrated. The main aim of the computing system is to support the work of the sensors in low-sensing mode. Thus, it contains specific solutions for range data denoising, upscaling and 2D/ToF data fusion. Three-stage denoising approach has been proposed. The first stage employs a version of the NLM method, modified to work with complex-valued spatio-temporal data. The second stage performs denoising in the point cloud by making use of the specific nature of depth data. The third stage utilizes structural information from the aligned color sensor and refines the upscaled depth at the stage of non-uniform resampling. As a result, the 2D+depth data provides 3D scene reconstruction with quality as high as if the range sensor was working in normal sensing mode. The performance of the system has been validated by experiments aimed at preprocessing for typical biometric tasks such as face detection, segmentation, surface mapping and normalization. The denoising improves the signal to noise ratio by more than 14 dB thus providing 2.5 D face data good enough for the subsequent stage of feature extraction and recognition. ICP-based face normalization works also fine on the denoised depth maps. In this case, it turns out that noisier data can lead to even better denoising results due to the amplitude component, which implicitly controls the distances between similar patches in the complex-valued modification of the NLM method. While demonstrated for face recognition tasks, the 3D capture and reconstruction framework is perfectly applicable to other biometric tasks where the availability of dynamic 3D scene is required. Such tasks may include body segmentation and tracking for e.g. human behavioral analysis.

References

1. Schimbinschi, F., Wiering, M., Mohan, R.E., Sheba, J.K.: 4D unconstrained real-time face recognition using a commodity depth camera. In: 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 166–173 (2012)
2. Ebers, O., Plaue, M., Raduntz, T., Barwolff, G., Schwandt, H.: Study on 3D face recognition with continuous-wave time-of-flight range cameras. Berlin, Germany (2011)
3. Ruiz-Sarmiento, J.R., Galindo, C., Gonzalez, J.: Improving Human Face Detection through TOF Cameras for Ambient Intelligence Applications. In: International Symposium on Ambient Intelligence (ISAmI), pp. 125-132 (2011)
4. Kim, J., Yu, S., Kim, I., Lee, S.: 3D Multi-Spectrum Sensor System with Face Recognition. In: IEEE SENSORS, vol. 13(10), pp. 12804-12827 (2013)
5. Van den Bergh, M., van Gool, L.: Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction. In: IEEE Workshop on Applications of Computer Vision, Kona, USA, pp. 66–72 (2011)
6. Bleiweiss, A., Werman, M.: Fusing Time-of-Flight Depth and Color for Real-Time Segmentation and Tracking. In: DAGM Workshop on Dyn. 3D Imaging, Germany, (2009)
7. Mirante, E., Georgiev, M., Gotchev, A.: A fast image segmentation algorithm using color and depth map. In: 3DTV(2011)
8. Kolb, A., Barth, E., Koch, R., and Larsen, R.: Time-of-flight cameras in computer graphics. Computer Graphics Forum, vol. 29(1), pp. 141-159 (2010)
9. Lindner, M., Kolb, A., Hartmann, K.: Data-fusion of PMD-based distance-information and high-resolution RGB-images. In: Symposium on Signals Circuits and Systems (ISSCS), pp. 121-124 (2007)
10. Linarth, A., Penne, J., Liu, B., Jesorsky, O., Kompe, R.: Fast fusion of range and video sensordata. In. Advanced Microsystems for Automotive Applications, pp. 119-134 (2007)
11. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A noise-aware filter for real-time depth upsampling. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, European Conference on Computer Vision (ECCV) (2008)
12. Richardt, C., Stoll, C., Dodgson, N., Seidel, H., Theobalt, C.: Coherent Spatiotemporal Filtering, Upsampling and Rendering of RGBZ Videos. In: Computer Graphics Forum (Proceedings of Eurographics), vol. 31 (2012)
13. Kopf, J., Cohen, M., Lischinski, D., Uyttendaele, M.: Joint Bilateral Upsampling. In: Special Interest Group on Comp. Graphics and Int. Techniques(SIGGRAPH)(2007)
14. Kim, Y.M., Chan, D., Theobalt, C., Thrun, S.: Design and calibration of a multi-view ToF sensor fusion system. In: CVPR W. on Time-of-flight Computer Vision (2008)
15. Zhang, C., Zhang, Z.: Calibration between depth and color sensors for commodity depth cameras. In: Multimedia Expo (ICME), Barcelona, Spain, pp. 1–6 (2011)
16. Herrera, C., Kannala, J.: Joint depth and color camera calibration with distortion correction. In: IEEE Trans. on Patt. Anal. and Machine Intell., vol. 34, pp. 2058–2064 (2012)
17. PMDTechnologies GmbH, PMD[Vision] CamCube 2.0., in Siegen, Germany (2010)
18. Lenzen, F., Kim, K. I., Nair, R., Meister, S., Schäfer, C., Theobalt, C.: Denoising strategies for tof data. In: ToF Imaging: Algorithms, Sensors and Applications (2013)
19. Frank, M., Plaue, M., Hamprecht, F.: Denoising of Continuous-wave Time-of-flight Depth Images Using Confidence Measures. J. of Optical Engineering, vol. 48(7) (2009)

20. Georgiev, M., Gotchev, A., Hannuksela, M.: Joint denoising and fusion of 2D video and depth map sequences sensed by low-powered ToF range sensor. In: ICME(2013)
21. Georgiev, M., Gotchev, A., Hannuksela, M.: Denoising of distance maps sensed by Time-of-Flight devices in poor sensing environment. In: ICASSP(2013)
22. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: International Conference on Computer Vision (ICCV), pp.839-847 (1998)
23. Buades, A., Morel, J.: A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 60-65 (2005)
24. Sankaran, H., Georgiev, M., Gotchev, A., Egiazarian, K.: Non-uniform to uniform image resampling utilizing a 2D farrow structure. In: SMMSP (2007)
25. Strohmer, T.: Efficient methods for digital signal and image reconstruction from nonuniform samples. In: PhD thesis, University of Vienna (1993)
26. Chuchvara, A., Georgiev, M., Gotchev, A.: A speed-optimized RGB-Z capture system with improved denoising capabilities, In: (SPIE), vol. 9019 (2014)
27. Viola, P., Jones, M.: Robust real-time face detection, In: Int. J. of Comp. Vision, vol. 57(2) (2004)
28. Georgiev, M. Gotchev, A., Hannuksela, M.: Real-Time Denoising of ToF Measurements by Spatio-Temporal Non-Local Mean Filtering. In: Hot3D Workshop, pp 1-6 (2013)
29. Rusinkiewicz, S., Levoy, M.: Efficient variants of the ICP algorithm. In: IEEE on 3-D Digital Imaging and Modeling, pp. 145–152 (2001)
30. Besl, P., McKay, N.: A method for registration of 3-D shapes. In: IEEE Trans. Pattern Anal. Mach. Intell., pp. 239–256 (1992)
31. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. In: Pattern Anal. Mach. Intell.,vol. 31, pp. 607–626 (2009)
32. Yin, L., Wei, X., Longo, P., Bhuvanesh, A.: Analyzing facial expressions using intensity-variant 3D data for human computer interaction, In: ICPR., vol. 1, pp. 1248–1251 (2006)
33. Mpiperis, I., Malassiotis, S., Strintzis, M.: Bilinear models for 3-D face and facial expression recognition. In: IEEE Trans. Inf. Forensics Secur., vol. 3 (3) , pp. 498–511 (2008)
34. Pomerleau, F., Colas, F., Ferland, F., Michaud, F.: Relative Motion Threshold for Rejection in ICP Registration. In: Field and Service Robots, pp. 229-238 (2009)
35. Boev, A., Georgiev, M., Gotchev, A., Daskalov, N., Egiazarian, K.: Optimized visualization of stereo images on an OMAP platform with integrated parallax barrier autostereoscopic display. In: European Signal Conference EUSIPCO (2009)