Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 57(1), 46–57(2016)

Marko Lelas, Tomislav Pribanić

Accurate stereo matching using pixel normalized cross correlation in time domain

DOI 10.7305/automatika.2016.01.1465
UDK [004.923:004.932.72]; 528.021.7

Original scientific paper

This paper proposes a novel method for stereo matching which is based on combination of active and passive stereo 3D reconstruction approaches. A laser line is used to scan the reconstructed scene and a stereo camera pair is used for the image acquisition. Each image pixel is scanned at a specific scan time so that the intensity time patterns of the correspondent pixels are highly correlated. This yield in highly confident and accurate disparity map calculation and also allows the reconstruction of poorly textured as well as the extremely textured surface which are very hard to deal with using the conventional passive stereo approaches. The occluded regions are also detected successfully. This method is not computationally intensive and can be used for turning the smartphone into the practical 3D scanner as presented in this work.

Key words: 3D reconstruction, Smartphone, Stereo matching, Temporal domain

Stereo uparivanje iz video isječka. Ovim radom predstavljena je nova metoda stereo uparivanja temeljena na kombinaciji aktivnog i pasivnog stereo pristupa. Rekonstruirana scena skenirana je laserskom linijom, dok se par stereo kamera koristi za akviziciju video isječka. Svaki slikovni element rekonstruirane scene skeniran je laserskom linijom u određenom trenutku stoga su profili intenziteta svjetline u vremenskoj domeni izrazito korelirani za slikovne elemente lijeve i desne kamere koji odgovaraju istom slikovnom element rekonstruirane scene. Stoga je rezultat predstavljene metode određivanje stereo parova slikovnih elemenata s visokom pouzdanošću. Nadalje, predstavljena metoda omogućuje rekonstruiranje izrazito slabo odnosno izrazito intenzivno teksturiranih scena što je često veoma teško postići korištenjem konvencionalnih metoda stereo 3D rekonstrukcije. Metoda je jednostavna te ju je moguće implementirati na sustavima ograničenih računarskih resursa, stoga je iznimno pogodna za primjenu na mobilnim platformama primjerice pametnim telefonima.

Ključne riječi: 3D rekonstrukcija, pametni telefoni, stereo uparivanje, vremenska domena

1 INTRODUCTION

Generation of accurate 3D models has been a goal in computer vision for some time, resulting in the numerous approaches and methods for achieving flexible and accurate systems for 3D reconstruction. The shape from stereo is one of the most powerful approaches which acquires 3D information of a scene using two or more images related to the different viewpoints. In a nutshell, the stereo system determines which point in one image corresponds to which point in another image and subsequently triangulates 3D position from such correspondence. A correspondence can be quantitatively expressed as a disparity value. There are passive stereo [1]and active stereo approaches, e.g. using a structured light (SL) concept [2]. Combining the active and passive stereo approaches shows a great potential in solving the issues such as occlusion detection, reconstruction of extremely low or high textured surfaces, which are very hard to deal with by using just active or just passive stereo approach individually. This insight motivated our work towards examining the possibilities of combining passive and active stereo approaches in the most effective way. Briefly, we propose a part of idea from a so called weakly SL approach [22]; therefore, taking advantage of all pros that SL offers, for example in the case of textures regions. At the same time, the proposed idea uses neither a projector (otherwise specific to SL approach) nor any standard way to retrieve SL code. Instead, a cheap laser source and a camera pair are utilized and well known passive stereo techniques are utilized, however, not in the spatial domain, but rather in the time domain. Moreover, we show an extension of the proposed idea for a smartphone implementation. Namely, despite the great enhancement of the computing resources which are available on the smartphones nowadays, they are still limited platform in the terms of computational power that can be used. Never-

theless, we demonstrate how the proposed method can be also efficiently implemented on the smartphone.

The rest of work is structured as follows: Section 2 presents a short overview of the state of the art work in the field of research while the proposed method is in depth explained in Section 3. Section 4 describes the method chosen to generate the ground truth data and numerically evaluate the results of the proposed method and presents the evaluation results. Section 5 gives the qualitative reconstruction results for the scene scenarios challenging for the conventional stereo methods, while Section 6 presents the results of the proposed method applied to HTC EVO 3D smartphone. Finally, Section 7 concludes with a discussion of the presented results and some sketches of the future work.

2 RELATED WORK

2.1 Passive stereo

The passive stereo methods can be further classified in two main groups called local and global methods [1].

The local methods attempt to match some image block (window) on the left (e.g. referent) image with the set of window candidates on the right image using a winner-take-all approach[3](best match wins).The simplest matching costs assume constant intensities at matching image locations, while more advanced costs implicitly or explicitly consider certain radiometric changes and noise. Common window-based matching costs are absolute or squared differences (SAD/SSD), normalized cross correlation (NCC), rank and census transforms. The main issue of the local algorithms is to determine the optimal support window for each pixel. An ideal support region should be bigger in the poorly textured regions and should be suspended at depth discontinuities. Fixed window size based algorithms are very fast, but yield poor results in poorly textured regions and near depth discontinuities. The more advanced window-based approaches, such as adaptive windows [4], shiftable windows [5] and compact windows [6], try to improve this issue. Local methods are relatively fast and easier to implement. However, they provide disparity solution for the certain window on the referent image, not considering simultaneously also solutions for other windows of the referent image.

Contrary to it global passive stereo methods provide disparities for the whole set of points on the referent image, minimizing a certain type of energy function [1]. The global optimization is usually performed by means of graph cuts, belief propagation or dynamic programming [7]. The acquired results typically outperform local methods but also require significantly more time to compute and an effort implementing it. In the case of the close viewpoints the passive stereo reconstruction results are quite good, but they deteriorate when increasing the viewpoints distance. Besides, basically all passive stereo methods are sensitive to texture less regions and occluded areas. Some methods try to overcome this by assuming the existence of significant features on the images. These methods use such points as reliable feature correspondences and expand them by using a growing-like process to obtain more point correspondences [8]. They are often called seed-growing or region-growing and they perform much better in large perspective distortions and increased occluded areas than the traditional ones. Similarly, other approaches are primarily concentrated on finding the accurate point correspondences as ground control points and they take the matched significant feature points as soft constraints [9]. The downside of this method is a high algorithm complexity resulting in high time needed to obtain an accurate disparity map. In [10] authors propose a two-step expansion based robust dense matching algorithm based on the aforementioned approach but with lower algorithm complexity and therefore a shorter computation time.Finally, we note the problem of reconstructing the sharp discontinuities which is targeted by some methods using the edge-aware filters like bilateral filter [11]. Although these methods advance the stereo vision, they are still highly affected by their local nature of traditional window-based cost aggregation and are vulnerable to the lack of the texture. Aiming to overcome this, semi-global algorithms are developed [12], as well as the non-local solution proposed in the [13].

2.2 Active stereo

To resolve the issue of dense 3D reconstruction of poorly textured surfaces, the active stereo methods introduce structured light techniques in a stereovision system. SL assumes a projection of controlled illumination of the scene through one or more projected patterns, commonly using DLP or LCD video projectors [14]. Projected light patterns have a characteristic structure which is a result of the way the pixels of the projected image are coded. Therefore, detecting the same code on the camera pixels quickly solves the correspondence problem between projector and camera pixels for a large number of points and leads to dense 3D surface acquisition. Based on the concept of coding the image pixels, three main approaches in SL can be identified: the coding techniques in spatial and temporal domain [15], and there is also coding in the frequency domain called Fourier Transform Profilometry (FTP) [16]. SL spatial coding often uses color which is sensitive to object albedo and the leakage between color channels [17]. Temporal coding will require a larger number of images, while FTP like approaches usually suffer from a correct extraction of a signal phase on the images. An attempt to combine an active and passive stereo ap-

proach was presented in [18]. The authors have used the single phase shifting approach (well-known active stereo method) and the traditional window based technique for passive stereo matching demonstratinga simplicity and effectiveness of dealing with occluded regions and poorly textured surfaces. In any case, SL requires the inclusion of a projector in 3D system and therefore contributes to its complexity and cost.

As an alternative of using a projector, in [19] author presented the idea of "weakly structured lighting". It is based on scanning the reconstructed scene by a shadow (projected using the lamp as a light source and a stick to project shadow). The scene pixels are coded in a way that is structured enough to enable the 3D reconstruction and yet the coding process and reconstruction method is much simpler and less computationally intensive then in the case of coded structured light stereo methods. Evidently, such alternative is low budget solution; still it achieves relatively good results compared to the high price and complexity stereo reconstruction systems.

2.3 Laser line scanning

Triangulation-based laser range finders and light-striping techniques are well-known since more than twenty years (e.g. [20], [21]). Beside other active techniques – like structured (coded) light, time of flight, Moiré interferometry, etc. (see e.g. [22] for an overview) – laser range scanners are commonly used for contactlessmeasuring of surfaces and 3D scenes in a wide range of applications. The field of application comprises computer graphics, robotics, industrial design, medical diagnosis, archaeology, multimedia and web design, as well as rapid prototyping and computer-aided quality control. Most commercial laser scan systems use a camera and a laser beam or laser plane.

In [23] authors present a simple hand-held laser scanner based on an analysis of a laser stripes in the camera images. The laser ray has to intersect two things at the same time: the (unknown) surface, and the a priori known reference geometry (usually the background). The visible intersection with the background is used to calibrate the laser, i.e. to calculate the exact 3D pose of the laser plane. The 3D point coordinates of the object's surface are calculated by intersecting the laser plane with the projecting rays. Certainly, the camera must have been calibrated so that its external and internal parameters are exactly known.

There are also more complex laser scanners aiming at high quality reconstruction results at the expanse of the system complexity and price. The laser scanner presented in [24] uses the principle of structured light. The PicoP® projector illuminates the scene with laser lines, a camera observes the line and a computer analyzes the image. The laser beam is steered by a micro-mirror steers to draw a video stream at 60Hz per image and 30kHz per line. The camera takes images with a very short shutter, about 30μs.In the evaluation section authors showed that the scanning laser line striper performs well outside while remaining eye-safe, even in direct sunlight. It is well suited to make 3D maps of natural objects like grass and man-made objects even at the distances of up to 2m.

The results of the reconstruction are highly dependent on the level of accuracy by which the laser line position is estimated. There are three noise sources which are found to influence the results: electrical noise, quantization noise and speckle. The first two are inherent to the image sensor and the last is caused by the nature of laser light. Authors in [25] present the FIR filter based filtering method which successfully deals with artefacts caused by scanning different surfaces with different noise levels.

In the [26] authors investigated the accuracy of the laser scanners giving a survey of error sources in those systems and an evaluation of the achieved reconstruction results.

2.4 Smartphone reconstruction

Field of 3D reconstruction on smartphones was all until recently relatively poorly investigated and therefore there are not many works which deal with this topic. Because of the emerging need for new applications and all the commercial potential which this area has, the situation is more and more changing. Klein and Murray [27] have presented the Parallel Tracking and Mapping (PTAM) system on a mobile phone which aims at enabling different augmented reality applications of the mobile phone. The goal to give the users the possibility to create and maintain the Augmented Reality (AR) content without the need for complex and expansive tools has also motivated authors in [28] to design a 3D reconstruction system on the smartphone based on the shape-from-silhouette approach. Won, Lee and Park [29] have presented the work where they implemented an active 3D shape acquisition method based on photometric stereo. They used a pair of smartphones which collaborated as the master and a slave where a slave was engaged to illuminate the scene (using the flash) from the appropriate viewing points while master recorded the images. In [30] authors have presented a complete on-device 3D reconstruction pipeline for mobile monocular hand-held devices, which generates dense 3D models. However, their method heavily relies on the smartphone embedded sensors such as accelerometer and gyroscope and the subsequent complex processing.

3 PROPOSED METHOD

The proposed method consists of three main parts: image acquisition, stereo correspondence computation and

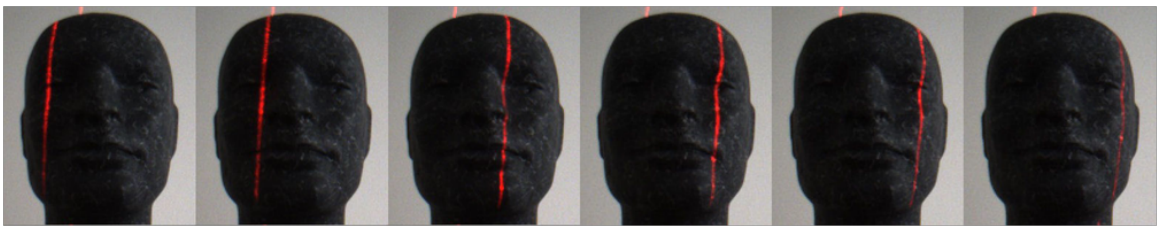

Fig. 1: Scanning the scene with the laser line

3D reconstruction and rendering. The image acquisition consists of simultaneously recording the reconstructed scene while it is being scanned with a laser line. This part could be considered as a part where ground idea of active stereo is being used because scanning the scene with a laser line gives each pixel a strong identification mark (similar to pixel coding in active stereo) which is later used for confident stereo correspondence estimation. Stereo correspondence computation is performed in the classical passive stereo approach manner where normalized cross correlation is used as a cost function. The 3D reconstruction is performed by triangulation between the identified correspondent pixels while 3D rendering is done using the Delaunay triangulation and texture fitting.

3.1 Image acquisition

The image acquisition consists of scanning the reconstructed scene with the laser line while stereo camera pair captures the images of the scene as it is shown on Figure 1. When the laser line crosses the image pixel it increases the light intensity value of the pixel. As the line is moving across the scene, each pixel gets illuminated in the specific frame. Observing the intensity values of the pixel through the consecutive frames gives an intensity value pattern which is specific for each pixel in the row. Figure 2 shows specific intensity patterns of two pixels in the central row of the image.

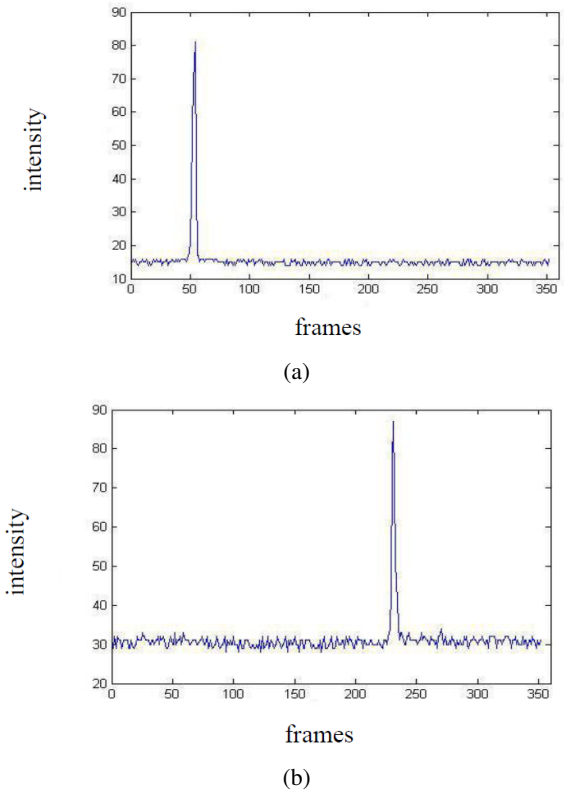

Fig. 2: Intensity patterns of the image pixels

It is easy to notice that for the pixel described by the intensity pattern a) in the Figure 2 the “laser-line time” is at the 50th frame, while for the pixel with intensity pattern b) the time is around the 240th frame. As the images obtained from the stereo camera pair are pre-processed by epipolar rectification, correspondence estimation is one dimensional problem. The matching pixel for a pixel in left image is the pixel in the same row in the right image which has the most correlated (ideally identical) intensity pattern. In other words, every pixel of the reconstructed scene gets an unique intensity value pattern in time domain and correspondence finding is nothing more than locating the pixel with (ideally) the same (intensity pattern. “Coding” pixels in the time domain enables finding correspondences even if the scene is not textured at all or is extremely highly textured because the spatial relationship of the pixels has no influence on the pixel matching process. Though, in our case, the laser line is estimated in time domain opposed to the case of most of other methods where it is estimated in the spatial domain, the FIR filter is designed as a preprocessing step to determine if the noise caused by the laser line influences the reconstruction results.

The Figure 3 shows the result of the laser line filtering and the comparison of the frequency spectrum of the original and filtered signal and the filter amplitude frequency characteristic. The signals marked by red dots represent the original signal and the signals marked by green line

represent the signal after filtering.

Filtering the signal further improved the matching results as the correspondent signals crosscorelation is even higher than in the case of nonfiltered signals. The comparison of the reconstruction results gained using the filtering preprocessing and without it is presented in section 4.

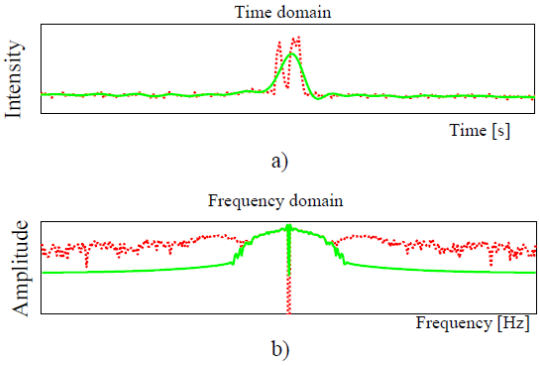

Fig. 3: Laser line filtering : a) time domain b) frequency domain

3.2 Stereo correspondence computation

As mentioned in the previous section, input to the stereo correspondence computation block are pixel intensity patterns. At this point, each pixel in the left (reference) and the right camera image is represented by the one-dimensional signal (intensity pattern in time domain). To find the correspondent pixels normalized cross correlation is calculated between the intensity patterns of the pixel in the left image and intensity patterns of all possible candidates in the right image. Let pl be the pixel intensity signal representing the pixel in the left image, pr the pixel intensity signal of the candidate pixel in the right image and N the number of frames which form the intensity pattern. The normalized cross correlation is calculated as shown in the equation below:

$$nc = \frac{\sum_{i=1}^{i=N} \left(p_l\left(i\right) - mean\left(p_l\right)\right) * \left(p_r\left(i\right) - mean\left(p_r\right)\right)}{std\left(p_l\right) * std\left(p_r\right) * N} \tag{1}$$

The example of the values obtained by calculating normalized cross correlation between the stereo matching candidates is presented in the Figure 4:

Finding cost function which deals with ambiguous matches in the satisfactory way is a very challenging task for passive stereo methods, which measure pixel similarity in the spatial domain. As it can be seen at part a)Figure 4 cross correlation in the time domain gives an extremely

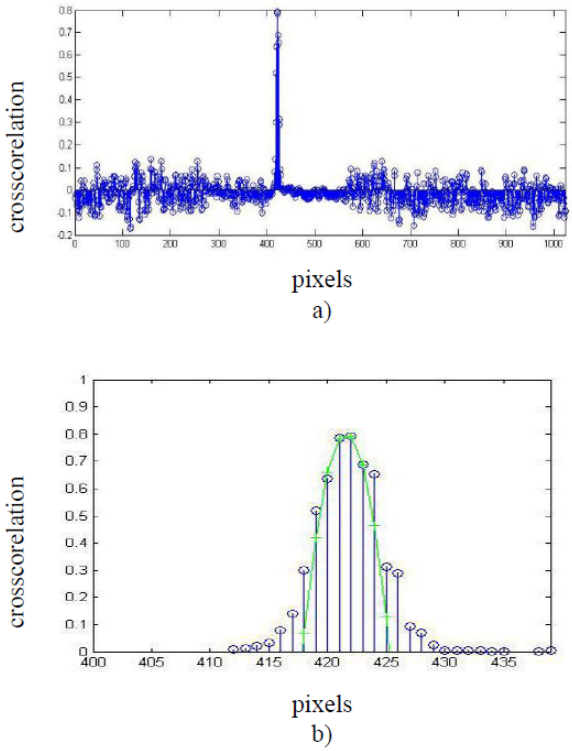

Fig. 4: Correlation between the pixel of the left image and possible candidates in the right image

strong matching candidates. The crossrelation values at the correspondent pixel position is very high while it is low at all other pixel positions. This way the ambiguous matching problem is suppressed. The part b) in Figure 4 shows only the nearest pixels to the strongest matching candidate. The position of the correspondent pixels is determined at the sub pixel (sub frame) position by fitting the parabolic function (marked green in the Figure 4 b)) through the point of the strongest candidate and one nearest point to the left and to the right of the candidate. The position of the maximum value of this parabola is considered the correspondent pixel coordinate. The explained correspondence finding method is also very effective in detecting the occluded pixels. Occlusion detection is performed in three steps. Firstly, the dynamic range of the pixel intensity signal is examined for the pixels of the left camera image. As it can be seen in Figure 2, the intensity value peak which is caused by the laser line passing the pixel is strongly distinguished. Therefore, examining the dynamic range of the intensity pattern gives the decision if the laser line has passed the pixel or not with high confidence. If the decision is made that the laser line did not pass the examined pixel, the pixel is considered occluded (camera did not

"see" the laser line passing through that pixel) and thus the ambiguous matches caused by the occlusion are eliminated even prior the correspondence estimation process.

Secondly, the dynamic range check is also performed for the potential matching candidates in the right camera image to detect the occluded pixels in the right camera view.

Finally, the normalize cross-correlation values are calculated for the pixels which are not occluded, the correlation values are compared to the empirically determined threshold so that only the really strong matching candidates are determined as correspondent matches.

3.3 3D reconstruction and rendering

The disparity map generated in the previous section is used for classical stereo triangulation. Images are captured by the PointGrey BumbleBee stereo camera pair. The camera manufacturer also provides the image processing library called the Triclops API. The supplied library functions were used for the image epipolar rectification and triangulation. Image rendering is done in Matlab using the Delaunay triangulation and texture fitting.

4 QUANTITATIVE EVALUATION OF THE PRESENTED METHOD

To verify the accuracy of the corresponding pixel estimation of the presented method a calibration pattern made of 23 rows and 33 columns of white markers on the black background is used. A function to calculate centers of the markers is applied to both left and right camera images as presented in the Figure 5:

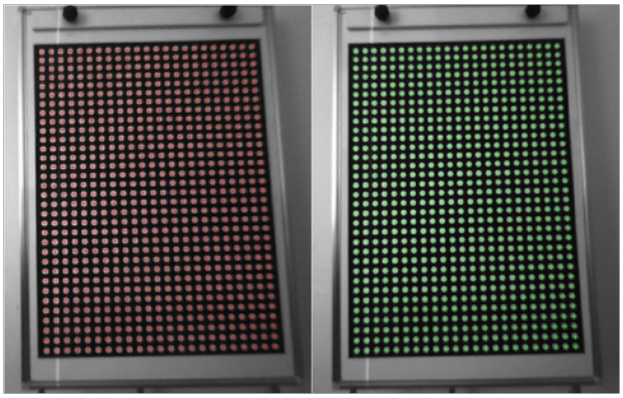

Fig. 5: Centers of markers on left and right camera images

Centers of markers on left camera image are marked by red crosses in the Figure 5 and are represented by pixel coordinates $(ux, vx)_L$, while corresponding pixels on the right camera image are marked by green crosses on the Figure 5and are represented by $(ux, vx)_R$. The correspondent homography matrix H is calculated in a way that satisfies the equation 2:

$$(ux, vx)_R = H * (ux, vx)_L \tag{2}$$

As the presented method assumes that input coordinates for the stereo matching are integer values a new (integer) set of coordinates is deducted according to equation 3:

$$(ux, vx)_{L1} = round((ux, vx)_L) \tag{3}$$

Also the new set of coordinates of the right image is calculated according to equation 4:

$$(ux, vx)_{R1} = H * (ux, vx)_{L1} \tag{4}$$

The coordinates $(ux, vx)_{L1}$ were also used as an input to the proposed stereo correspondence estimation method and thus the $(ux, vx)_{Rmatch}$ coordinates are generated. The calculated $(ux, vx)_{R1}$ values were used as ground truth data for the numerical evaluation of the proposed method. The mean value of the Euclidean distance between the ground truth data and the estimated correspondent pixels is calculated as a measure of accuracy of the proposed method according to the equation 5:

$$err = \sqrt{(ux_{R1} - ux_{Rmatch})^2 + (vx_{R1} - vx_{Rmatch})^2} \tag{5}$$

Due to finite precision of homography matrix calculation, there is a part of the error which is not caused by the matching process but is an inherited system error this error component is denoted err_H. The other component is caused by the matching process err_{match}. This is elaborated by the equation 6:

$$err = err_H + err_{match} \tag{6}$$

The system error can be evaluated as a mean value of the Euclidean distance between the coordinates $(ux, vc)_R$ and$(ux, vx)_{RH}$ which are generated by multiplying the $(ux, vx)_L$with the homography matrix H. System error is calculated following the equation 7:

$$err_H = \sqrt{(ux_R - ux_{RH})^2 + (vx_R - vx_{RH})^2} \tag{7}$$

Finally, the error caused by the matching process is given in the equation 8:

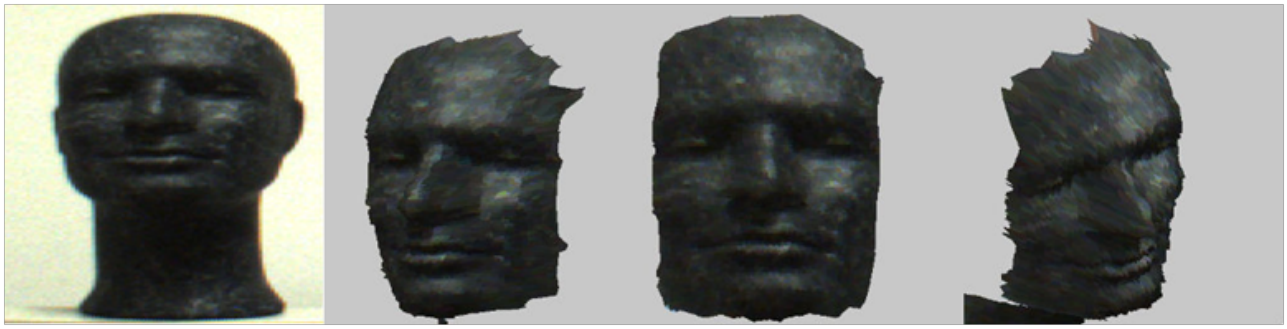

Fig. 6: Reconstruction results Scenario 1

$$err_{match} = err - err_H \tag{8}$$

The error caused by the finite precision of homography matrix calculation is equal:

$$err_H = 0.3792$$

Error in the case when filtering preprocessing is not implemented equals:

$$err_{match} = err - err_H = 0.5992 - 0.3792 = 0.22$$

When the preprocessing is implemented, the error is slightly smaller and equals:

$$err_{match} = err - err_H = 0.5741 - 0.3792 = 0.1949$$

Although the matching results are enhanced by implementing the filtering, in the case of method application on the mobile platform the performance advantage does not match up to the increased method complexity. Therefore, preprocessing is not implemented on the smartphone.

4.1 Left-right consistency check

To ensure that the ambiguous matches caused by occlusions and other sources are removed, the left-right consistency check is introduced at the end of the correspondence estimation step.

In the first iteration, for the pixels in the left image $(ux, vx)_L$, correspondent matches are found in the right image. Similarly to the previously explained evaluation process, values calculated by the matching process are rounded. Rounded values were used as input to another correspondence estimation process resulting in the set of pixels $(ux, vx)_{RL}$. The error calculated according to the eq. 9:

$$err_{LR} = \sqrt{(ux_L - ux_{RL})^2 + (vx_L - vx_{RL})^2} \tag{9}$$

Shows that the $err_{LR} = 0.0952$ thus confirming that the resulting pixels were consistent to the pixels taken as input in the step 1 of the consistency check process up to the measure of error caused by rounding and the finite calculation precision.

5 RECONSTRUCTION

The proposed method is verified in three scenarios which have in common that the conventional stereo reconstruction methods have trouble dealing with them. First reconstructed object is a model of a human head which is a very dark and poorly textured object. Finding stereo correspondences with most of the common passive stereo algorithms would result in high number of ambiguous matches. As it can be seen from the Figure 5, the proposed method works well in this scenario giving the satisfactory results.

Figure 6 shows the source image and the left, front and right view of the reconstruction results.

The second scenario is highly textured surface which would be extremely challenging for active stereo methods and also for passive stereo methods which utilize edge detection as a part of algorithm. Figure 7 shows that our method succeeds in reconstructing these kind of textured scenes successfully:

The last scenario is a very poorly textured surface such as a white painted wall and a white mannequin. As explained previously in the text, these kinds of surfaces are also very challenging for passive stereo algorithms. As Figure 8 and Figure 9 show, the proposed method deals with such a demand in a satisfactory way.

The right image in the Figure 9 shows results of 3D reconstruction of white mannequin produced with the 3D reconstruction demo application provided by the PointGrey.

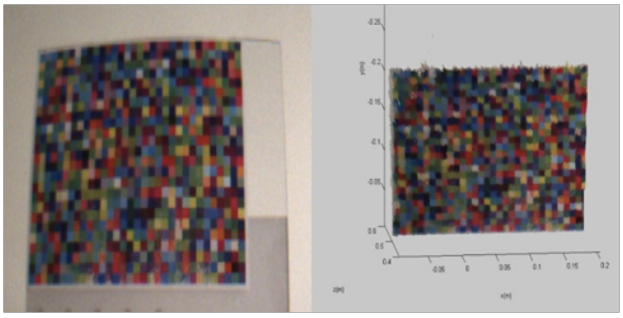

Fig. 7: Reconstruction results - highly textured surface

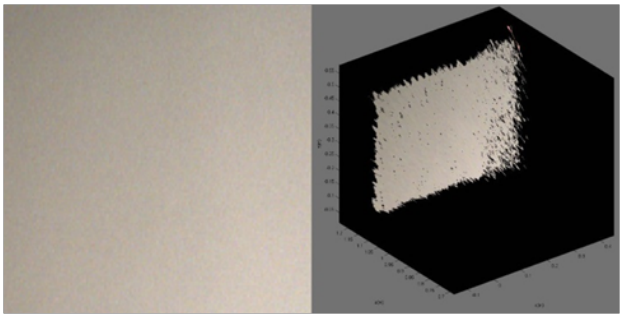

Fig. 8: Results of the reconstruction - poorly textured surface

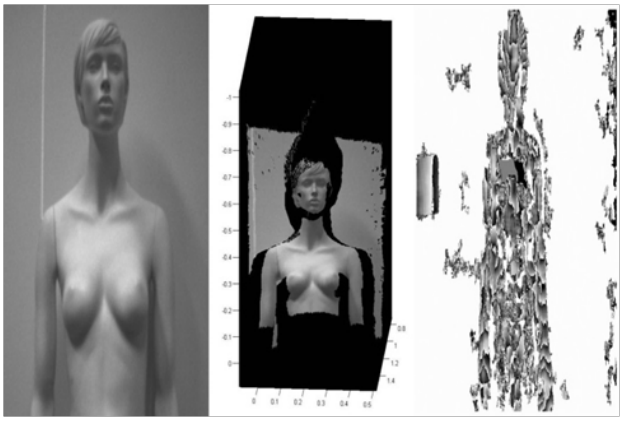

Fig. 9: Results of the reconstruction – mannequin

The white wall in the background cannot be reconstructed at all, while the mannequin is not reconstructed at the satisfactory way due to ambiguous matches caused by lack of scene texture. The central image in the Figure 9is the result of reconstruction using the presented method showing that both mannequin and background are reconstructed. The black gaps in the reconstructed results are caused by occlusion (the laser line was not visible by both cameras). However, occluded regions are well detected and there are no reconstruction errors caused by occlusion.

6 SMARTPHONE IMPLEMENTATAION

The HTC EVO 3D smartphone has two rear-facing cameras pointing our research towards implementing the proposed method on the smartphone. Having in mind that the smartphone also has a flash light source, the laser line is not needed. A simple stick can be used to cast a shadow scanning the reconstructed scene the same way as laser line does.

Without any prior knowledge about the orientation and position of the cameras which are coupled to make a stereo pair, the correspondence estimation is a complicated and computationally intensive process. To simplify the correspondence estimation process and turn it into one dimensional problem, an epipolar constraint is introduced. In our work, we have used a pinhole camera model. For completeness, we next shortly describe the model and the epipolar constraint.

6.1 Camera model

As depicted by Figure 10, the pinhole camera model is applied. The model is defined by its optical center C and its *image plane*. The 3D point is projected into an image point by intersecting the image plane with the line containing the optical center and the point. The line containing optical center and orthogonal to the image plane is called *principal axis*and its intersection with the image plane *R*is called *principal point.* The distance between the *optical center* and *image plane* is *focal length*.

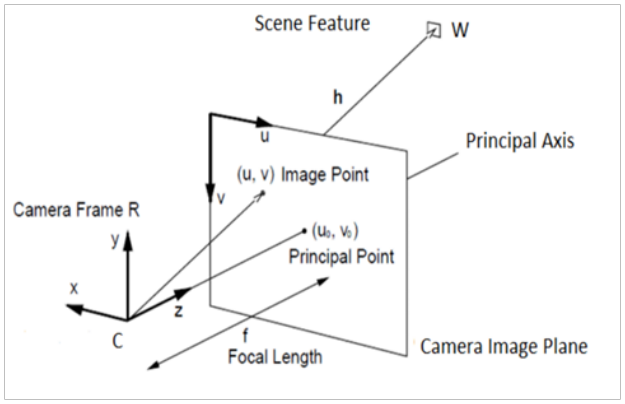

Fig. 10: Pinhole camera model

Let $W = [X\ Y\ Z]^T$ be the coordinates of a scene feature (3D point) in the world reference frame and $m = [u\ v]^T$ the coordinates of image point in the image plane.

The mapping from 3D to 2D is the perspective projection, which is represented by a linear transformation in *homogenous coordinates*. Let $m_h = [u\ v\ 1]^T$ and $W_h = [X\ Y\ Z\ 1]^T$ be the homogenous coordinates of m and W respectively. The perspective transformation is given by the *projection matrix P*:

$$m_h \cong P \cdot w_h \tag{10}$$

where $\cong$ means equal up to a scale factor.

The projection matrix can be decomposed, using the QR factorization, into the product:

$$P = K \cdot [R|t] \tag{11}$$

The matrix K depends on the intrinsic parameters only and has the following form:

$$K = \begin{bmatrix} f_u & \gamma & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

where f_u and f_v are the focal lengths in horizontal and vertical pixels, (u_0, v_0) are the coordinates of the principal point and γ is the skew factor that models the non-orthogonal $u - v$ axes (for the orthogonal axes $\gamma = 0$).

The extrinsic parameters (camera position and orientation) are encoded by the 3x3 rotation matrix R and translation vector t, respectively.

Estimation of intrinsic and extrinsic camera parameters is called camera calibration. There are number of methods to obtain the calibration. In our work we have used [31] and deeper explanation of this topic exceeds the scope of the work.

6.2 Epipolar rectification

Using the camera model explained in the previous section, a model which binds two cameras into a stereo rig is presented in the Figure 11:

Let C1 and C2 be optical centers of the left and the right camera. The points m1and m2 are the projections of the 3D point W onto image planes. m1 in the left image plane, its conjugate point in the right image is constrained to lie on a line called the epipolar line (of m1). Since m1 may be the projection of an arbitrary point on its optical ray, the epipolar line is the projection through C2 of the optical ray of m1. All the epipolar lines in one image plane pass through a common point (e1and e2, respectively) called the epipole, which is the projection of the optical center of the other camera. When C1 is in the focal plane of the right camera, the right epipole is at infinity, and the epipolar lines form a bundle of parallel lines in the right image. A very special case is when both epipoles are at infinity, that

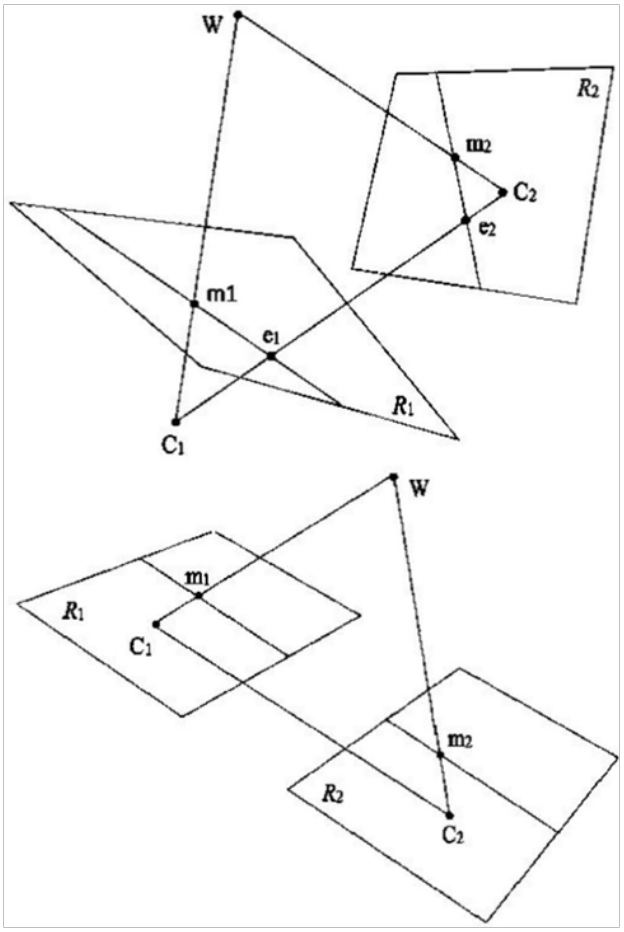

Fig. 11: Epipolar rectification

happens when the line C1C2 (the baseline) is contained in both focal planes, i.e., the retinal planes are parallel to the baseline. Epipolar lines, then, form a bundle of parallel lines in both images. Any pair of images can be transformed so that epipolar lines are parallel and horizontal in each image. This procedure is called rectification[32].

We assume that the stereo rig is calibrated, i.e., the the projection matrixes P1 and P2 are known. The idea behind rectification is to define two new matrices Pn1 and Pn2 obtained by rotating the old ones around their optical centers until focal planes becomes coplanar, thereby containing the baseline. This ensures that epipoles are at infinity; hence, epipolar lines are parallel. To have horizontal epipolar lines, the baseline mustbe parallel to the new X axis of both cameras. In addition,to have a proper rectification, conjugate points must have the same vertical coordinate. This is obtained by requiring that the new cameras have the same intrinsic parameters. Note that, being the focal length the same, retinal planes are coplanar too, as can be seen in the right part of the Figure 11.

Rectified images are input to the stereo matching process described in the Section 3.2Stereo correspondence computation.

6.3 Experimental results

In the presented work the white colored stick was used to ensure that the stick is not confused with the shadow in the cases when the stick is also in the cameras field of view.

The experimental setup is shown in the Figure 12:

Fig. 12: Smartphone reconstruction experimental setup

Using the shadow instead of a laser line has no influence on the presented method, so in this case occlusions are also very efficiently removed.

Figure 13 presents the results of the 3D reconstruction of a white (poorly textured) ball:

Figure 14 presents the 3D reconstruction results when a little figure of white angel is reconstructed:

As it is shown in the figures above, the implemented method is successful in reconstruction of poorly textured objects. It applies both to the case of a simple object such as a little white ball and a figure of a little angel which has detailed surface.

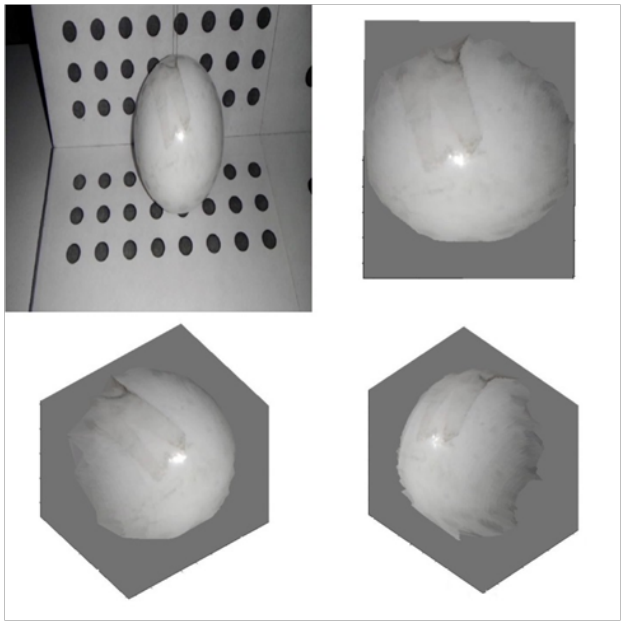

Fig. 13: 3D reconstruction of a white ball

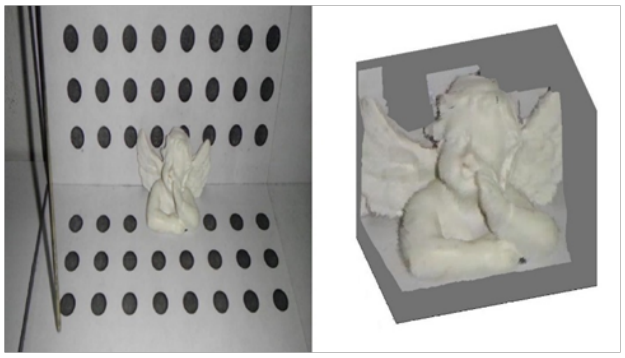

Fig. 14: 3D reconstruction of a small figure

7 CONCLUSION

The presented method shows that combining active and passive stereo approaches yields with satisfactory results in the scenarios where conventional approaches fail due to ambiguous matches. Poorly textured scenes are known to be extremely challenging for the passive stereo methods and this work has shown to provide good reconstruction results of such a scene scenario. The other scenario which is hard to handle with conventional methods is a highly textured scene where the object edges are hard to determine, has also shown not to be an obstacle for the presented method. Other than the ability to cope with the texture less or highly textured objects, the method has also shown a very good accuracy in stereo correspondence estimation. Processing the pixel intensity values in the time domain is shown to be a very simple yet efficient approach as the normalized cross correlation between the intensity signals yields with strong matching candidates enabling accurate stereo matching and providing the way to detect occluded regions with high confidence. Due to its simplicity, the method implementation imposes low requirements for processing resources and therefore it is suitable for application on the smartphone thus turning the smartphone into easy to use practical 3D scanner.

In the scope of the future work application of dynamic programming to refine the matching results is planned in

which case the 3D reconstruction results are expected to be enhanced by leveraging the advantages of global stereo matching approach.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," International Journal of Computer Vision, vol. 47, p. 7-42, 2002.

[2] J. Salvi, S. Fernandez, T. Pribanic and X. Llado, "A State of the Art in Structured Light Patterns for Surface Profilometry," Pattern Recognition, vol. 43, pp. 2666-2680, 2010.

[3] A. Geiger, M. Roser and R. Urtasun, "Efficient large-scale stereo matching," in 10th Asian Conference on Computer Vision, 2010.

[4] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 9, pp. 920-932, September 1994.

[5] A. Bobick and S. Intille, "Large occlusion stereo," International Journal of Computer Vision, vol. 33, pp. 181-200, 1999.

[6] O. Veksler, "Stereo Matching by Compact Windows via Minimum Ratio Cycle," ICCV, vol. 1, pp. 540–547, 2001.

[7] L. Valgaerts, A. Bruhn, M.Mainberger and J.Weickert, "Dense versus sparse approaches for estimating the fundamental matrix," International Journal of Computer Vision, vol. 96, no. 2, pp. 212-234, January 2012.

[8] H.Wu, Z. Song, J. Yao, L. Li and Y. Gu, "Stereo matching based on support points propagation," Proceeding of IEEE International Conference on Information Science and Technology, p. 23-25, March 2012.

[9] L.Wang, H. Jin and R. Yang, "Search space reduction for MRF stereo," Proceedings of the European Conference on Computer Vision, vol. 1, pp. 576-588, 2008.

[10] L. Wang, Z. Liu and Z. Zhang, "Feature Based Stereo Matching Using Two-Step Expansion," vol. 2014, 2014.

[11] Q. Yang, "Hardware-efficient bilateral filtering for stereo matching," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 5, pp. 1026-1032, 2014.

[12] H. Hirschmueller, "Stereo processing by semi-global matching and mutual information," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 328-341, 2008.

[13] Q. Yang, "Stereo Matching Using Tree Filtering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 4, pp. 834-846, April 2015.

[14] T.Pribanic, S. Mrvos and J. Salvi, "Efficient multiple phase shift patterns for dense 3D acquisition in structured light scanning," Image and Vision Computing, vol. 28, no. 8, pp. 1255-1266, 2010.

[15] J.Salvi, J.Pages and J.Batlle, "Pattern codification strategies in structured light systems," Pattern Recognition, vol. 37, no. 4, p. 827-849, 2004.

[16] X.Su and W.Chen, "Fourier transform profilometry : a review," Optics and Lasers in Engineering, vol. 35, no. 5, p. 263-284, 2001.

[17] K.Boyer and A.Kak, "Color-encoded structured light for rapid active ranging," vol. 9, no. 1, p. 14-28, 1987.

[18] T. Pribanic, N. Obradovic and J. Salvi, "Stereo computation combining structured light and passive stereo matching," Optics Communications, vol. 285, pp. 1017-1022, 2012.

[19] J. Bouguet, Visual methods for three-dimensional modelling. A Thesis Submitted in partial fulfilment of the requirements of the California Institute of Technology for the degree of Doctor of Philosophy. California Institute of Technology Pasadena, California, Pasadena, 1999.

[20] F. Pipitone and T. Marshall, "A wide-field scanning triangulation rangefinder for machine vision," International Journal of Robotics Research, vol. 2, no. 1, pp. 39-49, 1983.

[21] E. L. Hall, J. B. K. Tio, C. A. McPherson and F. A. Sadjadi, "Measuring Curved Surfaces for Robot Vision," Computer, vol. 15, pp. 42-54, 1982.

[22] F. Blais, "Review of 20 years of range sensor development.," J. Electronic Imaging, vol. 13, no. 1, pp. 231-243, 2004.

[23] S. Winkelbach, S. Molkenstruck and F. M. Wahl, "Low-Cost Laser Range Scanner and Fast Surface Registration Approach.," in DAGM-Symposium, 2006.

[24] C. Mertz, "Performance of a scanning laser line striper in outdoor lighting," in SPIE Defense, Security, and Sensing, conference 8741, 2013.

[25] J. Forest, J. Salvi, E. Cabruja and C. Pous, "Laser stripe peak detector for 3D scanners, a FIR filter approach," in 17th International Conference on Pattern Recognition, Cambridge, 2004.

[26] W. Boehler, M. B. Vicent and A. Marbs, "Investigating laser scanner accuracy," in XIXth CIPA Symposium at Antalya, 2003.

[27] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," in Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007.

[28] A. Hartl, L. Gruber, C. Arth, S. Hauswiesner and D. Schmalstieg, "Rapid reconstruction of small objects on mobile phones," in IEEE Computer Society Computer Vision and Pattern Recognition Workshops, 2011.

[29] J. H. Won, Lee, M. H., Park and I. K., "Active 3D shape acquisition using smartphones," in IEEE Computer Society Computer Vision and Pattern Recognition Workshops, 2012.

[30] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer and M. Pollefeys, "Live Metric 3D Reconstruction on Mobile Phones," in IEEE International Conference on Computer Vision, 2013.

[31] Z. Zhang, "A Flexible New Technique for Camera Calibration," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

[32] A. Fusiello, E. Trucco and A. Verri, "A compact algorithm for rectification of stereo pairs.," Mach. Vis. Appl., vol. 12, no. 1, pp. 16-22, 2000.

Marko Lelas graduate from the Faculty of electrical engineering at University of Zagreb in 2007. Since the beginning of 2008 he was working as FPGA engineer in Xylond.o.o. He has been working on the numerous industrial projects involving, among other things, parking assistance, pedestrian detection, automatic video quality enhancement etc. Since 2014 he is a Research Associate at University of Zagreb Faculty of Electrical Engineering and Computing. He is also enrolled in PhD program. His research is focused on computer vison methods and algorithms for 3D stereo reconstructions on mobile platforms.

Tomislav Pribanić is an Associate Professor at University of Zagreb Faculty of Electrical Engineering and Computing. He teaches several undergraduate and graduate courses in the field of algorithms and data structures, image processing, sensors and human motion analysis. His main research interests include computer vision and biomedical signal measurement and analysis. He has led a number of scientific domestic and international projects, collaborating with the researchers from EU and outside EU. He was a visiting researcher at INRIA Rhone-Alpes, Grenoble, France and Fraunhofer IGD, Darmstadt, Germany. Results of his research have been implemented in technological projects and he has received recognitions for innovations as well. He is a member of IEEE, IFMBE and a collaborating member of the Croatian Academy of Engineering.

AUTHORS' ADDRESSES

Marko Lelas,
Prof. Tomislav Pribanić, Ph. D.,
Department of Electronic Systems and Information Processing,
Faculty of Electrical Engineering and Computing,
University of Zagreb,
Unska 3, 10000 Zagreb, Croatia
email: marko.lelas@gmail.com, tomislav.pribanic@fer.hr

Received: 2015-09-03
Accepted: 2016-01-15