Multi-stereo Matching for Light Field Camera Arrays

Ségolène Rogge, Beerend Ceulemans, Quentin Bolsée and Adrian Munteanu
Department of Electronics and Informatics
Vrije Universiteit Brussel, Belgium
srogge@etrovub.be, bceulema@etrovub.be, qbolsee@etrovub.be, acmuntea@etrovub.be

***Abstract*—Light field cameras capture information about the incoming light from multiple directions, going beyond classical capturing of light intensity performed by regular RGB cameras. This enables the computation of more accurate depth maps compared to stereo methods based on conventional cameras. However, the very small angular resolution of light field cameras limits their practical use in 3D applications. In this paper, we introduce for the first time in the literature the use of light field camera arrays, with the aim of improving the depth maps while providing a wide field of view. In this context, a novel algorithm for multi-stereo matching based on light field camera arrays is proposed. The disparity maps for the sub-aperture images are computed based on light field camera pairs using a novel multi-scale and multi-window stereo-matching algorithm. A global energy minimization based on belief propagation is proposed to regularize the results. The resulting depth maps are efficiently fused by means of k-means clustering. The proposed approach demonstrates very promising results for accurate 3D scene reconstruction and free navigation applications.**

I. Introduction

Dynamic 3D scenes are conventionally acquired with arrays of RGB cameras possibly enriched with time-of-flight sensors. It is well-known that depth maps that are estimated using (multi-)stereo methods are subject to noise which impairs their use for advanced applications, such as free navigation, multi-view synthesis, and accurate 3D scene reconstruction. Despite of the stringent need for highly accurate depth information, it is still particularly challenging to extract depth maps of sufficient quality using current active or passive depth-sensing methods. In particular, objects with fine details are especially difficult for multi-stereo methods, as such features occupy small volumes and are only visible in a small portion of the available views.

Light field cameras record the light intensity in a scene but also the incoming direction of the light rays hitting the photo sensor. Light field images contain different depth cues, such as correspondence and defocus, which can be combined to compute a depth map [1]. The technique from [1] was modified by Wang et al. [2] to deal with occlusions, and by Tao et al. [3] who developed an algorithm for dense depth estimation that combines defocus and correspondence metrics, and refines fine object details using shading under the Lambertian assumption. The state-of-the art in depth estimation in light field cameras based on correspondence cues is the work of Navarro et al. [4]. In [4] disparity maps are computed on pairs of sub-aperture images, and subsequently fused to obtain a robust and dense depth map.

However, the low angular resolution of light field cameras often limits their practical use in 3D applications. Moreover, the extremely small baselines in light field cameras yield noisy depths maps in difficult areas. This calls for using light field camera arrays such that the 3D information can be captured within a much wider field of view. To the best of our knowledge, arrays of light field cameras have not yet been studied in the literature. In this work, we explore the combination of several light field cameras in order to compute more accurate depth maps compared to those extracted from only a single light field camera, hereby greatly improving the angular resolution. To achieve this, multiple depth maps are estimated using a multi-scale multi-window matching on different pairs of sub-aperture images [5]. We consider pairs of sub-aperture images in the same light field camera (intra-mode) as well as pairs across different cameras (inter-mode). Using the estimates from different images, the depth of the center view of each light field camera can be refined by multi-hypothesis prediction.

The paper is structured as follows. Section II details the proposed multi-stereo method for light field camera arrays. Section III presents the depth fusion algorithm, whereas section IV reports the experimental results obtained with the proposed method. Finally, section V draws the conclusions of this work.

II. Disparity map computation

Currently, depth maps can be determined using either time-of-flight cameras or RGB stereo-pairs. For the latter, two images of the same scene are captured from different locations. By determining corresponding pixels in the two camera views, one determines a disparity map which expresses the displacement of pixels from one image to the other. We know that the disparity of a pixel is inversely proportional to its depth, i.e. pixels closer to the camera appear to move more than pixels further away. If the cameras are calibrated, the depth of each pixel can be computed from the estimated disparity. This technique is known as *stereo-matching*.

The method proposed in this work builds on the basic principles of correspondence matching for light field cameras proposed in [4]. The approach of [4] demonstrates state-of-the-art performance for single light field camera depth estimation,

ISBN 978-90-827970-1-5

Fig. 1. Windows used in the stereo-matching algorithm.

which is due to performing multi-scale and multi-window stereo-matching [5]. However, in contrast to [4], which is essentially a local disparity estimation method, we propose a global optimization method for disparity selection, which brings robustness and improves the depth estimation results over [4]. Details are provided next.

A. Multi-window matching

This section explains the core of the stereo-matching algorithm, in which we attempt to find for every pixel the most likely disparity $d \in [minD, maxD]$. This disparity is the perceived horizontal displacement of a pixel as viewed in two different images, and is inversely proportional to the distance of that point to the camera. In case the images are rectified, this displacement is purely horizontal, otherwise it follows the epipolar line. To select an appropriate disparity, we compute the zero-mean SSD cost (ZSSD) of associating a pixel p in a reference image u to a pixel q in a search image v by minimizing:

$$C_W(p,q) = \frac{1}{|W|} \sum_{t \in W} \Big| \Big(u(p+t) - \overline{u_W(p)} \Big) - \Big(v(q+t) - \overline{v_W(q)} \Big) \Big|^2, \quad (1)$$

where W are windows centered in pixels p and q in images u and v respectively, $\overline{u_W(p)}$ and $\overline{v_W(q)}$ are the average intensities within the windows in images u and v, respectively, and $|W|$ is the size of W.

In order to find the best correspondence, different window-shapes are taken into account (Fig. 1). The usage of multiple windows improves the accuracy at objects boundaries and $3D$ surfaces that are not fronto-parallel to the camera plane.

B. Multi-scale refinement

An important aspect of the stereo-matching algorithm is to appropriately choose the range $[minD, maxD]$ of values the disparity can take. It has to be wide enough to be sure that the disparity of every pixel is considered. On the other hand, we want to keep this range as tight as possible to speed-up the estimation procedure and to prevent it from selecting wrong disparity values.

For this reason we apply a coarse-to-fine approach. We build a Gaussian image pyramid using logarithmic scale factors, and the disparity is estimated using the multi-window algorithm explained in Section II-A, the results being propagated from coarser to finer levels. If a pixel is marked as reliable, its disparity range at the next level is then reduced to the disparity values of that same pixel and the one of its neighbors in this level. Otherwise, the next scale considers the full disparity range for that same pixel. The different criteria to reject unreliable pixels are explained in more detail in Section II-C.

In practice, the true disparity of a particular pixel is rarely an integer value. For this reason, our method is able to search for matches at an arbitrary sub-pixel accuracy (half-pixel, quarter-pixel, etc.). This is particularly important in intra-mode where the disparity values are often less than a full pixel. Sub-pixel matching is implemented using Lanczos filtering and re-sampling [12]; this increases in principle the disparity search range, making it very important to keep tight upper- and lower-bounds of likely disparity values per pixel in order to maintain reasonable computational loads.

C. Rejection criteria

To determine whether the disparity-estimate for a pixel is reliable or not, four different rejection criteria are applied:

a) Fattening detection: We consider the neighborhood contained in the current window, selected among the ones of Figure 1, centred on the considered pixel p. In this neighborhood we select the pixel with the lowest match-cost and two random pixels. Then, the $3D$ plane formed by these three pixels and their estimated disparities is computed. Following a RANSAC-approach [11], this process is iterated a few times, and the plane fitting best the neighborhood of the central pixel is kept. Finally, the disparity $d(p)$ is rejected if it is too far away from this plane.

b) Match ambiguity: For every pixel, we find the cost of the best match within the same image [5]. If this cost is smaller than the cost of the best match in the other image, the pixel is rejected because it is likely part of a repetitive texture.

c) Left right (LR) consistency: This rejection criterion is the most often used in stereo-matching. It verifies that the disparity found for a given pixel x in the left view corresponds to corresponding pixel in the right image. Disparity differences larger than τ are marked as unreliable. Formally, the disparity for a pixel is rejected if:

$$|d_R(x) - d_L(x + d_R(x))| > \tau, \quad (2)$$

where in practice we choose τ equal to one.

d) Isolated matches: This last criterion rejects a pixel if it is isolated, meaning that all its 8-connected neighbors are either unreliable or have a disparity difference larger than a given threshold (which we set to 1). If this is the case, the pixel is likely to be a mismatch and we have the multi-window matching a the next level consider the full disparity range.

D. Optimal disparity selection using belief propagation

In [5], a winner-takes-all (WTA) approach is followed, selecting the disparity d of each pixel as the one having the smallest cost over all match-windows. However, such greedy selection methods are known to be prone to noise and methods relying on global optimization are known to give more robust results [6]. We have thus decided to use the *belief propagation algorithm* [7], [10] to compute the disparity maps. A more detailed explanation is given at the end of this section. Pseudo

ISBN 978-90-827970-1-5 © EURASIP 2018

code for the WTA and BP methods can be found in Algorithms 1 and 2, respectively.

In [5], a disparity map is computed for each window. To obtain a final disparity map, the algorithm of [5] selects for each pixel the disparity value that has the smallest cost over all windows. As this is a local method, we improve the final disparity selection by using a global optimization strategy. In contrast to this WTA approach, in our method the final disparity map is obtained by minimizing a global energy function over the entire image. The energy function balances the minimization of the local ZSSD cost per pixel with a first-order Markov term that promotes piece-wise smoothness of the final depth map. This is obtained by minimizing the energy:

$$E(l_1, l_2, \ldots, l_N) = \sum_{i=1}^{N} V_d(i, l_i) + \sum_{j \in N_i} V_s(l_i, l_j), \quad (3)$$

$$V_d(i, d) = \min C_W(i, i \pm d), \quad (4)$$

$$V_s(l_i, l_j) = \lambda * \min\left(\tau, abs(l_i - l_j)\right). \quad (5)$$

In these equations, the data term $V_d(i, d)$ is the cost of assigning a disparity d (Eq. (1)) to a given node i in the MRF. The smoothness $V_s(l_i, l_j)$ term assesses the compatibility of two adjacent nodes taking on labels l_i and l_j, respectively. As adjacent pixels are often part of the same object, they should be likely to have similar disparity values, so this situation is strongly encouraged by this energy function. τ is a truncation factor used to reject too different labels that can either be wrong of be of distinct objects, while λ is a smoothness factor.

Algorithm 1 Multi-window WTA algorithm of [5]

for each window w **do**
  $disp_w \leftarrow ZSSDMatching(u, v, dMin, dMax, w)$
  Update $disp_w$ applying Fattening detection
  Update $disp_w$ applying Match Ambiguity detection
  Update $disp_w$ applying LR criterion
  Update $disp_w$ removing Isolated Matches
end for
$disp \leftarrow$ Combine all the $disp_w$
Update $disp$ applying LR criterion
Update $disp$ removing Isolated Matches

Algorithm 2 Proposed multi-window algorithm using Belief Propagation

for each disparity $d \in [dMin, dMax]$ **do**
  $finalCost[d] \leftarrow matrix(maxCost)$
  for each window w **do**
    $cost(d, w) \leftarrow ZSSDCost(u, v, d, w)$
    $finalCost[d] \leftarrow min(finalCost[d], cost(d, w))$
  end for
end for
$disp \leftarrow BeliefPropagation(finalCost)$
Update $disp$ applying LR criterion
Update $disp$ removing Isolated Matches

To minimize the global energy function as defined in Eq. (3), we employ the max-sum variant of the belief propagation algorithm [7] on the 8-connected grid of the left and right images. The max-sum method is a global optimization method aiming here at finding a minimum total MRF energy solution for the final disparity map. Essentially, after performing our multi-scale multi-window disparity selection, each pixel is associated with a disparity range indicating possible disparities. Subsequently, the belief propagation will then select one optimal disparity value within this range. During each iteration of the algorithm, information is passed through the MRF by means of directional message passing alternating between horizontal, vertical and diagonal sweeping directions. The messages exchanged between neighbouring nodes i and j are:

$$m_{ij}(l) = \max_{l_j} \left[-V_d(i, l_j) - V_s(l, l_j) + \sum_{\substack{k \in \mathcal{N}_i \\ k \neq j}} m_{ki}(l_j) \right]. \quad (6)$$

This information represents the belief of a node in a given disparity. It is made of a data term V_d and a smoothness term V_s. After convergence, the disparity with the highest belief is assigned to each node:

$$D(i) = \arg\max b_i(d), \quad (7)$$

with:

$$b_i(d) = -V_d(i, d) - \sum_{k \in \mathcal{N}_i} m_{ki}(d). \quad (8)$$

III. Disparity map merging

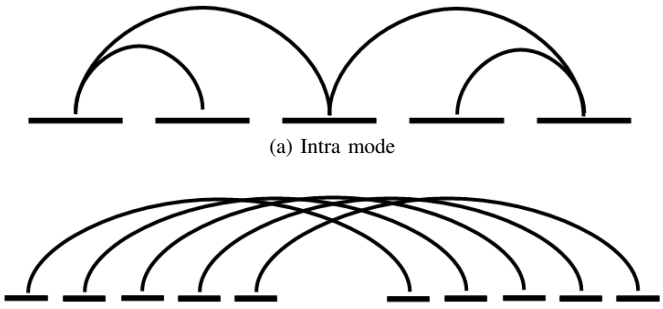
The goal of this multi-stereo-matching is to improve the robustness of the generated depth maps using light field cameras; these cameras produce much richer information compared to regular RGB cameras, but they suffer from small angular resolutions. We solve this problem by using light-field camera pairs, leading us to two groups of images for which we have to compute depth maps. That is, we perform (i) depth estimation within each light field camera, and (ii) multi-stereo matching in between light field cameras. The resulting depth maps need to be merged to obtain the final depth maps at the center of each light-field camera.

A. Image pairing

The stereo-matching algorithm explained in Section II was applied within a single camera - *intra mode* - and between sup-aperture images of different cameras - *inter mode* - to compute disparity maps at every image location. Pair selections for the intra- and inter-modes are shown in Figure 2.

These two modes have different properties. On one hand, the disparity maps are more difficult to compute for the *intra mode* because of the small baseline between the images, leading to disparity values often smaller than a pixel. On the other hand, the *inter mode* has the disadvantage of big occlusion zone in which the disparities cannot be computed. By computing the disparity maps at each image position using both modes,

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) Intra mode

(b) Inter mode

Fig. 2. Different stereo-matching modes used.

we are thus able to combine their advantages and to obtain accurate depth maps for every light-field camera in the array.

B. Occlusion-handling

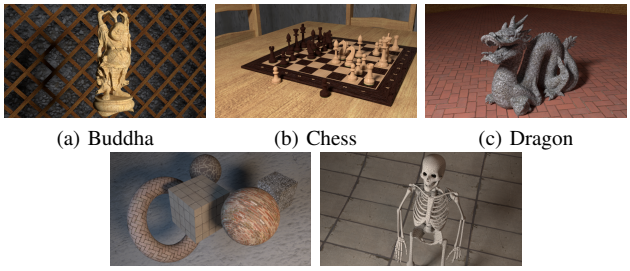
Both intra- and inter-mode depth maps are reprojected to the central view of each light field camera, ignoring pixels that were marked as unreliable. This gives, for every pixel of the central depth map, a set of candidate disparity values. In some regions, these disparity values are close to each other, however, they can differ significantly in other regions that were occluded in one or more sub-aperture images. As explained in [8], occlusions occur at object boundaries, hiding the background for certain view points. When re-projecting, the same pixel will then get values from the background and the foreground object, and the latter should not be kept.

To separate the foreground and the background disparity values, a k-means clustering [9] method was used. After outlier-removal, the cluster with the smallest centroid, corresponding to the background, was selected. If the other centroids were close enough, they were kept as well. Finally, an average was made, leading to a final disparity map at the camera center.

This procedure generates two disparity maps for each light field camera center: intra and inter. The accuracy of the inter-depth-estimate is higher than the intra one, but it contains more occlusions. Therefore, the last step was to use the intra disparity maps to fill-in holes of the inter disparity, leading to a final, accurate and complete disparity map.

IV. Experimental results

To the best of our knowledge, no previous work addresses arrays of light field cameras. In order to have ground-truth data for objective evaluation, we rendered a new dataset. The dataset simulates two light field cameras with a lenslet array of five by five sup-aperture images. This work is the starting point in the domain and only makes use of the five images of the central row in each light field camera. The scenes, were designed with different textures, shapes and backgrounds, and rendered in *Blender 2.78*. Central views of the different datasets are shown in Figure 3.



(a) Buddha (b) Chess (c) Dragon

(d) Geometric (e) Skeleton

Fig. 3. Dataset used in the experiments.

TABLE I
PSNR values of disparity maps using WTA and BP

	Dataset	Buades' method [5]	Proposed method (BP).
Intra	Buddha	33.7829	**34.0620**
	Chess	29.7416	**29.9529**
	Dragon	31.3294	**32.3040**
	Geometric	33.9428	**35.4944**
	Skeleton	36.7150	**38.4359**
Inter	Buddha	33.7983	**34.5498**
	Chess	30.2827	**30.3567**
	Dragon	33.3455	**33.4564**
	Geometric	36.7338	**36.8131**
	Skeleton	40.5818	**40.6898**

A. Belief propagation improvements

To assess the efficiency of the belief propagation method in the context of this stereo-matching algorithm, in Table I we report the PSNR results obtained using both the reference WTA method of [5] and the proposed BP method when operating in both intra- and inter- modes.

We note that our algorithm systematically outperforms the state-of-the art winner-takes-all (WTA) method of [5]. The differences are substantial in the intra case, with PSNR gains up to 1.72dB. This is an important improvement, demonstrating that the proposed BP method substantially improves performance over the original WTA method of [5].

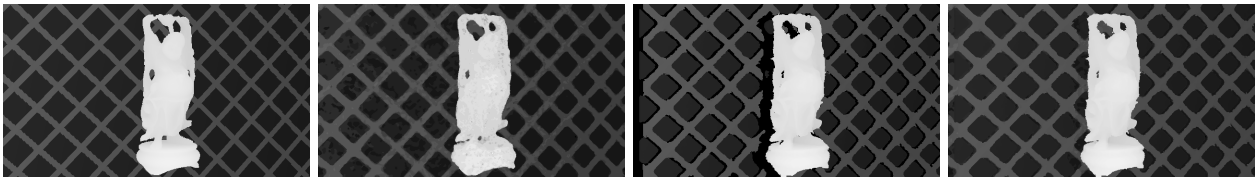
B. Merged depth maps

The final result of our work can be seen in Figure 4. As it was expected, we observe that the inter mode is more accurate than the intra mode, while having more occlusions. Those occlusions were filled-in using the intra mode, giving us our final hole-free disparity maps at each camera location.

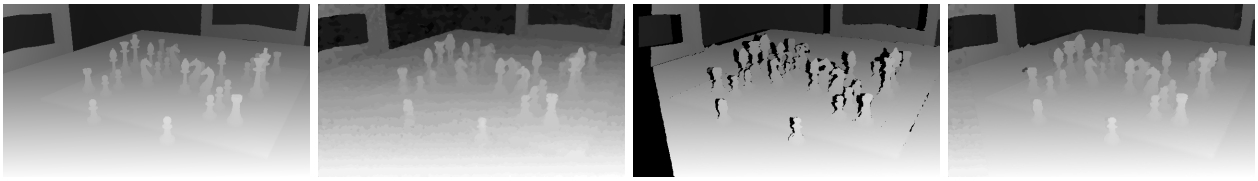
V. Conclusions

In this paper we have introduced a novel method to estimate depth from multiple light field cameras. The proposed method uses multi-window matching for accurate correspondence finding, multi-scale processing for both robustness and speed, and global energy minimization to select the optimal depth values for each pixel. We have shown that it is possible to obtain accurate depth maps within a wide field-of-view by optimally combining the disparity maps computed based on different pairs of sub-aperture images in the light field camera array.

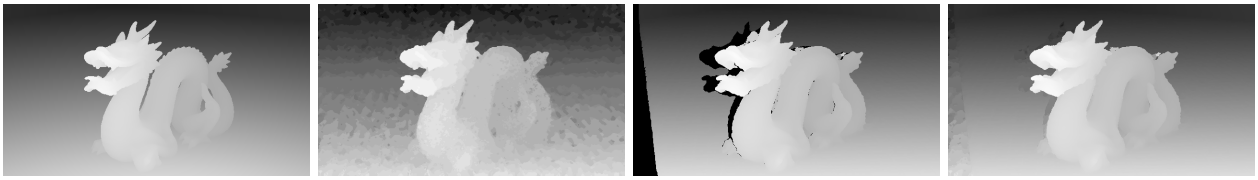
ISBN 978-90-827970-1-5 © EURASIP 2018



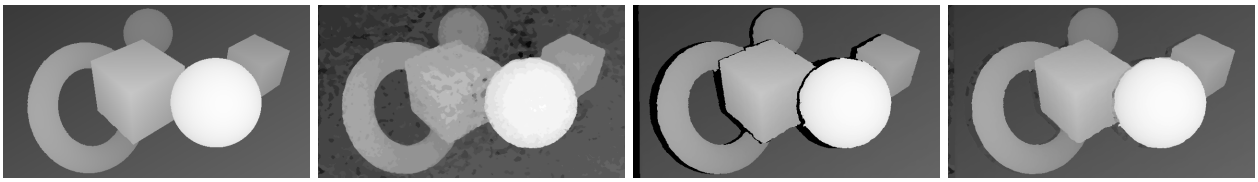
(a) Buddha



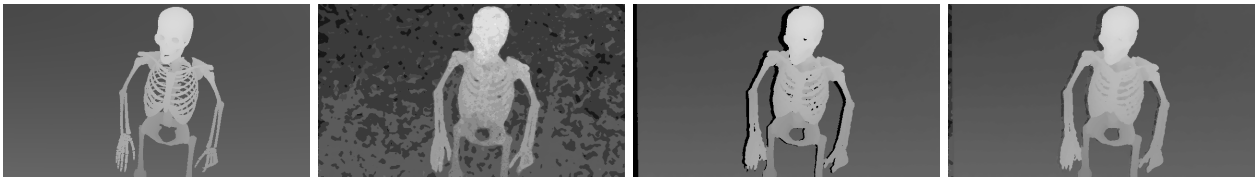
(b) Chess



(c) Dragon



(d) Geometric



(e) Skeleton

Fig. 4. Final results : from left to right, (i) ground truth, (ii) intra-, (iii) inter- and (iv) merged disparity maps.

ACKNOWLEDGMENTS

The first author is a FWO-SB PhD fellow funded by the Fund for Scientific Research-Flanders (FWO), project number 1S83118N.

REFERENCES

[1] M. W. Tao, S. Hadap, J. Malik and R. Ramamoorthi, *Depth from combining defocus and correspondence using light-field cameras*, IEEE International Conference on Computer Vision (ICCV), pp. 673-680, 2013.

[2] T.-C. Wang, A. A. Efros and R. Ramamoorthi, *Occlusion-aware depth estimation using light-field cameras*, IEEE International Conference on Computer Vision (ICCV), pp. 3487-3495, 2015.

[3] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik and R. Ramamoorthi, *Shape estimation from shading, defocus, and correspondence using light-field angular coherence*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 39, pp. 546-560, 2017.

[4] J. Navarro and A. Buades, *Robust and dense depth estimation for light field images*, IEEE Transactions on Image Processing (TIP), vol. 26, no. 4, pp. 1873-1886, 2017.

[5] A. Buades and G. Facciolo, *Reliable multiscale and multiwindow stereo matching*, SIAM Journal on Imaging Sciences, vol. 8, pp. 888-915, 2015.

[6] D. M. Nguyen, J. Hanca, S.-P. Lu, P. Schelkens and A. Munteanu, *Accuracy and robustness evaluation in stereo matching*, The international society for optics and photonics (SPIE) Optical Engineering + Applications, pp. 12, 2016.

[7] J. S. Yedidia, W. T. Freeman and Y. Weiss, *Understanding belief propagation and its generalization*, Exploring artificial intelligence in the new millennium, vol. 8, pp. 236-239, 2003.

[8] P. Buyssens, M. Daisy, D. Tschumperl and O. Lzoray, *Superpixel-based depth map inpainting for RGB-D view synthesis*, IEEE International Conference on Image Processing (ICIP), 2015.

[9] D. Sisodia, L. Singh, S. Sisodia and K. Saxena, *Clustering techniques: a brief survey of different clustering algorithms*, International Journal of latest trends in engineering and technology, vol. 1, 2012.

[10] J. Sun, N.-N. Zheng and H.-Y. Suhm, *Stereo matching using belief propagation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 25, pp. 787-800, 2003.

[11] S. Choi, T. Kim and W. Yu, *Performance evaluation of RANSAC family*, British Machine Vision Conference, 2009.

[12] B. N. Madhukar and R. Narendra, *Lanczos resampling for the digital processing of remotely sensed images*, International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN), pp. 403-411, 2013.

ISBN 978-90-827970-1-5 © EURASIP 2018