DEPTH ESTIMATION IN INTEGRAL IMAGES BY ANCHORING OPTIMIZATION TECHNIQUES

D. Zarpalas[a], I. Biperis[a], E. Fotiadou[a], E. Lyka[b], P. Daras[a] and M. G. Strintzis[b]

[a] Informatics & Telematics Institute,
1^{st} km Thermi-Panorama Rd, P.O. GR57001, Thessaloniki, Greece

[b] Information Processing Laboratory, Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, GR54124, Thessaloniki, Greece

ABSTRACT

This paper presents two algorithms for estimating depth from integral images, which capture a scene by using multiple lenses, offering anaglyph depictions. The first algorithm involves the 3-D integral imaging grid formed by casting rays inversely through the lenses used to capture the integral image. In this formulation, depth estimation is equivalent to finding correspondences on the ray-crossing points. The second algorithm follows the depth-through disparity approach. In this case, a stereo-like minimization problem is formulated which is handled by the graph cuts method. The novelty of the proposed paper lies in constraining the optimization procedures with the "anchor points". This results in enhanced estimation accuracy, while eliminating the optimization complexity. Anchor points is a set of reliable reference points, detected by applying a robust local image descriptor to viewpoint images, called self-similarity descriptor. The performance of both algorithms is evaluated on a synthetic integral image database in comparison with another state-of-the-art algorithm.

Index Terms— Integral image, depth estimation, image correlation, disparity, self-similarity descriptor, graph cuts.

1. INTRODUCTION

Following the modern trend towards 3-D viewing technologies, integral imaging, a photography technique dating back to 1908 [1], has attracted growing research interest over the last decade. The main idea involves an array of lenses over a film sheet or an electronic image sensor that captures a special 2-D recording of the viewed scene [2]. Multiple neighbouring lenses capture overlapping regions of the scene thus, allowing parallax information to be encoded into the integral image. Depending on the shape of the lenses, cylindrical or spherical, integral images are divided into unidirectional and omnidirectional respectively. Cylindrical lenses offer only horizontal parallax, while spherical offer full parallax information. Integral imaging offers many advantages over other 3-D sensing techniques, e.g. sensing 3-D in full parallax, in full natural colour, in real time and without calibration, that make it appealing to many fields where depth information is essential, e.g. biometrics, medical imaging, robotic vision, etc. However, explicit depth extraction from implicitly encoded parallax information into the integral image is not a trivial task.

Towards this direction, this paper presents two techniques for depth estimation in unidirectional integral images, though both are easily expandable in omnidirectional too. The concept of the first technique is exactly the inverse of the capturing concept (Figure 1). The intersections of rays that emanate from integral image pixels and are headed, through the lenses, to the viewed object form a 3D grid of possible object elements' positions (Figure 4). Pixels of the integral image are to be paired and through triangulation the actual 3-D positions are revealed. The problem can be mathematically formulated as a global optimization problem of seeking the subset of the capturing ray intersections that correspond to the object's outer surface. The second technique is based on the depth-through-disparity concept. Disparities are found between consecutive viewpoint images (*i.e.* images created by re-arranging the integral image pixels [2]). Actual disparities are found through an optimization procedure that is handled by the well known graph cuts approach [3]. For each pair of consecutive viewpoint images a partial depth-map is produced and the final complete is calculated by merging all of them. The difference among the two proposed techniques lies in the fact that in the first case the problem is treated as an actual 3-D problem, while in the second one as a merging of multiple stereo-like problems.

Both of the proposed techniques are based on optimization procedures, which often suffer from high complexity and more importantly they get trapped on local extrema. To overcome this, constraints are introduced, which refer to reliable

This work was supported by the EU funded project "3D VIVANT", GA-248420.

978-1-61284-350-6/11/$26.00 ©2011 IEEE

estimations of specific points' depths, based on a robust correspondence finding technique. This leads to enhanced depth accuracy in both cases, while the optimization burden gets downscaled. The employed point correspondence technique detects pairs of pixels in viewpoint images that correspond to the same object points, thus enabling the computation of their 3-D coordinates. These points, which are called anchor points, are detected with extremely high accuracy following an image matching approach based on the local "self-similarity" image descriptor [4]. As it will become evident in the following sections, the usage of the anchor points leads to high quality depth maps.

To evaluate the performance of the proposed algorithms and show their advantages, a small scale database of synthetic integral images was built and a comparative study with another state-of-the-art technique, proposed by Wu *et al.* [5], was performed.

This paper is organized as follows. Section 2 first describes the basics of integral image generation and viewpoint image extraction and then proceeds with a report on the related work. Section 3 presents the proposed algorithms and their contributions. Finally, the evaluation framework is presented in Section 4, where a description of the database used in the experiments, and quantitative and qualitative comparisons with the algorithm described in [5] are provided.

2. FUNDAMENTALS AND RELATED WORK

Integral imaging involves a simultaneous recording of a 3-D object through multiple lenses placed in a regular grid in front of a recording material, as shown in Figure 1. In the unidirectional case, the behaviour of a cylindrical lens can be modelled by a thin slot on the surface of the cylinder. The projection of a 3-D point $P = (X, Y, Z)$ to the image plane (Figure 2) is a composite projection including an orthographic projection along the Y axis and a projective projection along the X axis. If P lies inside the k-th lens' field of view, then the local coordinates (x_{e_k}, y_{e_k}) of its projection onto the image plane behind the k-th lens are given by:

$$x_{e_k} = -f\frac{X - d_{e_k}}{Z} \quad , \quad y_{e_k} = Y \tag{1}$$

where f stands for the focal length and d_{e_k} is the horizontal displacement of the k-th lens' slot from the origin.

The image formed behind each lens is called elemental image (EI). Re-arranging the columns of EIs leads to the formation of viewpoint images (VI). The first VI is formed by the first columns of the EIs, the second VI is formed by the second columns of the EIs and so on. By construction, the VIs can be considered as orthographic projections of the object onto successively rotated planes whose angles derive by lens geometry. Specifically, if L is the number of lenses, R is the horizontal pixel resolution of the lens, H is its vertical pixel resolution and p is its pitch, then there can be R VIs,

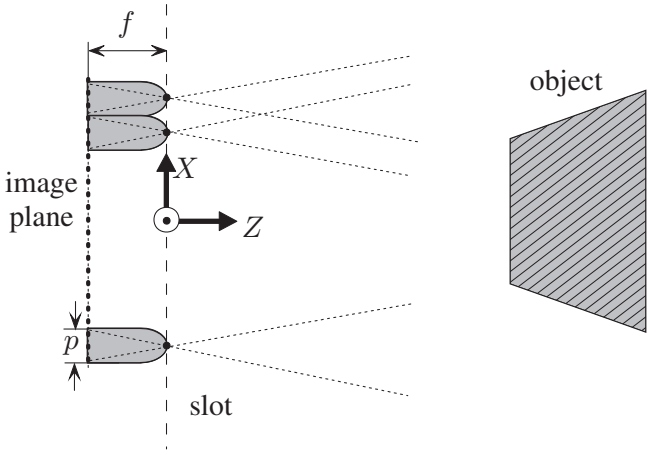

Fig. 1. Set-up for integral image generation. Top view of a cylindrical lens array.

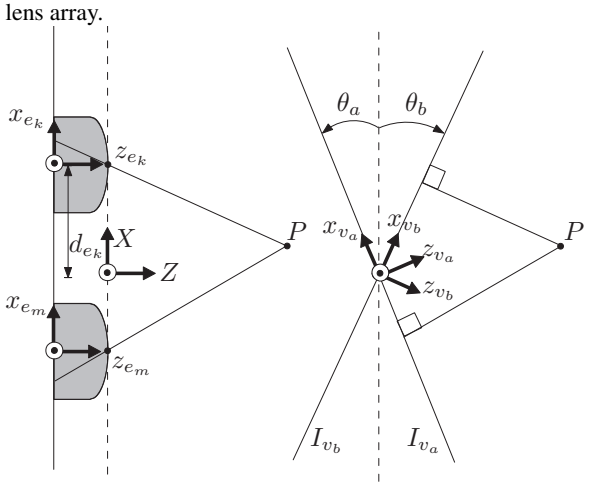

Fig. 2. Local coordinate systems associated with elemental and viewpoint images.

each with $L \times H$ resolution and angle θ with the image plane ranging in $[-atan(\frac{p}{2f}), atan(\frac{p}{2f})]$. VIs depict the object under various directions (Figure 2), hence their name. Point's P 3-D coordinates (X, Y, Z) and its "viewpoint" coordinates (x_{v_r}, y_{v_r}) in the local coordinate system of the r-th VI are related by:

$$x_{v_r} = X\cos\theta - Z\sin\theta \quad , \quad y_{v_r} = Y \tag{2}$$

Finally, using the above equations a useful formula for retrieving 3-D coordinates from VI pixels can be derived. If I_{v_a}, I_{v_b} denote two VIs forming angles θ_a, θ_b with the image plane respectively (Figure 2) and also pixels $p_a = (x_{v_a}, y_{v_a})$ in I_{v_a} and $p_b = (x_{v_b}, y_{v_b})$ in I_{v_b} correspond to point $P = (X, Y, Z)$, then the 3-D coordinates of P can be retrieved by:

$$X = \frac{x_{v_b}\sin\theta_a - x_{v_a}\sin\theta_b}{\sin(\theta_a - \theta_b)} \tag{3a}$$

$$Y = y_{v_a} = y_{v_b} \tag{3b}$$

$$Z = \frac{x_{v_b}\cos\theta_a - x_{v_a}\cos\theta_b}{\sin(\theta_a - \theta_b)} \tag{3c}$$

Though a really old idea, integral imaging attracted substantial research interest only recently. Early works on depth estimation [6, 7] were based on modelling the entire optical system by the point spread function of the recording and then casting depth estimation as an inverse problem. However this approach leads to an ill-posed problem formulation suitable for simulation purposes rather than practical applications.

Most recent works approach depth estimation as a stereo matching problem. The main line of work includes VIs extraction, disparity estimation between them and then straightforward depth computation from estimated disparities. Depth is a direct function of the disparity and therefore accurate disparity establishment is the key issue for high quality depth map generation.

Following this approach, in [8], Wu *et al.* test several correlation-based disparity estimation methods and show some preliminary results for one synthetic integral image. To improve disparity accuracy, in a later work [5], authors include a smoothness term in the formulation of the correlation-based disparity metric. Disparity between the blocks of a pair of VIs is now defined not only by the correlation metric between the blocks, but also by a weighted sum of the correlation metrics between their neighbouring blocks. The improvement brought by smoothness incorporation is demonstrated on a small set of synthetic and real integral images. Completing the limited literature, Park *et al.* [9] work along the same line and propose a method that uses a lens array consisting of vertically long rectangular lens elements. They also form VIs and apply a modified correlation-based multibaseline stereo algorithm that reduces quantization error in depth extraction.

3. PROPOSED DEPTH ESTIMATION ALGORITHMS

This section presents the two depth estimation algorithms, which are both primarily based on constraining an optimization procedure through the set of points, called anchor points. Thus, firstly the procedure followed to acquire anchor points is described, followed by the presentation of its twofold usage.

3.1. Anchor Point Detection

Capturing the same object by slightly different angles, obviously increases the integral image self-correlation. Thus, due to integral imaging nature, inherently there exists rich information that could be used for depth estimation and it should be exploited. Toward this, tractable correspondences among the VIs should be identified, to compute the associated points' 3-D coordinates. Despite that rich self correlated information, a series of experiments showed that simple correlation based metrics failed to robustly indicate true correspondences. Hence, for the proposed work, a more sophisticated local descriptor was chosen, namely the "self-similarity descriptor" [4].

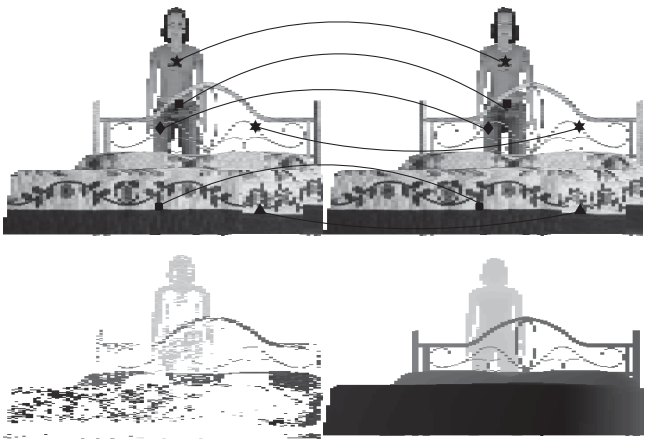

Fig. 3. An illustration of anchor point detection. Top row: A few samples from corresponding pixels on viewpoint images. Second row: Estimated depth for all anchor points (left) and real depth for entire object (right). The brighter a pixel, the higher its depth.

Anchor point detection starts with the computation of the self-similarity descriptor for every pixel of every VI. For each pixel, its surrounding patch is compared with every same-sized overlapping patch in its larger surrounding region. The output is a 2-D array listing the sum of squared distances for each patch comparison. After a normalization step, the SSD array undergoes a log-polar partition forming the local self-similarity descriptor.

Descriptors corresponding to large homogeneous image regions are non-informative and they are filtered out. In order to match the remaining informative pixels of the central VI[1], with pixels in the other VIs, the sigmoid on the L_1 distance between the descriptors is used. To identify the most reliable pairs of corresponding pixels, only the ones with high similarity are considered. In order a point to be regarded as an anchor point, it should have a chain of correspondences of high similarity in consecutive VIs. The chain should have length equal with the number of the VIs minus the two. This is to accommodate for points of the outermost VIs that are not depicted in the rest. Those chained pairs of correspondences should robustly lie in a very very narrow 3D region, otherwise the chain gets filtered out. The anchor point's 3-D coordinates are computed by least squares fitting in that region.

This procedure yields a set of anchor points whose 3-D position estimation is of very high accuracy and therefore they can be safely regarded as true object points (Figure 5). Figure 3 [2] depicts six anchor point correspondences in two consecutive VIs, and the depth-map produced by the complete set of anchor points. As it can be seen, anchor points lie on texture-rich regions. In this example the mean depth estimation rel-

[1]The central VI is the one who has zero angle with the image plane.

[2]For a coloured version of the figure, and other supplemental material, please visit `ftp.iti.gr/pub/Holoscopy`

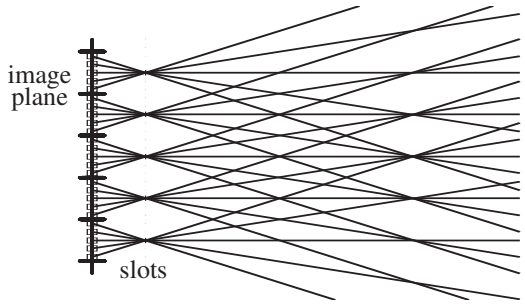

Fig. 4. The 3-D grid formed by casting rays from pixels through lenses' slots.

ative error of those anchor points, comparing with the actual, is 2.37% with a variance of 5.96%. The importance of the anchor points will become evident in the experiments sections, where their influence in both optimization problems will be quantitatively stated.

3.2. Depth Estimation Using the Anchored 3-D Integral Imaging Grid

In this approach, depth estimation is motivated by inverting the principle idea of capturing an integral image. Assuming that each pixel in the image plane is the projection of a single point lying on the object's surface, the corresponding ray of projection for each pixel can be defined (Figure 4). It is obvious that under this assumption the lens array can capture only 3-D points lying on these rays and that object's surface points are on their intersections. To estimate the 3-D coordinates of a point, which under this concept is a rays' intersection, at least two corresponding pixels should be identified. Therefore, the vertices of the 3-D grid formed by rays' intersections forms the set of all possible object points. The problem thus translates to identifying among all vertices, the subset of those that truly belong to the object. Defining a cost function C_v for each vertex, an optimization problem can be formed: the minimization of E, which sums the costs of the vertices that will be chosen, subject to the obvious fact that only one vertex can be chosen on each ray, i.e.,

$$E = \sum_{\forall v \in \mathcal{N}} x_v C_v, \quad \text{subject to:} \quad \sum_{j_r \in \mathcal{J}_r} x_{f(j_r)} = 1 \quad \forall r \in \mathcal{R}, \tag{4}$$

where $\mathcal{N}$ is the number of vertices, $x_v \in \{0, 1\}$ denotes if vertex v belongs to the object surface, $\mathcal{R}$ is the number of rays, $\mathcal{J}_r$ is the set of vertices on ray r and $f(j_r)$ is a re-indexing function. Anchor points are used on the grid to restrict it by offering their accurate position predictions as grid's intersections, thus a reliable starting point for the minimization procedure. Further, they do simplify the grid by deleting intersections, as rays can stop expanding once an anchor point is found.

By the construction of the grid, the set of pixels to which each vertex is projected is known. Under this projection model, it can be assumed that for a vertex belonging to the surface of the object, its associated pixels should record the same colour. However, it is likely that only few vertices will fall exactly on the object's surface, since in practice the surface will lie between the vertices. Also, in practice, pixels do not record just a point element on the object's surface, but actually a small patch of the surface around this point element. Considering these facts, the assumption should be relaxed so that for each vertex belonging to the object's surface, the image regions around its associated pixels should have similar colours. Mathematically, this can be expressed by a correlation-based metric between the neighbourhoods of the associated pixels. Motivated by this conclusion, each vertex v is assigned a cost C_v such that

$$C_v = \min_{(p,q) \in A_v} (1 - M_{p,q}) \tag{5a}$$

$$M_{p,q} = \frac{\sum_{w \in W} \left(I_p(w) - \bar{I}_p\right)\left(I_q(w) - \bar{I}_q\right)}{\sqrt{\sum_{w \in W} \left(I_p(w) - \bar{I}_p\right)^2} \sqrt{\sum_{w \in W} \left(I_q(w) - \bar{I}_q\right)^2}} \tag{5b}$$

where A_v denotes the set of pixels associated to vertex v, I_p and I_q denote the image regions of size W around pixels p and q respectively and $\bar{I}_p$, $\bar{I}_q$ the mean values of I_p and I_q. The notion of neighbourhood can be considered either in the elemental images or in the VIs. The latter were experimentally shown to yield better results.

An obvious disadvantage of this approach, however, is that even after inserting the anchor points, still a huge number of vertices has to be checked. In an effort to further reduce this number, only vertices lying on rays parallel to the Z axis (horizontal rays in Figure 4) are checked.

To improve depth map quality, a filtering post-processing procedure is also followed, taking advantage of the accurate depth estimations of the anchor points. In an effort to propagate this reliable information, pixels around anchor points whose depth estimations differs more than 50% from the neighbouring anchor point's depth are detected. For these pixels, depth is re-estimated by a very local interpolation.

3.3. Enhanced Disparity Estimation Using Anchored Graph Cuts

In this case, the depth-through-disparity approach is followed and a technique for establishing accurate disparity maps between VIs is presented. Disparity is estimated by graph cuts [3], a method used for establishing correspondence between stereo images. Graph cuts are an efficient method for approximating NP-hard problems that seek to assign bijectively a set of labels, disparities d_p in this case, to a set of pixels p with minimal cost, which should be of the form:

$$E = \sum_{\forall p} D_p(d_p) + \sum_{\forall p} \sum_{q \in N_p} V_{p,q}(d_p, d_q) \tag{6}$$

$D_p(d_p)$, called data term, is the cost of assigning disparity d_p to pixel p; q is a pixel in the neighbourhood N_p of p; and $V_{p,q}(d_p, d_q)$, called smoothness term, is the cost emerging when two neighbouring pixels p and q are assigned the disparities d_p and d_q respectively. The merit of graph cuts is that they can model interaction between first ring neighbouring pixels and also allow for piece-wise smooth solutions, perfectly suited in this case of depth map estimation, especially with the anchor points, since this way the reliable depth estimations of the anchor points will further propagate to their neighbourhoods.

To assert the rational assumption that corresponding pixels should have the same colour, the data term is defined as the norm of corresponding pixels' colours, i.e.

$$D_p(d_p) = \|I_{v_a}(p) - I_{v_b}(p + d_p)\| \tag{7}$$

VIs are related with a rigid transformation, which is based on the lens array arrangement. Taking this into consideration, neighbouring pixels with similar colour in all probability belong to the same surface patch of the object and therefore their disparities should be similar. To assert that fact, the smoothness term is defined as:

$$V_{p,q}(d_p, d_q) = K \min(T, |d_p - d_q|) \tag{8}$$

Experiments showed that the optimal performance is achieved for $K = 400$ and $T = 3$.

To exploit anchor points, the data term is modified when pixels corresponding to anchor points are considered. Since their disparities are already known, assigning the known disparity to such a pixel should yield zero data cost, while assigning any other disparity should yield an infinite cost (in practice a very high cost). Plugging disparities estimated by graph cuts to Equations 3a-3c, 3-D coordinates can be recovered.

Although, in principle, depth can be estimated by disparities between any VIs, we opt to estimate depth from successive VIs, taking advantage of the larger overlapping regions. Successive VIs give rise to partial, stereo-like depth maps corresponding to the parts of the object that are visible in the directions of the particular VIs. This way, the outermost, side parts of the object, which may not be visible from the central VI, can be also captured. However, a problem that arises now is that, for each VI, there is a vertical band on one of the image's sides without corresponding region to the next VI and therefore depth estimation there is faulty. The use of the anchor points is once more "catalytic"; the outermost, in both horizontal directions, anchor points define a wide band where disparity is reliable, since anchor points exist there and in between. The reliable bands of the partial depth maps, are then merged into a single one using a mean rule. Anchor points come again, to ease this registration, since due to their computation procedure, they have same 3-D positions in all partial depth maps. Thus, anchor points not only leverage on the optimization procedure, but also on obtaining a high quality multi-direction depth map.

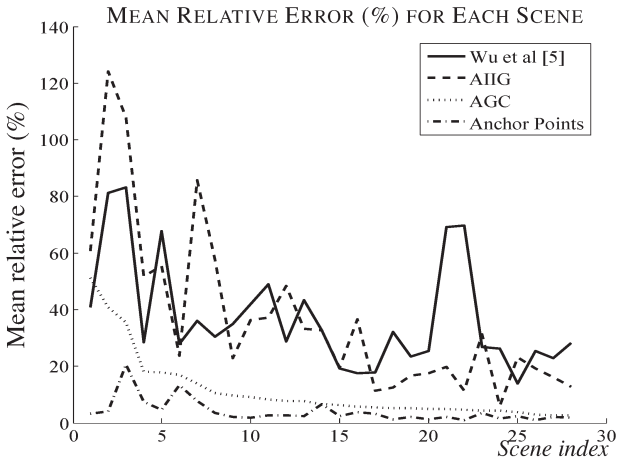

Fig. 5. Comparison of the two proposed methods versus Wu *et al.* [5] (continuous line) on the mean relative error (%) for each scene. "AIIG" stands for Anchored 3-D Integral Imaging Grid (dashed-line), while "AGC" for Anchored Graph Cuts (dotted line). The accuracy of anchor points is also displayed.

4. EXPERIMENTAL RESULTS

4.1. Synthetic Integral Image Generation

To evaluate the performance of both the proposed algorithms, a database[3] of synthetic integral images was built. The database consists of 28 images obtained by simulating a cylindrical lenses array. Each image corresponds to a composite scene composed of 3-D models, of varying texture richness, available in the internet. The objects were transformed, i.e. scaled, rotated and translated, so that they can fit the field of view of the virtual lens array and present a plausible orientation.

4.2. Depth Estimation Accuracy

The depth accuracy achieved using the presented algorithms is depicted in Figure 5. The mean relative error[4] in the entire database is 37.28% for the Wu *et al.* [5] algorithm, 36.85% for the Anchored 3-D Integral Imaging Grid (AIIG) algorithm and 11.32% for the Anchored Graph Cuts (AGC) algorithm. AIIG achieved a comparable performance with the algorithm of [5], while AGC clearly outperformed both.

Regarding the AIIG algorithm, apart from Pearson correlation ($M_{p,q}$ in Equation 5b), cost functions based on the sum of squared distances and the mean absolute error were also tested, but both resulted in lower estimation quality. Post filtering was shown to contribute significantly. This implies that

[3]The database is available at `ftp.iti.gr/pub/Holoscopy`.

[4]The relative error is defined as the absolute difference between estimated and real depth over real depth.

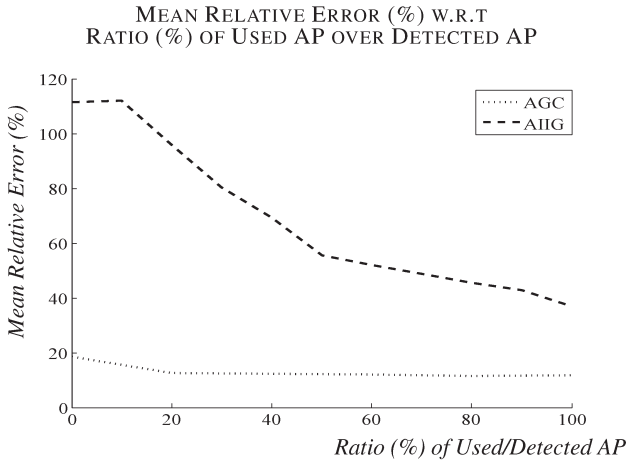

Fig. 6. Mean relative error in the entire database with respect to the number of anchor points (AP) used per image for the proposed methods.

neglecting the piece-wise smooth nature of the depth map, by letting each vertex to be estimated independently of its neighbours, is an important source of error, which is not the case in AGC where pixels are allowed to interact with their neighbours.

In the AGC algorithm, another experiment was also conducted to verify that using pairs of successive VIs rather than other pairs leads to best performance. In that case, the central VI was used as reference and then the disparity map to every VI was estimated. Hence, the depth of every pixel of the central VI could be estimated from multiple *<central VI, VI>* pairs. To achieve best accuracy, the VI leading to the minimum cost in the graph cuts sense (Equation 6) for a small window around the pixel was kept. Actually, this matching methodology was followed in [5]. However, in the AGC framework, this methodology was shown to result in inferior performance due to the fact that outermost object points are estimated by VIs with high angular distance, which contain small overlapping regions and thus more errors are produced.

In order to visualize and quantitatively show the importance of anchor points, how both optimization algorithms highly depend on them and benefit from them, and how efficient the blending of the concepts is, a graph showing their influence in depth estimation was prepared. In Figure 6 the mean relative error in the entire database with respect to the number of anchor points used per image is shown. It is clear that the more the anchor points, the less the error. This is due to the fact that the 3-D coordinates of the anchor points are estimated accurately and therefore they significantly bias the rest of estimations towards the exact values.

5. CONCLUSIONS AND FUTURE WORK

In this paper, an algorithm for detecting anchor points with high accuracy using a local image descriptor and two algorithms for depth estimation exploiting the detected anchor points were presented. The first depth estimation algorithm uses image correlation associated with the vertices of a 3-D grid refined by anchor points, while the second algorithm is based on disparity estimation between VIs through a constrained graph cuts method. The experiments showed that the graph cuts algorithm outperforms state-of-the-art algorithms such as [5], while the 3-D grid algorithm achieves a comparable performance. Regarding the graph cuts approach, the interaction between neighbouring pixels was shown to be a key feature to success. The disadvantage is that it is not clear how to combine optimally disparities between VIs. On the other hand, the combination of VIs is quite clear when the 3-D grid is used, but neighbouring vertices do not interact in this case, revealing a complementarity between the two methods. To improve depth accuracy, a possible fusion of the presented algorithms, i.e. the application of graph cuts to a 3-D grid, will be investigated in the future.

6. REFERENCES

[1] M. G. Lippmann, "La photographie integrale," *Comptes-Rendus de l'Acadmie des Sciences*, vol. 146, pp. 446–551, 1908.

[2] A. Aggoun, "3D holoscopic imaging technology for real-time volume processing and display," High-Quality Visual Experience, pp. 411–428. SpringerLink, 2010.

[3] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[4] E. Shechtman and M. Irani, "Recognition of hand gestures using range images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapoli, MN, USA, Jun. 2007, pp. 1–8.

[5] C. H. Wu, M. McCormick, A. Aggoun, and S.-Y. Kung, "Depth mapping of integral images through viewpoint image extraction with a hybrid disparity analysis algorithm," *Journal of Display technology*, vol. 4, 2008.

[6] S. Manolache, M. McCormick, and S.-Y. Kung, "Hierarchical adaptive regularisation method for depth extraction from planar recording of 3D-integral images," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001, pp. 1433–1436, IEEE Computer Society.

[7] S. Manolache, S.-Y. Kung, M. McCormick, and A. Aggoun, "3d-object space reconstruction from planar recorded data of 3d-integral images," *Journal of VLSI Signal Processing*, vol. 35, no. 1, pp. 5–18, Aug. 2003.

[8] C. H. Wu, A. Aggoun, M. McCormick, and S.-Y. Kung, "Depth extraction from unidirectional integral image using a modified multibaseline technique," in *Stereoscopic Displays and Virtual Reality Systems IX*, S. A. Benton M. T. Bolas A. J. Woods, J. O. Merritt, Ed., San Jose, CA, USA, 2002, vol. 4660, pp. 135–145, SPIE.

[9] J. Park, S. Jung, H. Choi, Y. Kim, and B. Lee, "Depth extraction by use of a rectangular lens array and one-dimensional elemental image modification," *Applied Optics*, vol. 43, no. 25, pp. 4882–4895, Sep. 2004.