FAST 2D TO 3D CONVERSION USING A CLUSTERING-BASED HIERARCHICAL SEARCH IN A MACHINE LEARNING FRAMEWORK

José L. Herrera *Carlos R. del-Blanco* *Narciso García*

ABSTRACT

Automatic 2D-to-3D conversion is an important application for filling the gap between the increasing number of 3D displays and the still scant 3D content. However, existing approaches have an excessive computational cost that complicates its practical application. In this paper, a fast automatic 2D-to-3D conversion technique is proposed, which uses a machine learning framework to infer the 3D structure of a query color image from a training database with color and depth images. Assuming that photometrically similar images have analogous 3D structures, a depth map is estimated by searching the most similar color images in the database, and fusing the corresponding depth maps. Large databases are desirable to achieve better results, but the computational cost also increases. A clustering-based hierarchical search using compact SURF descriptors to characterize images is proposed to drastically reduce search times. A significant computational time improvement has been obtained regarding other state-of-the-art approaches, maintaining the quality results.

Index Terms **—** 2D-to-3D conversion, fast conversion, 3D inference, machine learning, hierarchical search, SURF descriptors, database clustering

1. INTRODUCTION

In the last years, there has been a substantial increment in the quantity and variety of devices that are able to reproduce 3D content, such as TVs, laptops, projectors, smart-phones, and DVD/Blu-Ray players. However, the generation of 3D content has not followed the same trend, creating an inconvenience unbalance between the number of 3D displays and the offer of 3D contents. A plausible solution is to convert the huge existing stock of 2D contents into 3D by means of automatic or semi-automatic 2D-to-3D conversion algorithms.

The generation of 3D contents from 2D ones is generally accomplished in two stages: depth extraction from a single 2D image, and rendering of a stereo-pair using the previous depth data and the original color image. This paper is focused on the first stage, depth extraction from monocular 2D images, which is an open and challenging task, while for the second stage there is already a wide range of algorithms that yield high quality stereo-pairs.

In the last decade, a new family of depth estimation techniques from monocular color images has appeared that are based on a machine learning approach. These techniques make use of a training database of color and depth images to infer the depth map of a query color image. In this process, there is a transfer of knowledge from the structure correlations of the color and depth images in the database to the query color image, which allows to estimate its 3D structure. The core assumption is that color images that are photometrically and structurally similar will have analogous depth maps. Most of the existing works have focused on improving the quality of the estimated depth maps by means of high computational cost algorithms. In [1] a depth map is estimated from a single color image using a image parsing processing along with a Markov Random Field framework. Liu et al. [2] proposed the incorporation of semantic labels and a higher complex supervised learning model to achieve more accurate scene depth results. A similar approach is adopted in [3] that exchanges the transference of labels by directly depth data. In addition, a SIFT-based image registration stage is performed to align the structure of the color and depth images improving the accuracy of the results. An extension to work with video sequences that takes into account spatio-temporal information was presented by [4]. With the purpose of reducing the computational burden of the previous approaches, Konrad et al. [5] discarded the SIFT-based image registration, using instead a matching-based search framework based on HOG features to find images with similar structure. Although this approach has a less computational cost than the previous approaches, it is still too high for many practical applications. In addition, the computational cost is proportional to the size of the used training image database, which avoids its practical use with large databases that would increase the result accuracy. Konrad et al. [5] made an additional effort proposing a 2D-to-3D conversion algorithm with a substantial less computational cost. Depth maps were directly estimated from spatial, color, and motion features using a pixel-wise framework. However, the quality of the obtained results were also significantly worse than the previous approaches.

In this paper, a low computational cost automatic 2D-to-3D conversion algorithm is proposed, which maintain the quality results of other more complex approaches. The al-

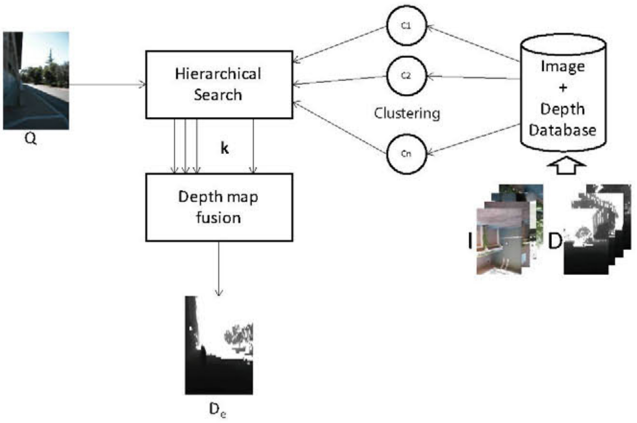

Figure 1. Block diagram of the 2D-to-3D conversion algorithm.

gorithm makes use of a color and depth image database to infer the depth structure of a query color image using a machine learning framework. First, a clustering-based hierarchical search is performed to efficiently find the most structurally similar color images in the database to one query color image. This search scheme allows to use huge databases, which in turn improve the quality of the obtained depth maps with a computational cost significantly lower that an exhaustive search approach. In addition, an efficient and fast descriptor is used to characterize the color images, achieving an additional speed performance. Instead of using costly and densely HOG or SIFT feature descriptors, a variation of the fast SURF descriptor is used to efficiently encode the image structure, which is called grid-based SURF descriptor. Finally, the depth maps of the previously selected color images are fused to obtain the 3D structure of the query image.

2. ALGORITHM DESCRIPTION

The proposed 2D-to-3D algorithm can be divided into three stages. The first stage can be considering a pre-processing that performs a clustering of all the color images in the training database to make clusters that contain similar images from a global structure point of view. The second stage carries out a hierarchical search that selects the of color images in the database that are the structurally closest to a given query color image Q. In the last stage, the depth maps related to the previously selected color images are fused to estimate the depth map D_e of the given query image Q. The Fig.1 shows a block diagram illustrating the different stages of the proposed 2D-to-3D conversion algorithm.

2.1. Clustering

The clustering process has the goal to organize the color images in structurally similar images that allows more efficient searches respect to a given query color image.

The clustering is performed over a compact feature-based representation of the color images in the training database. For this purpose, a new variation of the fast SURF descriptor [6], called grid-based SURF descriptor is used to encode the structure of every color image. The computation is as follows. First the color image is divided into $N \times N$ non-overlapping blocks. Then, a SURF descriptor is computed per block with the appropriate scale parameter so that all the pixels contribute to the calculation of the descriptor. Finally, all the SURF descriptors are stacked up to form a feature vector that represents the image structure.

The k-means algorithm along with the Euclidean distance is used to cluster the color images according to their the feature-based representation in such a way that color images in the same cluster are structurally similar.

The idea behind this clustering is to speed up the posterior search of structurally similar images in the database to a given query image Q. Therefore, it is more important to partition the images in the database in an adequate number that speed up that search than performing a optimal clustering of the feature space from an unsupervised learning point of view. Ideally, an over-partition with equal number of components for cluster would produce a more efficient search. The k-means algorithm approximates quite well this ideal situation.

2.2. Hierarchical search

The hierarchical search is accomplished in two steps. In the first one, the search is carried out between the grid-based SURF descriptor of the query image Q and the descriptors of the color images that are the representatives of every cluster. The search uses the Euclidean distance between descriptors to measure the concept of structurally similar images, likewise the clustering of color images in the database. As a result, the m nearest clusters are selected. The second step exhaustively searches among all the members of the previous m selected clusters to find the k most structurally similar color images to Q. This search scheme is much faster than an exhaustive search and it is almost mandatory in huge databases where the exhaustive search is not possible for computational cost restrictions.

The value of the parameter m is a tradeoff between computational cost and accuracy in the search. High values of m guarantee that the closest grid-based SURF descriptor in the database is found, but also increase the computational cost. Low values of m make the search faster, but does not ensure that the closest component is found. For the database used in the result section, a value of $m = 3$ has been considered as good tradeoff.

The value of the parameter k affects to the quality of the estimated depth map for the query image Q. Each author uses a different number k of images. For example, Konrad et al. [7] use a constant value of $k = 45$, while others like Karsch et al. [4] have chosen a value $k = 7$. In this paper, a value of $k = 10$ has been used, obtaining a quality of the estimated depth maps similar to the others.

2.3. Depth map fusion

A depth map D_e, representing the 3D structure of the query image, is computed by fusing the selected depth images obtained in the previous stage. In this fusion, there can be some selected images that are not really close to the actual structure of the query image, acting as outliers. The median operator has been used by Konrad et al. [5] to perform a robust depth image combination, although the accuracy will be less than the combination of all the images that really have a similar structure (inliers).

The following approach is proposed to obtain a depth estimation that is accurate and robust. The contribution of every selected depth image to the final estimation is proportional to the similarity measurement between its associated color image and the query image. This relies on the assumption that images structurally more similar will have a more similar depth map. As a result, the selected depth images that are potential outliers will have a meaningless contribution, minimizing the potential bias in the final depth map estimation. Specifically, the weighting/fusion process related to one image I_k can be expressed as

$$D_e = \sum_k c(k) D_k, \tag{1}$$

where D_e is the depth map estimation resulting from the fusion process; $c(k)$ is the weight value, which is obtained from the normalization between 0 and 1 of the inverse of the Euclidean distance values computed in the previous stage; D_k is the depth map associated to I_k; and k is the number of selected images in the previous stage. As a result, a robust and accurate depth map Q of the query image is obtained.

3. RESULTS

The proposed approach has been tested using the stereo RGB-D database [4] which consists in 4 subsets of images. The first one, composed by 8430 images, is used as training data and the other three ones, composed by 1414, 2518 and 1912 images, respectively, are used as test data. The results derived from the first subset of images have been obtained by using the testing procedure hold-one-out.

The resolution of the color images and depth maps is 579×430. Nevertheless, like Karsch [4], color and depth images have been resized to a 460×345 resolution for computational efficiency and for a straightforward comparison with their results.

The value of the number of clusters in our hierarchical searching affects the quality of the final estimation and the speed of the conversion. As can be shown in Fig 2, the computation time increases for low and high values of the number of clusters and is lower for intermediate values. Therefore, a value of 100 clusters have been set for the number of clusters in the k-means clustering stage, which is a good tradeoff between computational time and quality of 2D-to-3D conversion.

The metrics used by Karsch et al. [4] have been used to evaluate the performance, taking the logarithmic error ($\log_{10}$), the root mean square error (RMSE) and the Peak Signal to Noise Ratio (PSNR), as final scores:

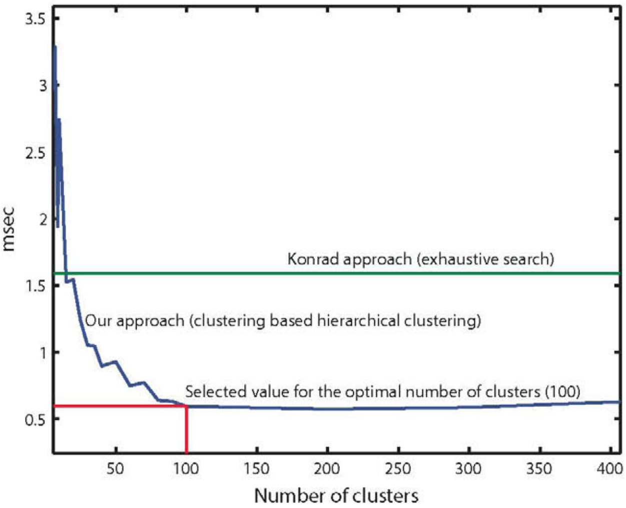

Figure 2. Computational times of the searching stage in msec for the Konrad's algorithm and the ours.

$$\begin{aligned} rel &= \frac{|D(i) - D_e(i)|}{D(i)}, \\ \log_{10} error &= |\log_{10}(D(i)) - \log_{10}(D_e))|, \\ RMSE &= \sqrt{\sum_i (D(i) - D_e(i))^2 / N}, \\ PSNR &= 20 \log_{10} \frac{\max(D)}{RMSE}, \end{aligned} \tag{2}$$

where D is the ground truth depth map, D_e is the estimated depth map, i refers to each pixel of the image, N is the amount of pixels in image, and $\max()$ is a function that return the maximum value.

The proposed approach has been compared with the Depth Transfer algorithm from Karsch et al. [4], and the HOG-based Depth Learning solution of Konrad et al. [5]. Both approaches have been runned in a dual core 2.67 GHz Intel i5. As can be observed in Table 1, the proposed approach outperforms some of the result of the other state-of-the-art methods and keeps close to those which not.

These results are achieved with a lower computational cost, as can be shown in Fig 2, The error measures are averaged for all images in the database. The improvement of the quality of the results is attributed to the use of the grid-based SURF features and the weighted combination of depth maps, while the improvement in the speed is achieved by the hierarchical search, that allows our algorithm to be faster than the approach proposed by Konrad which at the same time is faster than other algorithms in the state of the art [7].

Fig 3 shows two examples of the 2D-to-3D conversion process. The left column corresponds to the query image Q, the central column is the available ground truth of the query image, and the right column is the estimation, obtained by the proposed approach, of the depth map of Q. The upper row is an example of the hold-one-out algorithm, while the lower row corresponds to an image of a different database than the train image.

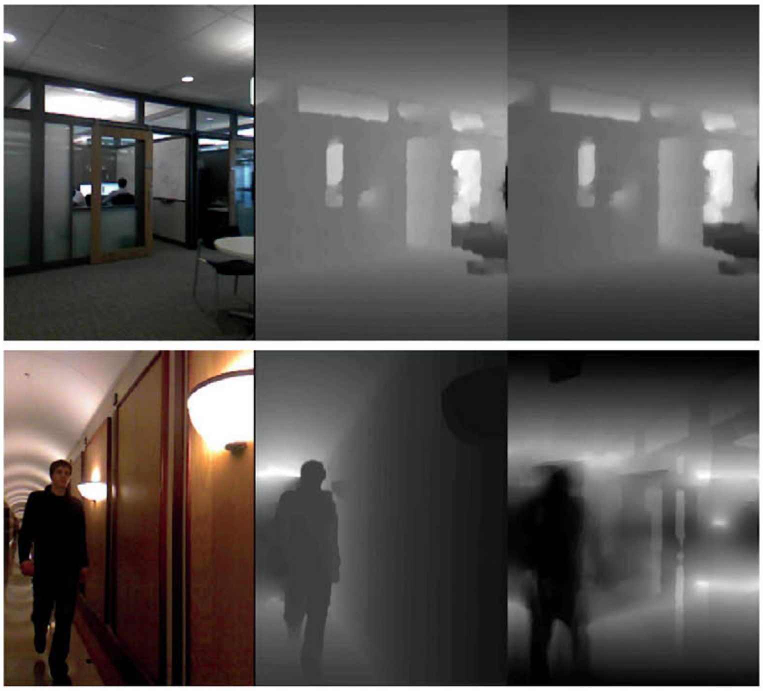

Figure 3. From left to right: query image Q, original depth map (ground truth), and estimated depth map D_e

Algorithm	rel	$\log_{10}$	RMSE	PSNR
Karsch [4] ... 2012	0.29	0.128	12.6	15.6
Konrad [7] ... 2012	0.39	0.096	7.5	26.4
Ours	0.30	0.110	7.6	34.0

Table 1. Evaluation of state-of-the-art algorithms using the rel, $\log_{10}$, RMSE and PSNR metrics in the stereo-RGBD database. The results are the average along all images.

4. CONCLUSIONS

Most of the existing 2D-to-3D conversion algorithms have an excessive computational cost that makes difficult their applications to real situations. To solve this problem, a novel and fast automatic 2D-to-3D conversion technique is proposed to drastically reduce the generation of 3D content with a quality similar to other approaches that require a high computational cost. A machine learning framework is used to infer the 3D structure of a query color image using a database of color and depth images. The key assumption is that photometrically similar images have analogous 3D structure. Thus, a convincing 3D structure can be estimated by searching the most similar color images in the database and adequately fusing the corresponding depth maps. A large database is desirable to achieve better results, however the computational cost also increase. For solving this problem, a clustering-based hierarchical search is proposed to drastically reduce the search time in the database. This strategy is combined with the use of SURF descriptors, which are efficient to compute and compactly characterize each color image to measure similarities among them. A significant improvement in the computational time has been obtained regarding other state-of-the-art approaches, maintaining the quality of the results.

5. REFERENCES

[1] A. Saxena, M. Sun, and A.Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[2] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *IEEE Conf. on Comput. Vis. and Pattern Recognit., 2009. CVPR 2009.*, June 2009, pp. 1972–1979.

[3] J. Konrad, M. Wang, and P. Ishwar, "2d-to-3d image conversion by learning depth from examples," in *IEEE Comput. Soc. Conf. on Comput. Vis. and Pattern Recognit. Workshops (CVPRW), 2012*, June 2012, pp. 16–22.

[4] K. Karsch, C. Liu, and S. Kang, "Depth extraction from video using non-parametric sampling," in *Computer Vision ECCV 2012*, 2012, vol. 7576 of *Lecture Notes in Computer Science*, pp. 775–788.

[5] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2d-to-3d image and video conversion," *IEEE Trans. on Image Process.*, vol. 22, no. 9, pp. 3485–3496, Sept 2013.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[7] J. Konrad, G. Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Automatic 2d-to-3d image conversion using 3d examples from the internet," *Proc. SPIE*, vol. 8288, pp. 82880F–82880F–12, 2012.