LEARNING 3D STRUCTURE FROM 2D IMAGES USING LBP FEATURES

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

ABSTRACT

An automatic machine learning strategy for computing the 3D structure of monocular images from a single image query using Local Binary Patterns is presented. The 3D structure is inferred through a training set composed by a repository of color and depth images, assuming that images with similar structure present similar depth maps. Local Binary Patterns are used to characterize the structure of the color images. The depth maps of those color images with a similar structure to the query image are adaptively combined and filtered to estimate the final depth map. Using public databases, promising results have been obtained outperforming other state-of-the-art algorithms and with a computational cost similar to the most efficient 2D-to-3D algorithms.

Index Terms— 2D-to-3D Conversion, Depth maps, Machine Learning, Local Binary Patterns, Bilateral Filtering

1. INTRODUCTION

The amount of devices, such as TVs, smart phones, DVD/Blu-Ray players, cinemas or video game consoles with 3D play ability has significantly grown in the last years. However, it has not been followed by a similar increment of the 3D content such as 3D movies or even 3D broadcasting, creating a gap between 3D displays and 3D contents. To alleviate this situation, different algorithms that automatically or semi-automatically convert 2D content into 3D one have appeared.

The 2D-to-3D conversion process usually consists of two main stages. The first one is the depth extraction from a single 2D image, and the second one is the rendering of a new image from the extracted depth map and the original image to form a stereo-pair. Since there are many algorithms that generate a good quality stereo-pair, this paper is only focused on the first step, which is more challenging.

In the last years, new learning-based methods have appeared as an interesting alternative for the automatic 2D to 3D conversion task. The key idea behind them is that images with high photometrical similarity will likely have a similar 3D-structure (depth). Saxena et al. [1][2] implemented a supervised learning strategy for estimating the scene depth from a monocular image using an image parsing approach and Markov Random Fields to infer 3D locations and orientations. In [3][4], better scene depth results are achieved by incorporating semantic labels and more sophisticated models. Konrad et al. [5] used a similar approach, but transferring depth data instead of labels, and using a SIFT-based matching and alignment stage to increase the accuracy of the computed depth map. Karsch et al. [6] extended the previous approach to work also with videos. Konrad et al. [7] also proposed a more computationally efficient approach that only performed a matching process based on HOG features to find the images with similar structure, discarding the SIFT-based alignment. The depth map was finally enhanced using a Joint Bilateral Filtering. An even less computationally demanding procedure was proposed in [8], which estimated the depth for each pixel independently using the spatial, color and motion information. However, the obtained results are significantly less accurate than the other alternatives.

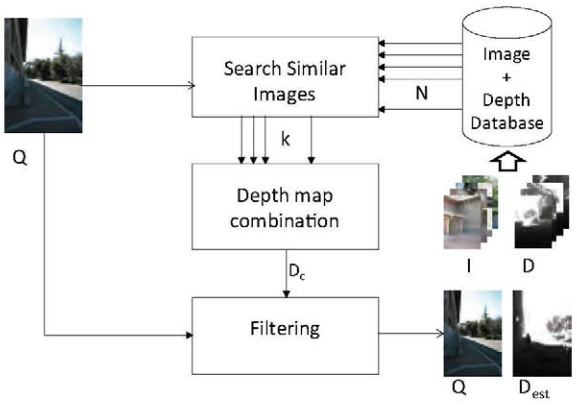

Fig. 1. Block diagram of the 2D-to-3D conversion method.

In this paper, an automatic 2D-to-3D conversion method based on learning approach is proposed. The new algorithm selects an adaptive number of images from a database, evaluating the similarity between them and the query image through Local Binary Pattern, and combines them to estimate the 3D structure.

2. ALGORITHM DESCRIPTION

Given a color-based query image Q and a database DB composed by pairs of color I and depth D images, the aim of the proposed algorithm is to estimate the depth map of Q. The algorithm can be divided in three stages. The first stage consists on searching in the database DB the color images that are photometrically closest to Q. This similarity is measured computing the correlation between the Local Binary Patterns of the images. In the second stage, the depth images associated to the found color images are combined using a correlation based weighting to obtain a preliminary depth map estimation D_c. Lastly, a Joint Bilateral Filtering is applied to remove spurious variations and enforce the consistency between the edges of D_c and Q. As result, a refined depth map D_{est} is obtained.

This method is an evolution of the work of Konrad et al.[7] incorporating the following three novel key contributions. The first one is to use LBP-based features to represent the structure of the color images. The second one is the adaptive selection in DB of the k most similar images to Q, reducing or even removing the outliers that can drift the posterior combination of depth maps. The last contribution is to use in the combination of depth maps a weighting scheme based on the correlation between the features of each selected color image and Q, which reduces even more the influence of potential outliers (i.e. depth maps that have a structure significantly different to Q) in the depth estimation of Q.

2.1. Search for similar images

Color images in the database DB with similar structure to the query image Q will be used in the depth estimation process. To find out which images are similar to a query image, the images are first characterized by a feature descriptor that represents the structure of the image. This image feature descriptor is based on LBP [9], which achieves a compact and efficient representation of the image structure. The image feature descriptor is computed by dividing the image into 4 x 4 blocks, and obtaining a LBP descriptor per block. Then, the descriptors of every block are stacked in a single vector F_I, which characterizes the whole image.

The structure similarity between Q and a candidate color image in DB is computed by means of the correlation between the corresponding image feature descriptor

$$c(n) = corr(F_I(Q), F_I(I_n)), \quad (1)$$

where $corr()$ is the correlation measure, $F_I(X)$ is the LBP-based image feature descriptor of an image X, and I_n is the n^{th} color image in DB.

The number of similar images in DB to a query image Q depends on the own database DB, the query image Q, and the similarity metrics. Some authors use a constant number k of similar images to build the estimated depth map. Karsch et al. [6] uses $k = 7$ and Konrad et al. [8] fix the value to $k = 45$. However, this approach does not guarantee that all the k images are really similar, existing very dissimilar images in the selected subset of images (outliers). To alleviate this problem, an adaptive method is used to select a variable number of k images for each query image Q. First, the images in DB are sorted using the previous correlation values $c(n)$, and those images with an associated correlation value greater than a predefined threshold p are selected. This strategy allows to select those images that really have a structure similar to Q, avoiding/reducing outliers in the selection. Therefore, the threshold p determines the number of images that will be used in the conversion process. In Section 3, the optimal value for this parameter will be discussed. Finally, the depth images associated to the selected color images are the ones that will be used in the depth estimation process.

Fig. 2. From left to right: query image and the three most similar images sorted by similarity value (descendent order).

Fig. 2 shows two examples of query images and the three most similar images to these ones in DB. As can be observed, the similar images capture the structure of the query images.

2.2. Depth map combination

The selected depth images obtained in the previous stage are combined to obtain the depth map of Q, capturing the best as possible its real 3D structure. Some authors [8] perform this combination by applying the median operator. Although this is a good option to remove outliers, the accuracy of the estimated depth map can be low. Under the assumption that images structurally more similar will likely have a more similar depth map, the following approach is proposed: the more similar images are, the higher the contribution of the corresponding depth maps is in the final depth estimation. This approach is consistent since, on the one hand, outliers have been removed or at least reduced by using only images with high similarity, and not a fixed number of them. On the other

hand, images are weighted according to their similarity, so the effect of potential outliers are also reduced. Specifically, each depth map is weighted by the correlation value computed in the previous stage as

$$D_c = \sum_k c(k) D_k, \tag{2}$$

where D_c is the result of the combination of the depth maps, $c(k)$ is the correlation value of one image I_k, D_k is the depth map associated to I_k, and k is the number of selected images in the previous stage. As result, D_c is obtained, which is a preliminary depth estimation of Q.

2.3. Filtering

After the depth map combination, a globally consistent depth estimation is obtained. However, the result presents local inconsistencies around the edges due to the smoothing generated by the weighted average filtering of the k most similar images. In order to maintain the global result, enhance the edges, and align them respect to the original edges of the query image Q, a Joint Bilateral Filtering is applied.

Joint Bilateral Filtering is a variant of bilateral filtering (an edge preserving smoothing filtering) where the Gaussian function is controlled by an external intensity image. In this case, the query image Q is used to control the smoothing. Moreover, Joint Bilateral Filtering reduces the noise in homogeneous areas, and enhance and align the edges of the estimated depth map regarding to the query image. Formally, it can be expressed as:

$$\begin{aligned} W(x) &= \sum_y h_d(x-y) h_Q(Y(x) - Y(y)) \\ D_{est} &= \frac{1}{W(x)} \sum_y D_c(y) h_d(x-y) h_Q(Y(x) - Y(y)), \end{aligned} \tag{3}$$

where D_{est} is the final estimated depth map, $h_d(x)$ and $h_Q(x)$ are Gaussian functions, and $Y(x)$ is the intensity value of pixel x in image Y. The Gaussian function over the position $h_d(x)$ is calculated over the depth map image, while the Gaussian function over the intensity $h_Q(x)$ is computed over the query image Q. As a result of this process, the depth map is generally smoothed, but preserving the edges of the query image.

3. EXPERIMENTAL RESULTS

The proposed approach has been tested using the Make3D dataset #1 [2]. It consists in 534 pairs of images and their associated depth maps, divided in one set of 400 training images and a set of 134 test images.

The resolution of the color images is 2272 x 1704 and the resolution of the depth maps is 55 x 305. Nevertheless, color and depth images have been resized to a 460 x 345 resolution

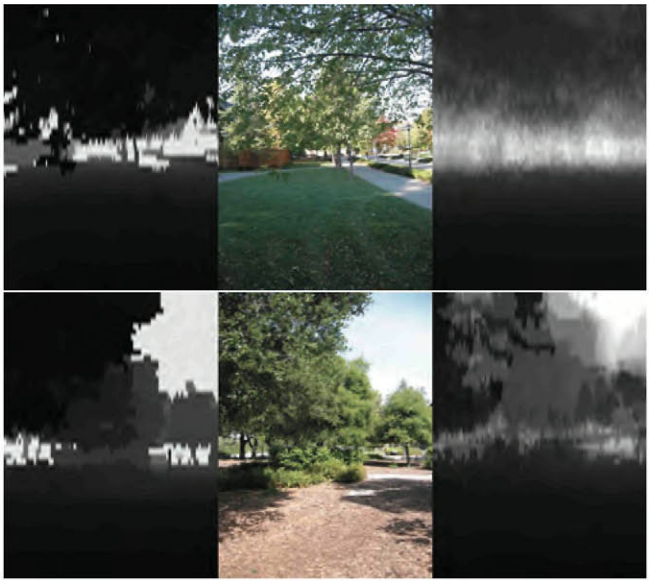

Fig. 3. From left to right: original depth map (ground truth), query image Q, and estimated depth map D_{est}

for computational efficiency and for a straightforward comparison with the results presented in Karsch work [6].

The proposed algorithm have been trained over the training set of 400 images and then the depth estimation has been computed for the 134 images of the test subset. The metrics used by Karsch et al. [6] have been used to evaluate the performance, taking the root mean square error (RMSE) and the peak signal to noise ratio (PSNR), as final scores:

$$\begin{aligned} RMSE &= \sqrt{\sum_i (D(i) - D_{est}(i))^2 / N}, \\ PSNR &= 20 \log_{10} \frac{max(D)}{RMSE}, \end{aligned} \tag{4}$$

where D is the ground truth depth map, D_{est} 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 value of p (see Sec. 2.1), responsible of the amount of depth images used in the 2D-to-3D conversion, has been fixed to 94.5%. This value has been chosen by evaluating the whole 2D-to-3D conversion process with different values of p, ranging from 60% up to 99% in increments of 0.5%, and taking the value which achieves the best depth estimation (minimum RMSE). A specific value of p determines the number k of images used in the conversion process. Fig. 4 shows the relation between the value of p and K. For the selected value of $p = 94.5\%$, an average value of $k = 25.11$ is obtained. This value is higher than the one used by Karsch et al. [6], but lower than the value employed by Konrad et al. [8].

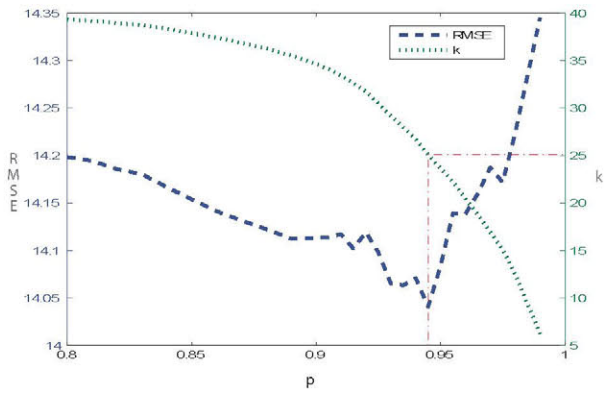

Fig. 4. Variation of the depth map error and evolution of the number of images k as a function of the parameter p.

In this case, the use of less images to compute the estimation of the depth map results in the possibility of a light increment in the efficiency of the method.

The proposed approach has been compared with the Depth MRF method of Saxena et al. [1], the Feedback Cascades algorithm presented by Li et al. [10], the Depth Transfer approach from Karsch et al. [6], and the HOG-based Depth Learning solution of Konrad et al. [8]. The results are shown in Table 1, where, as can be observed, the proposed approach outperforms the result of the other state-of-the-art methods, while keeping a similar computational cost. The error measures are averaged for all images in the database. This improvement of the results is attributed to the use of the LBP features, the adaptive number of images used in the combination process, and the weighted combination of depth maps.

Algorithm	RMSE	PSNR
Depth MRF [1] ... 2005	16.7	N/A
Feedback Cascades [10] ... 2012	15.2	N/A
Depth Transfer [6] ... 2012	15.1	34.4
HOG Based Depth Learning [7] ... 2012	14.9	34.8
A-LBP Based Depth Learning (ours)	**14.0**	**35.7**

Table 1. Evaluation of state-of-the-art algorithms using the RMSE and PSNR metrics in the Make3D database. The results are the average along the 134 test images.

4. CONCLUSIONS

An automatic method for estimating the 3D structure from a single 2D query image have been presented. A machine learning based approach have been adopted that infers the depth of the scene using a database composed by pairs of color and depth images. Our method uses LBP-based features to estimate those images in the database that are more similar to a given query image. Then their depth maps are combined using a weighting scheme that achieves a higher accuracy in the depth estimation than other methods in the state of the art, while keeping the computational cost equal or below than the most efficient algorithm of this family.

5. REFERENCES

[1] A. Saxena, H. Chung Sung, and Y. Ng Andrew, "Learning depth from single monocular images," in *In NIPS 18*. 2005, MIT Press.

[2] 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.

[3] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR), 2010*, June 2010, pp. 1253–1260.

[4] 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.

[5] 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.

[6] 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.

[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.

[8] 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.

[9] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[10] C. Li, A. Kowdle, A. Saxena, and T. Chen, "Toward holistic scene understanding: Feedback enabled cascaded classification models," *IEEE Trans. on Pattern Anal. and Mach. Intell.,*, vol. 34, no. 7, pp. 1394–1408, July 2012.