LEARNING-BASED DEPTH ESTIMATION FROM 2D IMAGES USING GIST AND SALIENCY

José L. Herrera[1] *Janusz Konrad*[2] *Carlos R. del-Blanco*[1] *Narciso García*[1]

Grupo de Tratamiento de Imágenes, ETSI Telecomunicación, Universidad Politécnica de Madrid[1]
Department of Electrical and Computer Engineering, Boston University [2]
{jhc, cda, narciso}[1]@gti.ssr.upm.es jkonrad@bu.edu[2]

ABSTRACT

Although there has been a significant proliferation of 3D displays in the last decade, the availability of 3D content is still scant compared to the volume of 2D data. To fill this gap, automatic 2D to 3D conversion algorithms are needed. In this paper, we present an automatic approach, inspired by machine learning principles, for estimating the depth of a 2D image. The depth of a query image is inferred from a dataset of color and depth images by searching this repository for images that are photometrically similar to the query. We measure the photometric similarity between two images by comparing their GIST descriptors. Since not all regions in the query image require the same visual attention, we give more weight in the GIST-descriptor comparison to regions with high saliency. Subsequently, we fuse the depths of the most similar images and adaptively filter the result to obtain a depth estimate. Our experimental results indicate that the proposed algorithm outperforms other state-of-the-art approaches on the commonly-used Kinect-NYU dataset.

Index Terms— 2D-to-3D Image Conversion, Depth maps, GIST Descriptor, Saliency

1. INTRODUCTION

In the last decade, we have witnessed a significant growth in the availability of 3-D devices, such as TVs, cinema projectors, video game consoles, DVD/Blu-Ray players and even smartphones. However, the availability of 3D content, such as 3D movies or 3D broadcasting, has been lagging behind, thus creating a gap in the 3-D production chain. To rectify this situation, different automatic and semi-automatic algorithms have been developed that convert 2D content into 3D.

The process of 2D-to-3D conversion usually consists of two main stages. In the first one, the depth of a single 2D image is extracted, and in the second one, a new image is generated from the original one and the extracted depth to form a stereo-pair. In this paper, we are only focused on the first stage, which is more challenging.

Recently, several learning-based algorithms have been developed as an alternative to heuristics-based 2D-to-3D conversion methods, often employed in commercial products. The key idea behind these methods is that two images with a high photometric similarity will likely have a similar depth structure. Saxena et al [1] [2] developed a supervised learning approach for estimating the scene depth from a single image using an image parsing strategy and Markov Random Fields to infer 3D locations and orientations. Better depth estimation results were achieved in [3] [4] through the incorporation of semantic labels and more sophisticated models. A similar strategy, but transferring depth data instead of labels, was developed by Konrad et al [5]. Following this approach, Karsch et al [6] added a depth optimization step to assure its smoothness and consistency with candidate depth maps, and also extended the approach to handle videos. More recently, Konrad et al [7][8] presented a computationally-efficient approach by discarding the SIFT-flow based image alignment, using HOG features to find photometrically-similar images and enhancing the final depth map by means of Cross-Bilateral Filtering. They accomplished to reduce the processing time by several orders of magnitude primarily due to skipping the SIFT-flow alignment that turned out to bring minimal quality gains while incurring very high computational cost. A new approach, based on LBP features and using an adaptive number of similar images in the conversion was introduced by Herrera et al [9]. Since the computational cost of such methods is proportional to the size of the dataset used, the approach becomes impractical on very large datasets where the number of similar images to a query is very large. To alleviate this problem, Herrera et al [10] presented a clustering-based hierarchical search that improves the efficiency of the search process. However, the above algorithms assign the same importance to the whole query image when searching for similar images in a dataset without taking into account the different visual attention that different parts of an image attract.

In this paper, we propose a new automatic 2D-to-3D image conversion approach based on machine learning. Instead of using HOG or LBP features to assess the similarity between a query image and database images, we use the GIST descriptor [11], which provides a holistic representation of the scene by measuring its global properties. Furthermore, due

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under projects TEC2010-20412 (Enhanced 3DTV) and TEC2013-48453 (MR-UHDTV).

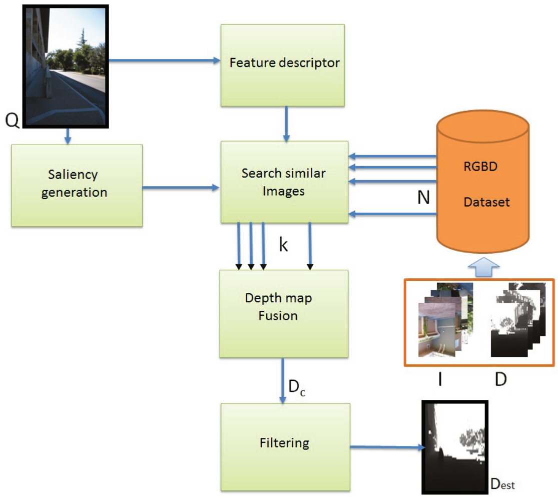

Fig. 1. Block diagram of the proposed 2D-to-3D conversion method.

to the varying visual importance of different regions in the query image, we propose to focus the algorithm on visually-important areas by exploiting the saliency of the query.

2. ALGORITHM DESCRIPTION

Given a color query image Q and an RGBD database DB, composed of a set of color images I and their corresponding depth maps D, the purpose of the approach is to obtain an estimate of the depth map of Q. This algorithm is divided into five main modules (see Fig 1).

1. GIST-based descriptors: The first module is the computation of GIST-based descriptors of the images, that represent the structure of the images. This is an online process for each image Q but an offline process for images in the dataset DB, so they can be computed before the conversion process starts.
2. Saliency-based weights: The second module of the algorithm is the computation of saliency-based weights of Q that will be used to focus the effort in the regions that require higher visual attention.
3. Search for similar images: Here, a search is performed in order to find the closest images to Q in database DB from a photometrical point of view. For this purpose, a saliency based weighted Euclidean distance is computed between GIST descriptors of Q and each image I, and the k most similar images are selected.
4. Depth fusion: In this module, we fuse the depth maps of the most similar images selected in the previous step to obtain a preliminary depth map estimate D_c. To accomplish this, we apply a weighted depth averaging across the selected depth maps using a set of weights derived from the distance metric of the third module.
5. Filtering: In this last module, we refine the fused depth map by applying a cross-bilateral filter (CBF) to remove spurious depth variations and to force the alignment of the edges between the depth estimate Dc and the query image Q. As a result, we obtain a refined depth map estimate D_{est} for query image Q.

Details of each stage are provided below. This approach is an extension of our previous work [9] with two main contributions. The first contribution is the use of a GIST-based descriptor as a representation of structure in color images. The second contribution is the use of saliency-based weights in order to focus the search for similar images on regions where visual attention is higher.

2.1. Feature descriptor

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 in the dataset are similar to the query image, we characterize the images by a feature descriptor that represents the structure of the image. This image feature descriptor is based on GIST [11], which provides a compact representation of the image structure. The overall descriptor is computed by dividing the image into 16 tiles (4 horizontally and 4 vertically), and obtaining a GIST descriptor per tile. Then, for image I, the descriptors of every tile are stacked in a single vector $F(I)$, which characterizes the whole image:

$$F(I) = [\overline{GIST(t_1)} \quad \overline{GIST(t_2)} \quad ... \quad \overline{GIST(t_{16})}], \quad (1)$$

where $\overline{GIST(t_i)}$ is the GIST descriptor of the tile i of the image

These descriptors are pre-calculated off-line for the whole dataset DB before the beginning of the conversion process, while for the query image Q this task is computed online at the beginning of the process.

2.2. Saliency weights

In parallel with the feature descriptor computation, saliency-based weights are first computed and then used in the search for similar images by assigning more importance to those areas that require higher visual attention. First of all, a saliency map of Q is computed using the approach of Harel et al [12]. Then, this saliency map is divided into 16 tiles (4 horizontally and 4 vertically) as was done for the GIST-descriptor calculation in the previous section, and the average of saliency in each tile is computed to obtain $S(t)$. These average saliency values will be used to weight the distance between GIST descriptors. Fig. 2 shows some examples of the saliency maps generated.

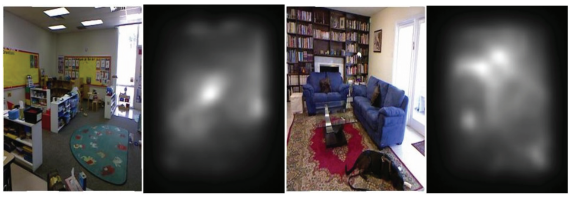

Fig. 2. Two examples of query images and their saliency maps

2.3. Search for similar images

The structure similarity between Q and a candidate color image in DB is computed by the Euclidean distance of the feature descriptors weighted by the saliency weights $S(t)$ as follows:

$$\rho(n) = \frac{\sum_i S(t_i) \| \overline{GIST_Q(t_i)} - \overline{GIST_{I_n}(t_i)} \|^2}{\sum_i S(t_i)}, \quad (2)$$

where $\rho(n)$ is the resulting weighted Euclidean distance, $\overline{GIST_Q}$ is the GIST-based feature descriptor of the query image Q, $\overline{GIST_{I_n}}$ is the same descriptor for image I_n from database DB, and $S(t_i)$ is the saliency weight for tile r_i discussed in the previous section.

After computing the distance $\rho(n)$ in (2) for all images I_n from database DB, k images with the lowest value of ρ are considered the most similar ones and are selected for further processing. Fig. 3 shows two examples of query images and its three most similar images to these ones in DB. The number of images k is a key parameter and the selection of its value is described in the experimental results section.

2.4. Depth fusion

The depth maps of k images selected in the previous stage are combined to obtain the initial depth map of the query Q capturing the 3D structure of the scene depicted in Q. The more similar an image is to the query Q, the higher should be its depth contribution to the final depth estimate. Specifically, each depth map is weighted by the inverse of the weighted distance value computed in the previous stage as follows

$$D_c = \sum_{n=1}^{k} \frac{1}{\rho(n)} D_n, \quad (3)$$

where D_c is the result of of depth fusion, D_n is the depth map associated with image I_n and k is the number of images selected by similarity search in the previous stage 2.3. The resulting D_c is a preliminary depth estimate of Q.

This approach is consistent since, on one hand outliers have been removed or at least reduced by using only images with high similarity to query image Q. On the other hand, images are weighted according to their similarity, so the effect of potential outliers is also reduced.

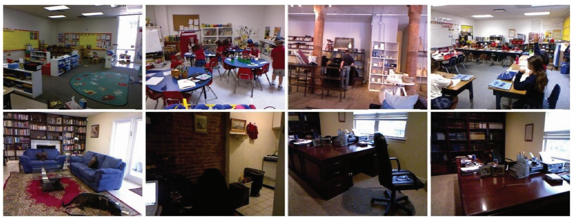

Fig. 3. First column: query image; next three columns: k=3 most similar images to the query sorted by similarity value in descending order.

2.5. Filtering

After depth fusion, a globally consistent depth estimate is obtained. However, this preliminary depth estimate contains local inconsistencies around the edges due to the smoothing generated by the weighted average filtering the depth maps of k most similar images. In order to enhance edges and align them with respect to the original edges of the query image Q, while preserving a globally-consistent depth of the preliminary estimate, we apply cross-bilateral filtering.

Cross-bilateral filtering is a variant of bilateral filtering (an edge preserving smoothing filtering) where the Gaussian function is controlled by an external intensity image [13]. In this case, the query image Q is used to control the smoothing. Moreover, cross-bilateral filtering reduces the noise in homogeneous areas, and enhances and aligns the edges of the estimated depth map with respect to the query image.

Formally, it can be expressed as:

$$\begin{aligned} D_{est} &= \frac{1}{W(x)} \sum_y D_c(y) g_d(x-y) g_Q(Q(x) - Q(y)) \\ W(x) &= \sum_y g_d(x-y) g_Q(Q(x) - Q(y)), \end{aligned} \quad (4)$$

where D_{est} is the final estimated depth map, $g_d(x)$ and $g_Q(x)$ are Gaussian functions, and $Q(x)$ is the intensity value of pixel x in query image Q. The Gaussian function $g_d(x)$ is calculated over positions in the depth image, while the Gaussian function $g_Q(x)$ is computed over intensities of the query image Q, thus enforcing directional smoothing. As a result of this process, the depth map is generally smoothed, while preserving the edges of the query image.

3. EXPERIMENTAL RESULTS

The proposed approach has been tested using the Kinect-NYU dataset [14] . It consists of 1449 pairs of images and their corresponding depth maps. The resolution of the color images and depth maps is 640 x 480 pixels. However, they have been resized to 320 x 240 for computational efficiency and for a straightforward comparison with the results presented by previous works.

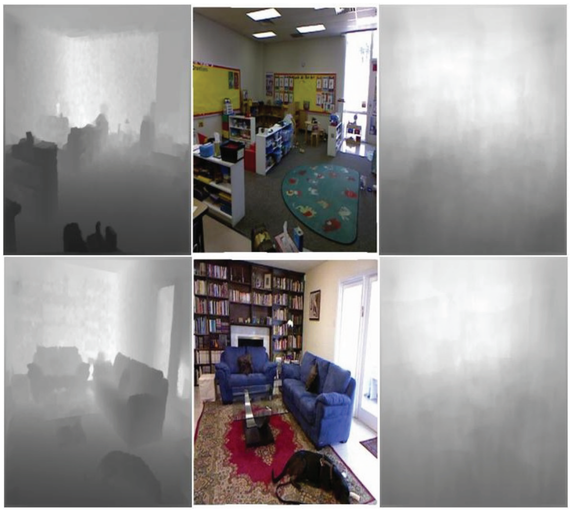

Fig. 4. From left to right: ground truth depth, query image and depth estimate computed by the proposed algorithm.

To quantitatively evaluate the performance of the proposed approach, we applied leave-one-out cross-validation (LOOCV) as follows. We chose one image+depth pair from the NYU database as our query, and treated the remaining pairs in DB as the 3D image repository. We applied the proposed algorithm to every image+depth pair in this repository. Fig. 4 shows two examples of the result of this process.

As the quality metric, we employed normalized cross-covariance between the estimated depth and the ground truth depth defined as follows:

$$C = \frac{\sum_x (D_{est}[x] - \mu_{D_{est}})(D_Q[x] - \mu_{D_Q})}{N\sigma_{D_{est}}\sigma_{D_Q}}, \qquad (5)$$

where N is the number of pixels in D_{est} and D_Q (ground-truth depth of the query image Q), $\mu_{D_{est}}$ and μ_{D_Q} are the empirical means of D_{est} and D_Q, respectively, while $\sigma_{D_{est}}$ and σ_{D_Q} are the corresponding empirical standard deviations. The normalized cross-covariance C takes values from -1 to +1 (values close to +1 indicate that the depth maps are very similar an values close to -1 suggest they are complementary).

A key parameter of this algorithm is the number k of depth maps used in depth fusion 2.4. This parameter has been selected by running the LOOCV test for each image in the dataset for different values of k, and then averaging the results of the cross-covariance across all tests. As can be seen in Fig. 5, the maximum value of C is achieved for $k = 30$. Nevertheless, the exact value of k is not critical since very similar values of C are obtained for a wide range of k values (in this case for $k = 20 - 50$, the value of C stays close to the maximum). The method seems robust to variations of this parameter.

We compute the average and median value of metric C obtained in the LOOCV test across all images in the Kinect-NYU dataset and compare the proposed approach with the Depth Transfer approach by Karsch et al. [6], the HOG-based Depth Learning solution of Konrad et al. [8] and the LBP-Based Learning algorithm of Herrera et al [9]. The results are shown in Table 1, where, as can be observed, the proposed approach outperforms the results of the other state-of-the-art methods for both the average the median of the metric C. This improvement of the results is attributed to the use of the GIST features, and the saliency based weights used to select the k most similar images. It is worth noting that the proposed approach outperforms even the depth transfer approach [6], that is significantly more complex (it includes an image rectification step).

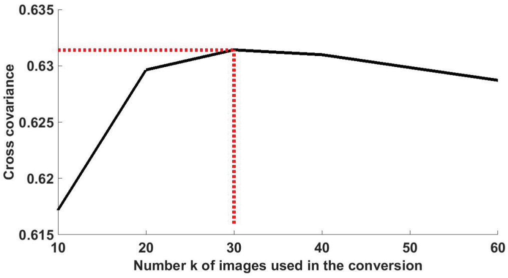

Fig. 5. Variation of the quality metric C as a function of the number k of images used in the conversion.

Algorithm	C (average)	C (median)
HOG-Based Depth Learning [8]	0.55	0.60
Depth Transfer [6]	0.62	0.67
LBP-Based Depth Learning[9]	0.61	0.67
GIST-Saliency Based (ours)	0.63	0.69

Table 1. Evaluation of state-of-the-art algorithms using the average the median of metric C in the Kinect-NYU database. The results have been computed across 1449 images in LOOCV test.

4. CONCLUSIONS

In this paper, an automatic method for estimating the depth of a scene from a single 2D query image has been presented. A machine learning inspired approach has been adopted that infers the 3D structure of the scene using a database composed of pairs of color and depth images. Our method uses GIST-based features, and saliency-based weights, to estimate those images in the database that are most similar to a given query image. Then their depth maps are combined and filtered to obtain the final depth estimate. Experimental results on the Kinect-NYU dataset demonstrate an improved performance over state-of-the-art methods.

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] 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.

[4] 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.

[5] 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.

[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, 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.

[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] J.L. Herrera, C.R. del Blanco, and N. Garcia, "Learning 3D structure from 2d images using lbp features," in *International Conference on Image Processing (ICIP), 2014*, October 2014, pp. 2022–2025.

[10] J.L. Herrera, C.R. del Blanco, and N. Garcia, "Fast 2D to 3D conversion using a clustering-based hierarchical search in a machine learning framework," in *3DTV-Conference*, July 2014, pp. 1–4.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*. 2007, pp. 545–552, MIT Press.

[13] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, jul 2002.

[14] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011*, Nov 2011, pp. 601–608.