SINGLE IMAGE DEPTH ESTIMATION FROM IMAGE DESCRIPTORS

[1,3]*Yu-Hsun Lin,* [3]*Wen-Huang Cheng,* [2]*Hsin Miao,* [2]*Tsung-Hao Ku, and* [3]*Yung-Huan Hsieh*

[1]Graduate Institute of Networking and Multimedia, National Taiwan University
[2]Dept. of Computer Science and Information Engineering, National Taiwan University
[3]Research Center for Information Technology Innovation, Academia Sinica
Taipei, Taiwan

ABSTRACT

With the rapid emergence of 3D displays, we can enrich the user's viewing experiences by adding depth information to the widely existing 2D contents. However, effectively inferring the associated depth from a single 2D image is still a challenging problem. By taking benefits from the recently appeared image descriptors, we proposed the use of an SVM based framework for addressing the single image depth estimation. One advantage is its direct extension to incorporate the recent researches of large scale classification via SVM to meet the upcoming cloud computing paradigm. Our experimental results showed that the proposed framework outperforms the state-of-the-art approaches in performance, even the ones using more complex graphical models like MRF. Also, we made a brief investigation on the individual effectiveness of a set of commonly used image descriptors and found that spatial descriptors (e.g. texture) would be more effective than frequency ones (e.g. DCT coefficients).

Index Terms— Depth Estimation, Single Image, SVM, Cloud Computing

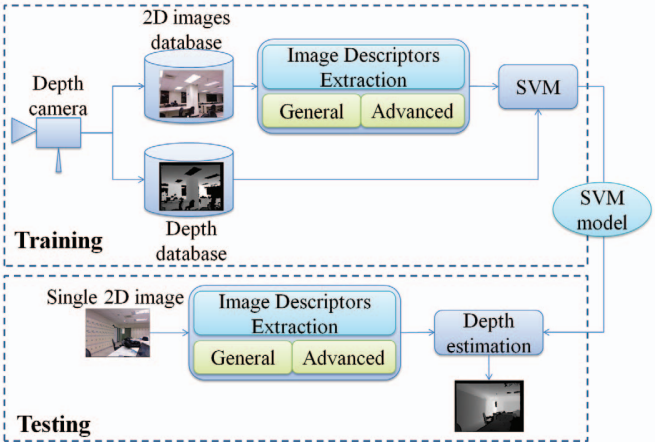


Fig. 1. The proposed SVM based framework for single image depth estimation.

1. INTRODUCTION

With the worldwide success of 3D movies (e.g., Avatar), the entire supply chain of 3D multimedia, from the 3D content creation to the 3D displays, has acquired a lot of attention from both industry and academic research. It can be found that the 3D displays are coming down fast in price with the improvement of the display technologies and have been widely employed in a variety of consumer electronics, such as Fujifilm's W3 digital 3D camera. On the other hand, however, the number of created 3D contents are growing at a much slower pace as compared with the advance of 3D display technologies. One way to bridge the gap is to convert the large amount of existing 2D contents into the corresponding 3D formats. For example, recovering the depth information from a video sequence [1] can utilize the relation among multiple frames. However, estimating the depth information from a given single image only is much difficult since the problem is naturally ill-posed and the visual cue is much less and implicit than a video sequence [2–4].

The research work in [4] presents the relation between the Fourier spectrum and the mean depth value for a given image. Estimating the depth information or the 3D structure from a single image has been addressed by Markov Random Field (MRF) in [3, 5, 6] and shows promising performances. The importance of semantic labels for depth estimation is revealed in the study [2] which can achieve the satisfactory performances with simpler models.

In the literature, it has been shown that in general the performance in dealing with computer vision based problems can be improved with suitable image descriptors (e.g., color, texture, and SIFT) [7]. However, for specific tasks, such as the single image depth estimation, the relative contribution of individual descriptors is less addressed. In addition, the effectiveness of the newly appeared image descriptors is lacking of further investigation. For example, stereo matching is for computing the disparity map between the two associated images of a stereo image pair. We argue that the adopted image descriptors suitable for stereo matching can also improve the accuracy of depth estimation from single image. We then consider in this work a recent advanced image descriptor DAISY [8] which is an efficient dense descriptor for wide-baseline stereo.

By taking all the adopted image descriptors as feature vectors (as detailed in Section 2), we use Support Vector Machine (SVM), instead of employing commonly used graphical models likes MRF which also take into account the influence of neighboring sites [3,5], to deal with the single image depth estimation for two main reasons. First, we can focus on evaluating the estimation effectiveness of individual image descriptors by the SVM classifiers. In some sense, our estimation problem is to be casted as a multi-class recognition problem. Second, the proposed framework is suitable for the upcoming cloud computing paradigm. Observing the recent booming of commercial available depth cameras (e.g., Microsoft's Kinect, which has been sold by more than 10 million units), the depth image of an object and its associated color image can be simultaneously captured

Fig. 2. Texture filters (See Section 2.1 for details).

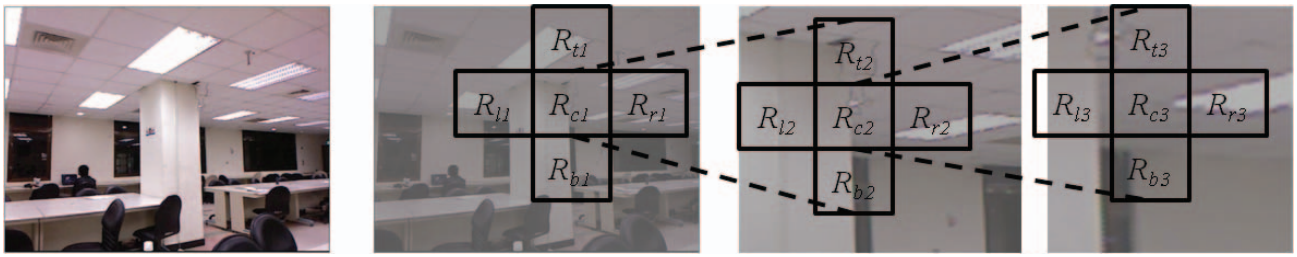


Fig. 3. The extraction flow of image descriptors at the multiple scales.

with ease. It is a cheaper and quicker alternative than using conventional 3D laser scanners to collect a huge amount of needed ground truth data [3, 6]. We can expect the related research issues on depth image will become a large scale problem in the near future. It can be found that the large scale classification via SVM has been addressed by some research effort like [9].

In the rest of this paper, Section 2 describes the proposed framework and the image descriptors involved during depth estimation. Section 3 presents the experimental results and Section 4 concludes this work and gives possible future works.

2. THE FRAMEWORK OF SINGLE IMAGE DEPTH ESTIMATION

Figure 1 illustrates the proposed framework for single image depth estimation based on the adopted image descriptors. The training phase collects color images and the depth information by an off-the-shelf depth camera. The image descriptors of the color images are extracted as the feature vector for the purpose of SVM training. The corresponding depth information is given to be the classification target, i.e. the depth labels. During the testing phase, the depth information of a given 2D image is estimated based on the image descriptors. The image descriptors can be divided into two categories, i.e. the general image descriptors and the advanced image descriptors, as explained in Section 2.1 and 2.2, respectively.

2.1. The General Image Descriptors

The general descriptors can be further classified into three categories: Texture, Color, and Frequency spectrum, which are widely used in the depth estimation literatures [3–6]. We divide an image into $m \times m$ blocks. In the following, the extraction process of the image descriptors is in the basis of a given $m \times m$ block:

Texture Figure 2 shows the 15 filters F_n we used to extract the texture descriptors. The first nine are Laws' masks and the remaining six are used for oriented edges. The extraction process is similar to [3], where the texture descriptors are computed from the cross region R_i of a given image. Figure 3 presents the cross regions at multi-scale and the texture descriptors are generated from each region by convoluting it with the filters. The texture descriptors $P_{Texture}$ are computed as $P_{Texture}(n) = \sum_{(x,y)\in R_i} |I(x,y) * F_n(x,y)|^k$, where $k = 2, 4$ gives the energy and kurtosis, respectively. As a result, for a given $m \times m$ block, the length of $P_{Texture}$ is $15 \times 5 \times 3 \times 2 = 450$.

Color The color descriptors P_{Color} are computed from the color components (Cb and Cr) for a given $m \times m$ block. Similar to the texture descriptors, the color descriptors are also computed from the cross regions at multi-scale. Instead of using the average value of the color components (Cb and Cr) directly, we take the energy and the kurtosis as the color descriptors. For a given $m \times m$ block, the length of P_{Color} is $5 \times 3 \times 2 \times 2 = 60$.

Frequency The relation between frequency spectrum and the mean depth value for a given single image is studied in [4]. We use the DCT coefficients as the frequency descriptors P_{DCT} because of its importance in the image compression standards, such as JPEG and MPEG-4. We apply 2D DCT for a given $m \times m$ block with the length of P_{DCT} being $m \times m$.

2.2. The Advanced Image Descriptors

The advance of image descriptors brings new opportunities for the computer vision researches. For example, the emergence of SIFT effectively improves the performance of image based object detection [7]. One observation is that such advanced image descriptors might also assist the stereo matching to generate high quality disparity maps [8]. Hence the effect of the advanced image descriptors for single image depth estimation is worthy studying. One candidate is the recently appeared DAISY descriptors [8]. In comparisons to SIFT, it shows more promising performances for the wide-baseline stereo and can estimate dense instead of sparse depth map from the stereo images. We then adopt DAISY descriptor, P_{DAISY}, for that we can obtain a corresponding descriptor for every pixel so as to make sure that there will be no block with no descriptor within it. Therefore, for a given $m \times m$ block, we take the P_{DAISY} of the central pixel as a representative one for that block. The DAISY descriptor has eight orientations associated with 25 dimensions on each direction, as a result, the length of P_{DAISY} for a given block is $8 \times 25 = 200$. The detailed computation of DAISY descriptors can be referred to [8].

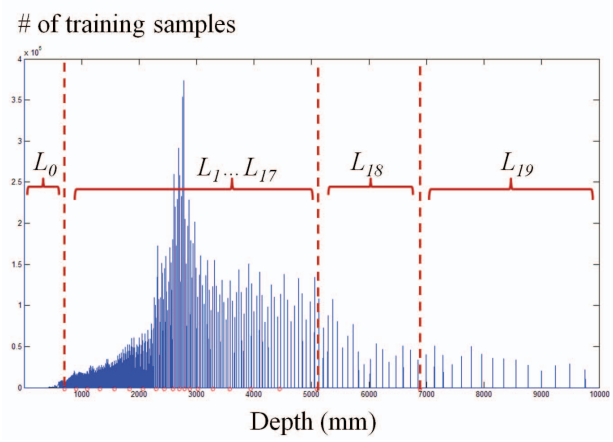


Fig. 4. The histogram of the total recorded depth values in our dataset. Each label L_i contains the same amount of training samples.

3. EXPERIMENTAL RESULTS

3.1. Data Collection

We collect the set of color images with the associated depth information by the Kinect depth camera which uses the infrared light to compute the distance from the target objects. Since the Kinect depth camera is originally designed for the interactive gaming purpose, its best operating range is for objects in the depth from 0 to 10 meters. Therefore, we focus on indoor scenes for experiments and currently capture 90 color/depth image pairs from our campus, such as a reading room (cf. Figure 3) and lockers (cf. Figure 5) in our school library. The captured depth d from Kinect depth camera is the ground truth and the estimated depth is denoted by d_E. In our current settings, the color image is in size of 640×480 and the block size is 8×8 (i.e., $m = 8$) during both the training and the testing process. We can find that a feature vector constituted by all the adopted image descriptors can be in a very high dimension, i.e. 774 (450 + 60 + 64 + 200), and the number of feature vectors in total is 432,000 (90 $\times$ (640 $\times$ 480) / (8 $\times$ 8)). We then address the single image depth estimation as a large-scale classification problem. For reducing the complexity of classification, we quantize the real-valued continuous depth values into 20 labels L_i. Also, in order to avoid the problem of unbalanced training data known in the machine learning literatures, we chosen non-uniform partitioning strategy to make each label L_i contain the same amount of training samples. That is, as shown in Figure 4, the covering range of depth values for the labels L_i from "near" to "far" is gradually increasing. It corresponds to the fact that the depth discrimination in human vision decreases as the object distance becoming further [10]. In addition, the quality of the estimated depth d_E is measured by the following metrics:

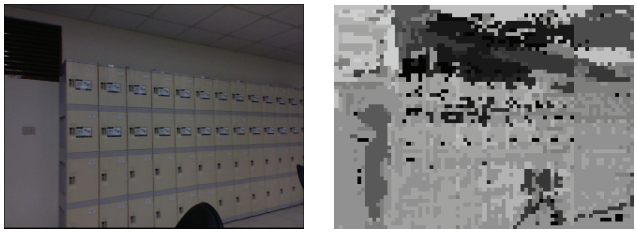
Depth Error $e_{log} = |\log_{10} d - \log_{10} d_E|$

Relative Depth Error $e_{REL} = \frac{|d - d_E|}{d}$

Section 3.2 evaluates the effectiveness of each individual image descriptor for single image depth estimation. Section 3.3 presents the performance of mixed image descriptors and shows promising performances. The distribution of the classification errors are showed in form of the confusion matrix as described in Section 3.4.

Table 1. The accuracy of 5-fold cross validation by using individual image descriptor alone.

Image descriptor	5-fold cross validation
$P_{Texture}$	27.2%
P_{Color}	19.8%
P_{DAISY}	**29.4**%
P_{DCT}	10.8%



(a) (b)

Fig. 5. (a)A sample 2D image, and (b) the estimated depth information by our approach.

3.2. Effectiveness of Single Image Descriptor

Table 1 shows the accuracy of 5-fold cross validation for the depth classification by using the individual image descriptor alone. Based on the experimental results, we can find that P_{DAISY} outperforms the other general image descriptors. Although $P_{Texture}$ achieves the highest accuracy among the three general descriptions, yet it is 2% less than P_{DAISY} in the accuracy. This fact shows the potential of utilizing the advanced image descriptor for the depth estimation. That is, the high accuracy of applying P_{DAISY} would confirm that choosing suitable image descriptors can improve the performance of depth estimation. An interesting phenomenon is the obvious performance gap between $P_{Texture}$ and P_{DCT}. Conventionally, the use of DCT coefficients is a common way for detecting texture patterns, but the performance of P_{DCT} achieves 10.8% only, far less than the $P_{Texture}$'s 27.2%. It might imply that spatial features could be more effective than frequency features for the depth estimation problem.

3.3. Effectiveness of Mixed Image Descriptors

Table 2 presents the accuracy of 5-fold cross validation for the mixed image descriptors. Here the results are obtained by first fusing all the image descriptors and then by removing one of them once at a time. It can be found that the accuracy of utilizing all image descriptors can achieve 44.1% which is going to be 1.5 times higher than the best results by applying a single image descriptor only (i.e. P_{DAISY}), cf. Table 1. Also considering the fact that the training samples are feature vectors of high dimension as described in Section 3.1, the resultant performances can be said to be not bad. This fact shows that most of the adopted image descriptors would be complementary to each other and their mixture is quite beneficial for depth estimation. One exception is the use of P_{DCT}. In Table 2, it is worthy to note that the accuracy of "All without P_{DCT}" is comparable to and even getting a bit better than that of the one using all the descriptors. Following the discussions in Section 3.2, it might further suggest that P_{DCT} would not be helpful and will even damage the performance of depth estimation.

Table 3 illustrates the quality of the estimated depth in terms of

Table 2. The accuracy of 5-fold cross validation for mixed image descriptors.

Mixed image descriptors	5-fold cross validation
All	44.1%
All without $P_{Texture}$	30.6%
All without P_{Color}	39.8%
All without P_{DAISY}	33.8%
All without P_{DCT}	**44.3**%

Table 3. The average depth error and relative depth error for the mixed image descriptors from the 5-fold cross validation.

Mixed image descriptors	Average e_{log}	Average e_{REL}
All	0.194	**0.232**
All without $P_{Texture}$	0.287	0.360
All without P_{Color}	0.228	0.282
All without P_{DAISY}	0.261	0.331
All without P_{DCT}	**0.193**	0.233

e_{log} and e_{REL}. Also, a sample result of the estimated depth map by our approach is showed in Figure 5. The performances by using all the adopted image descriptors would outperform those of the state-of-the-art approaches [2, 3, 6]. For example, the results in [3] are $e_{log} = 0.187$ and $e_{REL} = 0.37$ by the PP-MRF approach, although the experimental settings are different in between, e.g. the data collectors and the used datasets. From the practical perspective, however, it relies on which algorithm will have smaller depth errors and would be more preferred for applications. Therefore, our performances could not be conclusive but are quite encouraging, especially in terms of e_{REL}.

3.4. Analysis of the Error Distribution

Figure 6 presents our depth estimation results among the depth labels L_i, in form of a confusion matrix where the results are obtained by using all the adopted image descriptors. The confusion matrix shows the classification results concentrate along the diagonal entries. It means that the value of estimated depth d_E's are mostly correct. Even if an error occurs, the erroneous value tends to be close to its true value. For the application of adding depth into a given 2D image, this property is pretty valuable, i.e. the user can still have a comfort 3D viewing experience even the d_E is not exactly correct.

4. CONCLUSIONS AND FUTURE WORK

This work proposed the use of an SVM based framework for single image depth estimation, which utilizes the recently appeared image descriptors. The progress of the image descriptors brings new opportunities for addressing the depth estimation problem. In the experiments, we showed that, by utilizing the adopted image descriptors, the proposed framework in terms of objective evaluation metrics (i.e. e_{log} and e_{REL}) outperforms those of the state-of-the-art approaches, even the ones with more complex models like MRF. This work would make one step forward to compensate for the shortage in generating 3D contents. Furthermore, one advantage of our framework is that it can easily adapt to the upcoming cloud computing paradigm. For example, the proposed framework can directly benefit from the recent researches of large scale classification via SVM. In the future, we will make more extensive comparisons and investigations, e.g. by using public depth datasets like Middlebury Stereo [11] to further validate our framework.

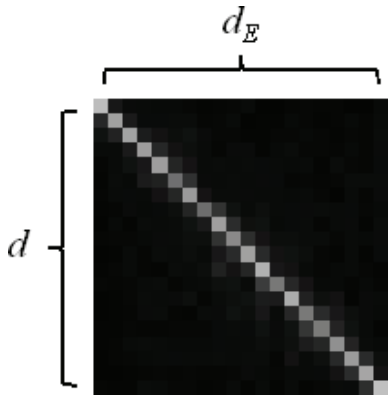


Fig. 6. The confusion matrix of our depth estimation, where the level of intensity indicates the corresponding number (See Section 3.4 for details).

5. REFERENCES

[1] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao, "Consistent depth maps recovery from a video sequence," *IEEE TPAMI*, vol. 31, no. 6, pp. 974 –988, 2009.

[2] Beyang Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," *CVPR '10*, pp. 1253–1260.

[3] A. Saxena, Min Sun, and A.Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE TPAMI*, vol. 31, no. 5, pp. 824–840, May 2009.

[4] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE TPAMI*, vol. 24, no. 9, pp. 1226–1238, Sept. 2002.

[5] Ashutosh Saxena, Sung H. Chung, and Andrew Ng, "Learning depth from single monocular images," *Advances in Neural Information Processing Systems 18*, pp. 1161–1168, 2006.

[6] A. Saxena, Min Sun, and A.Y. Ng, "Learning 3-d scene structure from a single still image," *ICCV '07.*, pp. 1–8.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[8] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE TPAMI*, vol. 32, no. 5, pp. 815–830, May 2010.

[9] Hsiang-Fu Yu, Cho-Jui Hsieh, Kai-Wei Chang, and Chih-Jen Lin, "Large linear classification when data cannot fit in memory," *ACM SIGKDD '10*, pp. 833–842.

[10] B Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*, Focal Press, 2009.

[11] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.