An automatic cluster-based approach for depth estimation of single 2D images

Muhammad Awais Shoukat[1], Allah Bux Sargano*[1], Zulfiqar Habib[1], and Lihua You[2]

[1]Department of Computer Science, COMSATS University Islamabad, Lahore, Pakistan
[2]National Centre for Computer Animation, Bournemouth University, United Kingdom

***Abstract*—In this paper, the problem of single 2D image depth estimation is considered. This is a very important problem due to its various applications in the industry. Previous learning-based methods are based on a key assumption that color images having photometric resemblance are likely to present similar depth structure. However, these methods search the whole dataset for finding corresponding images using handcrafted features, which is quite cumbersome and inefficient process. To overcome this, we have proposed a clustering-based algorithm for depth estimation of a single 2D image using transfer learning. To realize this, images are categorized into clusters using K-means clustering algorithm and features are extracted through a pre-trained deep learning model i.e., ResNet-50. After clustering, an efficient step of replacing feature vector is embedded to speedup the process without compromising on accuracy. After then, images with similar structure as an input image, are retrieved from the best matched cluster based on their correlation values. Then, retrieved candidate depth images are employed to initialize prior depth of a query image using weighted-correlation-average (WCA). Finally, the estimated depth is improved by removing variations using cross-bilateral-filter. In order to evaluate the performance of proposed algorithm, experiments are conducted on two benchmark datasets, NYU *v*2 and Make3D.**

***Index Terms*—Depth estimation, transfer learning, 2D to 3D conversion, K-means clustering.**

I. Introduction

Digital images are mostly represented in 2D image plane which is different from the 3D world coordinates, as 3D world coordinates also contain the depth/distance information of every element. Humans have the natural ability to perceive this 3D world but the simple image capturing devices lack this functionality. When conventional 2D cameras capture the pictures, the depth information is lost as a result of projection of the scene onto the 2D image plane. However, estimation of this lost dimension for recovering 3D geometry of the scene has become an important research problem due to its wide range of applications such as 2D to 3D image/video conversion [1], robot vision [2], classify and recognize the objects [3].

For depth estimation of a single 2D image, different semi-automatic and automatic approaches have been presented in literature. The semi-automatic approaches allow interference of an expert operator for providing some rough guess of depth, which assist in producing dense depth maps. For example, a diffusion-based method was proposed in [4] to create a dense depth map from sparse values assigned by the expert operator. Another approach, used cross bilateral filter over sparse depth values for depth estimation [5]. A similar approach was proposed in [6], where an edge-based filtering is used instead of cross-bilateral filtering. Likewise, a more efficient and simplified algorithm was presented using random-walks and a graph-cut-strategy in [7]. Due to the dependency on the human operator, these methods are expensive and time-consuming [2, 8, 9]. Automatic approaches do not require any human assistance and can predict depth of the scene automatically. A variety of automatic algorithms have been presented in literature based on different cues like defocus, haze, motion and shading. These algorithms rely on different heuristic assumptions e.g., camera motion, and similarity in color and texture of images [10–13]. However, these algorithms have certain limitations and can not be used in complex scenarios [2, 8, 14].

In recent years, machine learning approaches have drawn much attention of the researchers for depth estimation. These approaches use both color images and corresponding depth maps as training data to estimate the depth of the unseen images. The key idea behind these algorithms is that color images having photo-metric resemblance are likely to possess similar depth structure. Different feature descriptors and their combination have been used for retrieving the best-matched images from the repository of RGB images and their depth. For example, a GIST feature descriptor over saliency map is proposed to analyze the photometric resemblance of images in [15]. The authors assumed that all the regions of the query image do not require an equal visual consideration, they assigned more weights to regions with high saliency in the feature comparison step. In [16, 17], histogram of oriented gradients (HOG) was used as feature descriptor for the selection of candidate images. After that, the candidate depth images were globally warped to estimate depth map of the query image. Similarly, local-binary-patterns (LBP) was introduced as feature extractor followed by a correlation parameter for the selection of variable number of candidate images and removal of the potential outliers in [18]. Likewise, the depth of the query image was estimated by modeling the correlation between color and depth images using a set of parameters in [1, 19].

In another study, a pre-processing step was introduced to enhance the visual appearance of images through multi-scale

*Corresponding author: allahbux@cuilahore.edu.pk

978-1-7281-2741-5/19/$31.00 ©2019 IEEE

Retinex in [20]. The authors claimed that the addition of multi-scale Retinex with HOG or LBP feature descriptor improved the accuracy. In [21], authors argued that the global fusion of depth maps is not sufficient for accurate prediction of depth and introduced an algorithm to integrate the local and global characteristics of images for the accurate prediction of depth. Then, this algorithm was further improved through a multi-level learning model based on different monocular depth cues like color, texture, blurriness, light and relative height in [2].

Most of the existing methods, perform an exhaustive search over the dataset for retrieving similar images. Due to which the computational complexity of such methods is dependent on the size of dataset [14, 22]. To overcome this, a cluster-based hierarchical search was performed to fetch the photo-metrically similar images using SURF feature descriptor in [22]. Then, depth of the query image was estimated through the weighted combination of candidate depth images. Another method, was proposed using combination of features (HOG, LBP, GIST, SURF) and aligning candidate depth images through image registration in [8]. An extension of this method was proposed by learning the parameter of optimal number of candidate images along with a segmentation-based filtering to improve the accuracy in [14].

Learning based method discussed above employed hand-crafted feature descriptors to fetch photo-metrically similar images from the RGB-D database. One of the major limitation of these approaches is that these is no universal feature descriptor available [14] that can be used for all types of scenes. In image recognition domain, this imitation has been addressed through deep learning approach by extracting high-level features directly from the input data [23, 24]. However, deep learning approaches require huge amount of data for training, while in depth estimation from single 2D images, required amount of data may not be available for training the deep learning model from scratch. To overcome this, some researchers employed data-augmentation techniques to increase the size of training images [25–29], which is not an ideal solution since it increases the computational complexity and model may not generalize well due to less number of discriminative features.

In order to minimize the need for a huge amount of data to train deep models, the concept of transfer learning has been employed in this research work to get accurate results with limited amount of dataset. The idea has been taken from the activity recognition technique based on transfer learning [23]. After extraction of high-level features with transfer learning using pre-trained deep learning model i.e., ResNet-50 [30]. In order to make the search process efficient, the dataset is categorized into clusters using the K-means clustering algorithm. Further, an innovative idea is implemented within clustering to replace the high dimension feature vector (feature vector with 4x4 tiles) of cluster images with low dimension feature vectors (feature vector with 2x2 tiles). This step has improved the computational complexity of the algorithm without effecting the accuracy. After then, a KNN-Search is performed on clusters to fetch the similar images to query image. Once similar images are fetched, then weighted-correlation-average is performed to estimate prior depth of query image. Finally, estimated depth map is refined through an edge-preserving filter. The contribution of the proposed research work is two fold. The use of transfer learning within depth estimation problem, and embedding an intelligent step within clustering of dataset images to improve the computational complexity of the algorithm. The rest of the paper is organized as follows: proposed methodology is presented in section 2, experimentation and results are discussed in section 3, and paper is concluded in section 4.

II. PROPOSED METHODOLOGY

Given an input image, and RGB-D Repository (NYU *v*2 or Make3D), the algorithm is aimed at estimating the depth map of the input image. The visual representation of the proposed methodology is presented in Fig. 1, and major steps are elaborated as follows:

- Split the color images of the repository into 4x4 tiles for preserving objects positions present in the image.
- Feature extraction of the newly generated dataset using pre-trained deep learning model i.e., ResNet-50. The feature vector F_i representing the i^{th} image is designed by concatenating the corresponding tiles of i^{th} image as shown in

$$F_i = [T_{i,j==1}\ T_{i,j==2}\ T_{i,j==3}\ \cdots\ T_{i,j==n}], \quad (1)$$

  where $T_{i,j}$ represents the feature vector of j^{th} tile and i^{th} image.
- Categorization of similar images into clusters ($C_{1\rightarrow K}$) using high-level features extracted in the previous step followed by K-means clustering algorithm. After that, cluster central features C_{avg} are computed as

$$C_{avg}[i] = \frac{1}{n}\left[\sum_{j=1}^{n} F_{C_i}[j]\right], \quad (2)$$

  $F_{C_i}[j]$ represents feature vector of j^{th} image present in i^{th} cluster, where *i* ranges from 1 to *k* (number of clusters) and *j* ranges from 1 to n (number of images present in i^{th} cluster).
- Computing correlation coefficient C_{coef} between features of input image and each cluster's central features, to retrieve similar structure images of the input image from the best matched clusters. To reduce the impact of number-of-clusters *k*, more than one cluster images are selected as candidate images. This step is generalized as

$$\begin{aligned} C_{coef}[i] &= Corr(F_{query}, C_{avg}[i]), \\ [C_{val}] &= Corr(F_{query}, F_{C_s}[l]), \end{aligned} \quad (3)$$

  where *i* ranges from 1 to *k*, *s* is the updated index of C_{coef} array after sorting which represents the number-of-candidate-clusters, while range of *l* is directly dependent on number-of-images available in cluster *s*.

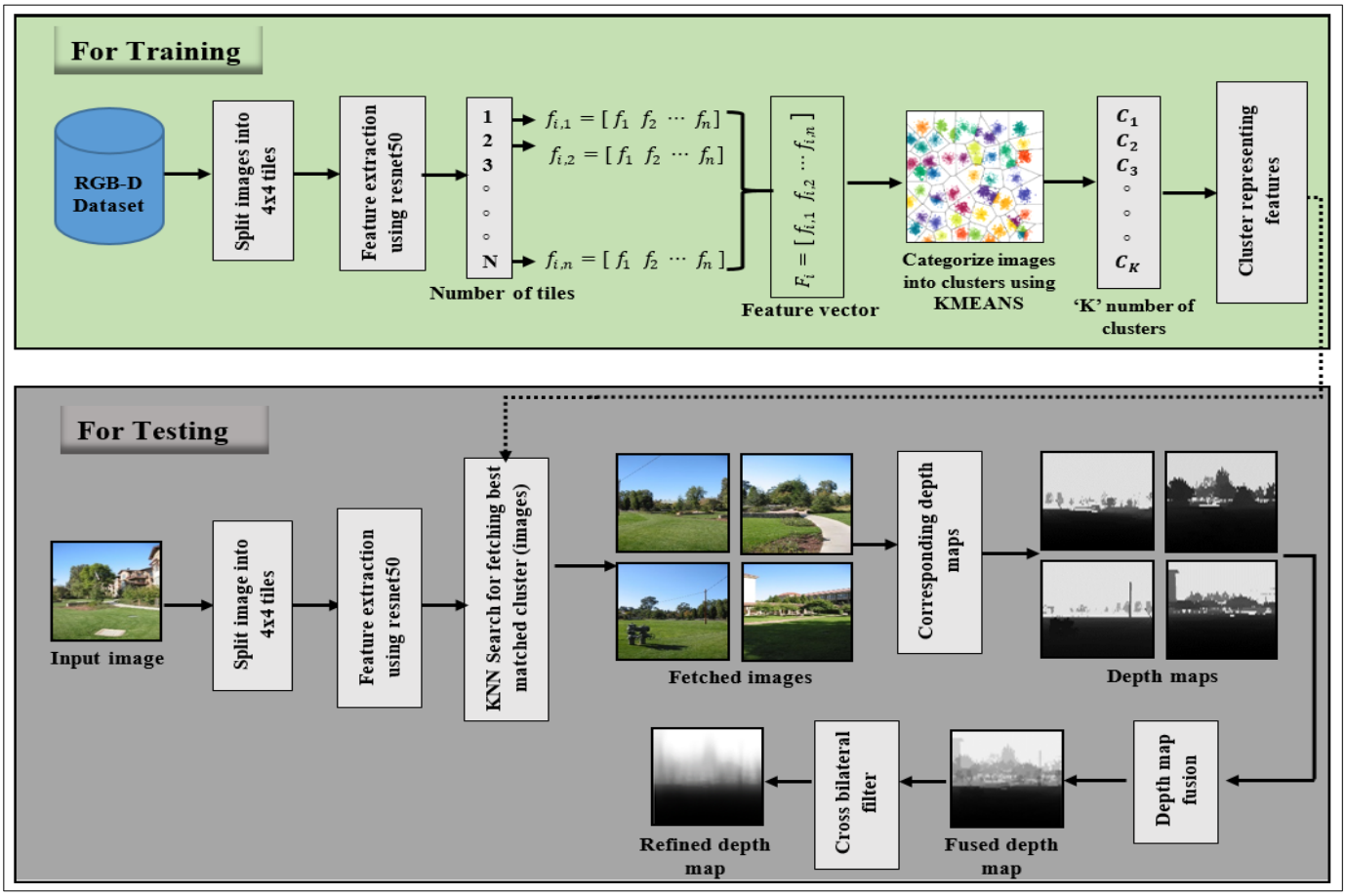

Fig. 1. Block diagram of proposed system.

Fig. 2. 2D query image (Column-1) and nearest neighbors (Columns 2-5) selected by presented algorithm on Make3D dataset.

Fig. 3. 2D query image (Column-1) and nearest neighbors (Columns 2-5) selected by presented algorithm on NYU *v*2 dataset.

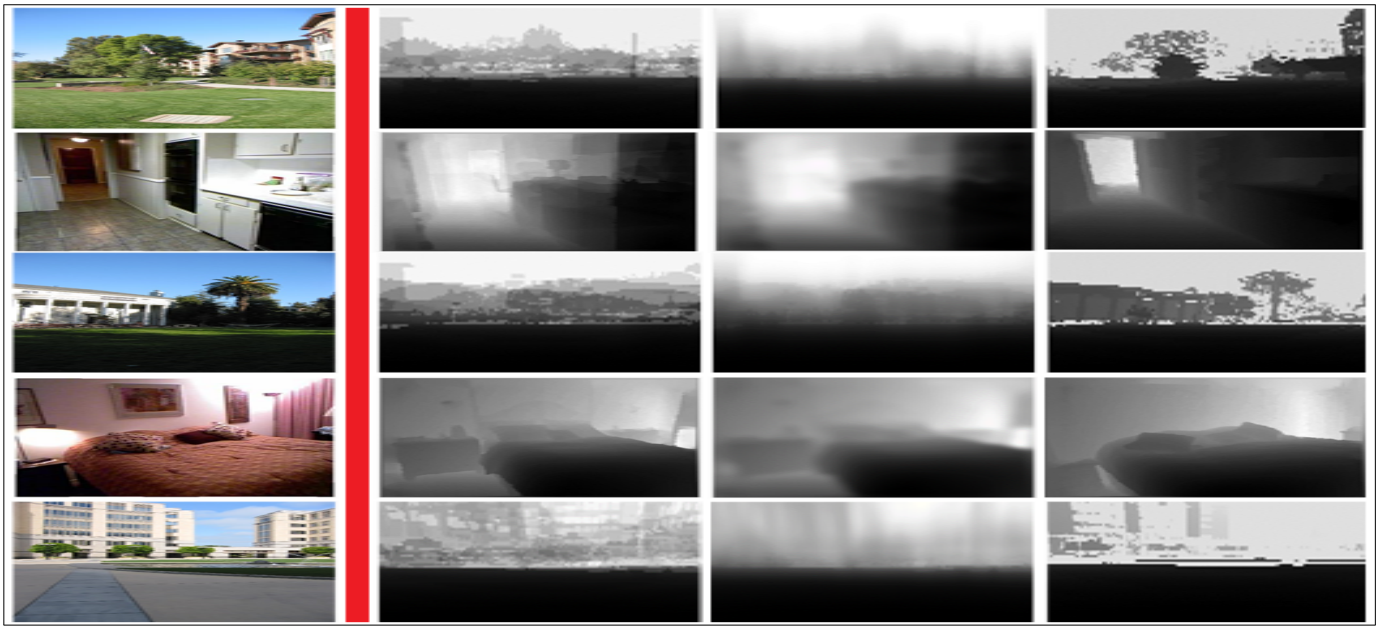

Fig. 4. 2D query image, estimated depth by proposed algorithm, refined depth map and actual ground truth depth map (Columns 1-4 respectively).

- The candidate images having a correlation with the input image less than 0.60 ($C_{val} < 60\%$ similarity) will be discarded in this step. Fig. 2 and 3 shows examples of candidate images selected by the proposed algorithm.
- The initial depth map of the query image is estimated through weighted-correlation-average (D_{prior}) on candidate depth maps. Fig. 4 shows some examples of generated depth maps by proposed algorithm. If D_{prior} is denoted for initial depth map of query image, $C_{val}[i]$ for correlation value between i^{th} candidate image features and query image features, and $D_{cand}[i]$ for associated depth map of i^{th} candidate image. The fusion process can be generalized as

$$D_{prior} = \frac{1}{\sum_i C_{val}[i]} \left[\sum_i C_{val}[i] \times D_{cand}[i] \right]. \quad (4)$$

- Finally, the estimated depth map is improved using an edge-preserving smoothing filter (Cross-Bilateral-Filter).

III. EXPERIMENTATIONS AND RESULTS

To assess the quality of the proposed algorithm, the experiments have been conducted on two different datasets (Make3D and NYU *v*2). The NYU data set is comprised 1449 RGB images of indoor scenes along with ground truth depth images. While Make3D dataset contains 534 outdoor scenes along with ground truth depth images. Some author reported the results on combination of both datasets (Make3D + Nyu (v2)), comparsion with those models is also presented in Table-II Following measures are used to evaluate the quality of the proposed algorithm. If P is the number of pixels, M estimated depth map, and M^* ground truth, the error measures can be mathematically expressed as

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum_i (M^*[i] - M[i])^2}{P}}, \\ PSNR &= 20 \log_{10} \frac{Max(M^*)}{RMSE}, \\ REL &= \frac{1}{P} \sum_i \left[\frac{|M^*[i] - M[i]|}{M^*[i]} \right], \\ log10 &= \frac{1}{P} \sum_i [\log_{10}(M^*[i]) - \log_{10}(M[i])]. \end{aligned} \tag{5}$$

TABLE I
COMPARISON OF STATE OF THE ART ALGORITHMS ON D1 (MAKE3D DATASET) AND D2 (NYU *v*2 DATASET) USING STANDARD TRAIN-TEST SPLIT. BEST RESULTS ARE REPRESENTED THROUGH BOLD CHARACTERS.

Algorithm	PSNR		RMSE		REL		Log10	
	D1	D2	D1	D2	D1	D2	D1	D2
Proposed	14.3	**14.18**	16.25	**1.04**	0.56	0.35	0.173	**0.13**
Multiple Features[14]	14.10	-	-	-	0.407	-	**0.140**	-
TRNN [2]	-	-	**13.43**	1.12	0.407	0.39	0.145	0.14
DEPT [19]	-	-	16.9	1.11	0.489	0.353	0.182	0.13
ML based [8]	-	-	14.7	-	0.505	-	0.18	-
Local-Global (Avg.) [21]	-	-	15.08	-	0.447	-	0.163	-
Local-Global (Med.) [21]	-	-	16.58	-	**0.351**	-	0.163	-
Depth Perception [31]	-	-	-	1.10	-	**0.34**	-	-
Mid-level vision [32]	-	-	-	1.20	-	0.40	-	-
Depth Transfer [33]	**14.56**	-	15.1	1.12	0.361	0.35	0.148	0.13
LBP Based [18]	14.06	-	16.8	-	0.384	-	0.156	-
Weighted median [34]	-	-	15.94	1.28	0.376	1.30	0.161	0.29
HOG based [16]	13.4	-	18.3	-	0.432	-	0.18	-
Batra et al [35].	-	-	15.8	-	0.36	-	0.168	-
Semantic label [36]	-	-	-	-	0.379	-	0.148	-
Make3D [37]	-	-	-	1.21	0.37	0.35	0.187	-

TABLE II
COMPARISON OF STATE OF THE ART ALGORITHMS ON D2 (NYU *v*2 DATASET) AND COMBINATION OF BOTH DATASETS D3 (MAKE3D + NYU *v*2 DATASET) WITH LEAVE-ONE-OUT-STRATEGY. BEST RESULTS ARE REPRESENTED THROUGH BOLD CHARACTERS.

Algorithm	PSNR		RMSE		REL		Log10	
	D2	D3	D2	D3	D2	D3	D2	D3
Proposed	**14.42**	**14.75**	**1.03**	**4.99**	**0.339**	0.418	**0.131**	**0.144**
Multiple Features[14]	13.86	14.00	-	-	0.352	**0.413**	0.137	0.164
Multiple Features (no filt.) [14]	13.90	14.05	-	-	0.407	0.541	0.140	0.182
Depth Transfer [33]	13.57	12.48	1.2	-	0.374	0.559	0.134	0.196
LBP Based [18]	13.74	13.81	-	-	0.422	0.568	0.153	0.192
HOG based [16]	12.90	13.58	-	-	0.539	1.01	0.183	0.245
Learning based [17]	-	-	1.3	-	0.371	-	0.137	-

TABLE III
TIME COMPARISON WITH AND WITHOUT CLUSTERING (IN SECONDS).

Algorithm	Make3D	NYU *v*2
Proposed	1.84	2.11
Without clustering	2.07	3.64

The above mentioned error measures are used to check the quality of the estimated depth map generated by our algorithm with actual ground truth depth map. Table I-II shows the comparison results of the proposed method with state of the art. The results of the other models have been taken from their publications or cited by other papers. The proposed approach shows improved results than most of state-of-the-art algorithms. The best results have been represented through bold values. The higher values of PSNR represent high quality, while lower the results of REL, Log10 and RMSE is better. The sign '-' indicates that the result of the cited paper is not available for that particular measure. The proposed approach falls into the category of algorithms where results depend on the selection of depth-wise similar images. With high-level feature of transfer learning more accurate candidates are presented, which reduces the error rate. Hence, better quality results are achieved through this novel approach. The value for number-of-clusters *K* has tuned on the datasets and best results are reported for Make3D and NYU *v*2 dataset. To minimize the effect of *K* on accuracy measures, more than one cluster images are selected as candidate images (as in our algorithm the value is 3). Fig. 4 shows some examples of generated depth maps by our proposed algorithm, first column of the figure contains 2D query image, the second one is the estimated depth by proposed algorithm, the third one is refined depthmap and the last one is actual ground truth depth map.

The results of the TABLE I-II shows that the proposed algorithm outperforms the state of the art methods, but due to involvement of deep learning layers (Resnet50) the computational complexity of the algorithm is high as in TABLE-III. To reduce the computational complexity factor, a number of experiments have been conducted to revise the steps mentioned in the proposed methodology section. Some of the changes concluded in the algorithm are given as:

- ✓ Features are computed over the dataset by splitting the images into 4x4 and also into 2x2 tiles.
- ✓ Categorization of similar images into clusters ($C_{1 \to K}$) using high-level features extracted in the previous step followed by K-means clustering algorithm. After clustering the dataset images, the feature vector (4x4 tiles) of all cluster images are intelligently replaced with feature vector (2x2 tiles). Using this approach the computational complexity of the algorithm has been reduced. These changes has been reflected in the Fig. 5.

By introducing the skip connections in the algorithm, the computational complexity of the algorithm has been reduced $\left(\frac{2}{3}\right)^{rd}$ of the previous approach Fig. 1. Due to transfer learning with

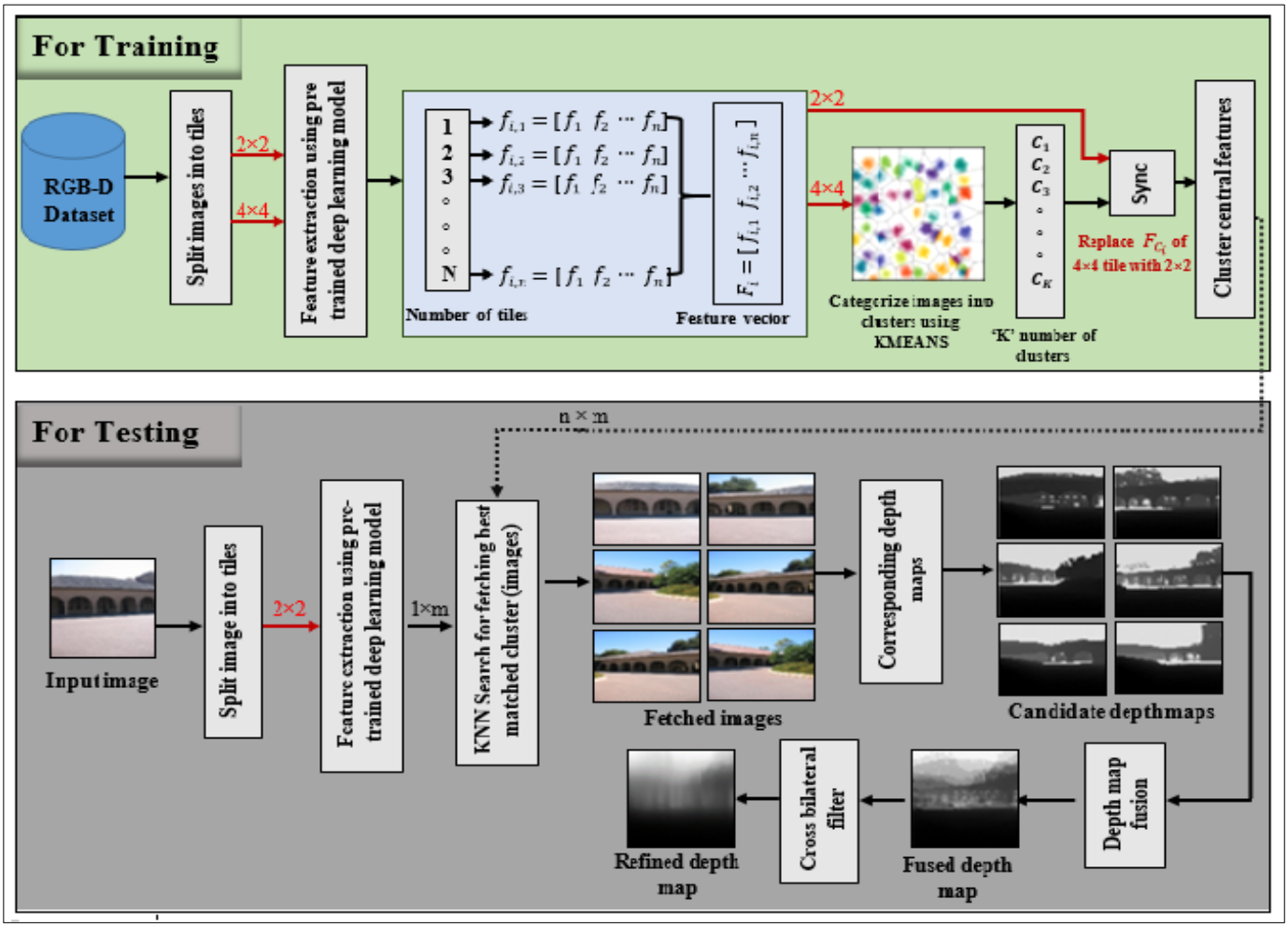

Fig. 5. Block diagram with skip connections to improve the time computation.

TABLE IV
COMPARISON OF STATE OF THE ART ALGORITHMS ON D1 (MAKE3D DATASET WITH 400-134 TRAIN-TEST SPLIT), D2 (NYU *v*2 DATASET WITH 795-654 TRAIN-TEST SPLIT), D2' (NYU *v*2 DATASET USING LEAVE-ONE-OUT-STRATEGY) AND D3 (COMBINATION ON BOTH DATASETS MAKE3D + NYU *v*2 DATASET USING LEAVE-ONE-OUT-STRATEGY). BEST RESULTS ARE REPRESENTED THROUGH BOLD CHARACTERS.

Algorithm	Year	PSNR				RMSE				REL				Log10				Time (s)	
		D1	D2	D2'	D3	D1	D2	D2'	D3	D1	D2	D2'	D3	D1	D2	D2'	D3	D1	D2
Proposed	-	14.37	**14.12**	**14.42**	**14.75**	16.14	**1.04**	**1.03**	**4.99**	0.54	0.35	**0.339**	0.418	0.17	**0.13**	**0.131**	**0.144**	0.69	0.77
Multiple Features[14]	2018	14.15	-	13.86	14.00	-	-	-	-	0.375	-	0.352	**0.413**	0.153	-	0.137	0.164	0.58	0.62
Multiple Features[14] (no filt.) [14]	2018	14.10	-	13.90	14.05	-	-	-	-	0.388	-	0.407	0.541	0.161	-	0.140	0.182	-	-
TRNN [2]	2018	-	-	-	-	**13.43**	1.12	-	-	0.407	0.39	-	-	**0.145**	0.14	-	-	-	-
DEPT [19]	2017	-	-	-	-	16.9	1.11	-	-	0.489	0.353	-	-	0.182	0.13	-	-	-	-
ML based [8]	2017	-	-	-	-	14.7	-	-	-	0.505	-	-	-	0.18	-	-	-	6.9	7.2
Local-Global (Avg.) [21]	2016	-	-	-	-	15.08	-	-	-	0.447	-	-	-	0.163	-	-	-	-	-
Local-Global (Med.) [21]	2016	-	-	-	-	16.58	-	-	-	**0.351**	-	-	-	0.163	-	-	-	-	-
Depth Perception [31]	2016	-	-	-	-	-	1.10	-	-	-	**0.34**	-	-	-	-	-	-	-	-
Mid-level vision [32]	2015	-	-	-	-	-	1.20	-	-	-	0.40	-	-	-	-	-	-	-	-
Depth Transfer [33]	2014	**14.56**	-	13.57	12.48	15.1	1.12	1.2	-	0.36	0.37	0.350	0.559	0.148	0.134	0.131	0.196	92.67	98.65
LBP Based [18]	2014	14.06	-	13.74	13.81	16.8	-	-	-	0.384	-	0.422	0.568	0.156	-	0.153	0.192	0.38	0.58
Weighted median [34]	2014	-	-	-	-	15.94	1.28	-	-	0.376	1.30	-	-	0.161	0.29	-	-	-	-
HOG based [16]	2013	13.4	-	12.90	13.58	18.3	-	-	-	0.432	-	0.539	1.01	0.18	-	0.183	0.245	**0.35**	**0.56**
Batra et al. [35]	2012	-	-	-	-	15.8	-	-	-	0.36	-	-	-	0.168	-	-	-	-	-
Learning based [17]	2012	-	-	-	-	-	-	1.3	-	-	-	0.371	-	-	-	0.137	-	-	-
Semantic label [36]	2010	-	-	-	-	-	-	-	-	0.379	-	-	-	0.148	-	-	-	-	-
Make3D [37]	2009	-	-	-	-	-	1.21	-	-	0.37	0.35	-	-	0.187	-	-	-	-	-

involvement of deep learning layers, the proposed algorithm as discussed in proposed methadology section showed high computational complexity, this problem has been addressed using the skip connections. The idea is to reduce the feature vector size to increase the computational speed, in the first step of the algorithm, the input images are split into 4x4 tiles to preserve the object position that helps to retrieve best matched images. If size of tiles reduced, the potential outliers are selected as candidate images. So, a mechanism to remove the potential outliers and also to reduce the size of tiles has been introduced in the form of skip connections. In this setup, firstly, we categorize the images with the 4x4 tiles, after that we replace the 4x4 tiles features with 2x2 tiles within each cluster images. The cluster central features are also computed on the basis of 2x2 tiles. Now, in testing process, only 2x2 tiles features are computed and compared with other save same dimension train set images. This process reduces the computational complexity of the algorithm, algorithm is well explained in the Fig. 5. The detailed comparison of proposed algorithm with state of the art is shown in Table-IV.

IV. CONCLUSION

This research work proposed an efficient cluster-based algorithm for depth estimation of single 2D images using transfer learning. In this regard, a pre-trained deep learning model, ResNet-50 was used for high-level feature extraction. This model was adapted according to the depth datasets, and then images were grouped into clusters using K-means clustering algorithm. Then, photo-metrically similar images, and their depths were retrieved from the best-matched clusters based on their correlation values. The depth for the query image was initialized using weighted-correlation-average (WCA) with matching depth from the dataset. Finally, the estimated depth was improved by removing depth variations through cross-bilateral-filter. The performance of the proposed algorithm was evaluated on benchmark datasets using different accuracy parameters. The results indicate that proposed method outperforms state-of-the-art methods. As a future research direction, we will extend this approach by embedding local information of the images which may be helpful to increase the accuracy.

ACKNOWLEDGEMENT

This research is supported by the PDE-GIR project which has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 778035.

REFERENCES

[1] Y. Wang, R. Wang, and Q. Dai, "A parametric model for describing the correlation between single color images and depth maps," *IEEE Signal Processing Letters*, vol. 21, no. 7, pp. 800–803, 2014.

[2] H. Mohaghegh, N. Karimi, S. R. Soroushmehr, S. Samavi, and K. Najarian, "Aggregation of rich depth aware features in a modified stacked generalization model for single image depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[3] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 9, pp. 1226–1238, 2002.

[4] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 136–142, IEEE, 2009.

[5] L. J. Angot, W.-J. Huang, and K.-C. Liu, "A 2d to 3d video and image conversion technique based on a bilateral filter," in *Three-Dimensional Image Processing (3DIP) and Applications*, vol. 7526, p. 75260D, International Society for Optics and Photonics, 2010.

[6] C.-C. Cheng, C.-T. Li, and L.-G. Chen, "A novel 2d-to-3d conversion system using edge information," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, 2010.

[7] R. Phan, R. Rzeszutek, and D. Androutsos, "Semi-automatic 2d to 3d image conversion using scale-space random walks and a graph cuts based depth prior," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 865–868, IEEE, 2011.

[8] J. L. Herrera, C. R. del Blanco, and N. García, "A novel 2d to 3d video conversion system based on a machine learning approach," *IEEE Transactions on Consumer Electronics*, vol. 62, no. 4, pp. 429–436, 2016.

[9] L. Zhang, C. Vazquez, and S. Knorr, "3d-tv content creation: automatic 2d-to-3d video conversion," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 372–383, 2011.

[10] T. Lindeberg and J. Garding, "Shape from texture from a multi-scale perspective," in *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pp. 683–691, IEEE, 1993.

[11] J. Malik and R. Rosenholtz, "Computing local surface orientation and shape from texture for curved surfaces," *International journal of computer vision*, vol. 23, no. 2, pp. 149–168, 1997.

[12] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[13] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining motion and lighting for 3d surface reconstruction," *International Journal of Computer Vision*, vol. 48, no. 2, pp. 75–90, 2002.

[14] J. L. Herrera, C. R. del Blanco, and N. García, "Automatic depth extraction from 2d images using a cluster-based learning framework," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3288–3299, 2018.

[15] J. L. Arciniegas Herrera, J. Konrad, C. R. d. Blanco Adán, and N. García Santos, "Learning-based depth estimation from 2d images using gist and saliency,"

in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 4753–4757, IEEE, 2015.

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

[17] J. Konrad, M. Wang, and P. Ishwar, "2d-to-3d image conversion by learning depth from examples," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 16–22, IEEE, 2012.

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

[19] H. Qin, X. Li, Y. Wang, Y. Zhang, and Q. Dai, "Depth estimation by parameter transfer with a lightweight model for single still images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 748–759, 2017.

[20] J. L. Herrera, C. R. del Blanco, and N. García, "Enhanced automatic 2d–3d conversion using retinex in machine learning framework," in *Consumer Electronics (ISCE), 2015 IEEE International Symposium on*, pp. 1–2, IEEE, 2015.

[21] H. Mohaghegh, N. Karimi, S. M. R. Soroushmehr, S. Samavi, and K. Najarian, "Single image depth estimation using joint local-global features," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 727–732, IEEE, 2016.

[22] 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: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2014*, pp. 1–4, IEEE, 2014.

[23] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 463–469, IEEE, 2017.

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition.," in *BMVC*, vol. 1, p. 6, 2015.

[25] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, 2018.

[26] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.

[27] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of CVPR*, vol. 1, 2017.

[28] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162–5170, 2015.

[29] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*, pp. 239–248, IEEE, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[31] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," in *Advances in Neural Information Processing Systems*, pp. 730–738, 2016.

[32] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman, "Learning ordinal relationships for mid-level vision," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 388–396, 2015.

[33] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2144–2158, 2014.

[34] Y. Kim, S. Choi, and K. Sohn, "Data-driven single image depth estimation using weighted median statistics," in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 3808–3812, IEEE, 2014.

[35] D. Batra and A. Saxena, "Learning the right model: Efficient max-margin learning in laplacian crfs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2136–2143, IEEE, 2012.

[36] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1253–1260, IEEE, 2010.

[37] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.