A New Data Preparation Methodology in Machine Learning-based Haze Removal Algorithms

Dat Ngo and Bongsoon Kang
Department of Electronics Engineering
Dong-A University
Busan, Korea
bongsoon@dau.ac.kr

***Abstract*—Haze removal is an intellectually challenging object of scientific study. There is a myriad of methods has been proposed hitherto, ranging from histogram-based, contrast-based to machine learning-based. Haze removal approaches founded upon machine learning require a large and reliable training database. Researchers are currently using the synthetic database due to the complexity of real database acquisition. To introduce the synthetic haze into the clear images, they assume that the depth map is drawn from the standard uniform distribution. In this paper, we present a new methodology for preparing the synthetic training database, in which the proposed equidistribution is employed instead of standard uniform distribution. The effectiveness of the proposed method has been verified by a large number of experiments.**

Keywords—machine learning, training database, haze removal, uniform distribution, equidistribution

I. INTRODUCTION

Haze is usually used to indicate the natural and anthropogenic aerosols, which are present in the atmosphere and are the major cause of the reduction in image clarity. Accordingly, haze removal holds many practical applications in various technology areas. Dehazing algorithms proposed heretofore are based upon histogram [1], image's contrast [2], or machine learning [3-4]. Machine learning has been studied since the 1970s but it is recently widely used due to the rapid advance of computing power [5]. Hence, in years to come, machine learning-based algorithms would continue improving on their performance, which already beat traditional methods' in some areas.

The general idea behind most machine learning is that an algorithm learns to handle a task by studying from a set of examples. Then, it will perform the same task with new data it has not encountered before [5]. Therefore, the training data is absolutely of crucial importance. However, for the haze removal problem, collecting a real database is extremely complicated. Thus, Q. Zhu, J. Mai, and L. Shao [3] as well as B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao [4] have suggested using the synthetic database to train the machine learning algorithm. In their methods, the depth map is assumed to follow standard uniform distribution but most of current pseudo random number generators do not guarantee the uniform distribution. In this paper, we, hereby, propose the novel methodology for preparing the training database, in which the equidistribution is employed to guarantee that the artificially-generated depth map is of truly uniform distribution. Experimental results show that algorithms retrained with our database outperform the original ones.

The rest of this paper is organized into four sections. Sections II summaries the training data preparation method proposed by Q. Zhu, J. Mai, and L. Shao. Section III describes our proposed methodology. Section IV provides the evaluation results to verify the superiority of our method. Conclusions are drawn in section V.

II. CONVENTIONAL METHODOLOGY

To provide an adequate explanation of conventional methodology for preparing the training database, we first briefly introduce the haze removal approach proposed by Q. Zhu, J. Mai, and L. Shao. The formation of hazy images are generally described by the atmospheric scattering model, as follows [3].

$$I(x,y) = J(x,y)t(x,y) + A[1\text{-}t(x,y)] \quad (1)$$

$$t(x,y) = exp[—\beta d(x,y)] \quad (2)$$

The pixel locations in the input hazy image I and output haze-free image J are denoted as x, y. A is the air-light and t is the transmission map, which is exponentially proportional to the product of scene depth d and scattering coefficient β. Equation (1) shows that haze removal is an ill-posed problem, since only the input image is known. By substituting the transmission map with (2), solving (1) for the haze-free image requires accurate estimates of air-light and scene depth. Through an in-depth analysis of vast number of images, Q. Zhu, J. Mai, and L. Shao have proposed a linear model called color attenuation prior describing the relation between scene depth and difference of image's brightness and saturation [3].

$$d(x,y) = \theta_0 + \theta_1 v(x,y) + \theta_2 s(x,y) + \varepsilon(x,y) \quad (3)$$

The image's depth, brightness, and saturation are denoted as $d(x,y)$, $v(x,y)$, and $s(x,y)$ respectively. θ_0, θ_1, θ_2 are unknown coefficients and ε is the random error of the model. Coefficients are estimated by supervised learning with synthetic training database. The training procedure is illustrated in Fig. 1. For each of collected haze-free images, a random depth map within which pixels are drawn from the standard uniform distribution on the open interval (0, 1) is generated. The air-light is also generated as a random number between 0.85 and 1. Thus, the hazy image can be synthesized according to (1) and (2). Then, by converting the Red-Green-Blue hazy image to Hue-Saturation-Value color space, we can obtain the input data of saturation and brightness. Since the depth map is already available, supervised learning can be applied to estimate the unknown coefficients.

III. Proposed Methodology

The standard uniform distribution is not always guaranteed to be uniform. In Fig. 2, the histogram of a sequence of 2,550 random numbers drawn from the uniform distribution on the open interval (0, 1) shows that the distribution is uniform but its standard deviation is relatively high. Therefore, in this section, we propose the equidistribution, which closely resembles the shape of theoretical uniform distribution. The idea behind our method is somewhat simple. It takes the pseudo uniformly distributed random sequence as the input, creates the histogram and sorts it into descending and ascending orders. Then, one of two sorted histograms is chosen to subtract from their average, which better resembles the shape of theoretical uniform distribution. The difference is rounded toward zero and is used to offset the histogram of input random sequence thereafter. In other words, the largest bins are used to offset the smallest bins in the histogram. Fig. 3 shows the example where our proposed algorithm is applied to a small random sequence and the detailed algorithm is provided in Algorithm text-box.

The procedure for preparing training database is almost identical to one illustrated in Fig. 1 with two major differences that the proposed equidistribution is utilized instead of standard uniform distribution to generate the depth map and we collected our own 500 haze-free images via Google Images and Flickr, an image/video hosting service. Haze-free images are selectively chosen to cover a wide variety of categories such as buildings, humans, animals, outdoor/indoor scenes and many others as well as possessing a good sense of depth.

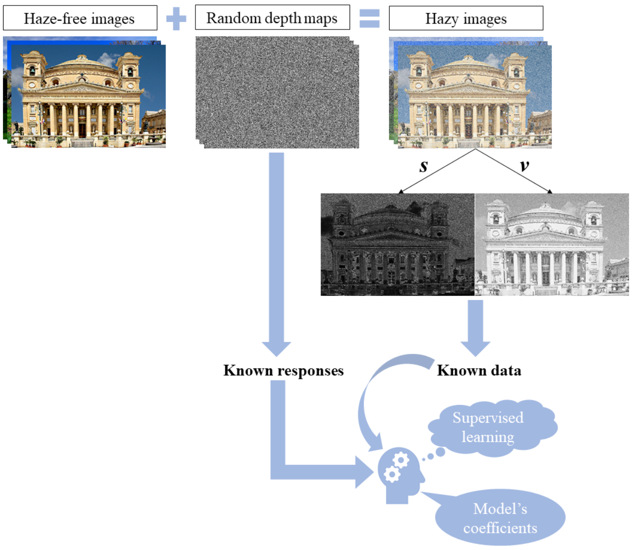

Fig. 1. Training procedure to estimate the model's coefficients.

```
Algorithm
Input: The input sequence $u$ and the number of histogram bins $nbin$
Output: The equidistributed sequence $e$
Auxiliary functions:
    1. $round(X)$: rounds each element of $X$ to the nearest integer toward zero
    2. $histcounts(X, N)$: partitions $X$ into $N$ bins and returns the count in each bin
    3. $length(V)$: returns the length of vector $V$
    4. $find(X)$: finds indices of nonzero elements in $X$
    5. $[S, I] = sort(X, op)$: sorts the elements of $X$ in order specified by $op$. It returns the sorted vector $S$ and corresponding index vector $I$
Begin
    $gu = u$
    $n = histcounts(u, nbin)$
    $[na, loca] = sort(n, 'ascend')$
    $[nd, locd] = sort(n, 'descend')$
    $avg = (na + nd) / 2$
    $diff = avg - na$
    $rdiff = round(diff)$
    for $i$ iterates from 1 to $length(rdiff)$ do
        if $rdiff(i) > 0$
            $loc = find(gu == locd(i))$
            for $j$ iterates from 1 to $length(rdiff(i))$ do
                $gu(loc(j)) = loca(i)$
            end for
        else if $rdiff(i) < 0$
            $loc = find(gu == loca(i))$
            for $j$ iterates from 1 to $length(rdiff(i))$ do
                $gu(loc(j)) = locd(i)$
            end for
        end if
    end for
End
```

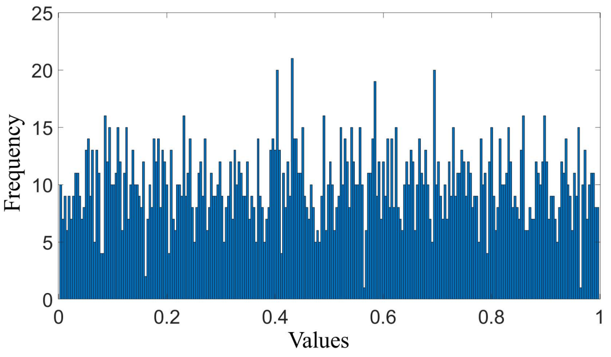

Fig. 2. Histogram of 2,550 uniformly distributed random numbers.

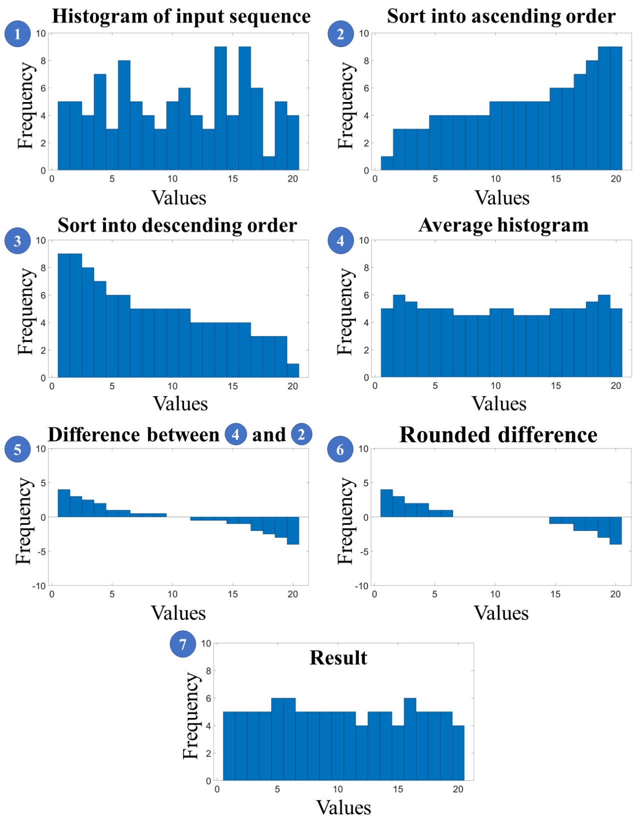

Fig. 3. An example of applying our proposed algorithm to a small random sequence.

IV. Evaluation

In order to verify the effectiveness of our proposed methodology, we retrain the color attenuation prior model with our prepared database and substitute the original coefficients for the newly estimated ones. Then, we evaluate the output dehazed images against those provided by the authors. 'OZhu' and 'NZhu' stand for the original algorithm and one retrained with our database, respectively. Two test image databases are employed in this paper. The first one is Waterloo IVC dehazed image database provided by K. Ma, W. Liu, and Z. Wang [6]. The second one is FRIDA2 image database provided by Tarel et al. [7]. Since the Waterloo IVC database does not contain the ground-truth haze-free images, two metrics for evaluation are the rate of new visible edges (e) and the quality of contrast restoration (r) [8]. In contrast, the FRIDA2 database does include the ground-truth haze-free images, thus, two evaluation metrics are the mean squared error (*MSE*) and tone mapped image quality index (*TMQI*) [9]. The experiment is conducted in MATLAB R2018a on a Core i7-6700 CPU (3.4GHz) with 8GB RAM.

According to the results summarized in TABLE I and TABLE II, all metrics are improved. e, r, *TMQI* increased by 2.56% (from 0.78 to 0.80), 2.54% (from 1.18 to 1.21), 1.3% (from 0.77 to 0.78) respectively, and *MSE* greatly decreased by 52.17% (from 0.23 to 0.11). This means that our proposed equidistribution is more appropriate for generating the synthetic depth map than the standard uniform distribution.

TABLE I. IVC Database Evaluation Results

Image No.	OZhu		NZhu	
	e	*r*	*e*	*r*
1	0.12	0.71	0.05	0.73
2	0.32	0.81	0.26	0.86
3	0.32	1.12	0.27	1.12
4	0.31	1.26	0.31	1.35
5	0.02	1.03	-0.03	1.05
6	0.49	1.37	0.54	1.43
7	0.31	1.03	0.88	1.01
8	1.02	1.67	1.64	1.70
9	1.11	1.15	0.37	0.92
10	1.18	1.14	1.40	1.29
11	1.90	1.29	2.11	1.21
12	0.00	0.90	0.20	0.73
13	0.01	0.67	-0.01	0.58
14	0.40	0.50	0.29	0.67
15	0.48	0.83	-0.48	0.59
16	1.07	1.00	1.02	0.93
17	-0.14	0.69	-0.09	0.63
18	0.27	1.09	0.28	1.19
19	0.10	0.97	0.15	0.95
20	2.85	1.78	3.23	2.00
21	1.62	2.20	1.73	2.48
22	1.91	1.73	2.52	1.87
23	1.22	2.06	1.46	2.31
24	0.73	1.54	1.29	1.65
25	1.84	0.85	0.68	1.06
Avg.	0.78	1.18	0.80	1.21

TABLE II. FRIDA2 Database Evaluation Results

Method	Haze type	TMQI	MSE
OZhu	Homogeneous	0.81	0.26
	Heterogeneous	0.79	0.26
	Cloudy homogeneous	0.76	0.20
	Cloudy heterogeneous	0.74	0.21
	Overall average	0.77	0.23
NZhu	Homogeneous	0.77	0.10
	Heterogeneous	0.79	0.12
	Cloudy homogeneous	0.80	0.11
	Cloudy heterogeneous	0.76	0.10
	Overall average	0.78	0.11

V. Conclusion

In this paper, a novel methodology for preparing the synthetic training database in machine learning-based haze removal algorithms is presented. In the conventional method, the depth map is drawn from the standard uniform distribution, which does not always guarantee to produce uniformly distributed numbers. Thus, we proposed the equidistribution capable of guaranteeing the uniformity. For evaluation, we retrain the linear model in OZhu algorithm and evaluate the results against the original's. After retraining with our database, the performance is boosted significantly. e, r, *TMQI* are increased and *MSE* is decreased. Therefore, it can be concluded that using our proposed methodology for preparing the training database is highly likely to benefit other machine learning-based haze removal algorithms as well.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01060427).

REFERENCES

[1] Z. Xu, X. Liu, and N. Ji, "Fog Removal from Color Images using Contrast Limited Adaptive Histogram Equalization," in 2009 2nd International Congress on Image and Signal Processing, 2009, pp. 1–5.

[2] R. T. Tan, "Visibility in bad weather from a single image," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[3] Q. Zhu, J. Mai, and L. Shao, "A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.

[4] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An End-to-End System for Single Image Haze Removal," IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[5] P. Louridas and C. Ebert, "Machine Learning," IEEE Software, vol. 33, no. 5, pp. 110–115, Sep. 2016.

[6] K. Ma, W. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 3600–3604.

[7] J. P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision Enhancement in Homogeneous and Heterogeneous Fog," IEEE Intelligent Transportation Systems Magazine, vol. 4, no. 2, pp. 6–20, Summer 2012.

[8] N. Hautière, J.-P. Tarel, D. Aubert, and É. Dumont, "BLIND CONTRAST ENHANCEMENT ASSESSMENT BY GRADIENT RATIOING AT VISIBLE EDGES," Image Analysis & Stereology, vol. 27, no. 2, pp. 87–95, May 2011.

[9] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," IEEE Transactions on Image Processing, vol. 22, no. 2, pp. 657–667, Feb. 2013.