<table>
<thead>
<tr>
<th>著者（英）</th>
<th>Laksmita Rahadianti, Fumihiko Sakaue, Jun Sato</th>
</tr>
</thead>
<tbody>
<tr>
<td>発表大会名</td>
<td>2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)</td>
</tr>
<tr>
<td>巻</td>
<td>238-243</td>
</tr>
<tr>
<td>年</td>
<td>2017</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://id.nii.ac.jp/1476/00006410/">http://id.nii.ac.jp/1476/00006410/</a></td>
</tr>
</tbody>
</table>

doi: 10.1109/ACPR.2017.100(http://doi.org/10.1109/ACPR.2017.100)
Abstract—3D scene understanding is important for many applications in the computer vision field. However, the majority of existing solutions commonly assume the images to be captured in clear media. In real-world situations, we may encounter less than ideal conditions, for example haze or fog. In these cases, the captured images will contain scattering and veiling effects that obscure the features of the scene. Many studies approach these images by first removing the scattering effects to obtain an approximate clear image. However, by studying the physical model of light propagation in scattering media, we have observed a relation between the captured image intensity and the distance from the camera. Therefore, as a contrast, we attempt to exploit these scattering effects to obtain 3D depth cues. In order to learn the relation between the scattering effects and the depth, we utilize deep networks to help extract and build high-level features. In this paper, we propose a novel classification approach for depth map estimation of hazy images using deep learning.

Keywords—depth, scattering media, deep learning

I. INTRODUCTION

3D depth estimation is one of the most important cues needed for 3D scene understanding. Some approaches have been proposed using either stereo cues from a multi-view camera system, monocular cues from single images, or a combination of both [1]. Although the stereo approach has been well-studied, the monocular approach has not been dealt with to the same extent. 3D depth estimation from a single image is an ill-posed problem because a singular view of a scene is very ambiguous.

Most 3D depth estimation methods assume that the scene in question is located in clear media, such as shown in Fig. 1(a). But in some real-world conditions, the surrounding media will contain micro-particles that interfere with the light propagation, for example fog or haze. This type of media can also be referred to as scattering media. Due to the light propagation in scattering media environments, an image captured in these environments will contain scattering and veiling effects such as shown in Fig. 1(b).

Most early works in this domain treat the scattering effects as noise, and focus on restoring the image to an approximated clear version before processing it further [2]. However, by studying the physical model of light propagation in scattering media images, we can observe a relation between the image intensity and the depth. Therefore, as a contrast, we aim to exploit the scattering effects in order to establish 3D depth cues.

To obtain depth information, prior knowledge about hazy scenes are often used, such as the Dark Channel Prior (DCP) [3]. Using prior information, it is possible to obtain the relative depth of the scene in the form of transmission [3] or time-to-contact [4], but not the explicit depth map of the scene.

Recently, deep learning approaches have become popular for 3D scene understanding from single images [5], [6]. These methods are able to estimate the 3D depth map of the scene, but require multiple networks and additional calculation models. Furthermore, these approaches assume that the scene was captured in clear media.

In this paper, we propose a deep learning approach to estimate the depth map of a single hazy image. Departing from the physical model of scattering media image formation, we establish a correlation between the scattering effects in the image to the depth of the scene. Therefore, in this paper, we aim to use the scattering effects to directly extract depth cues. From local patches of a hazy images, we can observe a distinct difference among them based on their corresponding distance from the camera. Hence, we use a classification approach to learn the relation between the depth and image intensity of local patches in scattering media images.

II. SCATTERING MEDIA IMAGES

The formation of an image depends on the amount of light that is captured by the camera sensors. In clear conditions, the microparticles contained in the media are minimal, enabling the light from the scene to travel relatively unhindered to the camera. However, in scattering media environments, the surrounding media contains a significant amount of microparticles. The light from the scene that is ultimately captured by the camera is therefore attenuated and saturated by additional scattered light. In order to understand the relation between scattering and distance, this section will cover the physical light propagation model in scattering media environments.
A. Scattering Media Image Formation Model

In scattering media environments, only a fraction of the light will be able to arrive at the camera [4], [7]. The amount of light that arrives at the camera is described by the transmission \( \tau \) of the medium. The transmission differs depending on the type of scattering media, and can be described as follows:

\[
\tau(\lambda) = e^{-c(\lambda)z}
\]

where \( c \) is the attenuation coefficient of the medium, \( z \) is the distance between the camera and the object, and \( \lambda \) is the wavelength of light.

In this paper we assume that the attenuation is wavelength-independent, meaning that the attenuation is equal across the spectrum. This is valid in scattering media images such as haze and fog. Therefore, \( \tau(\lambda) \) can then be written simply as \( \tau \).

The final image that is captured by the camera at each pixel can be separated into two main components, the direct component and the scattering component, as shown in Fig. 2. The direct component is the original intensity of the scene \( J \) that travels directly to the camera. In scattering media, this intensity is attenuated, and the amount that ultimately arrives at the camera is \( \tau J \).

The scattering component consists solely of scattering effects of stray light that is scattered by microparticles into the camera. This creates the color of the surrounding media, which can also be referred to as airlight. This component makes up the remaining part of the intensity captured by the camera, resulting in the final captured image as follows:

\[
I = \tau J + (1 - \tau)A \tag{2}
\]

where \( A \) is airlight.

B. Dark Channel Prior (DCP)

From the image formation model in Section II-A, it is clear that the depth information is encoded in the intensity itself. However, Eq.(2) is under-constrained for extracting depth information. Therefore, He et al. [3] proposed the Dark Channel Prior (DCP). The Dark Channel Prior states that for clear natural outdoor images, in local patches of non-sky regions, at least 1 channel in R,G,B will have a very low intensity, as follows:

\[
J_{DCP}(x) = \min_{y \in \omega(x)} \left( \min_{c \in \{R,G,B\}} J^c(y) \right) \approx 0 \tag{3}
\]

This dark channel will have a consistently low value for clear haze-free images. For hazy images, the intensity of the dark channel will increase due to the amount of transmitted airlight \( A \). Therefore, the DCP can give us a direct cue to depth. Based on Eq.(3), the transmission \( \tau \) can be estimated as follows:

\[
\tau = 1 - \min_{s \in \{R,G,B\}} \left( \min_{y \in \Omega(x)} \frac{I^s(y)}{A^s} \right) \tag{4}
\]

with the assumption that the transmission \( \tau \) in the local patch \( \Omega(x) \) is constant [3].

From this estimated transmission \( \tau \), it is possible to extract the depth \( z \) using Eq.(1) if the attenuation coefficient \( c \) is known. However, since the attenuation coefficient \( c \) is unique for every type of scattering media, this approach is not feasible for general use.

III. DEEP LEARNING FOR IMAGE UNDERSTANDING

As mentioned previously, 3D image understanding from monocular images is very ambiguous. In recent years, many studies have moved towards the use of deep learning networks to handle this problem. Multi-layered deep networks are becoming increasingly popular because they are able to learn high-dimensional mappings from a large set of examples. In this section, we will give a brief literature review on deep learning approaches for image understanding.

A. Convolutional Neural Networks

The convolutional neural network (CNN) is a deep network that has been proven to be powerful for computer vision problems such as image recognition and understanding [8], [9]. A convolutional neural network (CNN) encompasses both the feature extractor as well as the training network. As a trade-off, a convolutional neural network will require a large amount of training data to reach a stable state.

The CNN implicitly assumes an image as input, which will be represented as the initial layer of the network. The higher layers near the input then are connected to local areas, or receptive fields, of the image. This is performed all over the image, enabling CNN to encode both local and global information of the image. This process is essentially a convolution, hence the name convolution layer. The initial convolutions done on the input image can extract the elementary features of the image and the following layers will combine them into meaningful higher-level features [10].

B. Image Understanding Using Deep CNNs

In recent years, many studies have used deep networks to approach image understanding tasks such as 3D distance estimation from single images. Although these deep networks are very good for 2D vision problems such as classification and recognition, the 3D vision problem has fundamental differences that makes the problem more complex.

Liu et al. [6] proposed a method that first takes the individual image pixels as one input, and pairwise neighboring
pixels into another. Each part is processed with a separate deep network before the outputs are calculated into the depth map using a probability calculation. Eigen et al. [5] proposed an alternate method that processes the image in multiple scales. The first network estimates a coarse global depth from the input image, and the second network takes both the coarse depth and the original image to estimate the final depth map.

All these aforementioned methods involve multiple networks or additional calculations. More importantly, these approaches used clear images as their inputs. Meanwhile, few studies have used deep networks for hazy images. Many opt to remove the scattering effects by dehazing, to obtain a clear image [11]. However, a method for direct depth estimation on scattering media images using CNNs has not been proposed.

IV. PATCHWISE DEPTH CLASSIFICATION OF HAZY IMAGES USING DEEP CNNS

As described in Section III-B, it has been established that mapping an image to its corresponding depth is complex and difficult. As an alternative, we looked into the possibility of using a CNN as a classifier for depth estimation. Eq.(1) and Eq.(2) show that the intensity of a hazy image has a direct correlation to the distance from the camera. Based on the relationship of the image appearance with depth, we attempt to take local image patches and classify them based on their distance from the camera.

Fig. 3 shows local patches of a hazy scene taken from areas found at different distances from the camera. Fig. 3 shows patches found at distances of 1 meter, 3 meters, 5 meters, and 8 meters from the camera. We can observe distinct differences between the areas near and far from the camera, such as change in scattering intensity and the patch texture. In order to learn this apparent relationship between the patch appearance and its depth, we use a deep convolutional network to classify the image patches into different depth classes.

A. DepthNet Architecture

Various CNN architectures have been proposed for recognition tasks. LeNet is a well-known convolutional network used for handwritten digit recognition, and has shown good results for isolated character recognition using smaller sized input images [10].

Based on LeNet, we designed DepthNet to classify local hazy image patches based on their depths. The DepthNet architecture used in this paper is shown in Fig. 4. We used 2 pairs of convolution and pooling layers, with sizes adapted to our input patches of 50 × 50 pixels. The convolution layers are followed by 2 fully connected layers to output classes of depth.

To simplify the classification, we use a 9-class classification system, in which we round the pixel depths to the nearest meter. Therefore, each patch will be labeled based on discreet depth classes of \( d \in \{0, 1, 2, 3, 4, 5, 6, 7, 8\} \). Using these distances as the target classes for our classification, we create a 9-class DepthNet classifier as shown in Fig. 4.

B. Training of DepthNet

For the training process, the training images were 50 × 50 pixel patches taken from hazy images. The DepthNet network is then trained using stochastic gradient decent for the gradient updates, starting with a learning rate of \( 10^{-2} \).

Since it is difficult to obtain a large training dataset of hazy images, for this research we generate a synthetic hazy image dataset and use it for training our DepthNet. The synthetic hazy images are generated based on Eq.(2) using RGB-depth image pairs from the NYU Depth Dataset [12]. The NYU Depth Dataset consists of pixelwise pairs of indoor scenes and their corresponding depth maps. Lastly, in order to simulate different conditions, we used various attenuation coefficients of \( c = 0.1, 0.2, 0.3, 0.4 \). Fig. 5 shows an example of a scene, its corresponding depth map, and the simulated hazy images based on Eq.(2) using \( c = 0.2 \) and \( c = 0.4 \).

A full image from the NYU Depth Dataset has a size of 400 × 600 pixels and is simulated using 4 variations of the attenuation coefficient, therefore each image can provide up to 8064 patches. Furthermore, we also rotate each patch on its center by an \(-60^\circ \) to \( 60^\circ \) angle. The use of these simulated patches enables us to quickly build a large number of training images, which is necessary for the CNN to reach a stable state, as mentioned in Section III-A.

For the testing dataset, the images and the attenuation coefficients used were different than those used for the training images. For the testing images we used the coefficients \( c = 0.15, 0.25, 0.35 \). To estimate the depth map of a test image, we iterate through the image, taking 50 × 50 patches that are centered at every pixel. The depth at the position of that center pixel is then assigned with the classification result. In order to take into account the pixels at the edge of the image, we use a padded image which is mirrored on each edge by 25 pixels.
V. EXPERIMENTS AND RESULTS

In this section we will describe the experiments conducted using the DepthNet depth classifier described in Section IV-B, followed by the depth map estimation results and analysis.

A. Depth Map Estimation Compared to Other Methods

In this experiment, we trained the proposed DepthNet depth classifier with hazy image patches as described in Section IV-B and used it to estimate the depth map of testing images. For comparison, we also estimated the depth map using the dark channel prior (DCP) as described in Section II-B. The DCP based method requires the knowledge of the attenuation coefficient $c$. In this experiment, we supplied the coefficient to the estimation.

For further comparison, we also applied the 9-class DepthNet to clear, non-hazy images. We created a clear training set using the local patches from the NYU Depth Dataset [12] images without simulating any scattering media effects. The proposed DepthNet network is then trained using clear image patches.

For our experiments, we estimated the depth map of testing images consisting of hazy and clear versions of 50 scenes. Fig. 6 shows the depth map estimation of the testing images using the proposed method compared to the other methods.

Next, we calculated the mean absolute error (MAE) of estimation based on the ground truth depth maps from the original NYU Depth Dataset [12]. The MAE is expressed in meters. In addition to calculating the objective error, we also evaluated the estimated depth map based on its structural similarity (SSIM) compared to the ground truth. SSIM is a perceptual image quality metric that considers human visual system, and considers luminance, contrast and structure of the image [13]. SSIM values range from -1 to 1, with a higher value indicating a better estimation.

From Fig. 6 we can see that the local patches of clear images can not be differentiated by depth unlike hazy image patches. This is because the clear patches do not contain the scattering effects, and consequently do not contain the depth cues. Therefore, DepthNet can not learn to classify them. Among the hazy images, the estimated depth using the proposed Hazy DepthNet has a lower MAE and a higher SSIM value, compared to the estimated depth maps based on DCP.

B. Training Set Size Evaluation

In this experiment, we focus on the depth map estimation of hazy images, and created multiple hazy training sets with different sizes to compare their classification performance. The hazy training sets used in our experiments are shown in Table II. Each of these training sets are then used to train a 9-class DepthNet classifier as explained in Section IV-B.

Using the trained 9-class DepthNet networks, we estimate the depth map of 50 hazy testing images. The depth estimation results are shown in Fig. 7. Furthermore, Table III shows the average MAE and SSIM of all 50 images of the testing set. As we expected, the largest training set is able to obtain the lowest MAE, as well as the highest SSIM.

VI. CONCLUSION

In this paper, we have proposed a novel method to estimate depth maps of single hazy images using a deep
CNN. Earlier works using deep CNNs involved multiple networks or additional calculations, and handled clear images only. Furthermore, previous works commonly viewed depth estimation as a 3D vision problem, while in this paper we model it as a 2D vision classification problem instead.

Based on the observation of hazy images, we have observed that the appearance of local patches differ based on their distance from the camera. This is supported by the physical image formation model in scattering media, which shows that the scattering effects and intensity contain direct information about the depth of the scene. While scattering effects have previously been perceived as a nuisance, in this paper we exploit them instead. By making use of the relation between image intensity and depth in hazy images, we proposed a deep network architecture, DepthNet, which classifies local patches of hazy images into 9 different depth classes.

As a comparison, we then estimated the depth using the dark channel prior (DCP). Depth estimation based on DCP needs information about the attenuation coefficient of scattering media, making it impractical for general use. Ultimately, our experiments show that the depth map estimation results using the proposed DepthNet are closer to the ground truth compared to the estimation using DCP. Additionally, we attempt to use the proposed DepthNet with clear clear non-hazy images. The proposed DepthNet network depends on the intensity of local patches, hence it is not suitable for clear images due to the lack of scattering effects. Our experiments confirm that local patches of clear images can not be easily differentiated based on their depth compared to their hazy counterparts.

<table>
<thead>
<tr>
<th>Training Set</th>
<th>MAE</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAZY1</td>
<td>1.392</td>
<td>0.919</td>
</tr>
<tr>
<td>HAZY2</td>
<td>1.156</td>
<td>0.939</td>
</tr>
<tr>
<td>HAZY5</td>
<td>0.991</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Table III
EVALUATION OF THE DEPTH MAPS OF 50 TESTING IMAGES, ESTIMATED USING DEPTHNET TRAINED WITH TRAINING SETS OF DIFFERENT SIZES.
Our proposed DepthNet classifier is trained using local image patches as the input. Therefore, a single scene can provide us with a large amount of patches. By using patches instead of whole images, we are able to quickly build a large training set of images, which is crucial for training a stable CNN. To emphasize this point, we also conducted further experiments using various training set sizes and ultimately use a large training set consisting of 5000000 patches for the best results.

The main contribution of this proposed model is a novel idea to approach depth estimation of hazy images from a new viewpoint that has not been previously used. The DepthNet architecture in this paper was built based on the idealized image formation model, and still needs further work to improve its stability. Currently, the training data is simulated using a widely used light propagation model, which is an approximation of the real world phenomenon. Therefore, DepthNet still needs to be developed further using a larger and more diverse dataset of hazy images. To obtain a larger dataset, it is possible to use a more precise simulation model for more precise synthetic images. Furthermore, DepthNet can also be trained using real hazy images. At this point a real hazy image dataset with precise corresponding ground truths is difficult to find and capture, especially one with a large number of images.

For future work, we will continue to build a better hazy image dataset to train and develop the proposed DepthNet classifier. With a larger variety of training data, the depth classifier can hopefully be applicable to more general hazy images. Additionally, we will also investigate other network architectures for depth estimation on hazy images.

REFERENCES


