Journal of Computers Vol. 32 No. 1, 2021, pp. 102-119
doi:10.3966/199115992021023201009

A Wearable Embedded System for Assisting Cognition of Visually Impaired People by Street Scene Description

Feng-Cheng Lin[1*], Shou-Jin She[2], Huu-Huy Ngo[1,3], Chyi-Ren Dow[1], Fang-Rong Hsu[1]

[1] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 407, Taiwan, ROC
{fclin, P0564319, crdow, frhsu}@fcu.edu.tw

[2] Department of Urban Planning and Spatial Information, Feng Chia University, Taichung 407, Taiwan, ROC
sheshoujin@gmail.com

[3] Thai Nguyen University of Information and Communication Technology, Thai Nguyen 250000, Vietnam
nhhuy@ictu.edu.vn

Received 1 June 2019; Revised 1 November 2019; Accepted 31 December 2019

Abstract. Visually impaired people face many challenges in their everyday lives. Thus, many studies have been conducted to provide guidance and support that are more helpful for them. However, while existing studies mostly focus on navigation and obstacles avoidance, there is less focus on semantic information of a current scene, such as color, emotion, object description, etc. In this paper, we propose a wearable embedded system that aims to convey semantic information to visually impaired people based on an embedded module-based approach and deep learning scheme. Furthermore, the proposed system is designed to develop an automated assessment of urban streetscapes through landscape indexes. First, the current scene in front of visually impaired people is captured by the webcam of the system. After the captured image is segmented using a semantic segmentation model, we perform street scene description scheme based on combining image semantic segmentation and color-based emotion classification. Besides, the landscape indexes are calculated to produce the data describing the current scene. Finally, the descriptions of the current scene are delivered to visually impaired people by a headphone or speaker. Moreover, a system prototype is implemented to verify the feasibility of the proposed system. In total, street scenes on 27 locations are selected in this study, and each scene includes 8 streetscape images taken from 8 different directions (front, right front, right, right back, back, left back, left, and left front). After image semantic segmentation and landscape index calculation, these results provide a method to evaluate urban streetscapes.

Keywords: NVIDIA Jetson AGX Xavier, deep learning, emotions color, smart healthcare, visually impaired people

1 Introduction

In Taiwan, the barrier-free facilities and space planning for visually impaired people are not generally available, which impedes their independence when being outdoors. In today's era of information technology and globalization, embedded and multimedia computing are important and widely used in the field of communication, medicine, and health care. In addition, research on wearable navigation devices has achieved noteworthy results and the findings are mainly used to help visually impaired people identify the objects in their surrounding environment to avoid obstacles and navigate. Although the above

* Corresponding Author

applications have increased the safety of visually impaired people, it only allows them to receive the external characteristics of the surrounding environment, lacking the deep semantic features, for example, the semantic feature of a streetscape image and the emotions brought by the surrounding environment.

Streets are the basic but vital infrastructure for urban life. Thus, we use the street scene data taken from the perspective of human vision. Quantitative analysis of urban environment based on street scale can be adopted to understand the impact of street environment on human activities. Street greenery has an important impact on the environmental quality of a city and urban green space planning system adjusted by some cities can re-plan the entire urban structure. Hence, research on the street green view index is of great significance to urban planning and other fields. Junwei and Liang [1] analyzed the correlation between cross-sectional width/height ratio and street width of street space patterns. It was found that indicators such as sky openness index and visual entropy can effectively reflect the objective attributes of streetscapes and can be adopted as the main indicators for streetscape visual evaluation. Therefore, this paper uses computer vision methods to make an automated assessment of urban streetscapes, and the results are adopted to assist the visually impaired people to better experience their surrounding environment.

Semantic segmentation is the process of classifying each pixel in an image to a class label. Examples of these labels can include a tree, building, car, person, etc. Digital image processing is often applied to classify each pixel of an image (or just several ones) into an instance and each instance (or category) is corresponding to an object or a part of the image (road, sky, etc.). The quality of the segmentation result has a substantial impact on the result of image recognition, target detection, scene analysis and other operations. Thus, it is very essential to study image semantic segmentation schemes in the field of computer vision applications and research. Traditional image semantic segmentation techniques typically use methods such as Grab Cut or Normalized Cut to classify pixels, which are ineffective or cannot be fully automated. However, image semantic segmentation based on deep learning solves the problems that traditional image segmentation techniques face.

Visually impaired people face numerous barriers and difficulties in their daily lives. Therefore, many researchers have been striving to develop devices or systems that support visually impaired people to carry out everyday tasks. Lee and Medioni [2] proposed an indoor navigation system helping visually impaired people avoid obstacles. This system used a laptop as the central processing unit requiring users to wear it while using this system and however it has caused some inconvenience to users. To solve this problem, Wang et al. [3] proposed a wearable system, with the advantage of portability to aid visually impaired people in navigating or while on the move. However, these studies only support visually impaired people to navigate and avoid obstacles. They cannot provide visually impaired people the semantic information about the current scene, such as colors, emotion, object description, etc.

Deep learning models have been widely used in the field of image semantic segmentation and can be considered a distinct advantage over traditional image semantic segmentation. With the implementation of deep learning on an embedded system, especially NVIDIA Jetson systems, it makes our proposed solution portable and run while consuming less power. Therefore, NVIDIA Jetson AGX Xavier [4] enables the development of millions of new, small, and low-power artificial intelligence (AI) systems. It opens a whole new world of embedded IoT applications.

Considering the problems mentioned above, this study aims to help visually impaired people better perceive and understand their surrounding environment. This study presents a wearable embedded system assisting the cognition of visually impaired people by delivering street scene descriptions. This proposed system is constructed by an embedded module-based approach and deep learning scheme. First, users can easily carry it wherever they go because of its small size and lightweight. Second, the proposed system is built based on an embedded module, NVIDIA Jetson AGX Xavier. This developer kit is employed to develop embedded IoT applications with the use of deep learning models due to its outstanding GPU (graphics processing unit) capacity and robust performance. Third, the proposed system can generate semantic information about a current scene after it calculates and analyzes the input image. Furthermore, an automated assessment of urban streetscapes through landscape indexes is conducted in the proposed system. In brief, the image data of current scene in front of visually impaired people is collected by the webcam of the system. After the process of segmenting these images using a semantic segmentation model, we perform a street scene description scheme by combining image semantic segmentation and color-based emotion classification. In addition, the landscape indexes are calculated to generate the data describing the current scene. Finally, visually impaired people can hear the description

of the current scene by a headphone or speaker, which enables them to "see" the world through sound. The main contributions of this study are as follows.

(i) Proposing a street scene description scheme which extracts the semantic information of an image.

(ii) Implementing schemes on an embedded module, including the deep learning and the street scene description schemes to provide visually impaired people the semantic information.

(iii) Evaluating urban streetscapes by applying the GVI, SOI, and VE indicators.

The remainder of this paper is organized as follows. Section 2 examines recent researches and technologies that are relevant to this article. Section 3 explains the street scene description scheme. Section 4 represents street landscape indexes. Section 5 details the implementation of the prototype. Section 6 discusses the experimental results. Finally, conclusions and future research directions are offered in Section 7.

2 Related Work

This section quickly reviews the recent relevant studies, including the image semantic segmentation, emotion images, and smart healthcare.

2.1 Image Semantic Segmentation

Image semantic segmentation is an interesting research topic in computer vision, and it plays an essential role in image analysis. Deep neural networks are applied very efficiently in semantic segmentation [5-7]. There are many semantic segmentation approaches, such as region-based, fully convolutional network (FCN)-based, and weakly supervised approaches. Long et al. [8] and Li et al. [9] used the FCN-based approach for semantic segmentation. These methods achieved state-of-the-art segmentation results. Besides, with a straightforward concept: the lightweight, Zhang et al. [10] presented a new method that is compatible with existing FCN-based approaches. They introduced the context encoding module, which selectively highlights class-dependent feature map and captures contextual information. Chen et al. [11] proposed DeepLabv3+ model, which is extended from DeepLabv3 model [12]. This study applies the encoder-decoder structure and atrous separable convolution. This model is capable of encoding multi-scale contextual information and detects the object boundaries by a decoder module. Furthermore, some research is focused on a specific area of semantic segmentation. For example, semantic segmentation of street scenes is an attractive research topic [13-14].

2.2 Emotion Images

An image is a handy tool to convey emotions. Many studies have investigated emotions in images by using various approaches. Kim et al. [15] predicted the emotions in images based on the objects and the background, which are high-level features in emotion classification. They proposed a deep neural network to recognize the emotions in a given image. They also combined high-level features and low-level features to achieve better emotion recognition performance. In this method, the object recognition accuracy is vital and affects the method accuracy directly. Furthermore, several researches applied convolutional neural networks (CNN) to advance the emotion recognition [16-17]. Many studies were performed to determine the relationship between color and emotion, and the results showed that different colors in an image carry different emotions [18]. Liu et al. [19] presented an emotion classification network by using deep learning. This method can solve several problems in traditional image color editing, such as unnatural color and block color, and then generate semantic information.

2.3 Smart Healthcare

Recently, with the rapid development of medical and computer technologies, the healthcare systems are being actively studied [20-22]. To monitor and analyze the physical health of users, Chen et al. [20] proposed a smart healthcare system that combines cognitive computing and edge computing. This system can determine the corresponding health risk grade for a user under different health statuses. Besides, the internet of things (IoT) devices are widely applied and improve smart healthcare systems. Catarinucci et al. [23] developed an IoT-aware architecture for automatic monitoring and tracking of patients, nursing staff, and biomedical devices. This system can collect both environmental conditions and patients'

physiological parameters in real-time. Wearable devices have gained a lot of attention in recently years because of the advantages of compact size and ability to provide continuous and real-time information [24-25].

Visually impaired people face many difficulties that impede their independence. Therefore, there are research efforts devoted to supporting visually impaired people [2, 26]. Long et al. [27] proposed a K-band millimeter-wave (MMW) radar system to aid visually impaired people to avoid obstructions. The results showed that MMW radar has high accuracy and stability in measuring the range and velocity. Wang et al. [3] presented a wearable system that enables visually impaired people to gain situational awareness. This system is built based on the computer vision and emotion, and users can carry it comfortably wherever they go. Long et al. [28] used MMW radar and RGB-Depth sensor to build a system, which can perceive barriers at a distance and then avoid them. This study applies deep learning to detect the objects and the system obtains many types of information by combining several different techniques.

3 Street Scene Description Scheme

This section presents the street scene description scheme. Section 3.1 provides an overview of the system, and Section 3.2 discusses the image semantic segmentation model. Finally, the color-based emotion classification is described in Sections 3.3.

3.1 System Architecture

Fig. 1 shows an overview of the proposed system. The main module of the system is an embedded system, NVIDIA Jetson AGX Xavier. This module is connected to peripheral devices, such as webcam, headphone, speaker, monitor, mouse, and USB wifi adapter. In addition, it performs almost all the functions of the system, including images collection, analysis, images processing, and the control of the system. First, the webcam of the system is used to capture the current scene of the user. The image is then segmented through a semantic segmentation model on the system. NVIDIA Jetson AGX Xavier has a GPU, and it can implement deep learning models. Moreover, we implement street scene description scheme to get the semantic information of the image. The landscape indexes are also calculated. Finally, the user can hear the description of the current scene by a headphone or speaker.

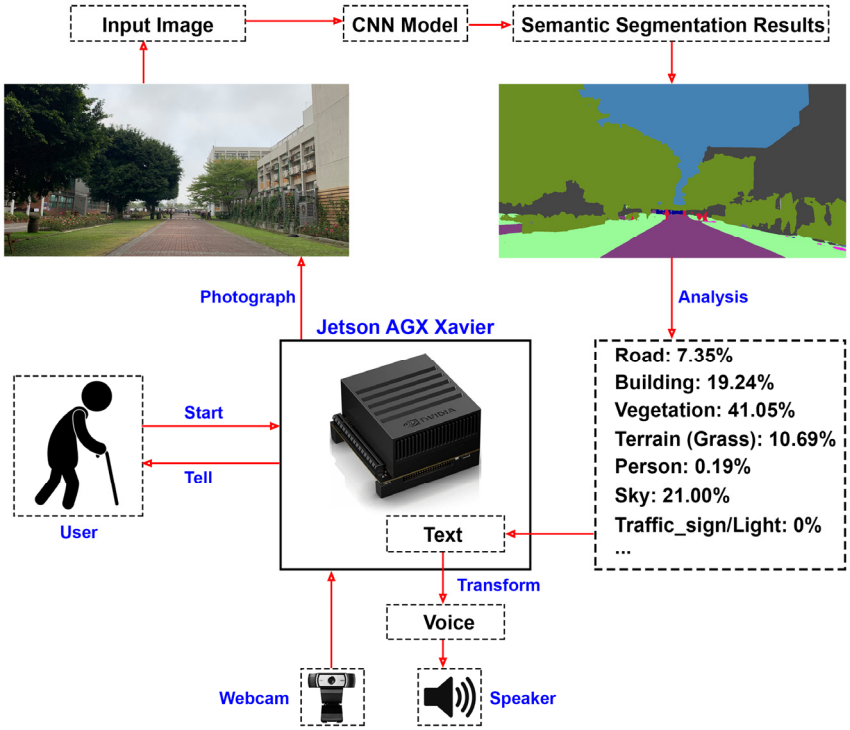

Fig. 1. System overview

3.2 Image Semantic Segmentation Model

LeNet [29] has enabled convolution neural network (CNN) to be applied in image semantic segmentation. After that, the emergence of AlexNet [30] proved that CNN-based image semantic segmentation results are better than traditional algorithms. Since then, CNN has been widely applied in image semantic segmentation algorithms. However, in CNN, pooling can increase the image space information while expanding the receptive field. Consequently, Yu et al. [31] used dilated/atrous convolution instead of pooling to solve this problem. Nevertheless, a network structure based entirely on hole convolution also has some problems, such as loss of continuity of information and relationships between objects of different sizes.

In 2017, atrous spatial pyramid pooling (ASPP) was proposed by Chen et al. [12]. ASPP connects the feature maps generated by atrous convolutions with different rates so that the neurons in the output feature map contain multiple receptive fields, which can encode multi-scale information and ultimately improve performance.

Therefore, Yang et al. [32] proposed DenseASPP, using densely convolved (Dense) hole convolution and ASPP to solve the problem of image segmentation in the field of autopilot, and achieved excellent results in the CityScapes dataset [33], as shown in Table 1. The model uses a 3×3 convolution kernel with an expansion rate of 6, 12, 18, 24 from the lower layer to the upper layer (Fig. 2), that is, the receptive field gradually increases from the lower layer to the upper layer. This article uses the DenseASPP161 model that has been trained on CityScapes to segment street scene images. The training parameters of the model are the same as those in Yang et al. [32]. Batch normalization is added before each weight layer to reduce the dimensions. Using the Adam optimizer with an initial learning rate of 0.0003 and a weight attenuation of 0.00001, the training has 80 epochs and the minimum batch size is 8. The test accuracy (mIOU) of the model on CityScapes is 79.5%.

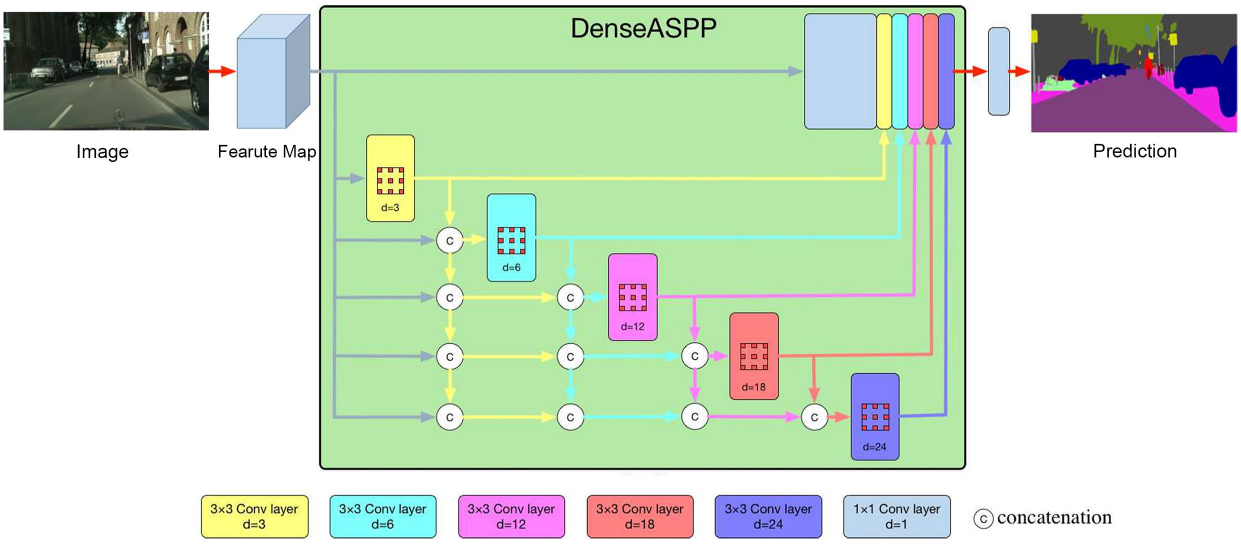

Fig. 2. The structure of DenseASPP [32]

Table 1. Performance of different models on the CityScapes dataset

Method	mIou cla.	mIoU cat.
FCN-8s	65.3	85.7
DeepLabv2-CRF	70.4	86.4
FRRN	71.8	88.9
RefineNet	73.6	87.9
PEARL	75.4	89.2
GCN	76.9	-
DUC	77.6	90.1
PSPNet	78.4	90.6
ResNet-38	78.4	90.9
DenseASPP	**80.6**	**90.7**

3.3 Color-Based Emotion Classification

Color-based emotion classification can be described as emotional feelings evoked by colors, and emotions classification is required in the process. Robert Plutchik [34] classified emotions into many categories, as shown in Fig. 3. In which, there are eight primary emotions, including trust, fear, surprise, sadness, disgust, anger, anticipation, and joy. Solli and Lenz [35] also classified the emotions based on colors. The color emotion metric uses three color emotion factors: activity, weight, and heat. Therefore, the conversion of RGB image to an emotion image with three channels: activity, weight, and heat is implemented.

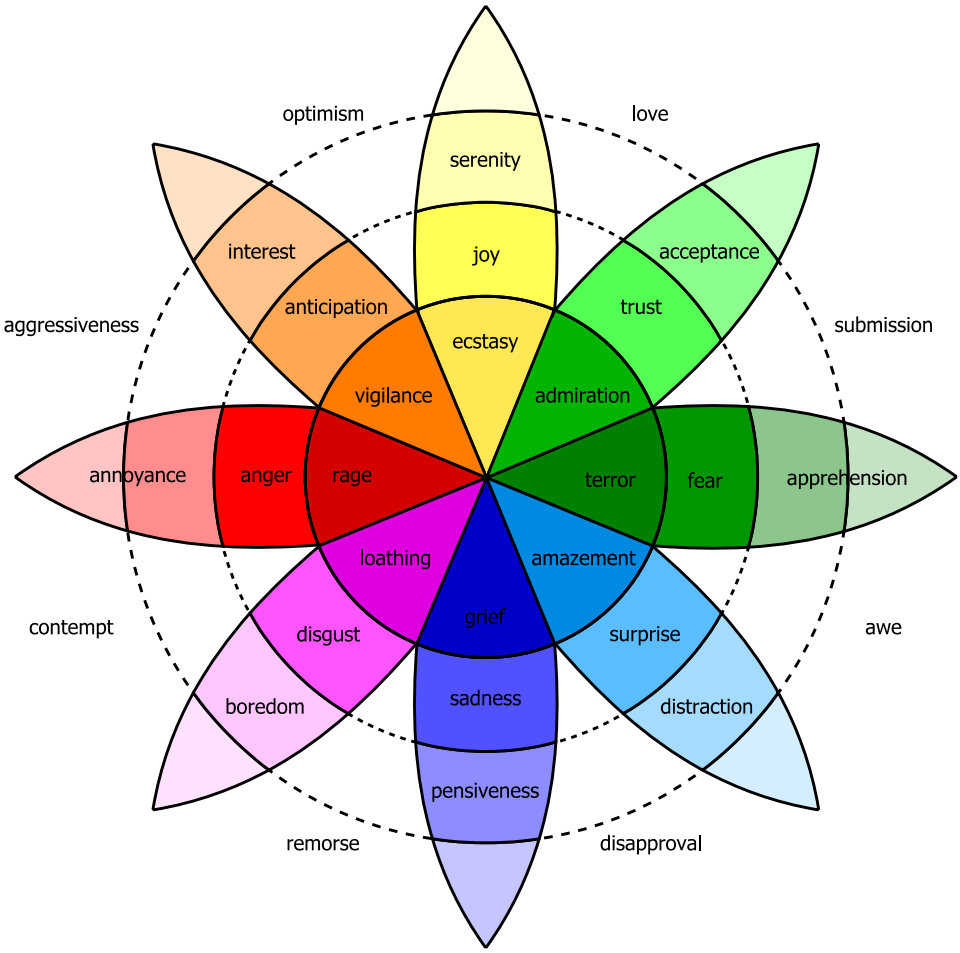

Fig. 3. Plutchik's wheel of emotions

In this study, the main color is determined after the image being analyzed. Subsequently, we can classify this main color to one of the eight color categories corresponding to eight primary emotions, as shown in Fig. 4. We use the RGB image directly and Euclidean distance to classify the color. First, we calculate the Euclidean distance from the main color, with dimensions R_m, G_m and B_m, to each color in eight primary emotions, with dimensions R_i, G_i and B_i. Then, we determine the minimum Euclidean distance to classify color class, as shown in (1).

$$d = min\left(\sqrt{(R_m - R_i)^2 + (G_m - G_i)^2 + (B_m - B_i)^2}\right), i = [1, 8] \quad \textbf{(1)}$$

Fig. 4. Eight basic emotions and colors corresponding

To get the semantic information of the image, we design a street scene description scheme based on the combination of image semantic segmentation and color-based emotion classification, as shown in Table 2. This semantic information is then transformed from text to sound that enables visually impaired people to hear the description of the current scene by a headphone or speaker.

Table 2. Pseudocode of street scene description scheme

Input: The street scene image.
Output: Get K main objects and dominant color/emotion.

1. **# Initialization**
2. basic_emotions = ["trust", "fear", "surprise", "sadness", "disgust", "anger", "anticipation", "joy"]
3. basic_colors = {(R_i, G_i, B_i), i=[1,8]} *# Eight basic colors and emotions corresponding*
4. objects = {} *# This variable contains the objects, one item is couple (obj_name: percentage)*
5. OBJ_Description = NULL *# This variable contains the object description of the input image*
6. set *K* value *# K main objects be described*
7. initialize parameters of DenseASPP scheme

8. **# Get *K* main objects**
9. img = read(input_image)
10. objects{(obj_name$_j$: percentage$_j$)} = inference DenseASPP(img)
11. sort objects{(obj_name$_j$: percentage$_j$)} following decrease percentage$_j$
12. OBJ_Description = OBJ_Description.add(obj_name$_j$), j=[1,K]
13. **# Get dominant color/emotion**
14. main_color(R_m, G_m, B_m) = get_color(img)
15. $Euclidean\ Distance = min\left(\sqrt{(R_m - R_i)^2 + (G_m - G_i)^2 + (B_m - B_i)^2}\right), i = [1,8]$
16. Classify main_color following Euclidean distance, main_color belongs to basic_colors(R_e, G_e, B_e)
17. Get emotion from color corresponding: basic_emotions[e]
18. OBJ_Description = OBJ_Description.add(basic_emotions[e])
19. **return** OBJ_Description

(Lines 1-7) First, the system initializes a group of parameters. The basic_emotions variable and basic_colors variable denote the set of eight basic emotions and colors, respectively. The objects variable represents a set of collected objects after image segmentation. Each item in this set is a couple (obj_name: percentage). obj_name indicates the object name, and percentage means the appearance frequency of this object in the image. The OBJ_Description variable denotes a set of object descriptions from the input image. The *K* variable represents the number of objects that are described.

(Lines 8-12) Next, the system performs the image semantic segmentation. The results are stored in the OBJ_Description, including *K* main objects.

(Lines 13-19) Finally, the system obtains the dominant color and corresponding emotion.

4 Street Landscape Indexes

This section presents the street landscape indexes. Section 4.1 introduces the green view index, and Section 4.2 discusses the sky-openness index. Visual entropy is described in Sections 4.3.

4.1 Green View Index

Yang et al. [36] obtained the green view index (GVI) by calculating the ratio of the green area to total area of the image taken at the intersection from four directions. Cheng et al. [37] filtered out the red and blue light and enhanced the green light. They then eroded to remove the false points, and finally recovered the image by morphological reconstruction to calculate the GVI. The captured images from eight directions are demonstrated as shown in Fig. 5. Through image semantic segmentation based on

deep convolutional neural networks, it is possible to accurately identify the areas of the street scene image that belong to the green area, thus the misclassification is ignored. The segmentation results of the street scene images in this paper include plants and terrain. In the statistics, the number of pixels in which the segmentation result is plant or terrain (grass) is first added to the number of green pixels, and then the GVI calculation formula is given as (2).

$$GVI = \frac{\sum_{i=1}^{8} Area_{g_{ij}}}{\sum_{i=1}^{8} Area_{t_{ij}}} \times 100\% \tag{2}$$

Where $Area_{g_{ij}}$ represents the number of green pixels from the segmentation result of image captured in the *i*-th direction among the eight directions, and $Area_{t_{ij}}$ represents the total pixel number of the image taken in the *i*-th direction.

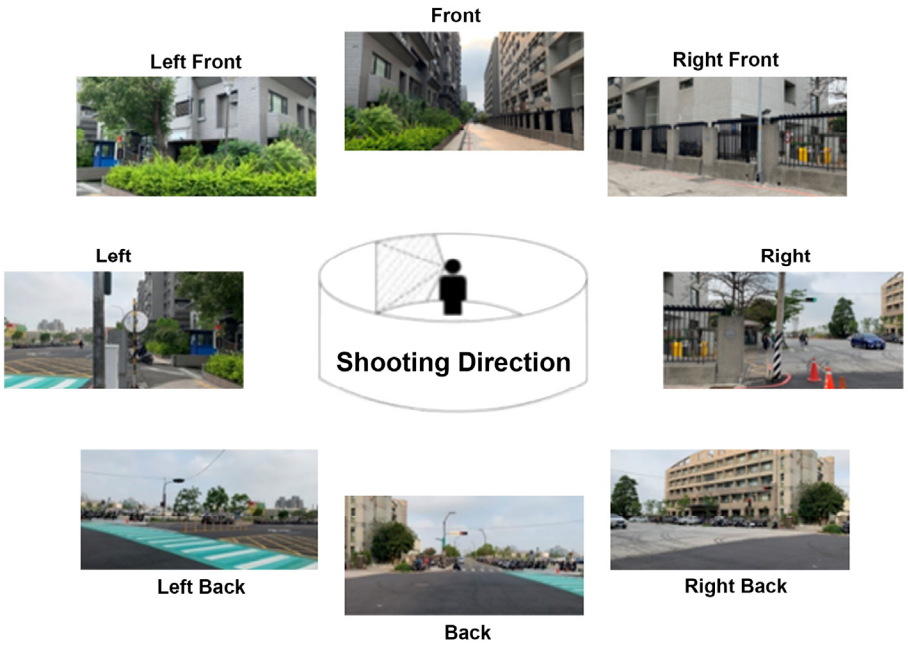

Fig. 5. Street scene shooting direction

4.2 Sky-Openness Index

The calculation of the sky-openness index (SOI) is similar to the GVI calculation, and defined as (3), where $Area_{s_{ij}}$ is the number of pixels in which the segmentation result of the sky image captured in the i-th direction.

$$SOI = \frac{\sum_{i=1}^{8} Area_{s_{ij}}}{\sum_{i=1}^{8} Area_{t_{ij}}} \times 100\% \tag{3}$$

4.3 Visual Entropy

Visual entropy (VE) is calculated based on the gray value of each pixel in the image, and it can reflect the complexity of the visual object and express subjective measurements of the visual information through human eyes. In this paper, the original RGB images are converted into grayscale images using the

COLOR_BGR2GRAY function in OpenCV-Python to calculate the visual entropy of each grayscale image according to (4).

$$VE = -\sum_{i=1}^{n} P_i \log P_i \tag{4}$$

Where $i \in [0, 255]$ is the gray value, P_i is the probability of gray value i in the image, and the visual entropy is $-P_i \log P_i$.

5 System Implementation and Prototype

A prototype is developed to verify the effectiveness of the proposed system. Details of the prototype are presented in this section. Technical specifications of NVIDIA Jetson AGX Xavier are shown in Table 3.

Table 3. Technical specifications of NVIDIA Jetson AGX Xavier

Device Name	Specifications
GPU	512-core Volta GPU with Tensor Cores
CPU	8-core ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3
Memory	32GB 256-Bit LPDDR4x | 137GB/s
Storage	32GB eMMC 5.1
DL Accelerator	(2x) NVDLA Engines
Vision Accelerator	7-way VLIW Vision Processor
Encoder/Decoder	(2x) 4Kp60 | HEVC/(2x) 4Kp60 | 12-Bit Support
Size	105 mm x 105 mm x 65 mm

In order to evaluate the effectiveness of this system, several devices are used, as shown in Fig. 6. The main module, NVIDIA Jetson AGX Xavier, which comes with a power bank is carried in a backpack. Owing to its small size and lightweight, users can carry it easily and use it in navigation. A Logitech C920 webcam is used to collect images, and it connects NVIDIA Jetson AGX Xavier directly through the USB port. This webcam is attached to the front of the backpack. In this way, the view of webcam can be used to represent the view of the user. Therefore, the objects in front of visually impaired people are described correctly. In addition, a Bluetooth headphone is connected to NVIDIA Jetson AGX Xavier by a Bluetooth audio transmitter adapter. This Bluetooth headphone is used to deliver the object description to its user.

Fig. 6. Implementation devices

6 Experimental Results

In this study, 5000 manual labeled images in the Cityscapes dataset [33] are used to train, test and verify our DenseASPP model and all the images have a resolution of 2048×1024. In total, there are 2975 images for the training stage, 500 images for validation, and 1525 images for tests. Furthermore, we use street scenes (216 images) captured from 27 locations in Feng Chia University and surrounding areas. We take 8 streetscape images from 8 directions (front, right front, right, right back, back, left back, left, and left front). We conduct image preprocessing by resizing those images to a resolution of 2048 x 1024 in order to fit the DenseASPP model, which is then used to perform image semantic segmentation. This section presents the experimental results. The first subsection analyzes the results of GVI, SOI, and VE. The second subsection presents the emotion classification results based on colors. The last subsection shows the results of street scene descriptions.

6.1 Analysis Results of GVI, SOI and VE

Fig. 7 shows the average GVI results of the 27 locations on the map. The larger the point in the figure, the higher the GVI of the street scene, similarly, the smaller the lower. According to the results in this figure, the GVI of the street scenes in the university, such as locations 10, 14, 16-21, 23, 27, are generally higher than those in other places. It can be clearly seen from the images that the trees basically cover the roads on the campus, and a large number of shrubs are around the buildings. At location 1, both sides of the road are lined with reinforced concrete buildings, lacking street greenery. Furthermore, the average GVI results of the 27 locations are explicitly presented in Fig. 8.

Fig. 7. Average GVI resulting map

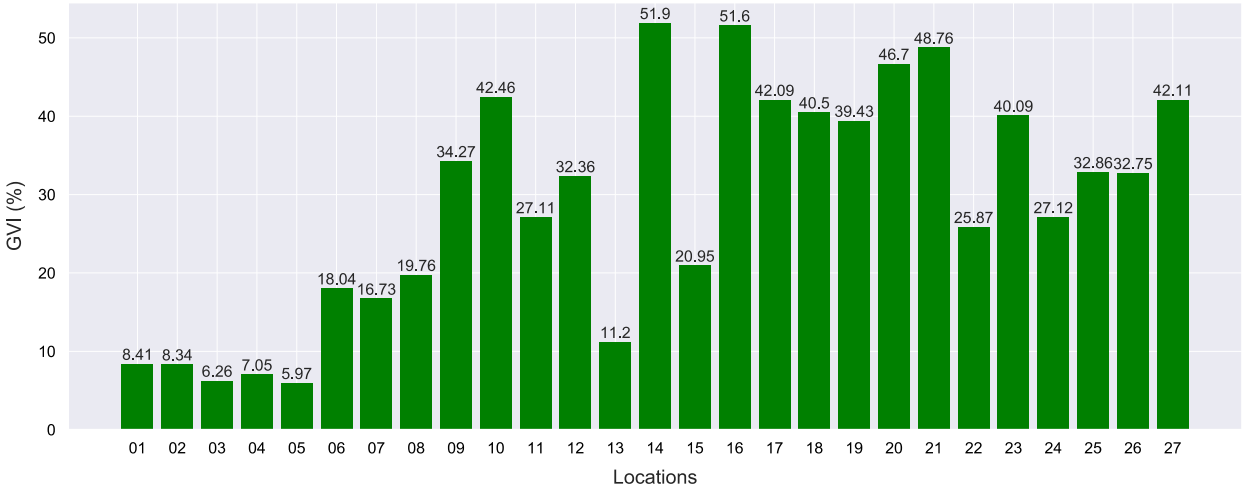

Fig. 8. Results of average GVI

Due to the coverage of the trees and the teaching buildings, the SOI results of the street scenes on the campus are generally low, as shown in Fig. 9, and even the SOI result of a certain viewing angle is approximately zero. However, location 26 on the campus has the highest SOI value because it is the playground on the campus. The SOI results of street scenes are relatively high at location 1, 7, 8, 25, which are 23.93%, 22.47%, 15.89%, 21.92% respectively as shown in Fig. 10.

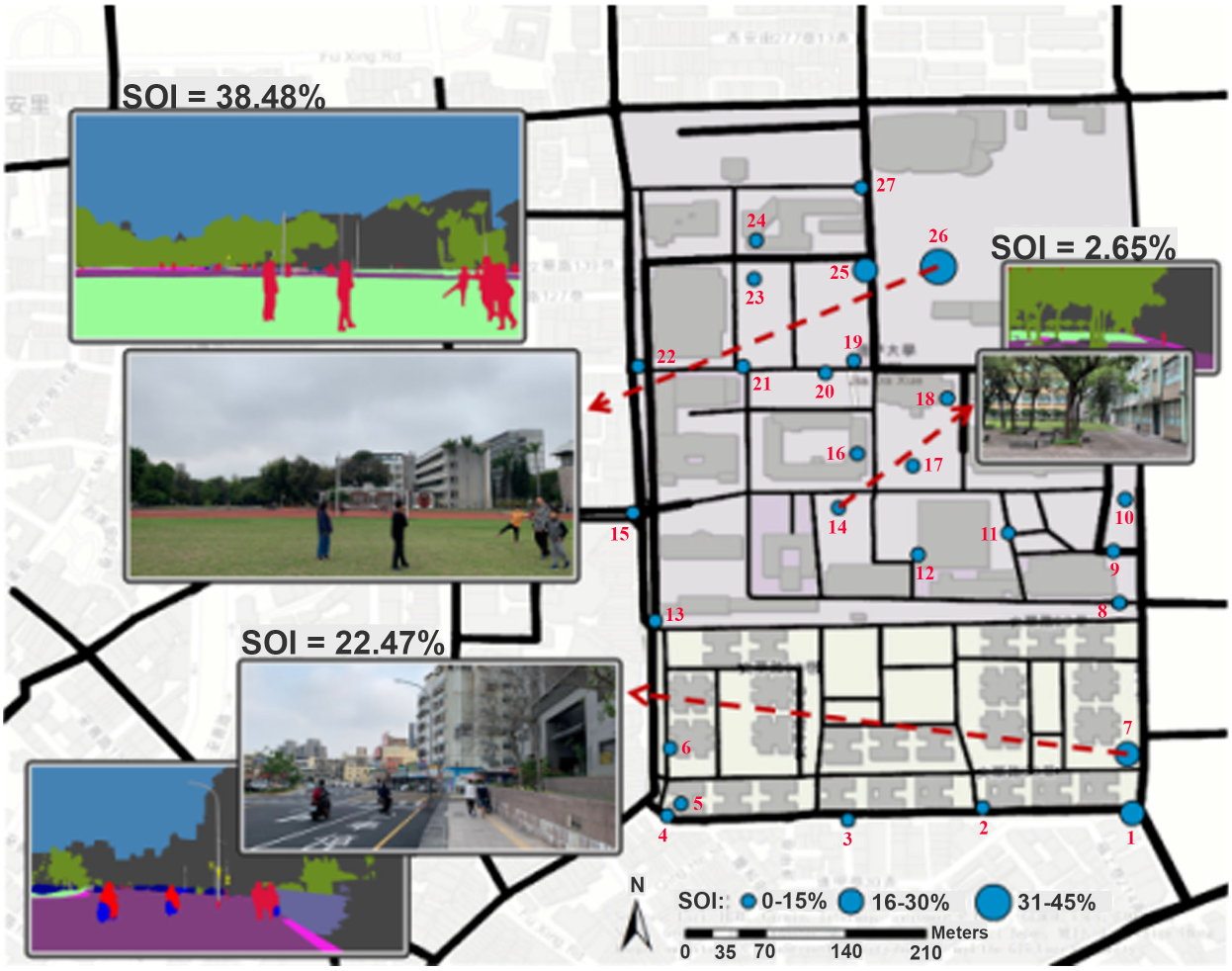

Fig. 9. Average SOI results on the map

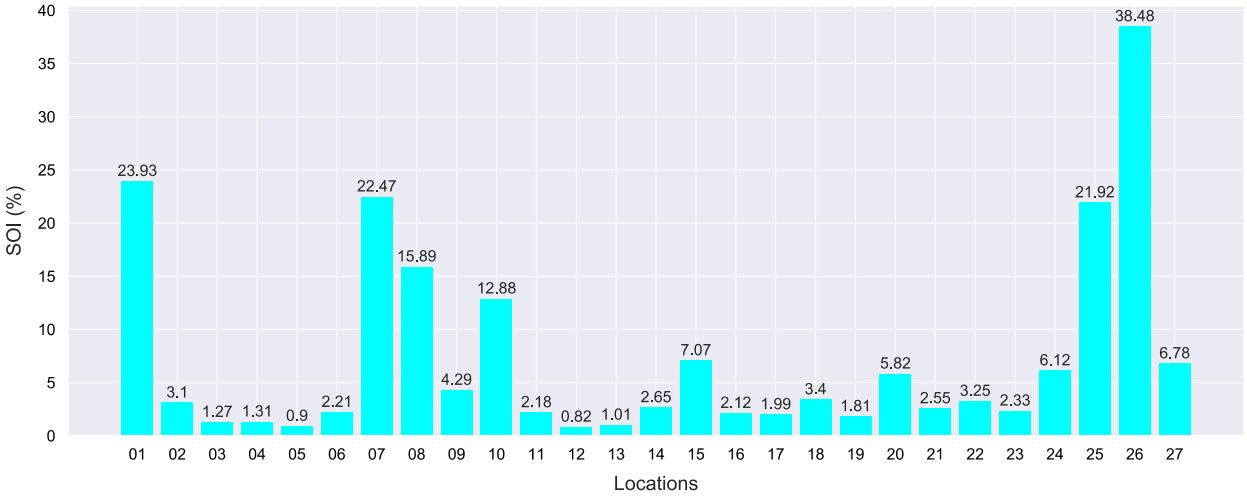

Fig. 10. Results of average SOI

Fig. 11 shows the VE results of the 27 locations on the map. Location 26, which is the playground on the campus, generates the lowest VE of street scene. On the street scene grayscale map, it can be clearly observed that the objects in the image only contain sky, grass and a small number of trees. The VE of street scene are relatively high in the majority of locations, as shown in Fig. 12.

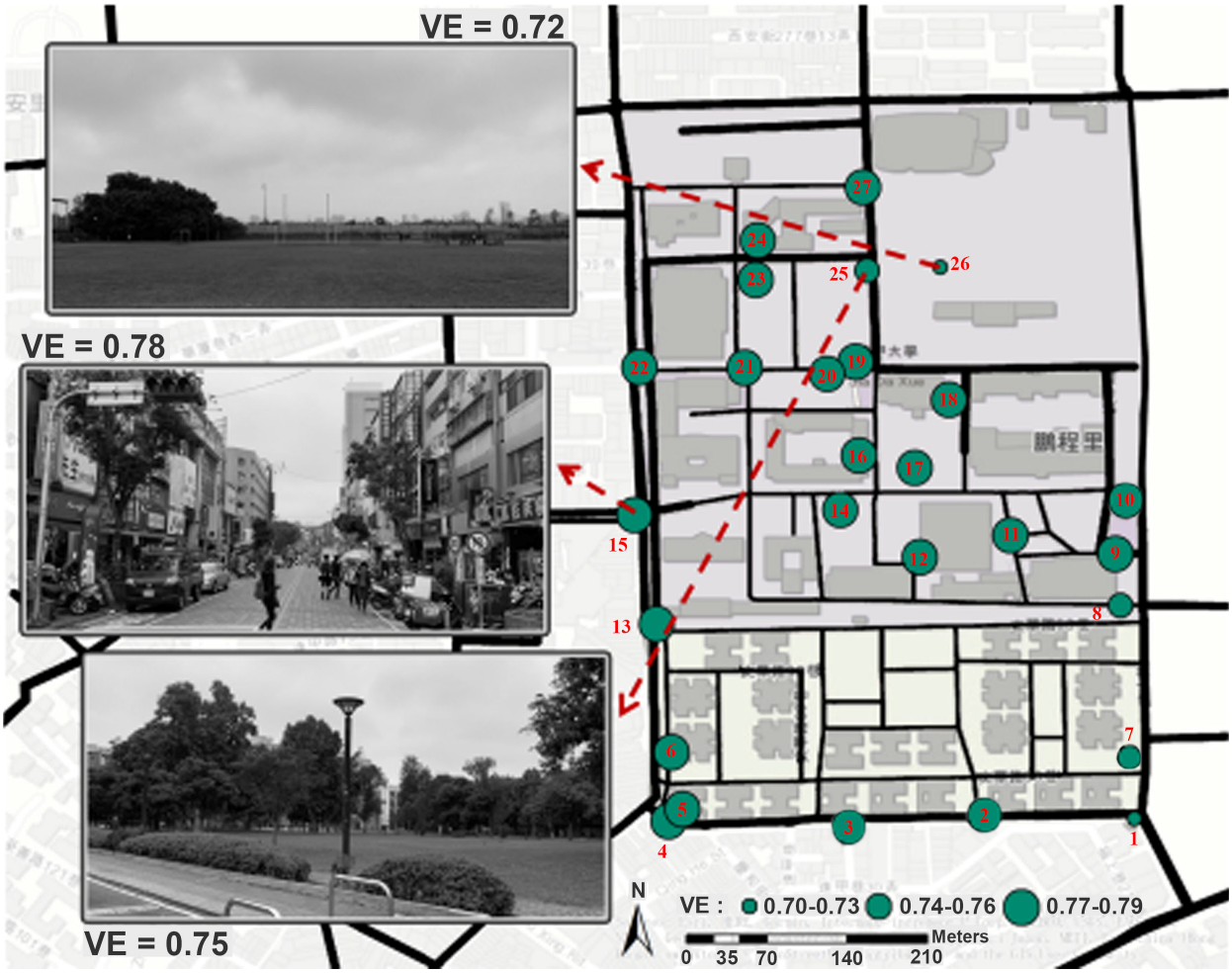

Fig. 11. VE results on the map

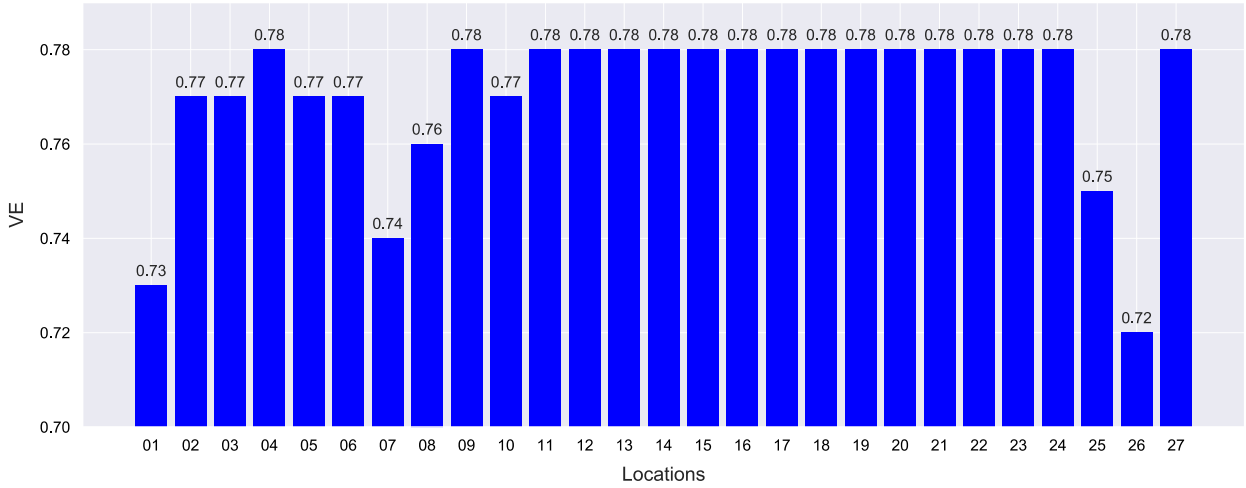

Fig. 12. Results of VE

6.2 Emotion Classification Results Based on Colors

We use the color thief method [38] to determine the dominant colors of these 216 images. Then, each collected color is classified into one of the following eight basic color categories corresponding to eight primary emotions, trust, fear, surprise, sadness, disgust, anger, anticipation, and joy. Fig. 13 shows the wheel results of emotion classification based on colors. We observe that emotions such as trust, fear, and surprise, demonstrate a high rate of 76.39% in terms of appearance frequency, while those of sadness, disgust, anger, and joy are zero. This results from the main objects in the street scene on the campus, including vegetation, sky, and sidewalk.

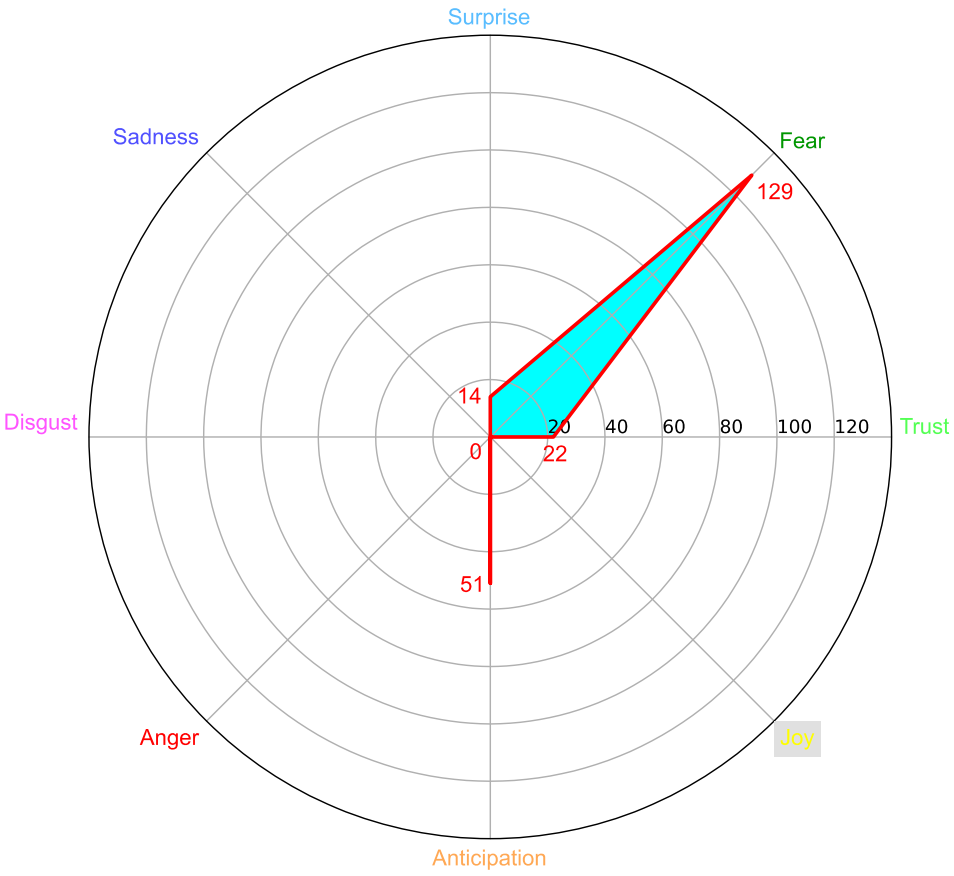

Fig. 13. The wheel results of emotion classification based on colors

6.3 Results of Street Scene Description

In order to get the semantic information of the image, we apply the street scene description scheme. We also select three different images to show different reactions. The first street scene is a campus area with lots of vegetation (Fig. 14); the second street scene is a narrow street lined with buildings (Fig. 15); the third street scene is covering a wide-open sky (Fig. 16). We extract the main colors from street scene images taken in Feng Chia University and surrounding areas. Further, those extracted main colors are used to generate the corresponding emotions (trust, fear, surprise, sadness, disgust, anger, anticipation, and joy) by referring to Fig. 4. An example of the street scene description is shown in Fig. 14. We first perform image semantic segmentation for the input image, as shown in Fig. 14a, b, respectively. Then, we analyze the semantic information, as shown in Fig. 14(c). In this instance, the top 5 most frequent objects are vegetation, terrain, building, sky, and sidewalk. Moreover, the dominant color of this image is determined to be green, corresponding to trust emotion. This semantic information, as shown in Fig. 14(d) is transformed from text to sound, and that allows visually impaired people to hear the description of the current scene by a headphone. While the same idea is applied to other images, as shown in Fig. 15 and Fig. 16, it is very clear that the proposed system provides different responses.

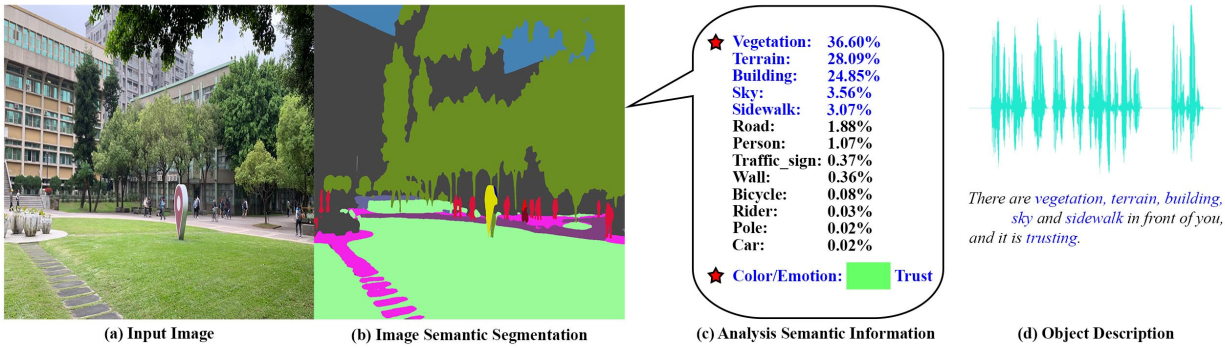

Fig. 14. Implementation of street scene description scheme for the first image

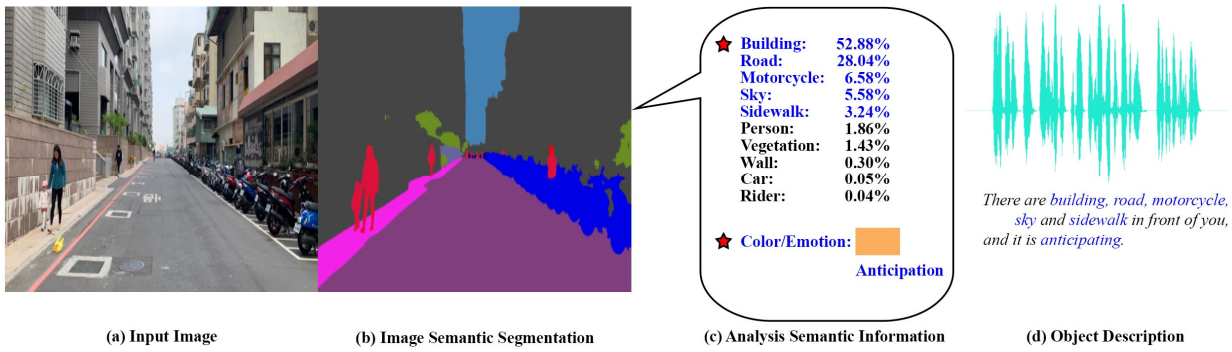

Fig. 15. Implementation of street scene description scheme for the second image

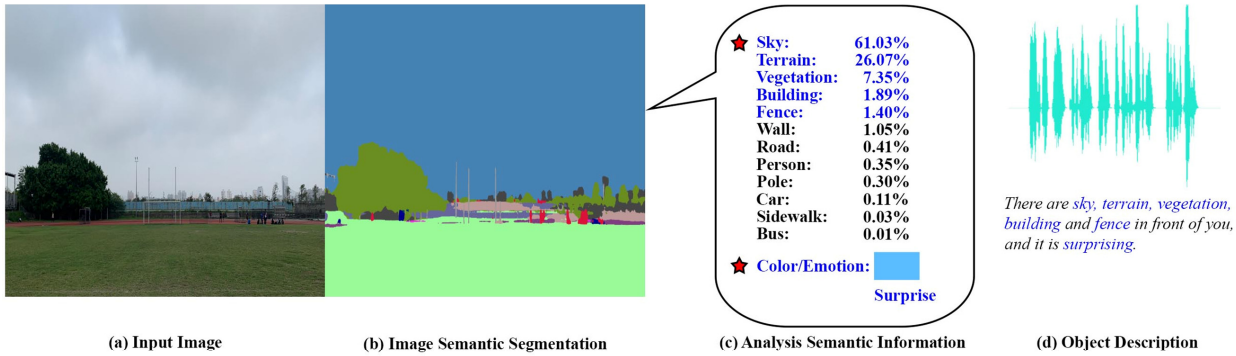

Fig. 16. Implementation of street scene description scheme for the third image

We implement the street scene description scheme at 27 locations and the results of the front view are shown in Table 4. We observe that the top 5 most frequent objects in each scene are building, vegetation, road, sky, and sidewalk, which are corresponding to the dominant colors of fear, anticipation, surprise, and trust. These results prove the accuracy and reliability of the above analysis results.

Table 4. Results of street scene description at 27 locations (Front view)

Locations	Top 5 Most Frequent Objects	Emotions
1	Sky, road, vegetation, car, building	Surprise
2	Building, motorcycle, road, rider, person	Anticipation
3	Building, car, road, vegetation, motorcycle	Fear
4	Building, road, car, sky, motorcycle	Anticipation
5	Building, road, car, vegetation, motorcycle	Anticipation
6	Building, vegetation, sidewalk, sky, person	Trust
7	Building, vegetation, sky, road, sidewalk	Fear
8	Building, vegetation, sky, road, sidewalk	Fear
9	Vegetation, building, sky, road, terrain	Fear
10	Vegetation, building, road, sky, terrain	Fear
11	Building, vegetation, road, sidewalk, person	Anticipation
12	Vegetation, building, road, sidewalk, terrain	Fear
13	Building, road, vegetation, person, terrain	Fear
14	Vegetation, building, terrain, road, person	Fear
15	Building, vegetation, sky, road, person	Fear
16	Vegetation, building, road, person, sidewalk	Fear
17	Vegetation, building, road, wall, person	Fear
18	Vegetation, sky, building, terrain, road	Fear
19	Vegetation, building, road, terrain, sky	Fear
20	Vegetation, building, road, terrain, sidewalk	Fear
21	Vegetation, sky, road, terrain, building	Fear
22	Vegetation, building, road, car, sky	Fear
23	Vegetation, building, person, road, sidewalk	Fear
24	Vegetation, building, sidewalk, person, road	Fear
25	Vegetation, road, building, sky, wall	Fear
26	Sky, terrain, vegetation, building, fence	Surprise
27	Vegetation, road, building, person, sidewalk	Fear

7 Conclusions

This study examines a wearable embedded system that aims to provide visually impaired people a better understanding of their surroundings by using a street scene description scheme. This proposed system is equipped with an embedded module-based approach and deep learning scheme. This paper provides a method to evaluate urban streetscapes by applying the GVI, SOI, and VE indicators. Furthermore, we design a street scene description scheme to extract the semantic information of an image. However, data synchronization of this system is complicated, and the processing results are not optimized. In addition, the results of the analysis are only described in simple text, and the streetscape image cannot be vividly conveyed to the visually impaired.

In the future, we intend to develop this system with a server connection, build a semantic information database, and use the power of big data to make our system more intelligent, which will enable the system to keep track of a user's journey and implement data synchronization automatically. Besides, the accuracy and processing speed of the image semantic segmentation model have a great influence on the results of the street scene images analysis. Therefore, using a lightweight CNN model with higher precision and faster speed can further improve the system performance. In this way, visually impaired people can access information as much as possible and broaden their experience of the surrounding environment.

Acknowledgements

This work was supported by Ministry of Science and Technology [grant numbers MOST107-2627-M-035-007 and MOST107-2119-M-035-006].

References

[1] H. Junwei, D. Liang, Quantitative indexes of streetscape visual evaluation and validity analysis, Journal of Landscape Research 8(3)(2016) 9-12.

[2] Y. H. Lee, G. Medioni, RGB-D camera based wearable navigation system for the visually impaired, Computer Vision and Image Understanding 149(2016) 3-20.

[3] H. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, D. Rus, Enabling independent navigation for visually impaired people through a wearable vision-based feedback system, in: Proc. 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017.

[4] NVIDIA, Jetson AGX Xavier, <https://www.nvidia.com/en-gb/autonomous-machines/embedded-systems/jetson-agx-xavier/> (accessed 15.04.2019).

[5] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, H. Lu, Stacked deconvolutional network for semantic segmentation, IEEE Transactions on Image Processing (2019) 1-13.

[6] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[7] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: Proc. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018.

[8] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proc. The IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[9] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, in: Proc. The IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[10] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proc. The European Conference on Computer Vision (ECCV), 2018.

[12] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv: 1706.05587 (2017) 1-14.

[13]L. Fan, W. Wang, F. Zha, J. Yan, Exploring new backbone and attention module for semantic segmentation in street scenes, IEEE Access 6(2018) 71566-71580.

[14] L. Fan, H. Kong, W. Wang, J. Yan, Semantic segmentation with global encoding and dilated decoder in street scenes, IEEE Access 6(2018) 50333-50343.

[15] H. Kim, Y. Kim, S. J. Kim, I. Lee, Building emotional machines: recognizing image emotions through deep neural networks, IEEE Transactions on Multimedia 20(11)(2018) 2980-2992.

[16] K.-C. Peng, T. Chen, A. Sadovnik, A. C. Gallagher, A mixed bag of emotions: model, predict, and transfer emotion distributions, in: Proc. The IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[17] J. Yang, D. She, M. Sun, Joint image emotion classification and distribution learning via deep convolutional neural network, in: Proc. The Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[18] S. Ryoo, Emotion affective color transfer, International Journal of Software Engineering and Its Applications 8(3)(2014) 227-232.

[19] D. Liu, Y. Jiang, M. Pei, S. Liu, Emotional image color transfer via deep learning, Pattern Recognition Letters 110(2018) 16-22.

[20] M. Chen, W. Li, Y. Hao, Y. Qian, I. Humar, Edge cognitive computing based smart healthcare system, Future Generation Computer Systems 86(2018) 403-411.

[21] N.-C. Hsieh, J.-F. Chen, H.-C. Tsai, Intelligent infection surveillance system to assist the control of healthcare-associated infections, Journal of Computers 27(2)(2016) 1-14.

[22] H. Yu, X. Miao, S. Wang, Y. Hu, Nursing robot safety path planning based on improved a star algorithm, Journal of Computers 30(3)(2019) 282-289.

[23] L. Catarinucci, D. de Donno, L. Mainetti, L. Palano, L. Patrono, M. L. Stefanizzi, L. Tarricone, An IoT- aware architecture for smart healthcare systems, IEEE Internet of Things Journal 2(6)(2015) 515-526.

[24] T. Wang, H. Yang, D. Qi, Z. Liu, P. Cai, H. Zhang, X. Chen, Mechano-based transductive sensing for wearable healthcare, Small 14(11)(2018) 1702933.

[25] J. Kim, A. S. Campbell, B. E.-F. de Ávila, J. Wang, Wearable biosensors for healthcare monitoring, Nature Biotechnology 37(4)(2019) 389-406.

[26] T. J. Eluvathingal, P. V. Misab, T. S. Vishnu, K. Anusree, Advanced walking stick for visually impaired, International Journal of Advance Research, Ideas and Innovations in Technology 4(2)(2018) 415-420.

[27] N. Long, K. Wang, R. Cheng, W. Hu, K. Yang, Low power millimeter wave radar system for the visually impaired, The Journal of Engineering (2019) 1-6.

[28] N. Long, K. Wang, R. Cheng, W. Hu, K. Yang, Unifying obstacle detection, recognition, and fusion based on millimeter wave radar and RGB-depth sensors for the visually impaired, Review of Scientific Instruments 90(4)(2019) 044102: 1-12.

[29] Y. LeCun, L. Bottou, Y. Bengio, D. Henderson, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11)(1998) 2278-2324.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 60(6)(2012) 1097-1105.

[31] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv:1511.07122 (2015).

[32] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, DenseASPP for semantic segmentation in street scenes, in: Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[33] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes dataset for semantic urban scene understanding, in: Proc. The IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[34] R. Plutchik, A general psychoevolutionary theory of emotion, in Theories of Emotion, Academic Press, 3-33, 1980.

[35] M. Solli, R. Lenz, Color based bags-of-emotions, in: Proc. Computer Analysis of Images and Patterns, 2009.

[36] J. Yang, L. Zhao, J. Mcbride, P. Gong, Can you see green? Assessing the visibility of urban forests in cities, Landscape and Urban Planning 91(2)(2009) 97-104.

[37] L. Cheng, S. Chu, W. Zong, S. Li, J. Wu, M. Li, Use of tencent street view imagery for visual perception of streets, ISPRS International Journal of Geo-Information 6(9)(2017) 265.

[38] Color Thief, <https://github.com/lokesh/color-thief> (accessed 15.04.2019).