Proposal of a system for the visually impaired to identify store front merchandise with no assistance

Tomoyuki Zaizen¹,*, Lifeng Zhang¹

¹Kyusyu Institute of Technology, Kitakyushu, 804-8550, Japan

Corresponding Author: q108055t@mail.kyutech.jp

Abstract

In recent years, the aging of the population in Japan has become a problem. Also, many older people are visually impaired, and the number of visually impaired people is increasing in the future. However, the lack of caregivers is also a problem in an aging society. The absence of caregivers working for the increasing number of visually impaired people is also going to serious in Japan in the future. Due to the shortage of assistants, elderly visually impaired people tend to avoid going out, and their safety confirmation may be delayed. That also becomes a big problem. Therefore, to avoid this happening, it is necessary to develop a system that enables the visually impaired to live without nursing care. This study proposes a system that enables the visually impaired to buy products without nursing care aided by image processing and machine learning.

**keywords:** Image processing, deep learning, visual impairment, shopping assistant.

1. Introduction

The definition of low vision is that the people whose corrected visual acuity is between 0.1 and 0.5 in the well-seeing eye, and the blindness is that the people whose corrected visual acuity is 0.1 or less in the well-seeing eye. Table 1 shows about 1.6 million visually impaired people in the total population. Fig. 1 shows that the percentage of the population aged 65 and over, indicated by the green and purple color areas, is increasing, and the population is getting older. Fig. 2 shows that 77% of the visually impaired persons are aged 60 years or older. As the population ages, the number of visually impaired people increases, and the number of caregivers is likely to be insufficient.

Currently, there are several ways to support the visually impaired in shopping, such as attaching Braille or electronic tags to the products themselves or the shelves. However, these have the disadvantage of costing manufacturers money and requiring time and effort when replacing products, making it...
difficult to introduce them. There is also research on applications that read product labels and detect objects to assist the visually impaired. However, visually impaired people may not know where the product name is printed on the product package because they cannot see their surroundings. The reflected light on the product surface will also cause the absence of information. Therefore, in this study, we propose a system that solves these problems and allows visually impaired people to identify the desired product correctly. This system detects only the products to be identified by creating an algorithm that extracts only moving feature points from the video of the waved products. Detecting the target product from the image in this way and using machine learning can help the visually impaired identify the product accurately.

2. Principle

2.1 optical flow

To identify products with higher accuracy, we cut out only the products using the optical flow from the video footage. The optical flow is a vector representation of the motion of objects in a digital image. It is mainly used for detecting moving objects and analyzing their operations. First, the basic principle of optical flow is described. Suppose that \( H \) and \( I \) of two consecutive images as shown in Fig. 3 are taken. In this case, if we want to investigate the motion of a moving object in the images, we need to find out how far the pixels of the moving object between image \( H \) and image \( I \). We assume that the following assumptions hold in images \( H, I \).

1. Brightness (pixel value) remains unchanged even if the pixels are moved
2. Images are smooth (spatially and temporally differentiable)

In this case, even if the image is moved from (1), the pixel value does not change. The following equation holds.

\[
I(x + u, y + v) = H(x, y) \quad (1)
\]

\((I(x + u, y + v)\) is the pixel value of a point \((x + u, y + v)\) on image \(I\), and \(H(x, y)\) is the pixel value of a point \((x, y)\) on image \(H\).\) Then, by Taylor expansion of \(I(x + u, y + v)\) from (2) to the first order term, we obtain the following equation.

\[
I(x + u, y + v) \approx I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v \quad (2)
\]

Substituting this equation (2) into equation (1), we obtain the following equation.

\[
\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v = -(I(x, y) - H(x, y)) \quad (3)
\]

\[
\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v = -\Delta I \quad (4)
\]

\[
I_xu + I_yv = -\Delta I \quad (5)
\]

Equation (5) is called the constraint equation for optical flow. The constraint equation is expressed as follows for all pixels whose (1) is true. The constraint equation can also be expressed as

\[
\nabla I \cdot \begin{pmatrix} u & v \end{pmatrix}^T = -\Delta I \quad (6)
\]

Equation (6) states that the inner product of the gradient of the edge and the vertical motion component of the movement vector relative to the edge is equal to the image difference. This formula means that the velocity component of the edge’s gradient direction can be determined, which means that the velocity component parallel to the edge can not be obtained. Although only one constraint equation for optical flow is available per pixel, one constraint equation has two unknowns (the pixel movement vector \(u\) and \(v\)). Therefore, solving the system of constraint equations does not allow us to find unique solutions to the multiple solutions of \(u\) and \(v\). So, among the multiple candidate solutions, the one closest to the correct solution \((u, v)\) must be estimated. There are two methods to estimate the motion vector \((u, v)\) of an object, the Lucas-Kanade method, and the Horn-Schunck method, and in this study, we use the Lucas-Kanade method. Next, we will elaborate on the Lucas-Kanade method.

2.2 Lucas-Kanade method

Since there is only one constraint equation per pixel, the motion vector cannot be uniquely determined. However, in the Lucas-Kanade method, the number of equations is increased. The solution \((u, v)\) is estimated by assuming that the neighboring pixels also have the same motion. If two adjacent points \(p1\) and \(p2\) in image \(I\) are assumed to have the
same motion, then the constraint equations \( u \) and \( v \) for these two pixels are the same as in the following equation.

\[
\begin{align*}
I_{x1}(p_1)u + I_{y1}(p_1)v &= -\Delta I_1 \\
I_{x2}(p_2)u + I_{y2}(p_2)v &= -\Delta I_2 \\
\vdots &
\end{align*}
\] (7)

Therefore, by solving these two equations in conjunction, we can uniquely find \( u \) and \( v \) since there are two unknowns and two equations. In fact, however, we assume that the motion is the same in a larger region than just the two adjacent points. For example, if all the motion of a \( 3 \times 3 (= 9) \) pixel is the same, we get the following nine equations.

\[
\begin{align*}
I_{x1}(p_1)u + I_{y1}(p_1)v &= -\Delta I_1 \\
I_{x2}(p_2)u + I_{y2}(p_2)v &= -\Delta I_2 \\
\vdots &
\end{align*}
\] (8)

In this case, we have two unknowns and nine equations, and the unknowns \( u \) and \( v \) that satisfy the nine equations cannot be found uniquely. Therefore, we use the least-squares method to estimate the kinetic vectors for \( u, v \) that satisfy the nine equations approximately. This is one of the basic principles of the Lucas-Kanade method. This is one of the basic concepts of the Lucas-Kanade method, which can be used to estimate kinetic vectors.

2.3 Machine Learning

Machine learning is to find a process to classify and recognize data through a process called "machine learning algorithm", which takes the input data as input values. By using the learned process, we can classify and discriminate the input data after the training. Machine learning is effective in various situations by combining the algorithms that are suitable for the usage scenarios. "Machine learning algorithms" enable us to tackle difficult tasks that fixed programs written by humans cannot solve. There are various ways to classify machine learning, and there is an idea to divide the algorithms into three types.

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

In this study we conducted supervised learning.

2.4 Supervised Learning

"Supervised learning" is an algorithm for learning the "input-output relationship" from pre-supplied data. The input data is given a predefined set of correct answers together with the input values. Since the correct answer is known, we call it "supervised learning". The teacher prepares a large amount of data, and then the program learns the relationship between input and output by feeding the data to the program. It is generally said that classification and prediction are better to use supervised learning. Once the relationship between input and output data is learned, it can be applied to unknown data, and the output can be predicted. For classification problems, this is called a classifier, and for regression problems, this is called a regression curve.

2.5 Deep Learning

In this study, we use deep learning. Machine learning does not require developers to program all actions but allows the AI to analyze the data and find laws and rules. In other words, it refers to artificial intelligence that can be trained to perform specific tasks. Deep learning is one of the methods in machine learning, which is developing a series of neural networks and related technologies that were studied in the late 20th century. Unlike conventional machine learning, deep learning refers to artificial intelligence that enhances data analysis and learning by layering layers of neural networks modeled on human nerves.

In this study, we have used Convolutional Neural Network (CNN) for deep learning neural networks shown in Fig. 4. CNN is used everywhere, including in image and speech recognition, and almost all of the deep learning methods in image recognition competitions are based on CNN. CNN is a good network for image classification that is capable of capturing spatial features and has been built on the inspiration of experiments on human visual cortex. CNN is a network of many layers. Concretely, the convolution layer and the pooling layer for size compression alternate with each other. From the processing of these layers, it is possible to reduce the image while maintaining the features of the input image. In general, the depth of the output increases with each successive layer and is eventually connected to a network of all the coupled layers.

![Network example by CNN](image-url)
3. Proposed Method

The proposed system does not use a classifier to classify the products as they are taken to improve product identification accuracy but pre-processes the product identification by extracting only the products from the taken images. To extract the products, we extract the feature points for each frame from the video images and track them to identify the products’ coordinates, and then trim them. Machine learning has two phases: the training phase and the discrimination phase. Finally, it provides audio output to convey the identification results to the visually impaired. The processing flow is shown in Fig. 5.

4. Experiment

4.1 Experiment 1

This research supposes using a smartphone takes the product photo. The smartphone was located in front of the user’s body by a neck strap to allow both user’s hands to be free. To identify a product accurately, we need to preprocess the images in an appropriate size and resolution. In other words, we need to crop the image as shown in Fig. 6. To realize it, we use optical flow to distinguish the products and the products’ background (i.e., detecting the products). Once we detected the products’ coordinates, the trimming process is performed. We attempted to trim the objects in motion by moving the products up and down and following the motion vector with the optical flow to decide the trimming area.

As a result, the successful trimmed images is about 70% in this research.

4.2 Experiment 2

After trimming the target product from the video using optical flow, study them with the deep learning, and then performed identification. Ten experiments were conducted for each of the six commodities. Also, the accuracy of the model used for identification is 100% for the fifth products of tea, but 63% for pizza potatoes. The verification results are shown in Table 2. The products packaged in boxes such as Toppo and PRETZ got 70% to 80% correct answers, and we succeeded in identifying them with high probability. The products in PET bottles, such as tea and juice, were identified with about 50% accuracy. The percentage of correct answers...
### Table 2. Results of Experiment 2

<table>
<thead>
<tr>
<th></th>
<th>Toppo</th>
<th>PRETZ</th>
<th>Pizza Potato</th>
<th>Popcorn</th>
<th>Tea</th>
<th>Juice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toppo</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PRETZ</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pizza Potato</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Popcorn</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tea</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Juice</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

for the products in the soft bags such as pizza potatoes was as low as 30%, while the percentage of correct answers for popcorn was 70%. Also, products with less than 10 total attempts are those that failed to be trimmed.

## 5. Consideration

### 5.1 Experiment 1

Reading one video frame at a time, updating the motion vector coordinates at each frame, and finally trimmed it at the position where the goods stopped. As a result, the success rate of the trimming was 70%. The failure in 30% of the cases is that the motion vectors are updated every frame. If the motion vectors were not detected correctly in the frame where the product is trimmed, the product’s coordinates could not be accurately determined.

### 5.2 Experiment 2

The percentage of correct answers was as high as 70-80% for box-type products such as Toppo and PRETZ is because the surfaces of the box-type products are flat and not deformed, such a priority making it easy to detect the feature points. The percentage of correct answers for the products in PET bottles, such as tea and juice, was 60% and 40%. The percentage of correct answers decreased from the box-type products because PET bottles’ feature points are challenging to detect. After all, the PET bottles have three sides visible from the front. Each side has a different angle and only can see a partial surface, making it difficult to detect the feature points. For pizza potatoes and popcorn, the percentages of correct answers were 30% and 70%, respectively. The training model’s discrimination accuracy is 63% for the pizza potatoes, while that of the popcorn is 100%.

## 6. Conclusion and Future Work

In this product identification for the visually impaired, we detected products by the optical flow, trimmed target moving object, and discriminated against them with a trained model. Experimentally, we found that the optical flow’s motion vectors help us narrow the product area and separate the products from the background. We also found that each product’s discrimination is possible using 1500 training images for each of the six products and creating a trained model for each product. However, in this study, only the six items used in the experiment were correctly identified, and there is room for improvement.

As for product detection, since only one outlier removal of misrecognized motion vectors can be performed at the present stage, we believe that the trimming accuracy can be improved if the number of outliers can be increased. Besides, we believe that product identification can be performed with high accuracy by learning from various images for training. Also, by training more kinds of products, we can handle more than six products.

## References

1. Number of Visually Impaired People in Japan, [https://code.kzakza.com/2018/05/gankaikai_popu/](https://code.kzakza.com/2018/05/gankaikai_popu/), Kzakza, access date: 2020.10.25
3. Introducing each type in a simple and easy-to-understand manner! Explain the difference from deep learning and DQN!(in Japanese), Udemy media, [https://udemy.benesse.co.jp/ai/machine-learning.html](https://udemy.benesse.co.jp/ai/machine-learning.html), access date: 2020.10.25